THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Discipline ou Spécialité : INFORMATIQUE

Présentée et soutenue par :
Ba-Duy DINH
Le mercredi 26 septembre 2012

Titre :
Accès à l'information biomédicale : vers une approche d'indexation et de recherche d'information conceptuelle basée sur la fusion de ressources termino-ontologiques

---

ED MITT : Image, Information, Hypermedia

Unité de recherche :
IRIT - UMR 5505

Directeur(s) de Thèse :

Mme. Lynda Tamine-Lechani (Professeur à l'Université Paul Sabatier, Toulouse)

Rapporteurs :

Mme. Brigitte Grau (Professeur) et Mme. Catherine Berrut (Professeur)

Autre(s) membre(s) du jury :

M. Mohand Boughanem (Président), M. Patrick Ruch (Examinateur),
et Mme. Nathalie Souf (Examinatrice)

Résumé

La recherche d'information (RI) est une discipline scientifique qui a pour objectif de produire des solutions permettant de sélectionner à partir de corpus d'information celle qui sont dites pertinentes pour un utilisateur ayant exprimé une requête. Dans le contexte applicatif de la RI biomédicale, les corpus concernent différentes sources d'information du domaine : dossiers médicaux de patients, guides de bonnes pratiques médicales, littérature scientifique du domaine médical etc. Les besoins en information peuvent concerner divers profils : des experts médicaux, des patients et leurs familles, des utilisateurs néophytes etc.

Plusieurs défis sont liés spécifiquement à la RI biomédicale : la représentation "spécialisée" des documents, basés sur l'usage des ressources terminologiques du domaine, le traitement des synonymes, des acronymes et des abréviations largement pratiquée dans le domaine, l'accès à l'information guidé par le contexte du besoin et des profils des utilisateurs.

Nos travaux de thèse s'inscrivent dans le domaine général de la RI biomédicale et traitent des défis de représentation de l'information biomédicale et de son accès.

Sur le volet de la représentation de l'information, nous proposons des techniques d'indexation de documents basées sur : 1) la reconnaissance de concepts termino-ontologiques : cette reconnaissance s'apparente à une recherche approximative de concepts pertinents associés à un contenu, vu comme un sac de mots. La technique associée exploite à la fois la similitude structurelle des contenus informationnels des concepts vis-à-vis des documents mais également la similitude du sujet porté par le document et le concept, 2) la désambiguïsation des entrées de concepts reconnus en exploitant la branche liée au sous-domaine principal de la ressource termino-ontologique, 3) l'exploitation de différentes ressources termino-ontologiques dans le but de couvrir au mieux la sémantique du contenu documentaire.

Sur le volet de l'accès à l'information, nous proposons des techniques d'appariement basées sur l'expansion combinée de requêtes et des documents guidées par le contexte du besoin en information d'une part et des contenus documentaires d'autre part. Notre analyse porte essentiellement sur l'étude de l'impact des différents paramètres d'expansion sur l'efficacité de la recherche : distribution des concepts dans les ressources ontologiques, modèle de fusion des concepts, modèle de pondération des concepts, etc.

L'ensemble de nos contributions, en termes de techniques d'indexation et

d'accès à l'information ont fait l'objet d'évaluation expérimentale sur des collections de test dédiées à la recherche d'information médicale, soit du point de vue de la tâche telles que TREC Medical track, CLEF Image, Medical case ou du point de vue des collections de test telles que TREC Genomics.

Abstract

Information Retrieval (IR) is a scientific field aiming at providing solutions to select relevant information from a corpus of documents in order to answer the user information need. In the context of biomedical IR, there are different sources of information : patient records, guidelines, scientific literature, etc. In addition, the information needs may concern different profiles : medical experts, patients and their families, and other users ...

Many challenges are specifically related to the biomedical IR : the document representation, the usage of terminologies with synonyms, acronyms, abbreviations as well as the access to the information guided by the context of information need and the user profiles.

Our work is most related to the biomedical IR and deals with the challenges of the representation of biomedical information and the access to this rich source of information in the biomedical domain.

Concerning the representation of biomedical information, we propose techniques and approaches to indexing documents based on :

1) recognizing and extracting concepts from terminologies : the method of concept extraction is basically based on an approximate lookup of candidate concepts that could be useful to index the document. This technique expoits two sources of evidence : (a) the content-based similarity between concepts and documents and (b) the semantic similarity between them.

2) disambiguating entry terms denoting concepts by exploiting the polyhierarchical structure of a medical thesaurus (MeSH - Medical Subject Headings). More specifically, the domains of each concept are exploited to compute the semantic similarity between ambiguous terms in documents. The most appropriate domain is detected and associated to each term denoting a particular concept. 3) exploiting different termino-ontological resources in an attempt to better cover the semantics of document contents.

Concerning the information access, we propose a document-query matching method based on the combination of document and query expansion techniques. Such a combination is guided by the context of information need on one hand and the semantic context in the document on the other hand. Our analysis is essentially based on the study of factors related to document and query expansion that could have an impact on the IR performance : distribution of concepts in termino-ontological resources, fusion techniques for concept extraction issued from multiple terminologies, concept weighting models, etc.

Our contributions, in terms of indexing techniques and information access, have been experimentally evaluated on tests collections devoted to biomedical IR, especially the TREC Genomics collections.

Remerciements

Je tiens à remercier chaleureusement :

- Mme. Lynda-Tamine Lechani, Professeur à l'Université Paul Sabatier, ma directrice de thèse, pour sa responsabilité professionnelle, sa générosité qui m'ont aidé tout au long de ces quatre années de thèse. Sans son encadrement bienveillant, ses conseils et ses idées, ce travail n'aurait pu voir le jour.

- M. Mohand Boughanem, Professeur à l'Université Paul Sabatier, pour m'avoir accueilli au sein de l'équipe SIG à l'Institut de Recherche en Informatique de Toulouse et pour avoir accepté de présider mon jury.

- M. Frédéric Courbon pour m'avoir permis de travailler à l'Institut Claudius Régaud (ICR) dans le cadre de la collaboration entre l'IRIT et l'ICR afin de collecter les données anonymisées des patients.

- Mme. Armelle Bonenfant, M. Martin Strecker, Mme. Karen Pinel-Sauvagnat, M. Riad Mokadem, M. Alain Croquette dans le cadre de mes activités d'enseignement à l'Université Paul Sabatier.

Je remercie très sincèrement Mme. Brigitte Grau, Professeur en Informatique à l'Ecole nationale supérieure d'informatique pour l'industrie et l'entreprise (ENSIIE), Mme. Catherine Berrut, Professeur à l'Université Joseph Fourier, Polytech'Grenoble pour avoir accepté d'être rapporteures de thèse, M. Patrick Ruch, Professeur à la Haute école de gestion de Genève (HEG) et Mme. Nathalie Souf, Maître de Conférence à l'Université de Toulouse, pour être examinateurs à la soutenance de ma thèse.

Je remercie également Mme. Cécile Laffaire ainsi que les membres du service Informatique pour le support technique durant les années de thèse.

Mes remerciements vont également à Mme. Josiane Mothe pour sa responsabilité et ses efforts d'améliorer les conditions de travail, notamment la présentation des ressources et des services disponibles au laboratoire, le confort des doctorants dans les salles de machine.

Je n'oublierai pas de remercier Mme. Catherine Stasiulis et Mme. Chantal Morand pour m'avoir aidé dans la démarche administrative.

Je tiens à remercier également mes collègues et amis avec qui j'ai passé de bons moments ensemble ainsi que des périodes dures de la thèse.

Enfin, mes remerciements sont destinés à ma famille pour son soutien, son aide et son espoir qui m'ont donné la confiance d'aller plus loin et encore...

Sommaire

Chapitre I Contexte et contributions de la thèse **1**

1 Contexte et problématique 2

2 Contributions 3

3 Publications dans le cadre de la thèse 6

4 Organisation de la thèse 8

Partie I

Indexation et Recherche d'Information : Application au domaine biomédical

Chapitre II *Indexation et Recherche d'Information : de la RI classique à la RI sémantique* **12**

1 Introduction 14

2 Principes et concepts de base de la RI 15
- 2.1 Concepts de base 15
- 2.2 Processus général de la RI 17
  - 2.2.1 Indexation des documents 18
  - 2.2.2 Appariement requête-document 21
- 2.3 Aperçu des modèles de RI 22
  - 2.3.1 Modèle(s) booléen(s) 22
  - 2.3.2 Modèle(s) vectoriel(s) 25
  - 2.3.3 Modèle(s) probabiliste(s) 27
- 2.4 Reformulation de la requête 32
  - 2.4.1 Reformulation par réinjection de la pertinence 32
  - 2.4.2 Reformulation par pseudo-réinjection de la pertinence . . . 34
- 2.5 Évaluation des performances de la RI 35
  - 2.5.1 Protocole d'évaluation de TREC 36
  - 2.5.2 Mesures d'évaluation 37

3 Fondements de la RI sémantique 40
- 3.1 Ressources termino-ontologiques 42
  - 3.1.1 Notions de base 42
  - 3.1.2 WordNet 43

3.1.3 Open Directory Project - ODP 44
3.1.4 Yet Another Great Ontology - YAGO 46
3.2 Principe de la RI sémantique 47
3.2.1 Les ressources exploitées pour l'indexation sémantique . . . 48
3.2.2 Désambiguïsation pour l'indexation sémantique 49
3.3 Aperçu général de travaux de désambiguïsation et d'indexation sémantique en RI 51
3.3.1 Désambiguïsation manuelle pour la RI sémantique 51
3.3.2 Désambiguïsation automatique pour la RI sémantique . . . 53
4 Conclusion 56

Chapitre III *Indexation et Recherche d'Information Biomédicale* 57

1 Introduction 59
2 Typologie des informations biomédicales 60
2.1 La littérature biomédicale 61
2.2 Les dossiers médicaux du patient 61
3 Ressources termino-ontologiques biomédicales 64
3.1 Typologie des ressources termino-ontologiques 65
3.1.1 Terminologie 65
3.1.2 Classification 65
3.1.3 Nomenclature 66
3.1.4 Thésaurus 66
3.1.5 Ontologie 67
3.2 Quelques ressources termino-ontologiques du domaine biomédical 67
3.2.1 Thésaurus MeSH 67
3.2.2 Nomenclature SNOMED 70
3.2.3 Ontologie de gènes - GO 71
3.2.4 Méta-thésaurus UMLS 72
4 Extraction des concepts biomédicaux 74
4.1 Principales approches d'extraction de concepts 75
4.1.1 Approche basée sur des règles linguistiques 75
4.1.2 Approche basée sur l'apprentissage automatique 76
4.1.3 Approche basée sur la recherche dans un dictionnaire . . . 80
4.1.4 Approche basée sur des mesures statistiques 81
4.2 Principaux outils d'extraction de concepts 83
4.2.1 PubMed ATM 83
4.2.2 MetaMap 85
4.2.3 MTI (Medical Text Indexer) 86
4.2.4 MaxMatcher 87
5 Indexation mono-terminologique *vs.* multi-terminologique de documents biomédicaux 88

5.1 Historique de l'indexation en RI biomédicale 88
5.2 Synthèse des travaux d'indexation des documents biomédicaux . . 90
5.2.1 Indexation mono-terminologique 90
5.2.2 Indexation multi-terminologique 94

6 Techniques et modèles d'appariement document-requête en RI biomédicale 98
6.1 Reformulation de requêtes 99
6.1.1 Reformulation conceptuelle de requêtes 99
6.1.2 Reformulation de requêtes par pseudo-réinjection de pertinence 104
6.2 Expansion conceptuelle de documents 105
6.3 Appariement basé sur l'identification de patrons de besoins cliniques (modèle PICO) 107

7 Évaluation de recherche d'information biomédicale 110
7.1 Campagne d'évaluation CLEF 110
7.2 Campagne d'évaluation de TREC 112
7.2.1 TREC Genomics pour la RI de la littérature biomédicale . 112
7.2.2 TRECMed pour la RI biomédicale des comptes-rendus médicaux de patients 115

8 Conclusion 118

Partie II

Proposition et Évaluation des Modèles d'Indexation et d'Accès à l'Information Biomédicale

Chapitre IV ***Résolution de l'ambiguïté des termes MeSH orientée domaine et son impact sur un processus de RI*** **123**

1 Introduction 124

2 Problématique et motivations 125

3 RI basée sur les domaines des termes MeSH 128
3.1 Indexation sémantique basée sur les domaines MeSH 130
3.1.1 Algorithme de désambiguïsation 1 (Left-to-Right TSD) . . 131
3.1.2 Algorithme de désambiguïsation 2 (Cluster-based TSD) . . 134
3.2 Appariement sémantique document-requête 136

4 Évaluation expérimentale 139
4.1 Cadre d'évaluation 139
4.2 Résultats expérimentaux 141
4.3 Discussion 145

5 Conclusion 145

Chapitre V ***Extraction de concepts biomédicaux : approche basée sur la pertinence et la corrélation des contextes documentaires et terminologiques*** **147**

1 Introduction 149

2 Problématiques et motivations 150

3 Extraction de concepts biomédicaux basée sur la combinaison du score thématique et du score de corrélation d'ordre de mots 153
3.1 Représentation des concepts de la terminologie 153
3.2 Calcul du score de pertinence des concepts candidats 155
3.2.1 Calcul du score thématique des concepts candidats 155
3.2.2 Calcul du score de la corrélation d'ordre de mots 156
3.3 Illustration de l'extraction des concepts par des exemples concrets 161

4 Évaluation expérimentale 163
4.1 Cadre d'évaluation de TREC Genomics 163
4.1.1 Description de la collection de documents 163
4.1.2 Description de l'ensemble de requêtes 164
4.1.3 Ressources termino-ontologiques des concepts biomédicaux 166
4.1.4 Acronymes et variants des noms de gènes 167
4.2 Scénarios d'évaluation 167
4.2.1 Évaluation de l'efficacité des méthodes d'extraction de concepts sur les documents 167
4.2.2 Évaluation de l'efficacité des méthodes d'extraction de concepts sur les requêtes 169
4.3 Mesures d'évaluation des performances de la RI 171
4.4 Résultats expérimentaux 172
4.4.1 Évaluation de notre méthode d'extraction de concepts sur les documents 172
4.4.1.1 Impact du nombre de termes préférés utilisés pour l'expansion de documents 172
4.4.1.2 Paramètres du modèle de reformulation de requêtes 173
4.4.1.3 Évaluation de l'efficacité de notre méthode d'extraction de concepts via l'expansion de documents et de requêtes 174
4.4.2 Résultats expérimentaux des méthodes d'extraction de concepts appliquées sur les requêtes 175
4.4.2.1 Évaluation de l'efficacité de la RI basée sur des méthodes d'extraction de concepts via l'expansion conceptuelle de requêtes 176
4.4.2.2 Évaluation de l'efficacité de la RI basée sur la combinaison de l'expansion conceptuelle et la reformulation de requêtes basée sur la technique PRF . . . 178
4.5 Évaluation comparative 180

5 Conclusion 184

Chapitre VI ***Indexation multi-terminologique pour la RI biomédicale*** **185**

1 Introduction . . . 186

2 Problématiques et motivations . . . 187

3 Architecture générale de notre approche de RI multi-terminologique . 189

4 Indexation multi-terminologique basée sur des techniques de vote . . . 191
4.1 Extraction mono-terminologique de concepts . . . 191
4.2 Extraction de concepts multi-terminologique . . . 193

5 Appariement multi-terminologique basé sur la combinaison des contextes document et requête . . . 198
5.1 Expansion conceptuelle de documents . . . 199
5.2 Combinaison de l'expansion documentaire et la reformulation de la requête par la méthode PRF . . . 200

6 Évaluation expérimentale . . . 202
6.1 Objectifs d'évaluation . . . 202
6.2 Cadre d'évaluation . . . 202
6.2.1 Collections de TREC Genomics . . . 203
6.2.2 Protocole d'évaluation . . . 205
6.2.3 Schémas d'appariement document-requête . . . 207
6.2.4 Modèles de reformulation de la requête par la méthode PRF 208
6.2.5 Mesures d'évaluation des performances de la RI . . . 209
6.3 Résultats expérimentaux . . . 210
6.3.1 Entraînement des modèles de pondération et modèles de reformulation de requêtes . . . 210
6.3.2 Évaluation de l'efficacité de l'indexation mono-terminologique 211
6.3.3 Évaluation de l'efficacité de l'indexation multi-terminologique 218
6.4 Discussion . . . 221

7 Conclusion . . . 223

Chapitre VII BioSIR - système prototype de RI biomédicale **224**

1 Introduction . . . 225

2 Extraction de concepts . . . 228

3 Expansion conceptuelle de documents . . . 231

4 Évaluation de requêtes . . . 235

5 Outils d'évaluation . . . 239

Conclusion générale

Bibliographie

Liste des tableaux

II.1 Racinisation *vs.* lemmatisation 20
II.2 Degrés d'appartenance de termes aux documents 25
II.3 Notations des mesures d'évaluation 38
II.4 Quelques statistiques de YAGO 46

III.1 Les différentes catégories ou domaines du MeSH 69
III.2 Les onze axes de la SNOMED 70
III.3 Description d'un terme dans GO 72
III.4 Règles de formation des termes composés 75
III.5 Illustration des concepts extraits par ATM 84
III.6 Exemple de l'expansion conceptuelle de la requête dans (Aronson *et al.*, 1997) 101
III.7 Les informations extraites du modèle PICO 108
III.8 Exemples de requêtes dans la tâche de recherche des cas de patients dans ImageCLEF 2011 111
III.9 Statistiques des collections ImageCLEF (Case-Based IR) . . . 112
III.10 Un exemple de documents utilisés dans TREC Genomics 2004-2005 114
III.11 Résumé des tâches de RI dans le cadre de TREC Genomics . . 115
III.12 Exemples de requêtes dans TRECMed 2011 116

IV.1 Nombre de (sous-)domaines partagés par les concepts MeSH . 127
IV.2 Exemple d'analyse lexicale en utilisant TreeTagger 129
IV.3 Exemples de documents de la collection OHSUMED 140
IV.4 Description statistique de la collection test 140
IV.5 Exemples de requêtes de la collection OHSUMED 141
IV.6 Les performances de RI (P@5, P@10 et MAP) obtenues sur la collection OHSUMED 143

V.1 Le concept "Minisatellite repeats" (C0242827), issu du thésaurus MeSH, vu comme un document qui est constitué par ses termes d'entrée 151
V.2 Illustration de la corrélation en termes d'ordre de mots entre le texte et les entrées des concepts. 152
V.3 Exemples des concepts du thésaurus MeSH vus comme les documents d'une collection de concepts 154

V.4 Calcul de la corrélation de positions des mots entre un terme d'entrée du concept désigné par le terme "Colorectal cancer" et la fenêtre d'un fragment de texte délimitée par le premier et le dernier mot du terme d'entrée. 158
V.5 Résultats de l'extraction de concepts basée sur la corrélation de Spearman en comparaison avec ceux obtenus par la méthode basée sur la mesure Cosinus en RI. 162
V.6 Statistiques de la collection TREC Genomics 2004 164
V.7 Exemples de requêtes de TREC Genomics 165
V.8 Un extrait des noms de gènes stockés dans un dictionnaire . . 168
V.9 Performances de la RI (P@5, P@10, MAP) pour $N = 0..25$ . . 173
V.10 Les résultats MAP obtenus par la reformulation de requêtes par la méthode PRF 174
V.11 Performances de la RI $P@5, P@10, MAP$ (% taux d'amélioration) de la combinaison des contextes documentaires et de requêtes en comparaison à la base d'évaluation de référence. . 175
V.12 Performances des méthodes d'extraction de concepts pour l'expansion conceptuelle de requêtes en comparaison avec les résultats de la base d'évaluation de référence BM25. 177
V.13 Performances de la RI (MAP, P@5, P@10) basée sur la reformulation PRF en combinaison avec l'expansion conceptuelle de requêtes. 181
V.14 Comparaison aux résultats de TREC Genomics 2004 182

VI.1 Exemples de concepts MeSH enregistrés dans un dictionnaire . 194
VI.2 Exemple de l'extraction de concepts mono-terminologique . . . 195
VI.3 Un documents biomédical sous le format TREC 195
VI.4 Listes de concepts extraits à partir des documents 196
VI.5 Résultats de l'extraction multi-terminologique de concepts . . 198
VI.6 Un exemple de citation ou document dans MEDLINE 203
VI.7 Quelques statistiques sur les requêtes de TREC Genomics . . . 205
VI.8 Résultats en terme de MAP de notre approche de RI basée sur une mono-terminologie sur la collection TREC Genomics 2005. 214
VI.9 Listes des concepts extraits à partir du document 1002230 . . 216
VI.10 Comparaison des performances (MAP, P@10, P@20, Rappel) de l'indexation mono-terminologique aux performances de la base de référence d'évaluation (Baseline_QE) 217
VI.11 Performances de notre approche de RI multi-terminologique sur la collection TREC Genomics 2005. 219
VI.12 Comparaison des performances (MAP, P@10, P@20, Rappel) de l'approche de RI **mono-** *vs.* **multi-**terminologique 220

VII.1 Lignes de commande pour interagir avec OSIRIM 227
VII.2 Documents étendus par des termes préférés désignant les concepts extraits 232

Liste des figures

II.1 Processus en général d'un SRI 17
II.2 Version texte en cache d'un document original 21
II.3 Exemple de requête booléenne dans Google 23
II.4 Documents pertinents *vs.* non-pertinents vis-à-vis d'une requête 38
II.5 Exemple du réseau des concepts dans WordNet 44
II.6 L'interface Web de l'ontologie ODP 45
II.7 Visualisation des entités dans YAGO 47
II.8 Exemple de réseau sémantique construit à partir de concepts candidats. 55

III.1 Exemple d'une citation de MEDLINE 62
III.2 Exemple d'un compte-rendu de consultation 63
III.3 Extrait de l'arborescence C (domaine 'Maladie') de MeSH . . . 68
III.4 Regroupement des termes synonymes dans l'UMLS 73
III.5 Le réseau sémantique *Biologic Function* 74
III.6 Les paramètres d'un modèle de Markov caché 77
III.7 Illustration des vecteurs de support 78
III.8 Exemples de la représentation des matrices LLSF 79
III.9 Architecture générale du système NOMINDEX (Pouliquen, 2002) 91
III.10 Architecture générale du système MAIF 93
III.11 Extraction des concepts pour une indexation multi-terminologique avec MetaMap 95
III.12 Les composantes principales de MTI (Medical Text Indexer) . 96
III.13 Les relations inter et intra entre les concepts issus de différentes terminologies (Avillach *et al.*, 2007) 97
III.14 Exemple de document et requête annotés par des concepts UMLS 107
III.15 Exemple d'un document lié à un cas de patients dans ImageCLEF 2011 111
III.16 Exemple d'un compte-rendu du DMP dans TRECMed 2011 . 117

IV.1 Visualisation du concept "Pain" dans l'architecture poly-hiérarchique de la terminologie MeSH 126
IV.2 Variation de la spécificité des concepts MeSH 128

IV.3 Illustration de la première méthode de désambiguïsation 133
IV.4 Illustration de la deuxième méthode de désambiguïsation . . . 136
IV.5 Représentation du document par deux ensembles de mots simples et termes désignant les concepts biomédicaux 137
IV.6 Illustration d'un fragment de texte annoté avec les étiquettes correspondant aux sens des termes ambigus 137
IV.7 Courbe rappel-précision des requêtes basées sur *titre* 143
IV.8 Rappel-précision des requêtes basées sur *titre* et *description* . . 143
IV.9 Analyse de résultats en fonction de la spécificité de la requête 144

V.1 Une fenêtre délimitée par le premier et le dernier mot de chaque terme d'entrée du concept 157
V.2 Structure poly-hiérarchique du MeSH 166
V.3 Résultats MAP de la méthode PRF en combinaison avec la méthode d'expansion conceptuelle 179
V.4 Résultats P@5 de la méthode PRF en combinaison avec la méthode d'expansion conceptuelle 179
V.5 Résultats P@10 de la méthode PRF en combinaison avec la méthode d'expansion conceptuelle 180
V.6 Distribution de la densité par noyau des résultats MAP en comparaison aux résultats officiels de TREC Genomics 2004 . 183

VI.1 Architecture générale du système de RI conceptuelle multi-terminologique 189
VI.2 Un exemple de défaut d'appariement document-requête 199
VI.3 Expansion documentaire en utilisation les termes préférés . . . 200
VI.4 Expansion documentaire *vs.* expansion de la requête 201
VI.5 Distribution de la longueur du document dans les collections TREC Genomics 204
VI.6 Distribution de la MAP pour chaque modèle de pondération en combinaison avec chaque modèle d'expansion de la requête et vice-versa. 212
VI.7 Estimation par la méthode du noyau d'un échantillon de 300 (3x10x10) valeurs de MAP obtenues pour chaque modèle de pondération en combinaison avec une des **trois** modèles d'expansion de requêtes. 213
VI.8 Histogrammes des nombres de concepts MeSH, SNOMED, GO extraits à partir des documents 214
VI.9 Distribution des concepts communs entre les terminologies MeSH, SNOMED et GO extraits à partir des documents de la collection TREC Genomics 2005 215

VI.10 Visualisation des performances (P@10, Rappel, MAP) de la RI mono-terminologique en 3D en comparaison aux performances de la base de référence d'évaluation sans utiliser aucune terminologie. 217

VII.1 Plateforme de la RI biomédicale BioSIR (version prototype) . 225
VII.2 Interface Ganglia : outil de visualisation des ressources en temps réels de la plateforme OSIRIM 226
VII.3 BioSIR : Extraction de concepts à partir des documents 228
VII.4 BioSIR : Expansion documentaire conceptuelle 231
VII.5 BioSIR : Recherche d'information (mode interactif) 236
VII.6 BioSIR : Configuration des modèles de RI 236
VII.7 BioSIR : Recherche d'information conceptuelle biomédicale (mode d'évaluation) 237
VII.8 Interface Web de BioSIR (accessible en Intranet) 237
VII.9 Résultats de recherche avec BioSIR 238
VII.10BioSIR : Expansion de la requête par des concepts biomédicaux, noms de gènes ou de protéines 238
VII.11BioSIR : Évaluation des performances de la RI 239

Chapitre I

Contexte et contributions de la thèse

Sommaire

1 Contexte et problématique **2**

2 Contributions **3**

3 Publications dans le cadre de la thèse **6**

4 Organisation de la thèse **8**

"Do a little more each day than you think you possibly can."
–Lowell Thomas

1 Contexte et problématique

Le domaine de la recherche d'information est une branche de l'informatique qui porte sur l'acquisition, l'organisation, le stockage et l'accès à l'information. Depuis les années 1970 (Salton, 1970), ce domaine n'a cessé d'évoluer pour rationaliser les processus associés du point de vue fondamental notamment concernant la définition de modèles d'indexation, la spécification de modèles formels d'appariement, la spécification de cadres d'évaluation. Ces résultats de recherche fondamentaux sont exploités cependant avec des différenciations ou pas à divers domaines d'application généraux tels que la recherche d'information sur le web, la recherche d'information multilingue ou alors spécifiques comme la recherche d'information dans le domaine légal (Cormack *et al.*, 2010) ou la recherche d'information médicale (Hersh, 2008). C'est précisément dans le domaine de la recherche d'information médicale que s'inscrivent nos travaux de thèse.

De manière générale, les systèmes informatiques médicaux ont connu une grande évolution depuis ces deux dernières décennies tant du point de vue de leur architecture que de la qualité et de la diversité des services autour du stockage de l'information, l'accès à l'information pertinente pour une médecine basée sur des niveaux de preuve, l'aide à la décision pour l'amélioration de la qualité des soins (Hersh, 2004; Tamburis, 2006). Dans ce cadre général, l'information biomédicale utilisée comme support pour les tâches de recherche, d'extraction d'information et de connaissances concerne principalement la littérature médicale, les dossiers de patients, les ressources sémantiques en médecine. Les facteurs de diversité, de volumétrie, d'hétérogénéité combinée aux exigences de qualité et de sécurité des informations gérées par de tels systèmes, sont à l'origine d'un essor remarquable ces dernières années de travaux de recherche, particulièrement dans le domaine de la recherche et de l'extraction d'information. En particulier, dans le domaine biomédical, les services de production et d'accès à l'information ne cessent de se diversifier. À titre d'exemple, MEDLINE (Medical Literature Analysis and Retrieval System Online), qui est la base de données bibliographiques de premier ordre développée par la NLM (US National Library of Medicine), contient plus de 21 millions de références d'articles [1] ou citations en sciences de la vie, notamment de la biomédecine.

1. Ces chiffres sont recensés en Mai 2012

En faisant la recherche d'information biomédicale sur Internet, les utilisateurs rencontrent souvent des difficultés pour formuler la requête (Keselman *et al.*, 2008). Les moteurs de recherche généralistes comme Google [2], Yahoo [3] ou Bing [4] sont populaires pour la recherche des informations générales sur Internet tandis que les moteurs de recherche dédiés à l'information biomédicale, comme PubMed ou CISMeF, sont essentiels pour récupérer des informations stockées dans les bases bibliographiques comme MEDLINE ou des documents biomédicaux dans la littérature de manière générale. Les limitations concernant les outils de recherche disponibles peuvent être résumées comme suit : d'une part, les outils de recherche généralistes ne considèrent pas les caractéristiques concernant les informations biomédicales comme la synonymie, l'utilisation des acronymes, ou des abréviations dans les documents biomédicaux. D'autre part, les outils de recherche dans le domaine biomédical intègrent les ressources termino-ontologiques comme MeSH (Medical Subject Headings), SNOMED (Systematized NOmenclature of MEDicine), ICD-10 (International Classification of Diseases), GO (Gene Ontology), ... Cependant, de nombreux verrous restent à lever comme : l'extraction efficace de concepts termino-ontologiques à partir de textes médicaux, l'exploitation de la diversité des granules d'information qui se trouvent dans les ressources termino-ontologiques, l'optimisation des modèles d'appariement en RI dans le contexte précis de l'information médicale.

2 Contributions

Les travaux présentés dans ce mémoire se situent dans le contexte précis de l'accès à l'information médicale. Plus précisément, nos contributions portent sur les quatre volets suivants : (1) la reconnaissance de concepts termino-ontologiques dans les contenus documentaires : cette reconnaissance s'apparente à une recherche approximative de concepts pertinents associés à un contenu, vu comme un sac de mots. La technique associée exploite à la fois la similitude structurelle des contenus informationnels des concepts vis-à-vis des documents mais également la similitude du sujet porté par le document et le concept, (2) la désambiguïsation des entrées des concepts reconnus en exploitant la branche liée au sous-domaine principal de la ressource-termino ontologique, (3) l'exploitation de différentes ressources termino-ontologiques dans le but de couvrir au mieux la sémantique du contenu documentaire, (4) l'appariement requête-document basé sur l'expansion combinée des requêtes et des documents guidées par le contexte du besoin en information d'une part et des contenus documentaires d'autre part.

2. `http://www.google.com`
3. `http://www.yahoo.com`
4. `http://www.bing.com`

Nous donnons un aperçu de ces contributions selon les principaux axes suivants : ***Indexation de l'information biomédicale*** et ***Recherche d'information biomédicale***.

1. **Indexation de l'information biomédicale**

   (a) **Extraction de concepts termino-ontologiques :** nous étudions l'impact de l'utilisation des concepts extraits à partir du document et de la requête pour améliorer les performances de la RI. Nous avons proposé une méthode d'extraction de concepts termino-ontologiques à partir de texte en se basant sur un appariement rapproché concept-document. La spécificité de cette méthode réside dans la combinaison de mesures structurelles et de contenu pour la sélection de concepts pertinents. Nous avons donc évalué plusieurs algorithmes ou outils d'extraction de concepts biomédicaux de l'état-de-l'art, notamment les outils utilisés quotidiennement à la bibliothèque de la santé de médecine NLM aux États-Unis (PubMed ATM, MTI, MetaMap, etc.). Nous comparons les performances de la RI obtenues par l'expansion conceptuelle de la requête ainsi que l'expansion conceptuelle de documents en utilisant ces outils par rapport aux performances obtenues par notre meilleure méthode d'extraction de concepts.

   (b) **Désambiguïsation :** nous proposons une nouvelle approche d'indexation conceptuelle basée sur le domaine le plus adéquat des termes qui désignent les concepts, issu du thésaurus MeSH (Medical Subject Headings), identifiés à partir du document et de la requête. Nous exploitons la structure poly-hiérarchique du thésaurus MeSH pour désambiguïser les termes ambigus dans les documents et la requête de l'utilisateur. Nous définissons l'ambiguïté d'un terme MeSH par le fait qu'il est défini dans plusieurs sous-domaines parmi les 16 domaines dans cette structure poly-hiérarchique. Les informations conceptuelles que nous avons identifiées sont : le domaine le plus adéquat du concept identifié et sa spécificité qui correspond à sa profondeur dans l'arborescence. Il s'agit ici d'une combinaison de la représentation textuelle classique et la représentation conceptuelle du document et de la requête.

   (c) **Indexation multi-terminologique :** nous proposons une nouvelle approche d'indexation multi-terminologique en se basant sur plusieurs techniques de fusion de concepts issus de différentes terminologies. Nous l'intégrons par la suite dans un processus de RI basée sur une représentation conceptuelle des documents via l'expansion conceptuelle de documents. Ensuite, lors de la recherche, la requête est modifiée par la technique de reformulation de la requête basée sur la méthode *pseudo relevance feedback*. Notre hypothèse derrière l'expansion conceptuelle

de documents est basée sur le fait que les sujets sémantiques (c-à-d les concepts extraits) permettent de mettre en évidence les informations biomédicales dans le contenu textuel. Ceci permettrait de résoudre le défaut d'appariement lié à l'utilisation des variantes des termes comme la synonymie, l'abréviation, ... dans les textes biomédicaux.

2. **Recherche d'information biomédicale**

   (a) **Combinaison des contextes :** nous proposons ici de combiner le contexte global du document (i.e., les concepts extraits à partir des ressources termino-ontologiques) et le contexte local de la requête (i.e., les premiers documents étendus retournés lors de la première recherche) dans le but d'améliorer le taux de couverture entre la requête et les documents pertinents dans la collection. Nous pouvons distinguer les deux types de "contexte" : *contexte global vs. contexte local*. La notion "contexte" indique la source d'information d'où les termes les plus significatifs qui sont reliés à la requête ou au document peuvent être extraits. Le contexte global (e.g., thésaurus, ontologies, collection de documents ...) est déterminé indépendamment de la requête tandis que le contexte local (e.g., les k premiers documents retournés) est déterminé en fonction du contenu textuel. De plus, notre étude vise également à évaluer l'efficacité de la RI en combinant ces sources d'information pour améliorer les performances de la RI.

   (b) **Analyse expérimentale des résultats de RI :** nous évaluons plusieurs facteurs qui pourraient influencer les performances de la RI biomédicale comme : (1) les connaissances sur les documents, c-à-d les concepts extraits qui sont prédéfinis dans une ou plusieurs ressources termino-ontologiques et (2) les connaissances sur la requête de l'utilisateur, c-à-d les termes extraits à partir des premiers documents retournés lors de la première phase de recherche, (3) les différents modèles d'appariement de l'état-de-l'art, notamment les modèles de pondération BM25 (Robertson *et al.*, 1994) et les modèles DFR (Divergence From Randomness) (Amati, 2003).

3 Publications dans le cadre de la thèse

Articles publiés dans des revues internationales avec comités de lecture

- **Duy Dinh, Lynda Tamine** (2012). Towards a context sensitive biomedical information retrieval based on domain knowledge sources. *Journal of Web Semantics : Science, Services and Agents on the World Wide Web*, Elsevier, 12-13 :41-52

- **Duy Dinh, Lynda Tamine, Fatiha Boubekeur** (2012). A study on factors affecting the effectiveness of biomedical document indexing and retrieval based on terminologies. *Artificial Intelligence in Medicine, Elsevier, 2012* (*à paraître*).

Articles publiés dans des conférences internationales avec comités de lecture

- **Duy Dinh, Lynda Tamine** (2011). *Voting techniques for a multi-terminology based biomedical information retrieval*. Dans : the 13th Conference on Artificial Intelligence in Medicine, **AIME 2011**, Juillet 2-6, 2011, Bled, Slovenia, p. 184–193 ;

- **Duy Dinh, Lynda Tamine** (2011). *Combining global and local semantic contexts for improving biomedical information retrieval*. Dans : the 33rd European Conference on Information Retrieval, **ECIR 2011**, Avril 18-21, 2011, Dublin, Ireland, p. 375–386 ;

- **Duy Dinh, Lynda Tamine** (2011). *Biomedical concept extraction based on combining the content-based and word order similarities*. Dans : the 26th ACM Symposium on Applied Computing, **SAC 2011**, Mars 21-25, 2011, Taichung, Taiwan, p. 1159–1163 ;

- **Duy Dinh, Lynda Tamine** (2010). *Sense-based biomedical indexing and retrieval*. Dans : the 15th International Conference on Applications of Natural Language to Information Systems, **NLDB 2010**, Juin 23-25, 2010, Cardiff University, Cardiff, Wales, UK, p. 24–35.

Articles publiés dans des conférences nationales avec comités de lecture

- **Duy Dinh, Lynda Tamine** (2011). *Recherche d'information dans les documents biomédicaux : approche basée sur le sens précis des concepts*. Dans : Symposium sur l'Ingénierie de l'Information Médicale, SIIM 2011, Juin 9–10, 2011, Toulouse, France ;

- **Duy Dinh, Lynda Tamine** (2010). *Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients*. Dans : Conférence francophone en Recherche d'Information et Applications, CORIA 2010, Mars 18-21, 2010, Sousse, Tunisie, p. 325–336 ;

- **Duy Dinh, Lynda Tamine** (2010). *Recherche d'information dans les documents biomédicaux : approche basée sur le sens précis des concepts*. Dans : INFormatique des Organisations et Systèmes d'Information et de Décision, INFORSID 2010, Mai 25-28, 2010, Marseille, France, p. 261–274 ;

Participation à des campagnes d'évaluation internationales

- **Duy Dinh, Lynda Tamine**. *IRIT at TREC 2011 : Medical Record Retrieval Task*. Dans : Working notes for **TRECMed 2011**, Novembre 15–18, 2011, Gaithersburg, Md. USA.

- **Duy Dinh, Lynda Tamine**. IRIT at ImageCLEF 2011 : medical retrieval track. Dans : Working notes for **ImageCLEF 2011**, Septembre 19-22, 2011, Amsterdam, the Nertherlands.

- **Duy Dinh, Lynda Tamine**. *IRIT at ImageCLEF 2010 : Medical Retrieval Task*. Dans : Working notes for **ImageCLEF 2010**, Septembre 22–24, 2010, Padua, Italy.

4 Organisation de la thèse

Cette thèse est organisée en deux principales parties. La première partie, se composant de deux chapitres d'état-de-l'art, présente le contexte de nos travaux de recherche dans le domaine de la RI biomédicale : *Indexation et Recherche d'Information: de la RI classique à la RI sémantique*(chapitre II, page 14), *Indexation et Recherche d'Information Biomédicale*(chapitre III, page 59). La deuxième partie, subdivisée en quatre chapitres : *Résolution de l'ambiguïté des termes MeSH orientée domaine et son impact sur un processus de RI*(chapitre IV, page 124), *Extraction de concepts biomédicaux : approche basée sur la pertinence et la corrélation des contextes documentaires et terminologiques*(chapitre V, page 149), *Indexation multi-terminologique pour la RI biomédicale*(chapitre VI, page 186) et *BioSIR - système prototype de RI biomédicale*(chapitre VII, page 225), présente nos contributions dans le cadre de la RI biomédicale, à savoir nos approches pour l'indexation et la recherche d'information biomédicale ainsi que les expérimentations que nous avons menées. Le détail de cette organisation est donné comme suit :

- Le chapitre II, ***Indexation et Recherche d'Information: de la RI classique à la RI sémantique***, présente les principes et concepts de base de la RI classique (section 2) ainsi que les fondements de base de la RI conceptuelle/sémantique (section 3). Nous présentons en particulier dans la section 2.1 les notions et concepts de base en RI de manière générale. Ensuite, nous décrivons le processus général de la RI dans la section 2.2. Nous passons par la suite en revue les modèles de RI les plus représentatifs de l'état-de-l'art dans la section 2.3. Les techniques de reformulation de requêtes sont présentées dans la section 2.4. Puis, nous abordons le protocole d'évaluation ainsi que les mesures d'évaluation des performances de la RI dans la section 2.5. Nous passons à la présentation des fondements de la RI conceptuelle/sémantique (section 3), y compris les ressources sémantiques (section 3.1), les techniques d'indexation sémantique (section 3.2), les techniques d'appariement sémantique (section 3.3).

- Le chapitre III, ***Indexation et Recherche d'Information Biomédicale***, donne un aperçu sur la RI conceptuelle/sémantique dans le domaine biomédical. Nous commençons d'abord par une présentation de la typologie de l'information biomédicale (section 2). Ensuite, nous présentons les principales ressources termino-ontologiques les plus utilisées et disponibles dans le domaine biomédical (section 3). Nous présentons par la suite les approches d'identification des termes techniques qui désignent les concepts biomédicaux (section 4). Nous donnons quelques liens vers les outils d'extraction des concepts qui sont accessibles (téléchargeables ou exécutables via un service Web) (section 4.2). Nous catégorisons les

différentes approches d'indexation conceptuelle en RI biomédicale qui intègrent les méthodes d'extraction des concepts (section 5). Puis, les techniques d'appariement document-requête sémantique en RI biomédicale sont décrites de manière synthétique (section 6). Enfin, la section 7 donne quelques éléments sur l'évaluation des performances des systèmes de RI biomédicale.

– Le chapitre IV, ***Résolution de l'ambiguïté des termes MeSH orientée domaine et son impact sur un processus de RI***, présente notre première contribution sur l'indexation conceptuelle basée sur le sens précis des termes qui désignent les concepts dans le document ainsi que de la requête. La section 2 présente les problématiques de la RI conceptuelle et les objectifs principaux de notre contribution dans ce cadre. La section 3 décrit notre méthode de désambiguïsation et le processus d'indexation des documents biomédicaux basée sur le sens des termes désignant les concepts biomédicaux. Nous supposons que le sens d'un terme MeSH est défini par la détermination de son propre domaine dans MeSH car chaque domaine correspond à un sujet particulier de la médecine. Notre approche d'indexation et de recherche d'information conceptuelle exploite donc la structure poly-hiérarchique du thésaurus MeSH (Medical Subject Headings) pour désambiguïser les termes ambigus dans les documents et des requêtes. Dans MeSH, un terme est dit ambigu s'il est défini par plusieurs domaines dans la structure poly-hiérarchique. Une évaluation expérimentale est présentée et discutée dans la section 4. La section 5 conclut ce chapitre et annonce des perspectives.

– Le chapitre V, ***Extraction de concepts biomédicaux : approche basée sur la pertinence et la corrélation des contextes documentaires et terminologiques***, présente notre deuxième contribution dans le cadre de la RI conceptuelle concernant particulièrement l'évaluation des performances de la RI obtenues par différentes méthodes d'extraction de concepts à partir du texte (documents, requêtes). Nous présentons tout d'abord les problématiques et les motivations de notre recherche dans la section 2. Ensuite, nous détaillons notre méthode d'extraction de concepts pour la RI conceptuelle/sémantique dans la section 3. Les résultats expérimentaux obtenus sont présentés et discutés dans la section 4 en utilisant les collections biomédicales dans le cadre de la campagne d'évaluation TREC Genomics 2004-2005. Afin de montrer l'efficacité de notre approche de RI conceptuelle/sémantique basée sur notre méthode d'extraction de concepts termino-ontologiques, nous comparons nos résultats expérimentaux aux meilleurs résultats obtenus par les particpants dans TREC Genomics.

– Le chapitre VI, ***Indexation multi-terminologique pour la***

RI biomédicale, présente notre troisième contribution concernant l'exploitation des ressources termino-ontologiques pour améliorer les représentations des documents ainsi que de la requête via une combinaison contextuelle de l'information, c-à-d la combinaison de l'expansion conceptuelle de documents basée sur les ressources externes (dans un contexte global) et l'expansion basée sur le contexte local de la requête. Nous commençons par présenter les problématiques et les motivations concernant l'indexation de l'information biomédicale basée sur les ressources termino-ontologiques biomédicales dans la section 2. Nous décrivons dans la section 3 l'architecture générale de notre approche de RI conceptuelle basée sur les terminologies. La section 4 présente les modèles de vote dédiés à la fusion des concepts extraits à partir de plusieurs terminologies pour chaque document biomédical. La section 5 décrit notre approche de RI basée sur les terminologies biomédicales pour améliorer les performances de la RI biomédicale. Nous évaluons notre approche de RI conceptuelle multi-terminologique dans la section 6 en utilisant deux sous ensembles de résumés d'articles de journaux de MEDLINE, à savoir un sous ensemble de 48.753 documents issus de TREC Genomics 2004 pour l'apprentissage et 41.018 documents issus de TREC Genomics 2005 pour le test.

– Le chapitre VII, ***BioSIR - système prototype de RI biomédicale*** présente notre système prototype, intitulé BioSIR, dédié à la recherche d'information conceptuel dans le domaine biomédical. BioSIR intègre nos solutions adéquates pour la recherche d'information conceptuelle/sémantique développées dans le cadre de cette thèse, à savoir l'expansion conceptuelle de documents, l'extraction de concepts basée sur une mono-terminologie ou plusieurs terminologies, l'expansion de requêtes basée sur les mesures statistiques dans un contexte local de la requête, l'expansion conceptuelle de la requête en se basant sur les ressources termino-ontologiques ... Le système intègre plusieurs moteurs de recherche de l'état-de-l'art comme Terrier, Lucene et Lemur et l'outil d'évaluation trec_eval permettant d'évaluer les résultats de recherche de plusieurs modèles d'ordonnancement, d'extraire les mesures de performances comme MAP, P@5, P@10, etc. Pour une évaluation plus fine, BioSIR permet également d'évaluer les performances requête par requête entre les différents méthodes de RI.

En conclusion, nous dressons le bilan des travaux réalisés dans le cadre de la thèse, synthétisons des éléments originaux de nos contributions. Nous présentons ensuite les différentes pistes d'évolution de ces travaux.

Partie I

De la Recherche d'Information classique à la Recherche d'Information Conceptuelle/Sémantique

Chapitre II

Indexation et Recherche d'Information: de la RI classique à la RI sémantique

Sommaire

1 Introduction **14**

2 Principes et concepts de base de la RI **15**
- 2.1 Concepts de base 15
- 2.2 Processus général de la RI 17
  - 2.2.1 Indexation des documents 18
  - 2.2.2 Appariement requête-document 21
- 2.3 Aperçu des modèles de RI 22
  - 2.3.1 Modèle(s) booléen(s) 22
  - 2.3.2 Modèle(s) vectoriel(s) 25
  - 2.3.3 Modèle(s) probabiliste(s) 27
- 2.4 Reformulation de la requête 32
  - 2.4.1 Reformulation par réinjection de la pertinence 32
  - 2.4.2 Reformulation par pseudo-réinjection de la pertinence 34
- 2.5 Évaluation des performances de la RI 35
  - 2.5.1 Protocole d'évaluation de TREC 36
  - 2.5.2 Mesures d'évaluation 37

3 Fondements de la RI sémantique **40**
- 3.1 Ressources termino-ontologiques 42
  - 3.1.1 Notions de base 42
  - 3.1.2 WordNet 43
  - 3.1.3 Open Directory Project - ODP 44
  - 3.1.4 Yet Another Great Ontology - YAGO 46
- 3.2 Principe de la RI sémantique 47
  - 3.2.1 Les ressources exploitées pour l'indexation sémantique 48
  - 3.2.2 Désambiguïsation pour l'indexation sémantique 49
- 3.3 Aperçu général de travaux de désambiguïsation et d'indexation sémantique en RI 51

3.3.1 Désambiguïsation manuelle pour la RI sémantique . . 51
3.3.2 Désambiguïsation automatique pour la RI sémantique 53

4 Conclusion 56

"The eventual demarcation of philosophy from science was made possible by the notion that philosophy's core was "theory of knowledge", a theory distinct from the sciences because it was their foundation... Without this idea of a "theory of knowledge", it is hard to imagine what "philosophy" could have been in the age of modern science."

–Richard Rorty, Philosophy and the Mirror of Nature

1 Introduction

La recherche d'information (RI) est un domaine historiquement lié aux sciences de l'information qui ont pour objectif d'établir les représentations des documents ainsi que des requêtes de l'utilisateur dans le but d'en récupérer des informations (texte, son, images, ou données multimédia ...), à travers la construction d'index. La recherche d'information, qui est une branche de l'informatique, concerne essentiellement *l'acquisition, l'organisation, le stockage* et *la recherche de l'information*. L'association des professionnels de l'information et de la documentation (ADBS[1]) distingue la *recherche d'information* de la *recherche de l'information* comme suit :

- "la recherche d'information est l'ensemble des méthodes, procédures et techniques permettant de sélectionner l'information dans un ou plusieurs fonds de documents plus ou moins structurés."
- "la recherche de l'information est l'ensemble des méthodes, procédures et techniques ayant pour objet l'extraction des informations pertinentes à partir d'un document ou d'un ensemble de documents."

Un système de RI est un ensemble de logiciels assurant l'ensemble des fonctions nécessaires à la recherche de l'information. Il offre des techniques et des outils permettant de localiser et de visualiser l'information pertinente relativement à un besoin en information, exprimé par un utilisateur sous forme de requête. La RI est aujourd'hui un champ interdisciplinaire qui est devenu inséparable des questions et enjeux politiques, culturels, sociaux ...

Ce chapitre a pour but de présenter les principes et concepts de base de la RI classique (section 2) et les fondements de base de la RI conceptuelle (section 3). Plus spécifiquement, la section 2.1 est dédiée à la description des notions de base en RI. Ensuite, la section 2.2 décrit le processus général de la RI. Un aperçu sur les modèles de RI est présenté dans la section 2.3. Le protocole d'évaluation ainsi que les mesures d'évaluation sont présentés

1. `http://www.adbs.fr/`

dans la section 2.5. Nous passons à la présentation des fondements de la RI conceptuelle/sémantique dans la section 3. Plus spécifiquement, nous présentons les ressources sémantiques qui sont créées pour la RI conceptuelle/sémantique dans la section 3.1. Nous rapportons les techniques d'indexation conceptuelle/sémantique dans la section 3.2 et les techniques d'appariement conceptuel/sémantique dans la section 3.3.

2 Principes et concepts de base de la RI

2.1 Concepts de base

Nous présentons dans ce qui suit les notions fondamentales en RI qui sont utilisées tout au long de cette thèse.

- **Un document** est une entité d'information correspondant à un contenu singulier (selon le dictionnaire *Larousse*, 2011). Techniquement, on peut le définir comme un ensemble de données informatives présentes sur un support, sous une forme permanente et lisible. De ce fait, un document peut être un texte, un fragment de texte, une audio, ou une bande de vidéo, etc. Concernant les bases de données textuelles, il existe deux sources principales de documents : *référothèques* et *bibliothèques*.

  - **Une référothèque** contient un ensemble de références aux documents dans lesquels se trouve de l'information intégrale. Par exemple, les résumés d'articles ou les citations dans la base MEDLINE [2] constituent une base de données bibliographiques regroupant la littérature relative aux sciences biologiques et biomédicales.

  - **Une bibliothèque** contient le texte intégral des documents sources y compris les tableaux, les figures, les images, etc. En pratique, pour diminuer le volume de stockage, les bases de données ne stockent que le texte en gardant ou ignorant la structure des documents originaux.

  Le terme "collection de documents" (ou encore base documentaire, fond documentaire, corpus) constitue l'ensemble des informations exploitables et accessibles au travers du système de RI.

- **Une requête** : est une expression correspondant au besoin d'information de l'utilisateur. Il s'agit d'une description sommaire des documents ciblés par la recherche. Pour une recherche documentaire *ad hoc*, l'utilisateur

2. Medical Literature Analysis and Retrieval System Online

formule sa requête, souvent en langage naturel, en spécifiant les mots-clés ou une expression particulière.

– **La pertinence** en RI indique dans quelle mesure les documents retournés par le système de RI répondent au besoin d'information de l'utilisateur. Cette notion représente un critère majeur de l'évaluation des performances du système de RI (Rijsbergen, 1979; Salton, 1989). La pertinence, qui est l'objet principal de tout système de RI, constitue une notion fondamentale en RI. Elle peut être définie comme la correspondance entre un document et une requête ou encore une mesure d'informativité du document à la requête.

Dans ce qui suit, nous détaillerons quelles sont les entités d'indexation ou les unités descriptives que nous pouvons trouver dans un texte (document ou requête). Dans un premier temps nous préciserons, selon notre point de vue, quelles sont les entités qui composent un texte.

Tokens *vs.* Mots

De manière générale, un texte (document) est constitué de plusieurs paragraphes. Chaque paragraphe est un ensemble d'une ou de plusieurs phrases. Chaque phrase est formée d'une succession de *tokens* ou *mots* qui apparaissent dans un ordre défini. En informatique, un **token** (ou jeton) est utilisé pour désigner un identificateur ou identifiant permettant de représenter une donnée (symbole, abréviation, constante, variable, etc.). En général, un token est une chaîne de caractères alphanumériques reconnaissable par l'ordinateur ou par un programme informatique. En linguistique, un **mot** est une unité lexicale constituant des termes représentant les idées ou sujets du document. Un token devient un mot s'il a une signification précise traduisant le sens du mot. Par exemple, les tokens comme 'ordinateur', 'information', 'médecine' sont des mots avec une signification précise tandis que 'p53', 'fancd2' ne sont que des abréviations dont le sens est déterminé dans un contexte particulier.

N-grammes

Ce sont des groupes de n éléments (caractères, mots, ...) permettant une description statistique de la langue par apprentissage des probabilités d'apparition de chaque groupe dans un corpus de la langue étudiée (Manning *et al.*, 2008). En pratique, les unités descriptives uni-grammes (n=1), bi-grammes (n = 2) et tri-grammes (n = 3) sont souvent utilisées. Cette modélisation correspond en fait à un modèle de Markov d'ordre n où seules les n dernières observations sont utilisées pour la prédiction de l'élément suivant. Ainsi un

tri-gramme est un modèle de Markov d'ordre 3. Par exemple, les tri-grammes au niveau de caractères de l'expression "recherche d'information" sont : "rec", "her", "che", "d'i", "for", "mat", "ion".

2.2 Processus général de la RI

L'objectif d'un SRI est de fournir un ensemble de documents qui sont utiles pour aider son utilisateur à trouver de l'information pertinente en répondant à sa requête. Pour cela, le SRI met en œuvre un processus pour réaliser la mise en correspondance des informations contenues dans un fond documentaire d'une part, et des besoins d'information des utilisateurs d'autre part. Ces processus correspondent essentiellement à (1) l'*indexation des documents* dans la collection et (2) l'*interrogation* de l'information qui vise à apparier la représentation des documents et celle de la requête. La figure II.1 représente les étapes d'un processus général d'un SRI. Les parallélogrammes représentent les documents ou la requête, les rectangles représentent les étapes ou traitements du SRI et les cylindres représentent les ressources terminologiques (thésaurus, ontologies, dictionnaires, etc.) ou statistiques (sur la distribution des termes dans un ensemble ou sous-ensemble de documents observés). Nous détaillons ces étapes ci-dessous.

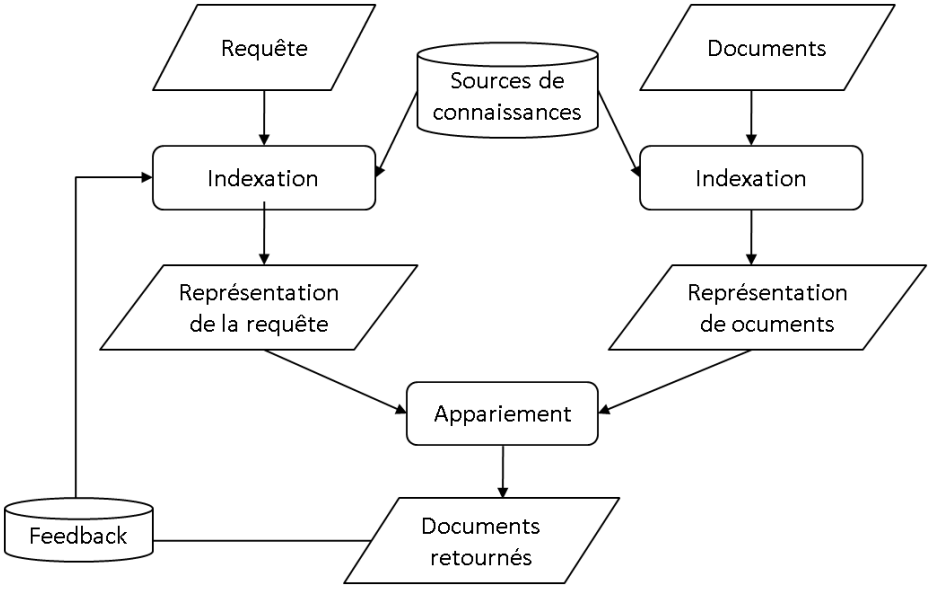


FIGURE II.1 – Processus en général d'un SRI

2.2.1 Indexation des documents

L'indexation des documents consiste à extraire, organiser, structurer, indexer, sauvegarder le contenu sémantique des documents dans des structures de données appelées *index*. L'indexation permet de retrouver rapidement les documents contenant un ou plusieurs termes donnés et éventuellement ses positions, ses fréquences d'occurrence, ... dans un document. L'indexation de la requête permet d'en extraire des mots-clés ou termes descriptifs exprimant le besoin en information de l'utilisateur. La différence entre l'indexation des documents et celle de la requête porte sur le fait que la première génère une structure d'index plus sophistiquée (lexique, index de documents, index inversé ...) tandis que la seconde génère une liste de termes normalisés qui sont utilisés pour l'interrogation de l'information dans des fonds documentaires.

On distingue deux types d'indexation : **indexation libre** *vs.* **indexation contrôlée**

1. **Indexation libre** : l'indexeur extrait les mots-clés d'un document ou les choisit librement sans l'aide de sources de connaissance. L'indexation peut être améliorée en filtrant les mots fonctionnels pour éliminer les non-descripteurs du document. L'avantage de cette indexation est que le traitement est plus facile et rapide en amont, mais le résultat présente de très grandes difficultés à l'interrogation : l'atomisation du vocabulaire (l'ensemble de descripteurs le constituant n'est pas connu *a priori*) ou l'ambiguïté de termes choisis par rapport aux sujets du document, ...

2. **Indexation contrôlée** : le vocabulaire ou le langage d'indexation est déjà défini *a priori* et son utilisation exclusive s'impose à l'indexeur. Ce vocabulaire peut se présenter sous deux formes : (1) liste alphabétique des descripteurs simples et (2) liste structurée de descripteurs reliés entre eux par des relations de hiérarchie, d'association ou d'équivalence. Les descripteurs choisis dans l'index ne sont pas nécessairement observables directement dans les documents à indexer. Dans le cas où il s'agit d'un domaine spécifique, le vocabulaire est appelé terminologie du domaine ou aussi terminologie de référence.

Notons que ces deux types d'indexation peuvent se réaliser manuellement ou automatiquement. L'indexation manuelle est une opération réalisée par un spécialiste ou un documentaliste qui consiste à recenser les sujets ou concepts dont traite un document et à les représenter à l'aide d'un vocabulaire contrôlé prédéfini. L'indexation automatique se base sur des méthodes statistiques et probabilistes permettant de repérer des éléments significatifs dans chaque document. Bien que les vocabulaires contrôlés soient souvent utilisés pour l'indexation manuelle, plusieurs travaux de recherche ont été abordés dans la lit-

térature pour une indexation contrôlée automatique en utilisant des ressources terminologiques (*cf.* chapitre III).

Les principaux traitements lors de l'indexation sont : tokenisation, racinisation/lemmatisation/troncature de mots-clés ou jetons, calcul des fréquences d'apparition des mots-clés, sauvegarde des positions de chaque mot-clé, compression du contenu de chaque document...

Tokenisation/Segmentation

La tokenisation consiste à segmenter un texte en plusieurs unités atomiques, appelées jetons ou tokens. La liste des tokens identifiés devient l'entrée pour des traitements ultérieurs comme l'analyse syntaxique ou des tâches plus ciblées comme la recherche d'information ou la fouille de texte (text mining). Les tokens peuvent être identifiés par des heuristiques simples ou par des expressions régulières comme suit :

- Tous les caractères alphabétiques minuscules [a-z] ou majuscules [A-Z] suivis d'un ou plusieurs chiffres [0-9] font partie d'un token. Par exemple, les chaînes de caractères 'A7', 'B1', 'alpha05' sont des tokens valides.
- Les tokens sont séparés l'un de l'autre par des espaces blancs [\t\r\n\v\f] ou par des ponctuations [!"#$%&'()*+,. :;<=>?@\^‘| -]. Par exemple, le symbol '@' est inclus dans une adresse de messageries.
- Les ponctuations et espaces blancs peuvent figurer ou ne pas figurer dans un token. Par exemple, l'espace blanc dans "New York" doit être considéré pour un nom de ville.

Notons que les tokens peuvent varier en genre ('chanteur' *vs.* 'chanteuse' ou 'agriculteur' *vs.* 'agricultrice'), en nombre ('maladie' au singulier *vs.* 'maladie**s**' au pluriel) ou selon le style de dactylographie ('USA' *vs.* 'U.S.A'). Pour ramener les variations lexicales à une forme canonique, la normalisation des tokens est nécessaire. Il s'agit d'un processus de transformation des tokens en une forme canonique afin que les appariements entre deux instances de textes (e.g., requête *vs.* document) se produisent correctement malgré les différences superficielles entre les tokens dans chaque instance (requête ou document). Il existe deux types de normalisation : racinisation *vs.* lemmatisation.

Racinisation *vs.* lemmatisation

La racinisation (désuffixation, ou stemming en anglais) est un procédé qui vise à transformer les flexions en leur radical ou *stemme*. La lemmatisation est

TABLEAU II.1 – Racinisation *vs.* lemmatisation

	Mot	**Racinisation**	**Lemmatisation**
Cas 1	maladie	malad	maladie
	maladies	malad	maladie
Cas 2	agriculteur	agriculteur	agriculteur
	agricultrice	agricultric	agriculteur
Cas 3	traiter	trait	traiter
	traitement	trait	traitement

une analyse lexicale du contenu d'un texte regroupant les mots d'une même famille et les transformant en leur forme canonique appelée *lemme*. La différence principale entre la racinisation et la lemmatisation porte sur le fait que la première focalise sur la normalisation d'un mot sans considération de son contexte. Un lemme est un mot ou un sigle qui est défini et compréhensible dans un langage spécifique tandis qu'un stemme est la plus petite unité lexicale qui peut générer un ou plusieurs mots de la même famille. Le tableau II.1 illustre trois exemples concrets de la normalisation où le nombre (cas 1), le genre (cas 2) et la catégorie grammaticale (cas 3) des tokens sont considérés. On peut observer que dans le premier et troisième cas, la racinisation effectue une désuffixation des mots et les transforme en une forme canonique quels que soient leur nombre et leur catégorie grammaticale. Si le token satisfait une expression régulière alors celui-ci est normalisé par une désuffixation correspondante. Dans le deuxième cas, le mot 'agricultrice' est transformé en 'agricultric' sans considération du genre féminin de ce mot. Dans ce cas, la lemmatisation permet de retrouver sa forme au masculin singulier qui est 'agriculteur'.

Traitements supplémentaires

Lors du processus d'indexation, plusieurs traitements peuvent être exécutés en chaîne ou en *pipeline* : décodage, tokenisation, suppression de mots vides (fonctionnels), racinisation et/ou lemmatisation, désaccentuation (pour les langues accentuées), normalisation par la synonymie, ... Chaque traitement peut être éventuellement ajouté ou enlevé du processus d'indexation afin d'optimiser les performances de l'indexation et de recherche d'information.

A part des principaux traitements présentés dans les sections précédentes, les traitements supplémentaires sont inclus lors du processus de l'indexation en fonction des caractéristiques de la collection. Le décodage permet de convertir les différents formats de fichiers (html, xml, pdf, doc, rtf, ...) en format texte géré par le SRI. La plupart des moteurs de recherche commerciaux s'occupent du décodage des documents originaux en stockant la version texte du document indexé dernièrement (*cf.* figure II.2). La suppression des mots vides

permet d'exclure les mots non descriptifs dans le contenu sémantique du document. Notons que cette opération est optionnelle car cela peut entraîner une dégradation des performances de la RI surtout pour des requêtes dont le sujet ou le thème principal concerne certaines notions qui sont exprimées par des mots figurant dans la liste des mots vides. Par exemple, le mot 'son' peut indiquer un adjectif possessif ou une sensation auditive engendrée par une onde acoustique.

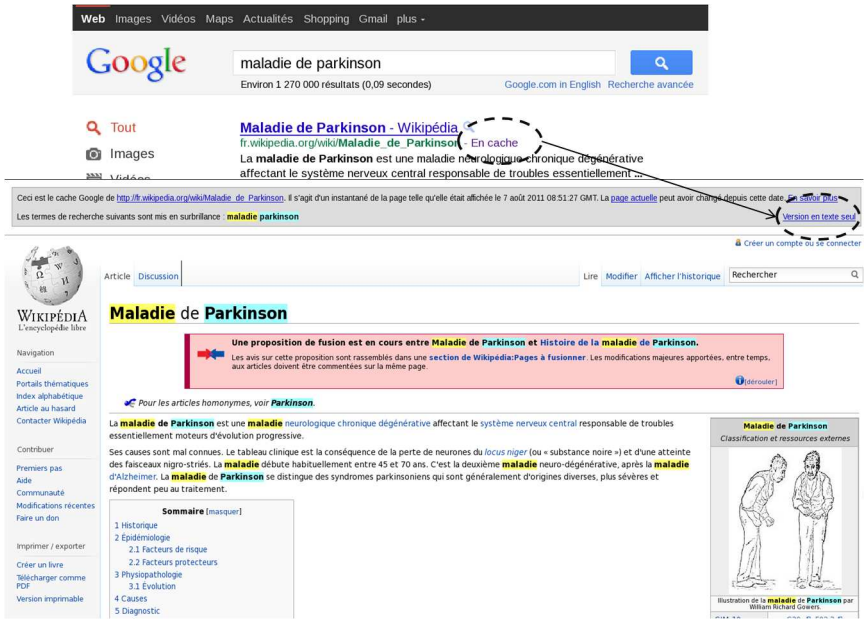


FIGURE II.2 – Version texte en cache d'un document original

2.2.2 Appariement requête-document

L'appariement requête-document consiste à calculer un poids de pertinence ou la similarité entre chaque document et la requête de l'utilisateur, notée $RSV(Q, D)$ (Retrieval Status Value), où Q représente la requête et D le document considéré. Ce poids est calculé grâce à un modèle d'ordonnancement en RI (*cf.* la section 2.3). Il permet donc d'ordonner les résultats renvoyés par le système de RI en fonction du poids calculé pour chaque document vis-à-vis de la requête. La plupart des systèmes de RI ordonnent les résultats dans un ordre décroissant de la valeur RSV qui traduit le degré de pertinence des résultats. Les documents les plus pertinents doivent être renvoyés en premier. Un système de RI idéal ne renvoie que les documents pertinents vis-à-vis de la requête, tous les documents non-pertinents ne sont pas retournés ou doivent être retournés après les documents pertinents.

2.3 Aperçu des modèles de RI

Depuis la naissance des premiers systèmes de RI dans les années 1960s, plusieurs modèles de recherche d'information ou modèles d'appariement document-requête ont été proposés dans la littérature. Sommairement, ils peuvent être classés en trois catégories principales : *modèles booléens*, *modèles vectoriels* et *modèles probabilistes*.

- les **modèles booléens** sont inspirés de la logique booléenne et de la théorie des ensembles pour modéliser l'appariement document-requête. Trois variations principales y sont distinguées : le modèle booléen classique, le modèle booléen étendu et le modèle booléen flou,

- les **modèles vectoriels** modélisent les documents et les requêtes comme des vecteurs de termes dans un espace multi-dimensionnel. Ils englobent le modèle vectoriel généralisé, le modèle LSI (Latent Semantic Indexing) et le modèle connexionniste,

- les **modèles probabilistes** ont été introduits pour modéliser la notion de pertinence. Ils englobent le modèle probabiliste général, le modèle de réseau inférentiel (Document Network), et le modèle de langue.

2.3.1 Modèle(s) booléen(s)

Les modèles booléens sont les premiers modèles utilisés en RI (Salton, 1970). Un des premiers modèles est le **modèle booléen standard** qui est basé sur la théorie des ensembles et l'algèbre de Boole. Dans ce modèle, chaque document est considéré comme un ensemble de mots ou termes séparés et chaque requête est exprimée par une phrase booléenne qui relie des termes par les trois connecteurs logiques "AND", "OR" et "NOT". La similarité entre la requête q et le document d, souvent appelée RSV (Retrieval Status Value), est donnée par la fonction booléenne suivante :

$$RSV(d,q) = \begin{cases} 1 & \text{si } d \text{ appartient à l'ensemble décrit par } q \\ 0 & \text{sinon} \end{cases} \qquad \text{(II.1)}$$

La figure II.3 illustre un exemple d'une requête booléenne en utilisant les connecteurs logiques pour indiquer que les documents retournés doivent contenir un ou plusieurs mots spécifiques ou pas. L'utilisateur dispose d'une interface graphique pour saisir des mots-clés et le SRI se charge de formuler la requête booléenne. Par défaut, les mots-clés se sont reliés par l'opérateur "AND", c-à-d les documents retournés doivent contenir tous les mots indiqués.

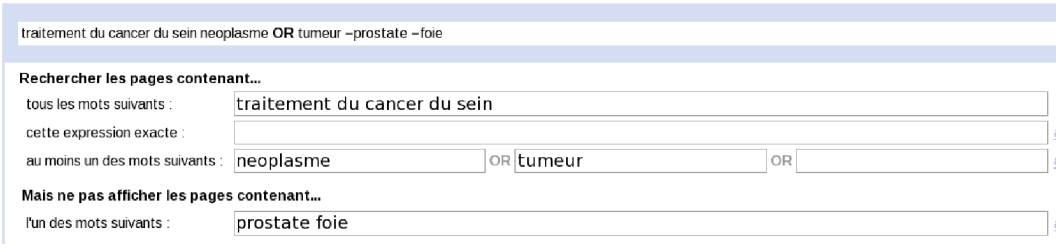


FIGURE II.3 – Exemple de requête booléenne dans Google

Le modèle booléen standard présente les principaux avantages de simplicité de mise en œuvre, d'expressivité et de clarté de l'expression de la requête grâce à des opérateurs logiques. Cependant, il possède les inconvénients suivants :

- la formulation de la requête est plutôt compliquée, non accessible à un large public,
- l'ordre des résultats n'est pas pris en compte car les premiers documents retrouvés sont présentés en premier,
- les opérateurs logiques (AND, OR, NOT) ne permettent pas de modéliser les notions 'imprécision', 'incertitude' de l'information. En plus, l'expression logique devient compliquée lorsque la requête est longue.

Afin de pallier les inconvénients du modèle booléen standard, une extension de ce modèle a été introduite par (Salton *et al.*, 1983). Le **modèle booléen étendu** tient compte du poids des termes dans les documents en se basant sur des statistiques dans la collection comme la fréquence d'un terme dans le document et la fréquence inverse de documents (IDF, Inverse Document Frequency). Cela permet d'ordonner les documents résultats selon leurs similarités à la requête. Dans ce modèle, chaque document est représenté par un vecteur de termes dans un espace multi-dimensionnel où chaque dimension correspond à un terme spécifique associé aux documents. Soit n le nombre de termes de la collection, le document d_j peut être représenté comme suit :

$$d_j = [w_{1j}, w_{2j}, ..., w_{nj}] \tag{II.2}$$

où w_{ij} est le poids du terme t_i dans le document d_j, calculé par une normalisation de sa fréquence dans le document :

$$w_{ij} = tf_{ij} * \frac{IDF(t_i)}{maxIDF(t_i)} \tag{II.3}$$

où tf_{ij} est la fréquence du terme t_i dans d_j, $IDF(t_i)$ est la fréquence inverse

de documents du termes t_i calculée par :

$$IDF(t_i) = \frac{N_c}{|\{d : t_i \in d\}|} \tag{II.4}$$

où N_c est le nombre total de documents dans la collection et $|\{d : t_i \in d\}|$ est le nombre total de documents contenant le terme t_i.

Pour augmenter la flexibilité des systèmes de RI basés sur le modèle booléen standard, le **modèle booléen flou** a été proposé dans l'objectif de modéliser des notions 'imprécision', 'incertitude' de l'information (Paice, 1984; Dubois et Prade, 1988; Bosc et Prade, 1996; Bordogna et Pasi, 2000) que le modèle booléen standard ne permet pas. Le modèle booléen flou qui est basé sur la théorie des ensembles flous ou la logique floue est essentiellement une extension du modèle booléen standard. Cette théorie permet de caractériser un élément par un *degré d'appartenance* à un ensemble flou et de représenter un document spécifique comme un ensemble flou de termes pondérés. Chaque terme (élément) dans un document donné (ensemble flou) est associé à un degré d'appartenance pour représenter l'imprécision ou l'incertitude de l'information. Les connecteurs logiques notamment l'intersection, l'union et le complément sont redéfinis en fonction des degrés d'appartenance ("membership function") entre les ensembles flous comme suit :

$$\begin{array}{lcl} \mu(a \textbf{ AND } b) & = & \textbf{min}(\mu(a), \mu(b)) \\ \mu(a \textbf{ OR } b) & = & \textbf{max}(\mu(a), \mu(b)) \\ \mu(\textbf{NOT } b) & = & 1 - \mu(b) \end{array} \tag{II.5}$$

où $x = \{a, b\}$ est l'ensemble de documents contenant le terme x ; $\mu(x)$ est le degré d'appartenance du terme x au document observé et à déterminer.

Les principaux avantages du modèle booléen flou portent sur le fait qu'il permet de réduire l'imperfection qui caractérise le processus de RI (Kraft *et al.*, 2007) et de contrôler l'imprécision de l'utilisateur lors de la formulation de sa requête (Bordogna et Pasi, 2001). Les limitations de ce modèle se posent au niveau de la pondération des termes en utilisant des fonctions d'appartenance floues :

- premièrement, la difficulté réside dans la détermination manuelle des valeurs d'appartenance des termes à un document donné qui exige toujours un coût lié au temps d'annotation et des expertises requises ;
- deuxièmement, les opérateurs booléens flous ne peuvent pas vraiment assurer un appariement approximatif. En effet, prenons comme exemple la requête suivante :

  Q = "Hexagone OR France".

TABLEAU II.2 – Degrés d'appartenance de termes aux documents

	$\mu(Hexagone)$	$\mu(France)$
d_1	0.8	0.8
d_2	0.9	0.0

On suppose que le degré d'appartenance de chaque mot-clé aux documents d_1 et d_2 ($d_1 \neq d_2$) soit donné par le tableau II.2. Le sujet principal du document d_1 concerne le pays France tandis que celui dans d_2 concerne un hexagone en géométrie. En appliquant les fonctions floues correspondantes pour réordonner ces deux documents, il est évident que d_2 est classé en premier car son poids est plus élevé que celui du document d_1 qui est plus pertinent pour la requête q.

$$\begin{aligned} RSV(d_1, Q) &= max\{\mu(hexagone), \mu(france)\} = max\{0.8, 0.8\} = 0.8 \\ RSV(d_2, Q) &= max\{\mu(hexagone), \mu(france)\} = max\{0.9, 0.0\} = 0.9 \end{aligned} \quad \text{(II.6)}$$

2.3.2 Modèle(s) vectoriel(s)

Les modèles vectoriels constituent une classe de modèles de RI dont plusieurs variantes ont été implémentées dans de nombreux systèmes de RI, notamment le système SMART (Salton *et al.*, 1983) qui est un des premiers travaux expérimentaux en RI. En général, dans les modèles vectoriels, chaque document d ainsi que chaque requête q de l'utilisateur sont représentés par des vecteurs dans un espace multidimensionnel dont chaque dimension correspond à un terme unique comme suit :

$$\begin{aligned} d &= < w_{d1}, w_{d2}, ..., w_{dN} > \\ q &= < w_{q1}, w_{q2}, ..., w_{qN} > \end{aligned} \quad \text{(II.7)}$$

où w_{di} (resp. w_{qi}) représente le poids du terme t_i dans le document d (resp. la requête q) ; N est la taille du vocabulaire ou le nombre total de termes d'index.

Le poids du terme dans le document est calculé en utilisant un schéma de pondération spécifique qui peut être éventuellement utilisé pour calculer le poids du terme de la requête. Le schéma de pondération de termes détermine une variante du modèle vectoriel. La fonction de pondération de termes la plus simple a été introduite par (Salton *et al.*, 1983). Cette fonction permet de modéliser le poids d'un terme en fonction de sa fréquence dans le document et

sa fréquence inverse de documents (IDF, Inverse Document Frequency) :

$$poids(t) = tf \times \left(log_2 \frac{N - n_t}{n_t} \right) \tag{II.8}$$

où N est le nombre total de documents de la collection ; n_t est le nombre de documents contenant le terme t.

Les modèles vectoriels permettent de mesurer la similarité entre le document d et la requête q en fonction du degré de corrélation vectorielle entre eux. Ce mécanisme assure un appariement partiel ou optimal (partial match, best match) entre les documents et la requête. Le score de pertinence du document d vis-à-vis de la requête q peut être calculée grâce à une des mesures suivantes :

- Produit scalaire : $RSV(d,q) = \sum_{i=1}^{N} w_{di}.w_{qi}$
- Mesure Cosinus : $RSV(d,q) = \sum_{i=1}^{N} \frac{\sum_{i=1}^{N} w_{di}.w_{qi}}{\sqrt{\sum_{i=1}^{N} w_{di}^2}.\sqrt{\sum_{i=1}^{N} w_{qi}^2}}$
- Mesure de Dice : $RSV(d,q) = \frac{2.\sum_{i=1}^{N} w_{di}.w_{qi}}{\sum_{i=1}^{N} w_{di}^2 + \sum_{i=1}^{N} w_{qi}^2}$
- Mesure de Jaccard : $RSV(d,q) = \frac{\sum_{i=1}^{N} w_{di}.w_{qi}}{\sum_{i=1}^{N} w_{di}^2 + \sum_{i=1}^{N} w_{qi}^2 - \sum_{i=1}^{N} w_{di}.w_{qi}}$
- Coefficient de superposition : $RSV(d,q) = \frac{\sum_{i=1}^{N} w_{di}.w_{qi}}{min(\sum_{i=1}^{N} w_{di}^2, \sum_{i=1}^{N} w_{qi}^2)}$

Les modèles vectoriels présentent plusieurs avantages par rapport aux modèles booléens classiques :

- la formulation de la requête est en langage naturel et plus proche de l'utilisateur,
- les termes dans le même document sont pondérés selon différents facteurs, notamment en fonction de leur fréquence d'occurrence et de leur fréquence inverse de documents, ce qui permet de mettre en évidence les sujets les plus importants du document,
- l'utilisation des mesures de corrélation vectorielle assure une correspondance partielle des résultats de la recherche par rapport à la requête,
- l'appariement document-requête permet de trier les résultats selon leur niveau de pertinence par rapport à la requête.

Les inconvénients majeurs des modèles vectoriels concernent l'hypothèse d'indépendance entre les termes dans le même document et dans la même requête. Ainsi, dans un texte, l'ordre des mots n'est pas pris en compte. Il ne prend pas non plus en compte les synonymes ou la sémantique des contenus.

2.3.3 Modèle(s) probabiliste(s)

Une autre catégorie de modèles de RI regroupe les modèles probabilistes où la théorie de la probabilité ainsi que des mesures statistiques ont été largement appliquées (Maron et Kuhns, 1960; Robertson, 1977; Salton *et al.*, 1983; Baeza-Yates et Ribeiro-Neto, 1999; Amati et van Rijsbergen, 2002; Croft *et al.*, 2009). L'idée de base des modèles probabilistes est de sélectionner les documents ayant à la fois une forte probabilité d'être pertinents et une faible probabilité d'être non pertinents à la requête de l'utilisateur. De manière générale, dans un modèle probabiliste, les documents sont sélectionnés selon le principe PRP (*Probability Ranking Principle*) (Robertson, 1977) qui représente la justification théorique des modèles probabilistes : pour optimiser les performances de la recherche, les documents doivent être sélectionnés et triés selon un ordre décroissant de la probabilité de pertinence à la requête. De ce fait, les documents pertinents doivent être classés avant ceux qui ne le sont pas. Le calcul de la probabilité que le document d soit pertinent à la requête q, notée $p(rel, d|q)$, peut être simplifié en adoptant l'hypothèse d'indépendance entre les termes d'index. D'après le théorème de Bayes, cette probabilité est donnée par :

$$p(rel, d|q) = \prod_{t \in q} p(rel, d|t) \tag{II.9}$$

où $p(rel, d|t)$ est la probabilité que le document d soit pertinent en observant le terme 't'. Formellement :

$$p(rel, d|t) = \frac{p(t, d|rel).p(rel)}{p(t|\overline{rel})} \tag{II.10}$$

Étant donné que $p(rel)$ est une constante et que l'événement "t,d" dans la probabilité $p(t, d|rel)$ signifie $t \in d$, la probabilité de pertinence du document d vis-à-vis de la requête est finalement proportionnelle à la distribution des termes de la requête dans les documents pertinents et inversement proportionnelle à leur distribution dans les documents non pertinents :

$$p(rel, d|q) \propto \prod_{t \in q \cap d} \frac{p(t|rel)}{p(t|\overline{rel})} \tag{II.11}$$

Dans ce qui suit, nous présentons deux variantes des modèles probabilistes de l'état de l'art : (1) *modèle BM25* et (2) *modèle de langue*.

Modèle de pondération BM25

Les modèles probabilistes sont basés sur la probabilité de pertinence associée à chaque document qui est susceptible d'être une bonne réponse à une requête spécifique. L'ordonnancement des documents en relation avec la requête de l'utilisateur est réalisé dans l'ordre décroissant de la probabilité de pertinence. Parmi plusieurs modèles probabilistes, le schéma de pondération de termes et la fonction d'appariement document-requête du modèle BM25 [3] est un des meilleurs modèles performants en RI (Robertson et Walker, 1994; Robertson *et al.*, 1998; Robertson et Zaragoza, 2009). Ce modèle prend en compte deux facteurs importants qui sont la fréquence d'occurrence des termes dans le document et la longueur du document par rapport à la longueur moyenne des documents dans la collection. Formellement, la similarité entre le document d et la requête q, notée $RSV(q,d)$, est donnée par la somme des produits du poids de chaque terme dans le document, noté w_{d_t}, par son poids associé à la requête, noté w_{q_t} :

$$RSV(d,q) = \sum_{t \in q \cap d} \underbrace{\frac{(k_1+1)\,.\,tf}{K+tf}}_{w_{d_t}} \,.\, \underbrace{\frac{(k_3+1)\,.\,qtf}{k_3+qtf}\,.\,w^{(1)}}_{w_{q_t}} \tag{II.12}$$

- k_1 et k_3 sont des paramètres qui déterminent l'importance de la fréquence du terme dans le document et dans la requête ;
- K est un facteur de normalisation de la longueur du document :

$$K = k_1\,.\,\left((1-b) + b\,.\,\frac{dl}{avg\_dl}\right) \tag{II.13}$$

  où le paramètre $b \in [0,1]$ permet de déterminer l'effet de la normalisation de la longueur dl du document par rapport à la longueur moyenne de documents avg_dl dans la collection ;
- tf et qtf sont respectivement la fréquence du terme dans le document et celle de la requête ;
- $w^{(1)}$ est la fréquence inverse de documents (IDF) :

$$w^{(1)} = log_2 \frac{N - N_t + 0.5}{N_t + 0.5} \tag{II.14}$$

  où N_t est le nombre de documents contenant le terme t et N le nombre total de documents dans la collection.

3. BM signifie 'Best Match'

Modèle(s) de langue

Un des premiers modèles de langue en RI a été proposé par (Ponte et Croft, 1998). En général, le mot "modèle" est utilisé en RI selon deux sens : le premier concerne une abstraction de la tâche de recherche. Par exemple, le modèle booléen en RI indique une tâche de recherche où les opérateurs booléens sont utilisés. Le deuxième sens fait référence à un modèle de données qui décrit de façon abstraite comment sont représentées les données. Le modèle de langue ici est défini comme une distribution probabiliste qui décrit adéquatement les caractéristiques statistiques d'un langage ou d'un sous-langage.

L'idée de base des modèles de langue est d'estimer un *modèle de langue*, noté M_d, pour chaque document d qui est susceptible de générer la requête q de l'utilisateur. Autrement dit, ces modèles visent à déterminer la probabilité que la requête q soit inférée ou générée par le modèle de langue M_d du document d. Cette probabilité est souvent calculée par différents types de distribution probabiliste sur laquelle se basent les modèles de langue.

Soit la requête $q = q_1 q_2 ... q_n$ et le document $d = d_1 d_2 ... d_m$ où n est le nombre de termes de la requête et m le nombre de termes du document d. Selon le principe du modèle de langue, la degré d'appariement entre le document d et la requête q, qui correspond à la probabilité d'observer d sachant q, est proportionnelle à la probabilité que q soit générée par le modèle de langue M_d :

$$RSV(q,d) = p(d|q) \propto p(q|M_d) \ . \ p(M_d) \tag{II.15}$$

La plupart des modèles de langue supposent que la requête q est une séquence de termes indépendants et que la croyance initiale (probabilité a priori) $p(M_d)$ est uniforme. Un exemple d'utilisation de la croyance initiale est de prendre en compte la longueur du document, ses auteurs, ou des liens entre documents (Miller *et al.*, 1999; Kraaij *et al.*, 2002). Si l'on adopte ces deux hypothèses, l'équation II.15 peut être simplifiée comme suit :

$$\begin{array}{rcl} RSV(q,d) & \propto & p(q_1 q_2 ... q_n|M_d) \\ & \propto & \prod_{i=1}^{n} p(q_i|M_d) \end{array} \tag{II.16}$$

L'appariement document-requête revient donc à estimer la probabilité que chaque terme q_i de la requête soit généré par le modèle de langue M_d, notée $p(q_i|M_d)$.

Les modèles de langue se différencient par le type de distribution probabiliste où on observe la présence ou l'absence des termes dans un document

et/ou dans la requête de l'utilisateur. Trois types de distribution probabilistes sont utilisés dans les modèles de langue : (1) *distribution de Bernoulli multiple*, (2) *distribution multinomiale*, et (3) *distribution de Poisson*. Dans ce qui suit, nous décrivons brièvement le modèle de langue basé sur la distribution Bernoulli multiple.

Modèle de langue Bernoulli multiple

Le modèle de langue Bernoulli multiple en RI se base sur une distribution Bernoulli multiple. Soit X une variable aléatoire, une distribution Bernoulli de X est une distribution discrète de probabilité qui est définie par la loi de probabilité suivante :

$$P(X=x) = \begin{cases} p & \text{si } x=1 \\ 1-p & \text{si } x=0 \\ 0 & \text{sinon} \end{cases} \tag{II.17}$$

où x est la valeur que X peut prendre et p est la probabilité que X prenne la valeur 1.

D'un point de vue statistique, la présence ou l'absence d'un terme t dans la requête peut être définie par une variable aléatoire binaire $X_i = x_i$ où $x_i = 1$ (ou $x_i = 0$) si l'on observe la présence (ou l'absence) du terme t dans la requête. En considérant tous les termes d'index (présents ou absents dans la requête), la similarité entre d et q est calculée par le produit de la probabilité que les termes présents dans q soient générés par le modèle de langue M_d du document d, notée $p_s(q|M_d)$, par la probabilité que les autres termes du vocabulaire d'index qui sont absents dans q ne soient pas générés par M_d, notée $p_u(\overline{q}|M_d)$. Formellement :

$$RSV(q,d) = \underbrace{\prod_{t\in q} p(t|M_d)}_{p_s(q|M_d)} \times \underbrace{\prod_{t\notin q} (1.0 - p(t|M_d))}_{p_u(\overline{q}|M_d)} \tag{II.18}$$

La plupart des modèles de langues Bernoulli multiple essaient de modéliser ces deux probabilités. La première peut être calculée par une estimation du maximum de vraisemblance (maximum likelihood estimation) qui tient compte de la fréquence d'occurrence des termes et de la longueur du document :

$$\widehat{p_{ml}}(t|M_d) = \frac{tf}{dl_d} \tag{II.19}$$

où tf est la fréquence relative du terme t dans le document d et dl_d est le nombre de termes ou tokens au total du document d.

Le problème se pose lorsqu'il existe au moins un terme de la requête qui n'est pas observé dans le document ($t \in q \cap \overline{d}$). En effet, celui-ci va recevoir une probabilité nulle car $tf = 0$. Afin de pallier ce problème, le lissage (smoothing) consiste à ajuster l'estimation du maximum de vraisemblance d'un modèle de langue en sorte que celui-ci soit plus efficace. (Ponte et Croft, 1998) ont proposé de calculer la probabilité qu'un terme t de la requête qui apparaît dans le document ($t \in q \cap d \Rightarrow tf > 0$) soit généré par M_d en se basant sur une estimation du maximum de vraisemblance qui est décrite dans l'équation II.19. Formellement :

$$\widehat{p_s}(t|M_d) = p_{ml}(t,d)^{1.0-\widehat{R}_{t,d}} \times p_{avg}(t)^{\widehat{R}_{t,d}} \tag{II.20}$$

où

- $p_{avg}(t)$ est la probabilité moyenne d'observer dans la requête le terme t qui figure dans les documents le contenant :

$$p_{avg}(t) = \frac{(\sum_{d \in E_t} p_{ml}(t|M_d))}{|E_t|} \tag{II.21}$$

  où E_t est l'ensemble de documents contenant le terme t et $|E_t|$ est le nombre de documents le contenant (document frequency).
- $\widehat{R}_{t,d}$ est un risque d'utiliser $p_{avg}(t)$ pour estimer $\widehat{p}(t|M_d)$. Ce risque est modélisé par une distribution géométrique :

$$\widehat{R}_{t,d} = \left(\frac{1}{1+\overline{f}_{t,d}}\right) \times \left(\frac{\overline{f}_{t,d}}{1+\overline{f}_{t,d}}\right)^{tf} \tag{II.22}$$

  où $\overline{f}_{t,d}$ est la fréquence moyenne du terme t dans les documents le contenant : $\overline{f}_{t,d} = p_{avg}(t) \times dl_d$

La deuxième probabilité concerne la distribution d'un terme t de la requête qui est absent du document d ne soit pas généré par M_d est ajustée par sa fréquence d'occurrence dans la collection :

$$\widehat{p_u}(t|M_d) = \frac{cf_t}{n_v} \tag{II.23}$$

où cf_t est la fréquence du terme t dans la collection et n_v est la taille du vocabulaire (nombre total de tokens dans la collection).

2.4 Reformulation de la requête

Pour retrouver de l'information pertinente en réponse à un besoin d'information particulier, l'utilisateur doit formuler sa requête et la soumettre au système de RI. La plupart des utilisateurs ont des difficultés pour formuler leur requête en sorte que les premiers documents retournés par le système de RI soient pertinents car le langage (les mots-clés) de l'utilisateur peut être différent du langage dans la collection. Cette difficulté suggère que la première requête devrait être considérée comme le premier essai de l'utilisateur pour récupérer les documents pertinents et que la requête devrait être ré-écrite ou reformulée afin de récupérer plus de documents pertinents en premier.

La reformulation de la requête est donc un processus ayant pour objectif de générer une nouvelle requête plus ciblée que celle initialement formulée par l'utilisateur. Ce processus consiste généralement à modifier la requête initiale de l'utilisateur via *la réinjection de la pertinence* (*relevance feedback*) ou l'expansion de la requête (*query expansion*). Pour être simple, nous appelons les techniques de modification (expansion, suppression des mots ou termes) de la requête initiale la reformulation de la requête. Nous distinguons deux types de reformulation de la requête : (1) *reformulation par réinjection de la pertinence* (*relevance feedback*) *vs.* (2) *reformulation par pseudo-réinjection de la pertinence* (*pseudo-relevance feedback* ou *blind feedback*).

2.4.1 Reformulation par réinjection de la pertinence

La reformulation par réinjection de la pertinence peut être effectuée de deux manières différentes : *explicite* ou *implicite*. Si l'utilisateur donne de manière explicite les informations sur la pertinence des documents retournés par le système de RI, on parlera de la réinjection explicite de la pertinence (*explicit relevance feedback*). En revanche, si l'utilisateur donne des informations impliquées dans la reformulation de la requête de manière implicite (par exemple les clics de souris sur les documents), on parlera de la reformulation par réinjection implicite de la pertinence (*implicit relevance feedback*).

Réinjection de la pertinence explicite

L'utilisateur clique sur les premiers document retournés par le système de RI en indiquant de manière explicite quels sont les documents pertinents ou non pertinents vis-à-vis de sa requête (Rocchio, 1971). L'idée de base de cette approche est que les termes ou expressions dans les documents qui ont été jugés pertinents sont utiles pour la reformulation de la requête. De même, les termes

ou expressions dans les documents non-pertinents peuvent être utilisés pour reformuler la requête mais avec les poids négatifs (Belkin *et al.*, 1995; Sumner *et al.*, 1998). Deux stratégies peuvent être subdivisées : (1) *l'expansion de la requête* où les nouveaux termes ajoutés à la requête sont issus des documents pertinents et (2) *la re-pondération des termes* où leur poids seront modifié en se basant sur les jugements pertinents de l'utilisateur (Baeza-Yates et Ribeiro-Neto, 1999). En général, les deux stratégies peuvent être combinées ensemble.

La reformulation de la requête par réinjection de la pertinence explicite a été introduite pour la première fois par (Rocchio, 1971). Les principales étapes de cette méthode sont comme suit :

1. les documents sont d'abord retournés par le système de RI en utilisant un modèle de pondération particulier, e.g., le modèle TF-IDF, lors de la première recherche pour la requête initiale Q_0.
2. l'utilisateur sélectionne un sous-ensemble de documents retournés, parmi les quels on peut avoir les documents pertinents ou non pertinents, dénotés respectivement R et S.
3. le système de RI génère la nouvelle requête Q_1, comme une combinaison linéaire de Q_0, R et S. Le poids du terme t dans la nouvelle requête, noté qtw_m, est donné par :

$$qtw_m = \alpha_1 qtf + \alpha_2 \frac{1}{n_1} \sum_{i=1}^{n_1} w_R(t) - \alpha_3 \sum_{i=1}^{n_2} w_S(t) \qquad \text{(II.24)}$$

où
- qtf est la fréquence (normalisée) du terme t dans la requête originale,
- $w_R(t)$ est le poids normalisé du terme t dans l'ensemble R,
- $w_S(t)$ est le poids normalisé du terme t dans l'ensemble S,
- α_1, α_2 et α_3 sont les cœfficients de la combinaison.

Réinjection de la pertinence implicite

Dans ce cas, le système de RI qui collecte les informations et interprète les comportements de l'utilisateur avec le système comme les sources d'évidence pour la reformulation de la requête sans demander d'autres actions supplémentaires ni d'efforts venant de l'utilisateur (Jung *et al.*, 2007). Contrairement à la réinjection de la pertinence explicite qui demande à l'utilisateur de juger la pertinence de chaque document, et puis ce dernier doit consulter les différentes parties du document (e.g., titre, résumé, URL, ...) et éventuellement lire tout le document, l'approche implicite est basée sur l'hypothèse que l'utilisateur donne ses jugements de pertinence des documents de manière "silencieuse" durant sa

recherche d'information. Les sources d'évidence implicites qui peuvent être exploitées sont : les cliques de souris sur les documents (Smyth *et al.*, 2005), le défilement sur une page Web (Claypool *et al.*, 2001), la sauvegarde des documents dans le marque-page (Oard et Kim, 1998), et la durée de consultation des documents (Kelly et Belkin, 2001; White *et al.*, 2002)...

2.4.2 Reformulation par pseudo-réinjection de la pertinence

La reformulation par pseudo-réinjection de la pertinence (*Blind Feedback* ou encore *Pseudo Relevance Feedback*, notée PRF) utilise des techniques de réinjection automatique à l'aveugle pour construire la nouvelle requête. Ainsi, l'idée de base de la méthode PRF est basée sur l'hypothèse que les k premiers documents sont pertinents (documents pseudo-pertinents) pour améliorer les performances de la RI (Croft et Harper, 1988; Robertson, 1991; Kwok, 1996; Xu et Croft, 1996). Le processus de la PRF consiste généralement à ajuster ou modifier le poids des termes (re-pondération) et ajouter les termes les plus pertinents qui sont extraits à partir des premiers documents retournés (considérés comme le contexte local de la requête).

L'application de la reformulation par PRF comme par exemple l'expansion de la requête dans une tâche de recherche *ad-hoc* a montré une amélioration des performances de la RI (Xu et Croft, 1996; Robertson *et al.*, 1998; Amati, 2003; Clinchant et Gaussier, 2010). La méthode PRF n'est efficace que si les requêtes initiales permettent de retrouver des documents pertinents. Dans le cas contraire, elle peut engendrer une dégradation des performances de la RI.

Dans le cadre de cette thèse, nous nous intéressons particulièrement à la méthode de reformulation PRF implémentée dans la plateforme DFR [4] (Amati, 2003). Les modèles de reformulation de requêtes PRF utilisent les statistiques de Bose-Einstein et Kullback-Leibler pour extraire les meilleurs termes qui sont susceptibles d'être pertinents pour générer la nouvelle requête Q^e obtenue par l'expansion de la requête originale Q. Formellement, le poids du terme candidat t de la nouvelle requête est donné par :

$$poids(t \in Q^e) = tfq_n + \beta * \frac{\text{Info}_{\text{DFR}}}{\text{MaxInfo}} \tag{II.25}$$

où

- $tfq_n = \frac{tfq}{\max_{t \in q} tfq}$: la fréquence normalisée du terme de la requête originale Q,
- $\text{MaxInfo} = \arg\max_{t \in q^e} \max \text{Info}_{\text{DFR}}$,
- Info_{DFR} est la fréquence normalisée du terme de la nouvelle requête induite par un modèle DFR. Formellement :

$$Info_{DFR} = -\log_2 Prob(f(t|M)|f(t|C)) \tag{II.26}$$

4. Divergence From Randomness

où Prob est la probabilité d'obtenir la fréquence $f(t|M)$ du terme observé t au sein de top m premiers documents retournés, appelés M, étant donnée sa fréquence $f(t|C)$ dans la collection C.

Dans les modèles DFR, cette probabilité est calculée en utilisant deux mesures statistiques appelées *Kullback-Leibler* (KL) et *Bose-Einstein* (Bo) Amati (2003).

Modèle Kullback-Leibler

La première mesure KL donne la normalisation de la fréquence du terme dans le document comme suit :

$$\text{Info}_{\text{KL}} = \frac{f(t|M)}{\sum f(M)} * \log_2 \frac{f(t|M)*\sum f(C)}{f(t|C)*\sum f(M)} \tag{II.27}$$

où la fonction $\sum f(X)$ correspond à la fréquence totale des tous les termes figurant dans l'ensemble de documents X.

Modèle Bose-Einstein

La seconde mesure Bo donne deux variantes de la fréquence normalisée du terme t dans le document comme suit :

$$\begin{aligned} \text{Info}_{\text{Bo1}} &= -\log_2 \text{Prob}(\text{f(t|M)}|\text{f(t|C)}) \\ &= \log_2(1+\tfrac{f(t|C)}{N}) + \text{f(t|M)} * \log_2(1+\tfrac{N}{f(t|C)}) \end{aligned} \tag{II.28}$$

et

$$\begin{aligned} \text{Info}_{\text{Bo2}} &= -\log_2 \text{Prob}(\text{f(t|M)}|\text{f(t|C)}) \\ &= \log_2(1+\tfrac{\sum f(M)*f(t|C)}{\sum f(C)}) + \text{f(t|M)} * \log_2(1+\tfrac{\sum f(C)}{\sum f(M)*f(t|C)}) \end{aligned} \tag{II.29}$$

2.5 Évaluation des performances de la RI

Les expérimentations en RI doivent aboutir à satisfaire les besoins d'informations de l'utilisateur. Cela signifie qu'un système de RI doit permettre aux utilisateurs de retrouver non seulement un nombre maximum de "meilleurs documents" vis-à-vis de la requête mais aussi les documents les plus pertinents parmi les premiers documents retournés.

L'évaluation des performances d'un SRI consiste à définir les critères, les indicateurs ou mesures permettant de quantifier la performance d'un SRI et de

comparer entre les modèles de RI ou entres les SRI. Plusieurs critères peuvent entrer en jeu dans le processus d'évaluation. (Cleverdon, 1970) a proposé six principaux critères d'évaluation de la performance d'un SRI : (1) la couverture du discours de la collection, (2) le temps de réponse, (3) la présentation des résultats, (4) l'effort de l'utilisateur pour récupérer de l'information pertinente, (5) la précision et (6) le rappel du SRI. Parmi ces facteurs, les deux derniers sont liés aux modèles de représentation de l'information du SRI.

Les premiers travaux d'évaluation des SRI ont été initiés par le projet Cranfield 2 (Cleverdon, 1967). Ces travaux ont fourni des ressources de base pour l'évaluation des performances des SRI, notamment le développement de la *collection test* qui contient trois éléments principaux : (1) une collection de documents, (2) un ensemble de requêtes (25 au minimum), et (3) un ensemble de documents jugés pertinents pour chaque requête. Ces travaux inspirent jusqu'à aujourd'hui les expérimentations d'évaluation de type laboratoire des SRI. Dans ce qui suit, nous présentons quelques mesures d'évaluation les plus utilisées dans les campagnes d'évaluation de RI.

2.5.1 Protocole d'évaluation de TREC

Le protocole d'évaluation consiste à définir comment les SRI sont évalués. Le protocole le plus utilisé pour évaluer les performances de la RI est celui de TREC [5] qui est un des fournisseurs principaux des collections *test* pour évaluer les performances des SRI. Née des expérimentations du projet Cranfield dans les années de 1957 à 1967 (Cleverdon, 1967), et initiée au tout début des années 90 comme une partie du programme TIPSTER, la campagne d'évaluation TREC est un projet international co-sponsorisé désormais par le NIST [6] et le DARPA [7]. Son objectif était de supporter la recherche dans le domaine de RI en fournissant une infrastructure nécessaire pour l'évaluation à grande échelle des méthodes de RI dans les bases documentaires. En particulier, la série d'ateliers TREC a pour objectif d'offrir un cadre d'évaluation uniforme pour mesurer les performances et comparer les résultats des SRI.

Les pistes explorées par TREC sont entre autres, la recherche d'information (RI) *ad-hoc*, le filtrage d'informations, la question-réponse, ... La RI *ad-hoc* concerne les documents dans des collections de vidéos (Smeaton *et al.*, 2006), blogs (Ounis *et al.*, 2006), et en particulier la RI biomédicale dans le cadre de TREC Genomics (Hersh et Voorhees, 2009)... La recherche *ad-hoc* est la tâche la plus importante visant à évaluer les résultats soumis par différents groupes de recherche dans le monde entier. Cette tâche concerne un ensemble

5. `http://trec.nist.gov/`
6. National Institute of Standards and Technology
7. Defense Advanced Research Projects Agency

de requêtes qui sont évaluées sur une collection de documents. Chaque groupe télécharge l'ensemble de documents et un ensemble de requêtes. Celui-ci indexe la collection et utilise un modèle d'ordonnancement particulier pour récupérer les résultats qui sont soumis à TREC pour l'évaluation.

Le protocole d'évaluation de TREC dans le cadre d'une recherche *ad-hoc* est décrit comme suit : chaque participant dispose d'une **collection de test** qui est essentiellement constituée de trois éléments suivants :

- un ensemble de documents (souvent très volumineux, environ des centaines milliers de documents ou plus),
- un ensemble de requêtes lié aux besoins d'information des utilisateurs,
- un ensemble de documents jugés pertinents vis-à-vis de chaque requête.

La méthodologie d'évaluation des performances des résultats renvoyés par les différentes méthodes de RI est essentiellement basée sur le modèle d'évaluation de Cranfield (Voorhees, 2002). Cette méthode d'évaluation se base essentiellement sur l'hypothèse que **les jugements de pertinence des documents vis-à-vis de chaque requête soient complets**, c-à-d que tous les documents de la collection sont jugés pertinents ou non-pertinents pour chaque requête à évaluer (Cleverdon, 1991). Selon le protocole d'évaluation défini dans TREC, une collection test contenant des documents originaux est à disposition de chaque groupe participant. Ce dernier indexe les documents dans la collection selon une ou plusieurs approches d'indexation particulières. Par la suite, le SRI retourne 1,000 premiers documents pour chaque requête. Les résultats finaux pour l'ensemble de requêtes sont soumis à TREC pour être évalués de manière officielle. Cependant, le jugement de la pertinence de tous les documents du corpus vis-à-vis d'une requête est quasiment infaisable pour une collection volumineuse. Pour cette raison, seulement les premiers documents (environ 100 documents pour chaque groupe) de chaque liste de résultats sont retenus pour le jugement des documents vis-à-vis d'une requête. Cette méthode est souvent appelée "pooling method" dans la littérature (Buckley et Voorhees, 2004).

2.5.2 Mesures d'évaluation

Considérons une collection de documents représentés par des cercles dans un rectangle (*cf.* la figure II.4). Un cercle blanc représente un document non-pertinent et un cercle noir représente un document pertinent vis-à-vis d'une requête. De manière virtuelle, pour n'importe quelle requête de l'utilisateur, les documents de la collection (tous les cercles dans le rectangle) peuvent être pertinents ou non-pertinents (séparés virtuellement par une ligne droite). Lors de la recherche, le SRI sélectionne les documents dans la collection en utilisant un modèle de RI spécifique, et les renvoie à l'utilisateur. Les documents (perti-

nents ou non-pertinents) retournés par le SRI sont représentés par des cercles noirs et blancs dans l'ovale. Pour mesurer la capacité d'un SRI à retrouver les documents pertinents et à rejeter ceux qui ne le sont pas, les mesures de *précision*, *rappel* ainsi que leur dérivées sont définies en utilisant les notations dans le tableau II.3.

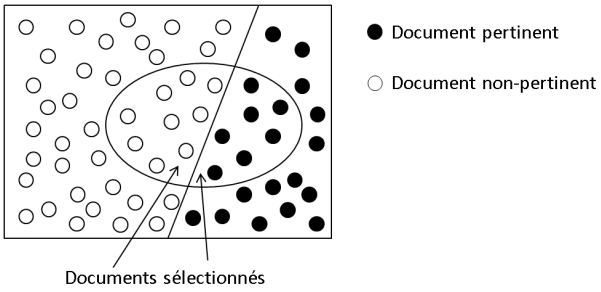


FIGURE II.4 – Documents pertinents *vs.* non-pertinents vis-à-vis d'une requête

TABLEAU II.3 – Notations des mesures d'évaluation

Notation	**Description**
q	La requête q
E^Q	L'ensemble de requêtes Q
$N_P(q)$	Nombre de documents pertinents vis-à-vis de q
$N_{PS}(q)$	Nombre de documents pertinents sélectionnés pour q
$N_S(q)$	Nombre de documents sélectionnés par le SRI pour q

Précision

La **précision** des résultats vis-à-vis d'**une requête** q, notée $Pr(q)$, indique la capacité d'un SRI à retourner les documents pertinents par rapport aux résultats renvoyés par le système. Cette mesure est équivalente à la proportion des documents pertinents sélectionnés (cercles noirs dans l'ovale) relativement à l'ensemble des documents sélectionnés (tous les cercles dans l'ovale). Formellement :

$$Pr(q) = \frac{N_{PS}(q)}{N_S(q)} \tag{II.30}$$

Précision à *k* documents

La qualité des résultats retournés est évaluée non seulement par le nombre de documents pertinents retournés mais aussi par le nombre de **premiers documents pertinents** retournés. La précision des résultats vis-à-vis de la requête q à k documents, notée $Pr(q, k)$, permet de mesurer la capacité d'un SRI

à retourner les documents pertinents parmi les k-premiers documents retournés en répondant à la requête q. Formellement :

$$Pr(q,k) = \frac{N_{PS}(q)}{k} \tag{II.31}$$

En général, k peut prendre la valeur dans l'ensemble $\{1, 2, 3, 4, 5, 10, 15, 20, 30, 50, 100, 200, 500, 1000\}$. Par exemple, si k est égal à 10, on a la précision à 10 premiers documents, notée $P@10$, qui correspond à la proportion de documents pertinents retournés parmi les dix premiers documents.

Rappel

Le rappel des résultats de la requête q, noté $R(q)$, indique la capacité d'un SRI à restituer un ensemble des documents pertinents répondant à la requête q. Cette mesure est équivalente à la proportion des documents pertinents sélectionnés (cercles noirs dans l'ovale) relativement à l'ensemble de documents pertinents de la collection (tous les cercles noirs dans le rectangle) vis-à-vis de la requête. Formellement :

$$R(q) = \frac{N_{PS}(q)}{N_P(q)} \tag{II.32}$$

Précision moyenne

Pour avoir un aperçu sur les performances d'un modèle de RI ou d'un SRI, l'évaluation des résultats doit être effectuée sur un ensemble de requêtes. En réalité, le jugement de la pertinence de tous les documents d'une collection volumineuse (e.g., des centaines milliers de documents) par rapport à une requête est quasi impossible ou infaisable. Pour cette raison, seule une partie de la collection de documents dédiée à une évaluation partielle est créée selon un protocole d'évaluation particulier. La précision moyenne d'un SRI ou la précision des résultats vis-à-vis d'un ensemble de requêtes peut être calculée de deux manières : (1) *interpolation* et (2) *non-interpolation*.

- **Précision moyenne interpolée**. Du fait que le taux de rappel r est varié selon chaque requête et le nombre de **documents pertinents retournés** correspondant, on retient dans la littérature 11 points de rappel dans l'intervalle [0..1] par pas de 0.1. Ainsi, les taux de précision doivent donc être interpolés en fonction d'un taux de rappel spécifique.
  La valeur interpolée de la précision des résultats vis-à-vis de la requête q à un niveau de rappel r_k, notée $Pr_{interp}(q, r_k)$, est la **précision maximale**

pour un niveau de rappel entre r_k et r_{k+1} :

$$Pr_{interp}(q, r_k) = max_{r_k \leq r \leq r_{k+1}} Pr(q, r) \tag{II.33}$$

La précision moyenne interpolée des résultats de l'ensemble de requêtes à un niveau de rappel r_k est donc définie par :

$$\overline{Pr}_{interp}(r_k) = \sum_{r \in [0..1]} \frac{Pr_{interp}(q, r)}{|E^Q|} \tag{II.34}$$

où $|E^Q|$ est le nombre total de requêtes traitées par le SRI.

- **Précision moyenne non-interpolée.** La précision moyenne non-interpolée d'un modèle de RI ou d'un SRI, souvent appelée MAP (Mean Average Precision) est la valeur moyenne des taux de précision moyenne non-interpolée obtenus pour chaque requête :

$$MAP = \frac{\sum_{q \in E^Q} \overline{Pr}(q)}{|E^Q|} \tag{II.35}$$

où $\overline{Pr}(q)$ est la précision moyenne des résultats vis-à-vis de la requête q calculée sur tous les k premiers documents retournés. Formellement :

$$\overline{Pr}(q, k) = \sum_{k=1}^{N_S(q)} \frac{Pr(q, k)}{N_{PS}(q)} \tag{II.36}$$

La précision moyenne à k premiers documents retournés, obtenus sur l'ensemble de requêtes, notée $\overline{Pr}(k)$ ou $P@k$ en général, est la valeur moyenne des taux de précision à k premiers documents de chaque requête :

$$\overline{Pr}(k) = \frac{\sum_{q \in E^Q} Pr(q, k)}{|E^Q|} \tag{II.37}$$

D'autres mesures d'évaluation proposées dans la littérature peuvent être retrouvées dans les ouvrages de RI (Baeza-Yates et Ribeiro-Neto, 1999; Boughanem et Savoy, 2008; Croft *et al.*, 2009).

3 Fondements de la RI sémantique

Depuis les années 90 du dernier siècle, la conception et le développement des ressources termino-ontologiques (e.g., ontologies, thésaurus, bases de données lexicales ...) sont devenus un des champs de recherche les plus populaires en Informatique, investi par les différentes communautés dont celle de l'intelligence artificielle (IA) et de la RI. En effet, les travaux sur les ontologies

ou les ressources sémantiques sont de plus en plus répandus dans les différentes communautés en Informatique comme : le Web, la bio-informatique ou les systèmes d'information géographiques. Par conséquent, les ressources termino-ontologiques sont de plus en plus disponibles, notamment dans le domaine biomédical. Une des raisons pour laquelle ces ressources sont devenues si importantes est due actuellement au besoin de définir les termes techniques utilisés en Ingénierie des connaissances pour standardiser la communication. On attend qu'elles jouent ce rôle de standardisation via la définition des concepts dans un domaine général comme WordNet (Miller, 1995), ou dans un domaine spécifique comme MeSH (Medical Subject Headings) ou UMLS (Unified Medical Language System).

De manière générale, les ressources termino-ontologiques peuvent être créées manuellement ou automatiquement à partir de textes. Pour avoir un aperçu sur la conception et le développement des ressources termino-ontologiques à partir de textes, nous faisons référence aux travaux de la communauté de l'Ingénierie des connaissances (Aussenac-Gilles *et al.*, 2000). L'objectif pour notre part, n'est pas de créer des ontologies à partir de textes, mais de proposer, dans un premier temps, un moyen pour les intégrer dans un processus de RI, appelé ***RI sémantique***. Ensuite, nous allons évaluer l'intérêt de leur utilisation en étudiant leur impact sur les performances des systèmes de RI. La RI sémantique est essentiellement basée sur une indexation sémantique des contenus d'information et l'exploitation de ces index, appelés *index sémantiques*, dans un modèle d'appariement sémantique qui intègre les informations sémantiques dans la requête et/ou dans les documents. L'indexation sémantique peut donc être définie comme un processus d'indexation classique qui tient compte des termes descripteurs qui sont les représentatifs des concepts (y compris leur sens, leurs variantes lexicales, leurs synonymes, etc.) dans les ressources termino-ontologiques.

Nous nous intéressons principalement au principe d'indexation sémantique qui est particulièrement novateur comparativement à une indexation classique. Nous citons d'abord quelques ressources termino-ontologiques qui sont les plus importantes (e.g., WordNet, ODP, YAGO ...) et puis nous présentons le principe général de la RI sémantique qui est essentiellement basée sur des méthodes de désambiguïsation. Enfin, nous allons présenter quelques applications de la désambiguïsation afin de l'intégrer dans un processus de RI sémantique.

3.1 Ressources termino-ontologiques

3.1.1 Notions de base

Termes

Un **terme** désigne un mot ou groupe de mots qui ne s'applique qu'à un et un seul objet ou une idée générale, et ce, dans un domaine donné. Autrement dit, il s'agit d'un mot qui désigne de façon univoque un objet ou un concept dans un domaine, qui, généralement est un domaine de spécialité. De ce fait, dans l'idéal terminologique, un terme n'a pas d'équivalent (ou synonyme) dans le domaine désigné. Un terme formé d'un seul mot (e.g., "hypertension", "traumatisme", etc.) est dit terme simple ou uniterme, alors que celui constitué de plusieurs mots est appelé terme complexe ou multi-terme ("cancer du poumon", "système nerveux central", etc.). Bien que le mot "terme" soit couramment employé dans les deux cas, on parle également d'unité terminologique.

Concepts

Un **concept** représente une idée générale et abstraite que se fait l'esprit humain d'un objet de pensée concret ou abstrait, et qui lui permet de rattacher à ce même objet les diverses perceptions qu'il en a, et d'en organiser les connaissances (Larousse 2012). Un concept, qui apparaît dans un contexte particulier, est exprimé par un terme simple ou complexe. Un terme peut être *préféré* ou *non-préféré* qui désigne un concept particulier selon **la définition par les linguistes**. Par exemple, le terme "Neoplasms" (en anglais) est préféré tandis que les termes comme "Cancer", "Tumors", "Benign Neoplasms" sont les termes non-préférés désignant le concept "Neoplasms". Chaque concept a un terme préféré unique qui est souvent le nom standard du concept et plusieurs termes non-préférés.

Lorsqu'un mot/terme (appelé communément *mot-clé*) participe à la description de la sémantique du document, il est considéré comme *descripteur* du document. Ces mots-clés permettent d'interpréter le contenu sémantique du document. L'ensemble des mots-clés reconnus par le SRI sont rangés dans une structure appelée **dictionnaire** ou **lexique** constituant le *langage d'indexation*. Ce dernier peut être *contrôlé* ou *libre*. Dans le premier cas, une expertise préalable sur le domaine d'application considéré, établit un vocabulaire exhaustif représenté dans une ressource termino-ontologique.

Sens

Le sens, ou encore *signification* en linguistique, désigne le contenu conceptuel d'une expression (mot, syntagme, phrase, énoncé...). Le sens est souvent défini dans un dictionnaire ou une source de connaissances (ontologie, thésaurus, terminologie, etc.). Un mot ayant plusieurs sens est appelé *polysémique*. Son sens dépend du contexte dans lequel il est utilisé. Un mot peut être utilisé dans son sens le plus courant (on parle alors de sens propre) ou dans un sens plus imagé (on dit qu'il est au sens figuré).

3.1.2 WordNet

WordNet est un réseau lexical électronique (Fellbaum, 1998) développé depuis 1985 à l'université de Princeton par une équipe de psycholinguistes et de linguistes du laboratoire des sciences cognitives, 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 pour la recherche.

WordNet couvre la majorité des *noms, verbes, adjectifs* et *adverbes* de la langue anglaise structurés en un réseau de nœuds et de liens. Chaque nœud, appelé **synset** (set of synonyms), est constitué d'un ensemble de termes synonymes. Cela signifie que les synonymes ayant le même sens sont groupés ensemble dans un nœud pour former un synset. Chaque synset représente un sens unique d'un mot particulier. Un terme peut être un mot simple ou une collocation (i.e. deux mots ou plusieurs mots reliés par des soulignés pour constituer un terme complexe correspondant).

Les synsets de WordNet sont reliés par des liens ou 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 relation de *subsomption* (**hyperonymie-hyponymie**), et la relation de *composition* (**méronymie-holonymie**). Ces relations sont formellement définies comme suit :

- **Hyperonymie** désigne une classe de concepts englobant des instances de classes plus spécifiques : Y est un hyperonyme de X si X est un type de Y. Par exemple, "fruit" est un hyperonyme de "pomme" et de "cerise".
- **Hyponymie** désigne un membre d'une classe de concepts : X est un hyponyme de Y si X est un type de Y. Par exemple, "France" est hyponyme de "pays", "cheval" est hyponyme de "animal".
- **Holonymie** est le nom de la classe globale dont les noms méronymes

font partie. Y est un holonyme de X si X est une partie de Y. Par exemple, "corps" est un holonyme de "bras", de même que "maison" est un holonyme de "toit".
– **Méronymie** est le nom d'une partie constituante, substance de ou membre d'une autre classe : X est un méronyme de Y si X est une partie de Y. Par exemple, "voiture" a pour méronymes "porte", "moteur", "roue", etc.

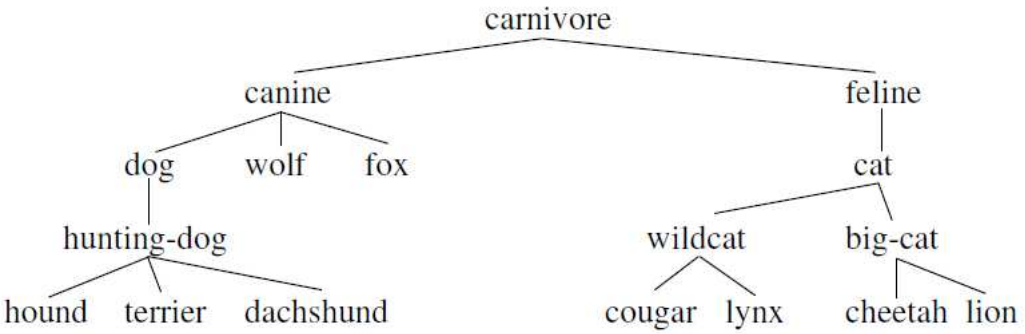


FIGURE II.5 – Exemple du réseau des concepts dans WordNet

La figure II.5 illustre un exemple de sous-hiérarchie de WordNet correspondant au concept "canine ". WordNet a été largement utilisée pour la tâche de désambiguïsation des mots (Navigli, 2009) ainsi que l'intégration de la désambiguïsation en RI (Voorhees, 1993; Baziz, 2005; Boubekeur, 2008). Les raisons de cette large utilisation sont dues au fait que cette base de données lexicale couvre de façon quasi-totale la langue anglaise, ce qui la place souvent en adéquation avec les données traitées en recherche d'information dans le cas général, qui sont de type presse (journaux et périodiques).

3.1.3 Open Directory Project - ODP

L'Open Directory Project, abrégé ODP, plus connu sous le nom de **dmoz** (Directory Mozilla qui donne son nom au site, www.dmoz.org), est un répertoire de sites web créé en 1998. L'ODP est considéré comme le plus grand et le plus complet des répertoires du Web édités par une vaste communauté d'éditeurs bénévoles provenant du monde entier. L'ODP utilise une structure hiérarchique pour organiser des listes de sites Web. Les sujets sémantiques ou concepts similaires sont regroupés en catégories qui peuvent inclure des sous-catégories.

L'ODP est construit manuellement par les éditeurs du Web qui associent les pages Web les plus similaires à une catégorie de concepts ou un thème spécifique dans l'ODP. Chaque concept qui se trouve dans cette architecture hiérarchique représente donc un thème ou un domaine d'intérêt des utilisateurs du Web. Chaque concept est défini par un **titre** et une **description** qui résume

le contenu des pages Web associées. Chaque page Web associée à une catégorie particulière possède également un titre et une description qui résument son contenu dans la page originale.

Les concepts de l'ODP sont reliés entre eux par les relations sémantiques de telles que "*est-un*", "*symbolique de* " et "*est relié à*".

- La relation "**est-un**" (is-a) : permet de définir une hiérarchie de concepts les plus génériques aux plus spécifiques. Par exemple, "système d'exploitation" est un "logiciel" qui peut être classé dans la catégorie "ordinateurs".
- La relation "**est symbolique de**" (symbolic link) : est un lien hypertexte qui établit une connexion à partir d'une page Web d'un répertoire à une autre page dans le même répertoire. Les liens symboliques permettent de créer des raccourcis entre les pages Web dans un répertoire, de les classer dans plusieurs catégories (multi-classification).
- La relation "**est relié à**" (related-to) : permet de pointer vers les autres concepts qui ont des thèmes sémantiquement reliés.

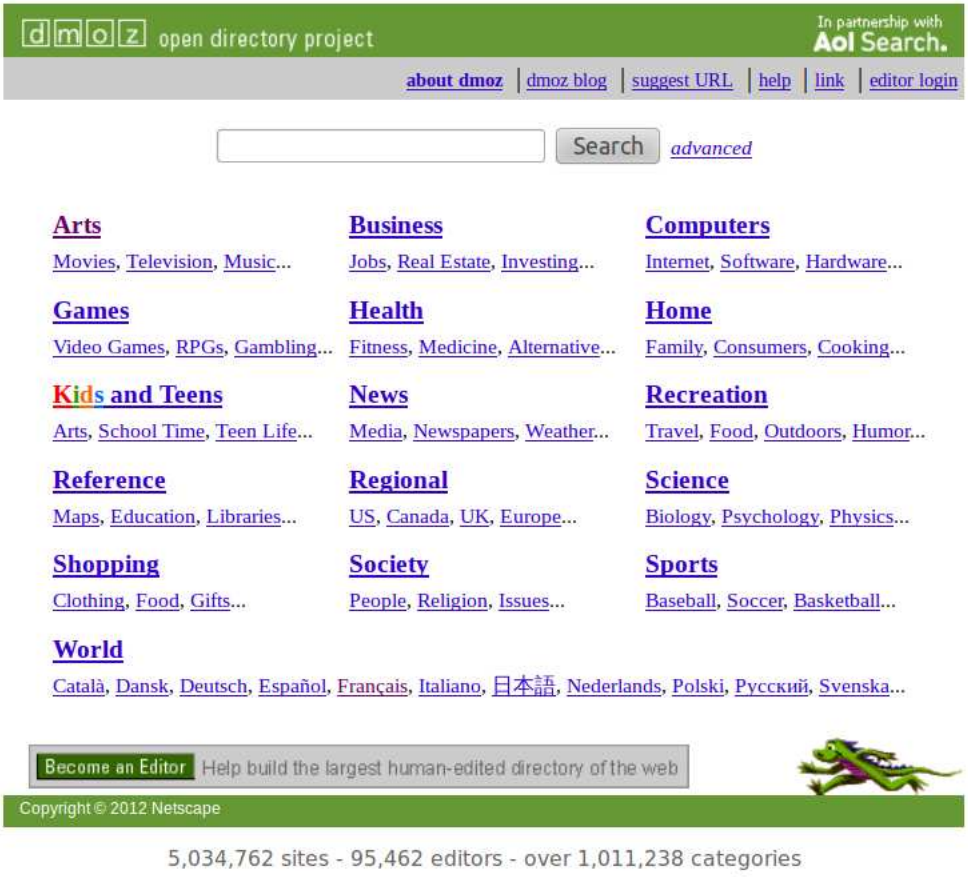


FIGURE II.6 – L'interface Web de l'ontologie ODP

La figure II.6 présente l'interface Web de l'ODP : les concepts génériques du plus haut niveau de l'ontologie sont mis en gras ; les concepts plus spécifiques

sont classés de manière adéquate dans chaque catégorie de concepts. L'ODP a été intégrée dans des systèmes de RI personnalisée (Chirita *et al.*, 2005; Ma *et al.*, 2007; Sieg *et al.*, 2007; Daoud, 2009) qui exploitent les concepts issus de cette ontologie pour représenter le profil de l'utilisateur. Par exemple, (Chirita *et al.*, 2005) ont défini le profil de l'utilisateur comme une liste de concepts dans l'ODP. Ensuite, le profil est intégré lors de la recherche par une combinaison linéaire du poids retourné par le système de RI et le poids basé sur la distance sémantique entre le profil et chaque résultat. Plusieurs travaux visant à intégrer les ontologies dans un processus de RI ont adopté cette technique de combinaison linéaire pour modifier les ordonnancements des résultats (Daoud, 2009; Trieschnigg, 2010). Cependant, il n'est pas évident de déterminer les cœfficients de la combinaison car il est difficile de quantifier le pourcentage du poids lié aux termes issus du document par rapport au poids lié aux termes issus du profil.

3.1.4 Yet Another Great Ontology - YAGO

Yet Another Great Ontology [8], abrégé YAGO, est une base de connaissances de l'humanité qui est extraite automatiquement à partir de Wikipedia [9], WordNet [10] et GeoNames [11]. Actuellement, YAGO contient plus de dix millions d'entités nommées (e.g., **personnes**, **organisations**, **villes**) et une centaine de millions d'informations au sujet de ces entités. YAGO dispose d'une interface Web permettant aux utilisateurs de poser des questions sous forme de requêtes. YAGO est en cours d'élaboration à l'Institut en Informatique de Max Planck. Le tableau II.4 montre quelques chiffres sur le nombre de catégories sémantiques, le nombre des entités stockées dans YAGO ainsi que les informations sur les relations entre elles. YAGO a été évaluée manuellement en utilisant 5,864 "faits" [12] avec une précision de 95.4% (Hoffart *et al.*, 2009).

TABLEAU II.4 – Quelques statistiques de YAGO

Type	YAGO sans GeoNames	YAGO avec GeoNames
Catégories	365.372	365.372
Entitées	2.648.387	9.756.178
Faits	124.333.521	447.470.256
Relations	104	114

La figure II.7 visualise l'entité "Albert Einstein" dans YAGO : chaque

8. `http://www.mpi-inf.mpg.de/yago-naga/yago/`
9. `http://en.wikipedia.org`
10. `http://wordnet.princeton.edu/wordnet/`
11. `http://www.geonames.org/`
12. Un fait est défini comme un triple d'une entité nommée (Sujet), d'une relation (Prédicat) et d'une autre entité nommée (Objet) reliée à la première par la relation.

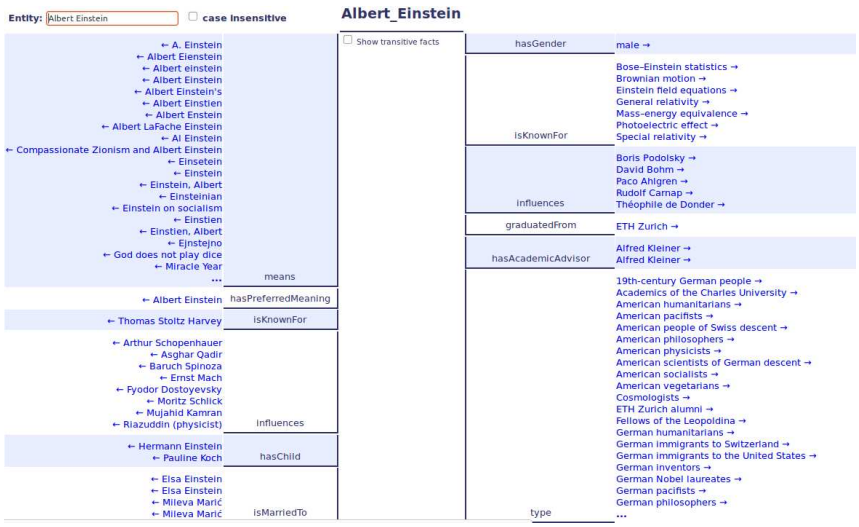


FIGURE II.7 – Visualisation des entités dans YAGO

entité peut être associée à une ou plusieurs relations avec d'autre entité, ce qui constitue un fait, par exemple ("Albert Einstein", "hasGender", "male"). YAGO a été utilisé pour catégoriser les résultats du Web (Ren *et al.*, 2009) ou pour personnaliser les résultats du Web (Calegari et Pasi, 2011). Par exemple, le travail de (Ren *et al.*, 2009) a pour but d'associer un concept issu de YAGO, appelé concept de catégorie (category concept), à chaque document résultat restitué par le moteur de recherche Google. La similarité concept-document est mesurée par la mesure TF IDF. Chaque document résultat (parmi les 100 premiers) est représenté par un vecteur de termes dans un modèle vectoriel (Salton *et al.*, 1983). Chaque concept est représenté comme un ensemble de vecteurs de faits du concept.

3.2 Principe de la RI sémantique

La RI sémantique est essentiellement basée sur une indexation sémantique qui intègre la sémantique, c-à-d le sens, des objets textuels (document, requête). L'indexation sémantique consiste, lors de l'analyse d'un document ou de la requête, à rattacher chaque mot à un concept sous-jacent qui représente le sens des mots. Ainsi, par exemple, pour le mot "jaguar", il faut déterminer s'il s'agit du félin, de la voiture ou de l'avion. Dans (Baziz, 2005), l'indexation sémantique a été définie comme une approche d'indexation basée sur le **sens des mots** (appelée *sense-based indexing* en anglais). Plusieurs travaux d'indexation sémantique ont été menés dans ce sens (Krovetz et Croft, 1992; Voorhees, 1993; Sanderson, 1994; Mihalcea *et al.*, 2004; Liu *et al.*, 2004; Baziz, 2005; Boubekeur, 2008).

De manière générale, le principe de base de l'indexation sémantique consiste

à extraire dans un premier temps, l'ensemble des mots les plus significatifs, appelés *descripteurs* ou *mots-clés* du document. Ces mots sont par la suite affectés par un sens qui lui correspond le plus. Si un mot-clé possède plusieurs sens, celui-ci est désambiguïsé pour retenir le sens le plus adéquat dans son contexte d'utilisation. Chaque mot-clé ambigu est associé à un ensemble de sens (synsets ou concepts) dans la ressource terminologique (souvent WordNet). Puis, les poids sont associés aux différents sens candidats. Le sens qui maximise le poids de similarité est retenu comme le sens le plus adéquat du mot-clé. Une fois que tous les mots-clés ambigus sont désambiguïsés, la représentation des textes se fait soit à partir des seuls sens (ou concepts) identifiés lors de l'étape de désambiguïsation, soit à partir d'une combinaison linéaire des mots-clés et sens associés. Les approches de RI sémantique de (Voorhees, 1993; Uzuner *et al.*, 1999; Mihalcea *et al.*, 2004; Khan *et al.*, 2004; Baziz, 2005) sont basées sur ce principe.

Deux caractéristiques suivantes concernent la plupart des méthodes d'indexation sémantique :

- L'usage des ressources externes (thésaurus, ontologies, dictionnaires, etc.) pour l'identification du sens des unités d'informations dans les contenus textuelles.

- Un processus de traitement qui consiste tout d'abord à identifier le meilleur sens associé à un passage de texte puis à exploiter le sens sélectionné comme élément d'indexation. Le processus de base appelé communément désambiguïsation est au cœur de l'indexation sémantique.

3.2.1 Les ressources exploitées pour l'indexation sémantique

Les ressources qui sont exploitées pour l'indexation sémantique peuvent être *structurées* ou *non-structurées*.

- Les *ressources structurées* telles que les thésaurus qui fournissent des informations sur les relations entre les mots comme la synonymie, les abréviations, ... (par exemple, Roget's International Thesaurus (Chapman, 1992)), les dictionnaires (par exemple, Collins English Dictionary (Sinclair, 1995)), les ontologies (par exemple, WordNet (Miller, 1995)). Ces ressources sont souvent définies pour l'utilisation dans un domaine général. Dans un domaine spécifique tel que le domaine biomédical, la ressource l'UMLS contient trois sources de connaissances principales, à savoir le Méta-thésaurus, le réseau sémantique et le lexique SPÉCIALISTE, fournissant une catégorisation des concepts médicaux (McInnes *et al.*, 2007).

– Les *ressources non-structurées* telles que des corpus de documents (avec ou sans l'annotation du sens), et des ressources de collocation, sont des ressources principales de la désambiguïsation basée sur des ressources non-structurées. Les corpus comme *Brown* Corpus (Francis et Kucera, 1979), *British National* Corpus (Lou, 1995), *Wall Street Journal* Corpus (Niwa et Nitta, 1994), fournissent les statistiques sur la distribution de mots. Les corpus annotés avec l'information sur le sens de mots comme *SemCor* Corpus (Miller *et al.*, 1993), la *ligne-hard-serve* Corpus (Leacock *et al.*, 1993), *interest* Corpus (Bruce et Wiebe, 1994), sont annotés avec les différentes sources de connaissances telles que Wordnet, LDOCE, ou HECTOR, etc. Les ressources de collocation comme [13], *Word Sketch engine* [14], ... définissent des restrictions sur la façon dont les mots peuvent être utilisés ensemble (contexte d'utilisation) en enregistrant la tendance de chaque mot à apparaître régulièrement avec d'autres.

3.2.2 Désambiguïsation pour l'indexation sémantique

La désambiguïsation de mots (ou *Word sense disambiguation* en anglais) (WSD) consiste à identifier le sens le plus adéquat de chaque mot ambigu dans un contexte d'utilisation de ce dernier. Cette tâche est considérée comme un problème le plus difficile à résoudre en intelligence artificielle (Mallery, 1988). En général, l'ambiguïté d'un mot est liée à ses multiples définitions dans une ou plusieurs ressources termino-ontologiques ou sources de connaissances de manière générale. Chaque définition correspond à un sens particulier dans le contexte d'utilisation donné.

Les approches de désambiguïsation s'appuient sur des ressources sémantiques comme les ontologies, les thésaurus et les dictionnaires pour déterminer le sens du mot dans son contexte. En outre, ces approches peuvent être (1) *supervisées* dans le cas où elles exploitent l'information sur le sens de mots dans un corpus manuellement annoté, ou (2) *non-supervisées* dans le cas où aucune annotation manuelle est fournie. La plupart des approches de désambiguïsation utilisent WordNet comme une ontologie d'un domaine général. Selon la source de données utilisée pour sélectionner le sens du mot ambigu dans son contexte, on peut distinguer deux catégories de désambiguïsation basée sur les sources de connaissances : (1) *désambiguïsation basée sur les graphes* et (2) *désambiguïsation basée sur les domaines*.

1. **La désambiguïsation basée sur les graphes** est essentiellement basée sur la localisation des sens des mots ambigus dans une architecture

13. `http://www.natcorp.ox.ac.uk/`
14. `http://www.sketchengine.co.uk/`

hiérarchique de l'ontologie. L'idée derrière est que les mots dans un même contexte ont les sens les plus proches au sein de la ressource ; par conséquent, le sens le plus probable du terme de chaque mot ambigu dans le contexte est choisi en fonction de sa proximité avec les autres sens des mots environnants.

Les travaux concernant la désambiguïsation basée sur les graphes s'appuient sur des mesures de distance de similarité entre les sens des mots ambigus dans un réseau sémantique. Ces mesures sont calculées en fonction des critères suivants :

- le chemin le plus court entre les nœuds représentant les sens des mots (Rada *et al.*, 1989; Sussna, 1993) et leur hyperonymes (Leacock et Chodorow, 1998) ;
- la densité des concepts liés aux sens des mots ambigus (Agirre et Rigau, 1996) ;
- l'information du contenu extrait à partir des corpus annotés en combinaison avec les mesures de similarité basées sur le sens (Jiang et Conrath, 1997).

D'autres mesures de similarité comme *Page Rank*, *Degree* ou *Closeness*, qui sont liées à la centralité des concepts, ont été également exploitées pour la désambiguïsation des mots (Mihalcea *et al.*, 2004). Plus précisément, la tâche de désambiguïsation est réalisée en trois principales étapes : (1) construire le graphe sémantique correspondant aux sens des mots dans leur contexte où chaque nœud dans le graphe représente un sens particulier d'un mot ambigu et une arête entre deux nœuds représente un lien sémantique entre eux. (2) calculer (itérativement) un poids centralité pour chaque nœud dans son contexte lié au graphe sémantique (3) les sens candidats sont classés en fonction de la centralité. Enfin, les meilleurs sens sont sélectionnés pour chaque mot ambigu grâce au poids maximal.

2. **Désambiguïsation basée sur les (sous-)domaines**. Cette approche exploite les informations sur les différents sous-domaines liés aux concepts dans l'ontologie (Gliozzo *et al.*, 2004; Buitelaar *et al.*, 2007). Le sens du mot ambigu est choisi en se basant sur la comparaison entre les domaines du mot en contexte et chaque domaine du mot ambigu. Pour cela, les domaines de WordNet sont souvent utilisés (Navigli, 2009). Les domaines de chaque mot sont représentés par un vecteur de domaine dont chaque composante représente un domaine. Étant donné le sens S d'un mot, le vecteur *synset* est défini comme $S = (R(D_1, S), R(D_2, S), ..., R(D_d, S))$, où D_i est le domaine i du mot courant ($i \in \{1, ..., d\}$) et $R(D_i, S)$ est

défini comme suit :

$$R(D_i, S) = \begin{cases} 1/\|Dom(S)\| & \text{si } D_i \in Dom(S) \\ 1/d & \text{si } Dom(S) = \{FACTOTUM\} \\ 0 & \text{sinon} \end{cases} \quad \text{(II.38)}$$

où $Dom(S)$est l'ensemble de domaines affectés au sens S dans WordNet et $FACTOTUM$ représente l'absence du domaine pertinent.

L'intérêt de la désambiguïsation basée sur les domaines porte sur le fait que les méthodes de désambiguïsation ne demandent pas un niveau élevé en terme de compréhension linguistique en se concentrant sur l'exploitation des domaines sémantiques.

3.3 Aperçu général de travaux de désambiguïsation et d'indexation sémantique en RI

La désambiguïsation des mots peut être intégrée dans un processus de RI sémantique via un schéma d'appariement entre la requête et les documents dans la collection. La désambiguïsation peut être effectuée et intégrée dans un processus de RI sémantique de différentes manières avec ou sans l'intervention de l'indexeur humain. On distingue ici la **désambiguïsation manuelle** *vs.* **désambiguïsation automatique**.

3.3.1 Désambiguïsation manuelle pour la RI sémantique

(Krovetz et Croft, 1992) ont évalué l'effet de la désambiguïsation sur les performances de la RI sémantique. Leurs évaluations expérimentales ont été effectuées sur deux collections de documents : CCAM (3,204 documents constitués d'un titre et d'un résumé) et TIME (324 documents courts). Leurs objectifs étaient de déterminer l'efficacité d'un schéma de pondération (appariement) sémantique en exploitant le(s) sens qu'un mot de la requête ou du document peut avoir et de déterminer l'utilité de l'identification du sens des mots en séparant les documents pertinents des documents non-pertinents.

Une *désambiguïsation manuelle* des mots figurant dans la requête ainsi que dans les dix premiers documents, qui sont restitués par un modèle de pondération probabiliste, a été effectuée en utilisant le dictionnaire LDOCE (Longman Dictionary of Contemporary English). Ce dernier contient essentiellement les informations suivantes pour chaque mot donné : catégories grammaticales (nom, verbe, adjectif, etc.), sous-catégories (e.g., verbe transitif *vs.* intransitif),

définition (sens du mot), exemples etc. Ils ont distingué les différents types de non-correspondance entre la requête et les premiers documents :

- la *non-correspondance de mots :* est liée au défaut de la normalisation des mots de la requête et du document par la racinisation, c-à-d que les formes radicales ne correspondent pas au même sujet, e.g., "arm"/"army", "passive"/"passing", "code"/"E.F. Codd".

- la *non-correspondance de sens :* apparaît seulement si le sens du mot de la requête ne correspond pas à celui du mot dans le document.

- la *non-correspondance des termes techniques :* est liée aux termes techniques dans la requête/document (e.g., "distributed system"/"probability distribution", "parallel between problems"/"parallel processing"). Cette caractéristique est plus souvent observée sur la collection CCAM mais rarement sur la collection TIME. Par contre, un mélange des **correspondances** et **non-correspondances** au niveau du **sens** est souvent observé sur la collection TIME mais rarement sur la collection CCAM.

Ils ont sélectionné/extrait 45 requêtes parmi 64 requêtes originales dans la collection CCAM et 45 requêtes parmi 83 requêtes originales de la collection TIME dans leur étude. Au total, 90 requêtes ainsi que les dix premiers documents restitués pour chaque requête ont été **manuellement désambiguïsés**. (Krovetz et Croft, 1992) ont examiné l'amélioration de l'efficacité de la recherche en **supprimant les documents** parmi les dix premiers documents qui représentent au moins une non-correspondance de sens avec la requête.

Leurs résultats ont montré qu'une proportion significative des non-correspondances (mismatches) du sens des mots sont due à la racinisation (stemming). La non-correspondance des mots entraîne systématiquement la non-correspondance de sens, c-à-d que le sens d'un mot particulier dans le document restitué ne correspond plus au sens du mot dans la requête car en réalité ce sont les deux mots différents mais ayant le même radical. Ils ont conclu que l'ambiguïté lexicale n'est pas un problème significatif si le document contient un "grand nombre" de mots communs entre celui-ci et la requête car la **collocation des mots** de la requête rend celle-ci plus claire en terme de sens.

Les inconvénients de leur méthode sont résumés comme suit :

- La désambiguïsation a été effectuée manuellement, ce qui demande à l'utilisateur de faire un grand effort pour formuler sa requête en spécifiant ou déterminant manuellement le sens pour chaque mot. De plus, il faut désambiguïser également les dix premiers documents renvoyés vis-à-vis de la requête, ce qui représente un temps additionnel pour la désambiguïsation.

– Le nombre de dix premiers documents impliqués dans la désambiguïsation n'a pas été justifié. De plus, la suppression des documents parmi les dix premiers documents dont le sens ne correspondent pas à la requête n'est pas une solution naturelle pour un moteur de recherche car il peut toujours y a voir ces types de documents après dix premiers documents.

– En se basant sur le dictionnaire LDOCE, il n'y aucun moyen pour traiter les termes techniques qui sont souvent des termes complexes tandis que LDOCE ne contient que les informations sur les mots simples.

3.3.2 Désambiguïsation automatique pour la RI sémantique

Travaux de (Voorhees, 1993). La première évaluation de la désambiguïsation automatique sur plusieurs collections à une échelle plus grande est le travail de (Voorhees, 1993) qui a exploité les relations de type "est-un" (is-a) dans WordNet, à savoir l'hyponymie et l'hyperonymie, et un ensemble de noms dans le texte pour assigner le sens à chaque mot polysémique. L'idée de leur méthode de désambiguïsation automatique est liée au fait que les mots apparaissant ensemble dans un contexte donné permettent de déterminer de manière appropriée leur sens sachant que ces mots sont individuellement ambigus s'ils sont considérés en dehors du contexte. Cette idée ressemble à l'observation de (Krovetz et Croft, 1992) concernant la collocation des mots dans un contexte spécifique. Par exemple, bien que les mots "*base*", "*bat*", "*glove*", "*hit*" possèdent chacun plusieurs sens, ils tendent à parler d'un thème commun "*baseball*" lorsqu'ils apparaissent ensemble.

Plus spécifiquement, la désambiguïsation a pour but de déterminer le "bon sens" d'un mot ambigu est basée sur degré de recouvrement entre d'une part, le contexte local de ce mot et d'autre part le voisinage de ce sens dans WordNet. Le contexte du mot courant peut être par exemple la phrase le contenant. Le voisinage du synset d'un mot a été défini comme "le plus large sous-graphe connexe contenant s et seulement les descendants d'un ancêtre de s et ne contenant aucun synset ayant un descendant qui inclut une autre instance d'un membre (mot) de s". Un synset peut être considéré comme un concept qui indique le sens du mot. Pour chaque mot non vide dans la phrase, le (ou les) synsets sont récupérés grâce à une recherche dans WordNet. Un mot ambigu correspond donc à plusieurs synsets dans WordNet. Pour déterminer le synset le plus adéquat dans une phrase, chaque synset de ce mot est classé en se basant sur le nombre de mots co-occurrents entre un voisinage (appelé *hood*) de ce synset et le contexte local du mot ambigu correspondant. Le synset le mieux classé est alors considéré comme le sens adéquat de l'occurrence analysée du mot ambigu.

Les évaluations ont été effectuées sur les collections CACM, CISI, CRAN-

FIELD 1400, MEDLINE, et TIME. Dans les expérimentations, (Voorhees, 1993) a crée les vecteurs documents/requêtes basés sur le sens de chaque mot identifié par la désambiguïsation. Ce vecteur est constitué de trois sous-vecteurs de différents types de concepts (ctypes) :

- le premier contient les radicaux (stems) des mots simples qui n'existent pas dans WordNet,
- le deuxième correspond aux synsets (synonymes) des noms désambiguïsés,
- le troisième correspond aux radicaux des mots (plus précisément des noms) désambiguïsés.

Le poids final de chaque document D vis-à-vis de la requête Q, noté $score(D, Q)$, est donné par une combinaison linéaire comme suit :

$$sim(D, Q) = \sum_{ctypei} \alpha_i sim_i(D_i, Q_i) \tag{II.39}$$

où sim_i est la similarité pour le type de concept i, α_i est un facteur déterminant le degré d'importance de chaque type de concept dans la représentation du document, D_i et Q_i sont les sous vecteurs de D et Q.

Travaux de (Baziz, 2005). Il s'agit d'une méthode de désambiguïsation par le réseau de concepts pour améliorer la représentation du document via un processus d'indexation sémantique des documents à base de concepts et de relations entre concepts. Les mots/termes (descripteurs) sont d'abord extraits du document par une approche classique d'indexation. Ils sont ensuite projetés sur l'ontologie linguistique WordNet afin d'identifier les concepts (ou sens) correspondants dans l'ontologie. Lorsqu'un descripteur correspond à plus d'un concept dans WordNet, il est dit ambigu. L'approche de désambiguïsation proposée dans (Baziz, 2005) 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. Leur approche consiste à affecter un poids à chaque concept candidat d'un terme d'indexation donné en sommant les valeurs de similarité entre celui-ci et les autres concepts candidats (correspondant aux différents sens des autres termes du document). Pour un terme T_i, le poids de son $k - ième$ sens est calculé comme suit :

$$C_score(C_k^i) = \sum_{l=1..m, l\neq i, j=1..n} P_{i,l}(C_k^i, C_j^l) \tag{II.40}$$

où m représente le nombre de termes du document et n le nombre de sens qui est propre à chaque terme T_i ; $P_{i,l}(C_k^i, C_j^l)$ est la valeur de similarité entre les deux sens (concepts) C_k^i et $C * j^l$.

Le concept candidat ayant le poids le plus élevé est retenu comme concept adéquat indiquant le sens du terme d'indexation associé. Finalement, le document est représenté comme un réseau de concepts et de liens entre concepts (*cf.* la figure II.8 (Baziz, 2005)). Les liens (arcs) entre les différents concepts sont pondérés par plusieurs mesures de similarité sémantique (Lesk, 1986; Leacock et Chodorow, 1998; Lin, 1998; Resnik, 1999).

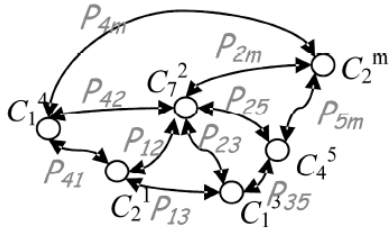


FIGURE II.8 – Exemple de réseau sémantique construit à partir de concepts candidats.

Travaux de (Liu *et al.*, 2004). Il s'agit d'une méthode de désambiguïsation des requêtes pour la RI sémantique via un processus d'expansion conceptuelle de requêtes. Plus spécifiquement, (Liu *et al.*, 2004) ont présenté une technique de désambiguïsation des requêtes courtes en utilisant WordNet. Les informations concernant les synsets d'un mot ainsi que les hyponymes de chaque synset dans WordNet ont été exploités pour la désambiguïsation. Après avoir identifié le sens le plus adéquat de chaque mot, la requête est étendue par les mots issus des synonymes, de la définition du synset, des hyponymes, ou d'un terme composé. Lors de l'appariement, chaque document est associé par un poids basé sur les groupes nominaux (noms propres, expressions dans un dictionnaire, expressions simples, expressions complexes) et un poids basé sur les termes simples. Plus spécifiquement, le premier est calculé par une fonction de similarité en se basant sur les expressions communes les plus significatives entre le document et la requête. Le deuxième poids est calculé en utilisant le schéma de pondération BM25 (Robertson *et al.*, 1998). Pour déterminer les expressions les plus significatives, ils ont définit une mesure de corrélation entre les termes d'une expression comme suit :

$$correlation(t_1, t_2, ..., t_n) = \frac{P(expression) - \prod_{t_i \in expression} P(t_i)}{\prod_{t_i \in expression} P(t_i)} \qquad \text{(II.41)}$$

où $t_1, t_2, ..., t_n$ sont les termes de l'expression, $P(expression)$ est la probabilité que le document contienne l'expression et $\prod_{t_i \in expression} P(t_i)$ est la probabilité que le document contienne tous les termes dans l'expression en supposant que ces termes soient indépendants.

4 Conclusion

Ce chapitre a porté essentiellement sur les principales notions et concepts de base qui font l'objet du domaine de la RI. Nous y avons décrit les principales étapes d'un processus de RI classique, à savoir, (1) l'indexation des documents dans la collection et (2) l'interrogation de l'information qui est basée sur les modèles d'appariement document-requête. Les différentes catégories de modèles de RI représentatifs ont été présentées ainsi que les caractéristiques de chaque modèle de RI ont été détaillées. La plupart des modèles de RI sont basés sur la représentation de l'information (documents et requêtes) par des sacs de mots qui sont basés sur l'hypothèse d'indépendance entre les mots (term independance assumption), ce qui représente la cause principale du défaut d'appariement (mismatch) entre les documents et la requête de l'utilisateur. Pour pallier à ce problème, une solution est de reformuler la requête et/ou les documents pour ajuster le degré de recouvrement (document-query overlap ou term overlap) entre les documents et la requête. Nous avons également présenté les concepts de base d'un processus d'indexation conceptuelle/sémantique en utilisant des ressources termino-ontologiques, notamment WordNet, dans un domaine général, les principes de base d'un processus d'indexation sémantique qui intègre le sens dans l'appariement document-requête. Quelques travaux représentatifs y sont également présentés. Les travaux d'indexation sémantique suggèrent avoir un algorithme de désambiguïsation "assez performante" pour pouvoir faire une correspondance entre les documents et la requête en fonction du sens des mots figurant dans les documents et dans la requête, une bonne stratégie pour intégrer le sens dans l'appariement sémantique. Dans un domaine spécifique, la désambiguïsation est moins importante car les concepts sont définis clairement pour un domaine.

Dans le chapitre suivant, nous présentons les travaux d'indexation et de recherche d'information conceptuelle dans le domaine biomédical. Les études menées nous permettent d'établir une base solide à partir de laquelle nous proposons des solutions et techniques adéquates pour améliorer les performances de la RI biomédicale.

Chapitre III

Indexation et Recherche d'Information Biomédicale

Sommaire

1 Introduction **59**

2 Typologie des informations biomédicales **60**
- 2.1 La littérature biomédicale 61
- 2.2 Les dossiers médicaux du patient 61

3 Ressources termino-ontologiques biomédicales **64**
- 3.1 Typologie des ressources termino-ontologiques 65
  - 3.1.1 Terminologie 65
  - 3.1.2 Classification 65
  - 3.1.3 Nomenclature 66
  - 3.1.4 Thésaurus 66
  - 3.1.5 Ontologie 67
- 3.2 Quelques ressources termino-ontologiques du domaine biomédical 67
  - 3.2.1 Thésaurus MeSH 67
  - 3.2.2 Nomenclature SNOMED 70
  - 3.2.3 Ontologie de gènes - GO 71
  - 3.2.4 Méta-thésaurus UMLS 72

4 Extraction des concepts biomédicaux **74**
- 4.1 Principales approches d'extraction de concepts 75
  - 4.1.1 Approche basée sur des règles linguistiques 75
  - 4.1.2 Approche basée sur l'apprentissage automatique . . . 76
  - 4.1.3 Approche basée sur la recherche dans un dictionnaire 80
  - 4.1.4 Approche basée sur des mesures statistiques 81
- 4.2 Principaux outils d'extraction de concepts 83
  - 4.2.1 PubMed ATM 83
  - 4.2.2 MetaMap 85
  - 4.2.3 MTI (Medical Text Indexer) 86
  - 4.2.4 MaxMatcher 87

5 Indexation mono-terminologique *vs.* multi-terminologique de documents biomédicaux **88**
5.1 Historique de l'indexation en RI biomédicale 88
5.2 Synthèse des travaux d'indexation des documents biomédicaux 90
5.2.1 Indexation mono-terminologique 90
5.2.2 Indexation multi-terminologique 94

6 Techniques et modèles d'appariement document-requête en RI biomédicale **98**
6.1 Reformulation de requêtes 99
6.1.1 Reformulation conceptuelle de requêtes 99
6.1.2 Reformulation de requêtes par pseudo-réinjection de pertinence 104
6.2 Expansion conceptuelle de documents 105
6.3 Appariement basé sur l'identification de patrons de besoins cliniques (modèle PICO) 107

7 Évaluation de recherche d'information biomédicale **110**
7.1 Campagne d'évaluation CLEF 110
7.2 Campagne d'évaluation de TREC 112
7.2.1 TREC Genomics pour la RI de la littérature biomédicale112
7.2.2 TRECMed pour la RI biomédicale des comptes-rendus médicaux de patients 115

8 Conclusion **118**

"We have a hunger of the mind which asks for knowledge of all around us, and the more we gain, the more is our desire, the more we see, the more we are capable of seeing."
–Maria Mitchell

1 Introduction

Depuis l'avènement d'Internet et des librairies digitales, les volumes d'informations évoluent de manière significative tant en volume qu'en qualité. D'autre part, l'orientation actuelle vers l'informatisation des systèmes d'information des organisations a fait augmenter l'information numérique stockée dans les bases documentaires. Dans le domaine biomédical, les services de production et d'accès à l'information ne cessent de se diversifier. À titre d'exemple la base de données bibliographique MEDLINE (Medical Literature Analysis and Retrieval System Online) contient plus de 21 millions de références d'articles en sciences de la vie, notamment de la biomédecine. Parallèlement, le stockage et l'échange de l'information dans les hôpitaux et les différents services de santé se font de plus en plus sur des supports numériques (Bringay, 2006). Ces informations, à l'origine non structurées, sont répertoriées, classifiées et stockées dans des bases de données sous une forme exploitable par l'ordinateur ou des programmes d'accès aux données dans le but de faciliter leur consultation ainsi que leur utilisation. L'indexation de ces sources de données permet de mieux organiser, de structurer et surtout de faciliter l'accès à l'information biomédicale.

Il existe à ce jour plusieurs terminologies pour indexer les documents biomédicaux. L'objectif de l'indexation en RI biomédicale est de faciliter l'accès à la littérature biomédicale en affectant à chaque document une liste de termes désignant les concepts issus d'une ou des terminologies biomédicales (Névéol *et al.*, 2006; Darmoni *et al.*, 2009). Les terminologies les plus utilisées pour l'indexation contrôlée sont : MeSH (Medical Subject Headings), SNOMED (Systematized Nomenclature of Medicine), ICD-10 (International Classification of Diseases), GO (Gene Ontology), UMLS (Unified Medical Language System) ... En effet, les documents de la base MEDLINE sont indexés par les descripteurs issus de MeSH qui sont sélectionnés avec soin par les experts de la bibliothèque nationale de santé aux États-Unis (NLM) (Aronson *et al.*, 2004b). En France, les documents biomédicaux dans le portail CiSMEF sont indexés par les descripteurs de MeSH en français (Névéol *et al.*, 2006).

Lors de la recherche d'information, les terminologies peuvent être utilisées pour modifier la requête initiale en ajoutant des synonymes ou des acronymes des concepts médicaux (Zhou *et al.*, 2007a; Stokes *et al.*, 2009; Lu *et al.*, 2009). Les termes issus du thésaurus MeSH sont les plus utilisés pour étendre la requête biomédicale en général tandis que les termes de l'Entrez Gene (Maglott *et al.*, 2005), HUGO (Eyre *et al.*, 2006) ou OMIM (Mckusick, 1998) sont utilisés pour traiter en particulier les requêtes liées aux gènes et aux protéines.

L'objectif de ce chapitre est de donner un aperçu sur la RI biomédicale. Nous présentons d'abord dans la section 2 une typologie de l'information biomédicale. Ensuite, la section 3 présente les principales ressources termino-ontologiques les plus utilisées dans le domaine. Dans la section 4, nous présentons les différentes approches d'extraction des termes techniques qui désignent les concepts biomédicaux ainsi que les outils d'extraction des concepts qui sont accessibles (téléchargeables ou exécutables via un service Web). La section 5 est consacrée à la présentation des approches d'indexation basée sur les ressources termino-ontologiques dans le domaine biomédical. La section 6 est dédiée à la description des techniques d'appariement sémantique en RI biomédicale. Enfin, la section 7 donne quelques éléments sur l'évaluation des performances des systèmes de RI biomédicale et la section 8 conclut le chapitre.

2 Typologie des informations biomédicales

(Hersh, 2008) classifie les informations biomédicales en deux catégories principales :

1. *L'information spécifique au patient* : ces informations visent à informer les fournisseurs de soins de santé, les administrateurs et les chercheurs de la santé et la maladie du patient. Les informations peuvent être présentes sous forme de résultats de laboratoire ou compte-rendus médicaux ;
2. *Les connaissances du domaine biomédical* : ces informations peuvent être subdivisées en deux catégories : *information primaire* et *information secondaire*. Telle qu'elle est décrite dans cet ouvrage, l'information primaire (aussi appelée la littérature primaire) est une information originale qui apparaît dans des revues, livres, rapports et autres sources. L'information secondaire apparaît dans des commentaires, des synthèses de la littérature primaire, par exemple, livres, monographies et articles de revue dans le journal et d'autres publications.

Dans le domaine de la RI biomédicale, nous nous intéressons à deux types d'information : (1) *littérature biomédicale* et (2) *dossiers médicaux personnels*.

2.1 La littérature biomédicale

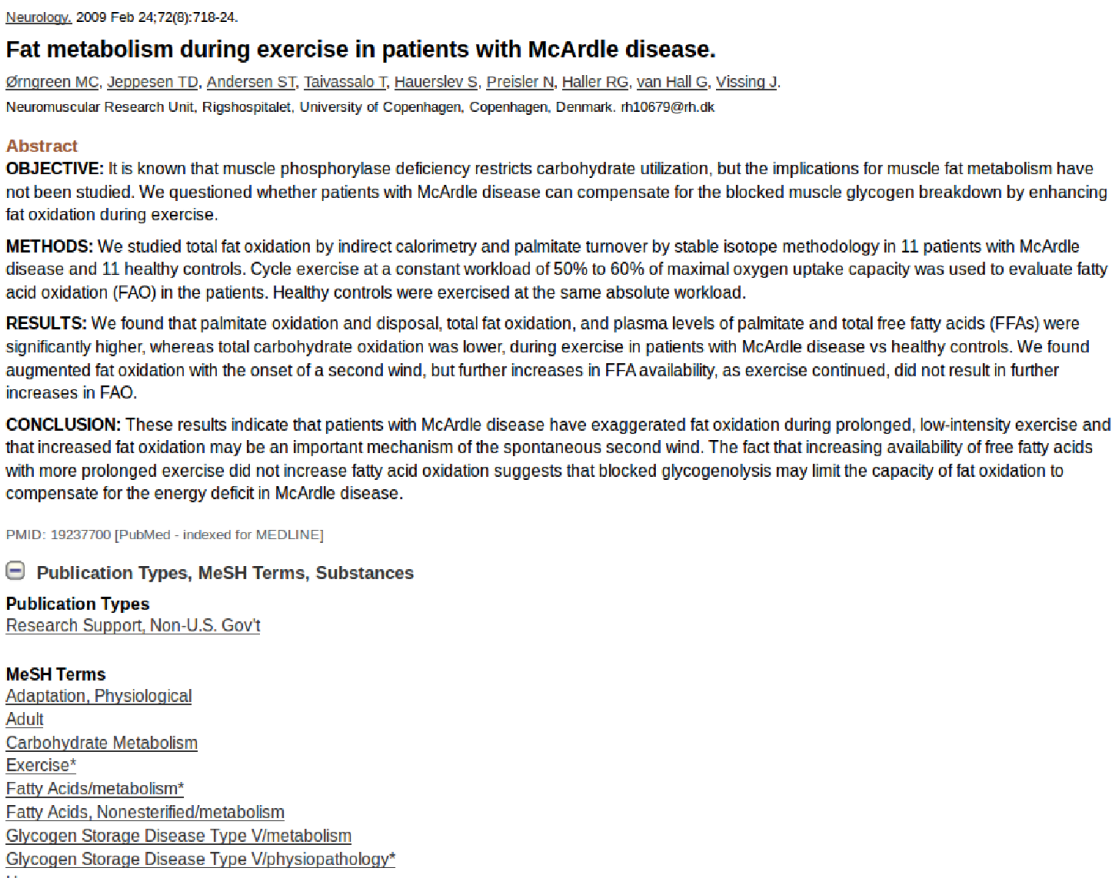
La littérature biomédicale est composée essentiellement des bases de données bibliographiques qui font références à des revues scientifiques et des comptes-rendus des conférences du milieu biomédical. MEDLINE est la base de données de référence du domaine. Cette base a été créée et gérée par la National Library of Medicine [1] (NLM) aux États-Unis regroupant des références d'une vingtaine de millions d'articles scientifiques indexés depuis 1966 jusqu'à aujourd'hui, principalement en langue anglaise. Elle fournit des informations dans les domaines de la médecine, des soins infirmiers, de la médecine dentaire, de la médecine vétérinaire, des paramédicaux et des sciences pré-cliniques. PubMed est le portail qui dispose d'un moteur de recherche pour accéder à l'information dans MEDLINE, à des résumés d'articles et des articles en texte intégral sur les sciences de la vie et biomédicales des sujets. La figure III.1 montre un exemple d'une citation (référence ou résumé d'articles de journaux) enregistrée dans la base bibliographique MEDLINE. Chaque document dans MEDLINE est annoté manuellement ou semi-automatiquement par des termes MeSH [2] (champ **MeSH Terms**). Les indexeurs humains lisent le texte intégral et affectent les termes MeSH les plus significatifs à chaque document. Pour la langue française, une large documentation francophone est également accessible à travers le portail CISMeF [3].

Les documents de MEDLINE ont été utilisés pour évaluer les performances des systèmes de RI biomédicale dans le cadre de TREC Genomics de 2003 à 2005. À partir de 2006, les documents en texte intégral obtenus à partir du site Highwire [4] ont été utilisés pour évaluer les performances des systèmes de type question-réponse (Hersh et Voorhees, 2009). Bien que PubMed propose essentiellement l'accès à des références d'articles de journaux (notamment les résumés d'articles), il est possible de récupérer des documents en texte intégral à partir du service PubMed Central [5] qui est une archive des journaux biomédicaux accessibles gratuitement sous la protection du copyright des auteurs.

2.2 Les dossiers médicaux du patient

Le Dossier Médical du Patient (DMP) est un dossier médical informatisé. Il constitue un noyau fondamental de la qualité des soins dans les structures hospitalières en s'inscrivant dans le cadre du parcours de soins et de la mise en place

1. `http://www.nlm.nih.gov/`
2. voir la section 3.2.1
3. `www.chu-rouen.fr/cismef/`
4. `http://highwire.stanford.edu/`
5. `http://www.ncbi.nlm.nih.gov/pmc/about/intro/`



Neurology. 2009 Feb 24;72(8):718-24.

Fat metabolism during exercise in patients with McArdle disease.

Ørngreen MC, Jeppesen TD, Andersen ST, Taivassalo T, Hauerslev S, Preisler N, Haller RG, van Hall G, Vissing J.

Neuromuscular Research Unit, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark. rh10679@rh.dk

Abstract

OBJECTIVE: It is known that muscle phosphorylase deficiency restricts carbohydrate utilization, but the implications for muscle fat metabolism have not been studied. We questioned whether patients with McArdle disease can compensate for the blocked muscle glycogen breakdown by enhancing fat oxidation during exercise.

METHODS: We studied total fat oxidation by indirect calorimetry and palmitate turnover by stable isotope methodology in 11 patients with McArdle disease and 11 healthy controls. Cycle exercise at a constant workload of 50% to 60% of maximal oxygen uptake capacity was used to evaluate fatty acid oxidation (FAO) in the patients. Healthy controls were exercised at the same absolute workload.

RESULTS: We found that palmitate oxidation and disposal, total fat oxidation, and plasma levels of palmitate and total free fatty acids (FFAs) were significantly higher, whereas total carbohydrate oxidation was lower, during exercise in patients with McArdle disease vs healthy controls. We found augmented fat oxidation with the onset of a second wind, but further increases in FFA availability, as exercise continued, did not result in further increases in FAO.

CONCLUSION: These results indicate that patients with McArdle disease have exaggerated fat oxidation during prolonged, low-intensity exercise and that increased fat oxidation may be an important mechanism of the spontaneous second wind. The fact that increasing availability of free fatty acids with more prolonged exercise did not increase fatty acid oxidation suggests that blocked glycogenolysis may limit the capacity of fat oxidation to compensate for the energy deficit in McArdle disease.

PMID: 19237700 [PubMed - indexed for MEDLINE]

Publication Types, MeSH Terms, Substances

Publication Types
Research Support, Non-U.S. Gov't

MeSH Terms
Adaptation, Physiological
Adult
Carbohydrate Metabolism
Exercise*
Fatty Acids/metabolism*
Fatty Acids, Nonesterified/metabolism
Glycogen Storage Disease Type V/metabolism
Glycogen Storage Disease Type V/physiopathology*

FIGURE III.1 – Exemple d'une citation de MEDLINE

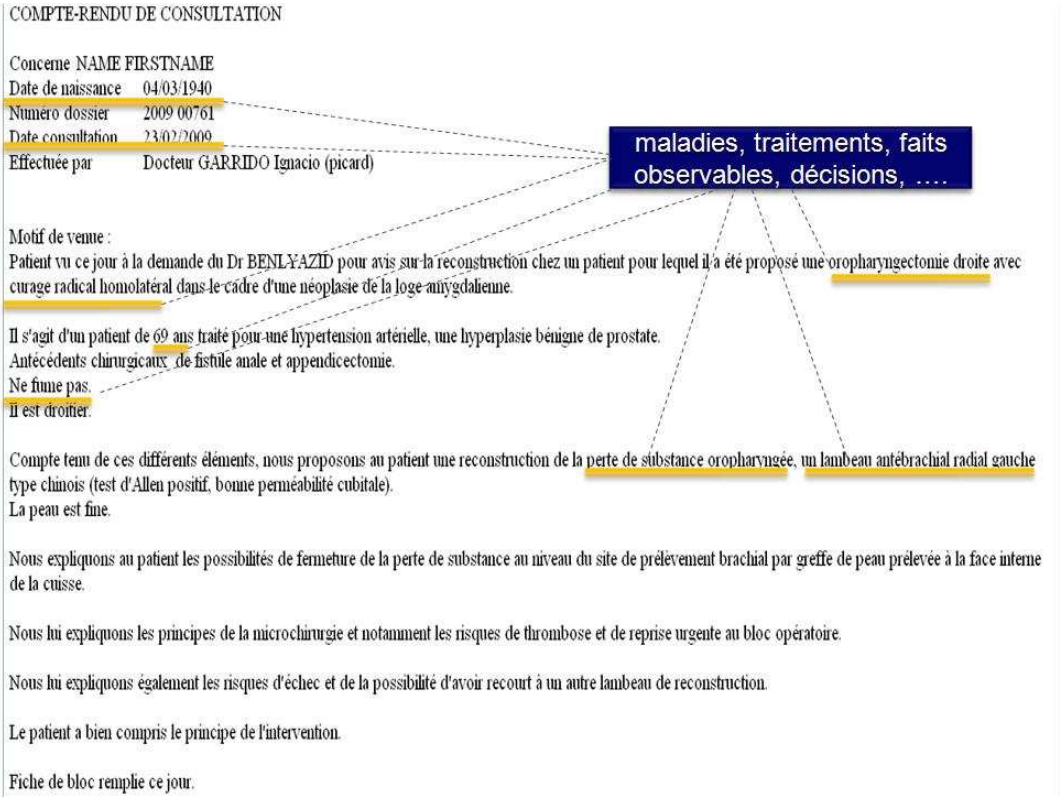
du traitement. À ce titre, il constitue un enjeu important pour l'avenir de l'assurance maladie en France. En effet, le DMP a été institué par la loi n°2004-810 du 13 août 2004 pour faciliter les échanges d'information entre professionnels de santé, éviter les actes redondants et agir contre les interactions médicamenteuses[6]. Face aux défis majeurs que représentent notamment le vieillissement de la population et le développement des maladies chroniques, le DMP est un outil moderne et performant qui permet d'améliorer la coordination, la qualité et la continuité des soins grâce à la traçabilité de l'information. Cela facilite la communication entre le médecin et le patient ainsi que la transmission des informations entre professionnels de santé.

Chaque DMP contient un enregistrement longitudinal électronique de renseignements sur la santé des patients produits suite à une ou plusieurs rencontres dans n'importe quel contexte de prestation des soins. Il comprend les données démographiques des patients, les notes d'évolution, les problèmes, les médicaments, les signes vitaux, antécédents médicaux, les vaccinations, les données de laboratoire, les rapports de radiologie, etc... Le DMP comporte essentiellement des données sur l'état médical du patient, les traitements qu'il a pris

6. `http://www.dmp.gouv.fr/web/dmp/patient/a-quoi-sert-le-dmp`

et les résultats obtenus (Ceusters et Smith, 2006). En pratique ces dossiers sont constitués d'un ensemble de documents qui sont hétérogènes de point de vue du contenu, du format et de la sémantique (Dinh et Tamine, 2010b) :

– **contenu** : Ces documents contiennent une riche source de données (figure III.2) qui peuvent être classées en trois catégories :



COMPTE-RENDU DE CONSULTATION

Concerne NAME FIRSTNAME
Date de naissance 04/03/1940
Numéro dossier 2009 00761
Date consultation 23/02/2009
Effectuée par Docteur GARRIDO Ignacio (picard)

Motif de venue :
Patient vu ce jour à la demande du Dr BENLYAZID pour avis sur la reconstruction chez un patient pour lequel il a été proposé une oropharyngectomie droite avec curage radical homolatéral dans le cadre d'une néoplasie de la loge amygdalienne.

Il s'agit d'un patient de 69 ans traité pour une hypertension artérielle, une hyperplasie bénigne de prostate.
Antécédents chirurgicaux : de fistule anale et appendicectomie.
Ne fume pas.
Il est droitier.

Compte tenu de ces différents éléments, nous proposons au patient une reconstruction de la perte de substance oropharyngée, un lambeau antébrachial radial gauche type chinois (test d'Allen positif, bonne perméabilité cubitale).
La peau est fine.

Nous expliquons au patient les possibilités de fermeture de la perte de substance au niveau du site de prélèvement brachial par greffe de peau prélevée à la face interne de la cuisse.

Nous lui expliquons les principes de la microchirurgie et notamment les risques de thrombose et de reprise urgente au bloc opératoire.

Nous lui expliquons également les risques d'échec et de la possibilité d'avoir recourt à un autre lambeau de reconstruction.

Le patient a bien compris le principe de l'intervention.

Fiche de bloc remplie ce jour.

FIGURE III.2 – Exemple d'un compte-rendu de consultation

  – Des informations textuelles et des termes médicaux décrivant l'historique médical du patient. Elles sont principalement exprimées par des phrases ou des expressions en langage naturel.
  – Des valeurs numériques relatives à la date de naissance, l'âge, des mesures d'analyses telles que le taux de cholestérol, de glucose ...
  – Des valeurs catégoriques telles que fumeur ou non fumeur, l'usage de l'alcool ...
– **format** : texte semi-structuré ou non structuré comme les comptes rendus d'intervention, de consultation, rapport anatomopathologie, images telles que les PET (Positron Emission Tomography), les IRM (Image de Résonance Magnétique) etc...
– **sémantique** : l'intelligibilité d'un document est dépendante de l'historique thérapeutique du patient contenu dans d'autres documents du même patient.

Cette hétérogénéité à triple facettes pose jusqu'à à ce jour deux principaux verrous que nous résumons dans ce qui suit :

- Paradoxe de la disponibilité/accessibilité : en effet, même si les données cliniques des patients sont disponibles (mais éparses) dans les dossiers, elles sont peu, voire pas, accessibles ni intégrables automatiquement pour servir de support à diverses applications telles que les études cliniques, les systèmes d'aide au diagnostic médical et de façon transversale, aux méthodologies d'extraction des connaissances (data mining) supportées essentiellement par des modèles d'analyse statistique des données cliniques,

- Absence d'une vision synthétique du profil du patient : en effet, seule l'analyse individuelle ou collégiale des praticiens experts permettent de "situer" le profil pathologique et thérapeutique des patients à partir de leurs dossiers médicaux. Cette vision ne peut être que partielle (toutes les données ne sont pas forcément intégrées et croisées), ce qui met à défaut le caractère prospectif, pourtant souhaité, pour les études cliniques.

3 Ressources termino-ontologiques biomédicales

Depuis les années 1950s, les technologies numériques ont donné lieu à une "explosion de l'information" qui est définie par l'accroissement considérable du nombre de publications, ce qui rend plus difficile le contrôle des bases documentaires et du repérage ou de la recherche de l'information. Pour faciliter l'accès à une telle quantité d'informations si volumineuse, plusieurs ressources terminologiques ont été conçues ayant pour objet la gestion et la représentation des connaissances. Ces ressources ont évolué d'une liste de termes techniques d'un domaine (une terminologie) à un ensemble structuré des termes et de concepts représentant le sens d'un champ d'informations (une ontologie). Dans ce qui suit, nous donnons quelques définitions de chaque notion élémentaire en ingénierie de connaissances et quelques vocabulaires contrôlés les plus utilisés dans le domaine biomédical.

3.1 Typologie des ressources termino-ontologiques

3.1.1 Terminologie

Une terminologie est un ensemble des termes, rigoureusement définis, qui sont spécifiques d'une science, d'une technique, d'un domaine particulier de l'activité humaine (Larousse, 2011). Une terminologie implique la normalisation des termes d'un domaine par la notion "*concept*". Le principal intérêt d'une terminologie est de réduire, voire supprimer, l'ambiguïté qui existe entre les termes d'un domaine et de pouvoir donc, mieux partager de l'information. En effet, puisque par définition, une terminologie de référence spécifie une norme pour un domaine donné, alors le sens de chaque terme est figé et il n'existe qu'une interprétation possible pour l'utilisateur.

La terminologie, considérée comme science, s'intéresse au recensement des concepts d'un domaine et des termes qui le désignent pour faciliter l'échange de connaissances dans une langue et d'une langue à l'autre. Il existe des terminologies de natures diverses adaptées aux différents objectifs de traitement de l'information : *classification*, *nomenclature*, *thésaurus*, et *ontologie*.

3.1.2 Classification

Une classification est une action de distribuer par classes, ou par catégories. Plus précisément, une classification est la répartition systématique en classes, en catégorie d'êtres, de choses, d'objets ou de notions ayant des caractères communs notamment afin d'en faciliter l'étude (Bourigault *et al.*, 2004). Suivant les objets considérés (les maladies, les documents, les requêtes...), se sont développés différents systèmes de classification. Les classifications sont utilisées dans tous les domaines d'activités humaines comme la biologie,la médecine, l'économie, l'astronomie, l'informatique... Les classifications portant sur un domaine limité sont généralement bien admises par les spécialistes du domaine. Les classifications à vocation universelle ne peuvent faire abstraction d'un point de vue et sont, de ce fait, l'objet de nombreuses critiques. Elles apportent cependant un éclairage sur la nature de la connaissance. La Classification Internationale des Maladies (CIM) et la Classification Commune des Actes Médicaux (CCAM) sont de bons exemples de classifications hiérarchiques dans le domaine médical bien qu'elles n'aient pas le même niveau de profondeur.

3.1.3 Nomenclature

Une nomenclature est un ensemble des termes en usage dans une science, un art, ou relatifs à un sujet donné, présentés selon une classification méthodique (Larousse, 2011). Elle désigne une instance de classification (code, tableau, liste, règles d'attribution d'identité...) faisant autorité et servant de référence à une discipline donnée. Il n'y a aucun agencement particulier des termes ni de définition explicite, l'objectif recherché étant l'*exhaustivité*. Les concepts d'un domaine sont décrits dans une nomenclature de manière complète sans se restreindre à un objectif spécifique. Une nomenclature importante dans le domaine médical est la Nomenclature Systématique des Médecines Humaine et Vétérinaire (SNOMED).

3.1.4 Thésaurus

Un thésaurus est un langage documentaire fondé sur une structuration hiérarchisée des termes désignant les concepts. Les termes y sont organisés de manière conceptuelle et reliés entre eux par des relations sémantiques. Plus concrètement, un thésaurus est un ensemble structuré de termes d'un vocabulaire, par exemple les termes techniques utilisés en médecine, représentés de façon normalisée par des descripteurs ou des mots-clés (Foskett, 1997). Il forme un répertoire alphabétique de termes normalisés pour l'analyse de contenu, le classement et donc l'indexation de documents d'information.

Un thésaurus fournit des informations sur chaque terme et ses relations ("synonyme de", "relié à") à d'autres termes. Un thésaurus indique également des termes plus spécifiques, des termes plus larges, ou des termes connexes.

En général, un thésaurus peut être construit manuellement par les experts (linguistes, documentalistes, bibliothécaires) ou automatiquement à partir d'une collection de documents. La construction manuelle d'un thésaurus doit garantir une bonne qualité des termes qui constituent des unités sémantiques, appelées *concepts*. La construction et le maintien du thésaurus par les experts sont des activités assez coûteuses notamment en termes d'expertise et de temps. Par contre, la construction automatique d'un thésaurus est moins coûteuse que la construction manuelle mais elle ne garantit pas la qualité des termes désignant les concepts ainsi que les relations entre eux car ses performances dépendent totalement de la qualité des méthodes du traitement du langage naturel et la qualité de la collection.

3.1.5 Ontologie

À l'origine, le terme "ontologie" fait référence à l'étude de l'être en tant qu'être, c'est-à-dire l'étude des propriétés générales de ce qui existe. Ce terme est repris en informatique et en science de l'information dans les années 90s. Sa première définition a été donnée par (Gruber, 1993) : "**une ontologie est une spécification d'une conceptualisation**". De cette manière, une ontologie est une description (comme une spécification formelle d'un programme) des concepts et des relations qui existent pour un objet ou un ensemble d'objets. L'objectif principal d'une ontologie est donc de partager et de réutiliser des connaissances propres à un domaine donné. De ce fait, une ontologie peut être conçue pour modéliser un ensemble de connaissances d'un domaine spécifique, comme par exemple le domaine de la médecine, le génie logiciel, l'aviation, etc. Une ontologie décrit généralement les notions suivantes considérées comme ses principaux constituants :

- *Un concept*, appelé également classe, peut représenter un objet matériel (par exemple, un comprimé de médicament), une notion (par exemple, la quantité) ou bien une idée,
- *Un attribut*, appelé également propriété, fonctionnalité, caractéristique ou paramètre, est associé à un tel ou tel objet,
- *Une relation* reflète les liens que les objets peuvent avoir entre eux,
- *Un événement* indique un ou des changements des attributs ou des relations.

3.2 Quelques ressources termino-ontologiques du domaine biomédical

3.2.1 Thésaurus MeSH

Afin d'indexer, classer et rechercher des documents (notamment ceux de la base MEDLINE), la NLM a créé en 1954 le thésaurus MeSH (Medical Subject Headings). Depuis, il a été régulièrement mis à jour. La traduction de MeSH vers le français est assuré par l'INSERM [7] qui met la version bilingue à la disposition de la communauté francophone. MeSH comprend essentiellement des termes qui désignent les concepts biomédicaux, des descripteurs, des relations et des qualificatifs. Nous détaillons par la suite ces éléments et leurs rôles.

– **Terme** : Un terme est un mot ou un ensemble de mots exprimant une

7. Institut Nationale de la Santé Et de la Recherche Médicale - `http://www.inserm.fr/`

notion particulière,

- **Concept** : Un concept comprend un ou plusieurs termes synonymes et porte le nom d'un de ces termes, dit *terme préféré*,
- **Relation** : Dans MeSH il existe deux types de relations entre les concepts : les relations hiérarchiques et les relations associatives (associé à). La hiérarchie dans MeSH est représentée par un code reflétant l'arborescence auquel le concept appartient (*cf.* la figure III.3) et peut véhiculer plusieurs sens :
  1. relation "*est une partie de*" (méronymie), par exemple le concept "*doigt*" (A01.378.800.667.430) *est une partie de* "*main*" (A01.378.800.667) .
  2. relation "*est un type de*" (hyponymie), par exemple le concept *pouce* (A01.378.800.667.430.705) *est un type de* "*doigt*" (A01.378.800.667.430).
  3. relation "*est sémantiquement proche de*" (aboutness), par exemple le concept "*sécurité*" (G03.850.110.060.075) *est sémantiquement proche de* "*accidents*" (G03.850.110).

C <maladies>
C04 <tumeurs>
C18 <métabolisme et nutrition, maladies>
C18.452 <métabolisme, maladies>
C18.452.090 <amyloïdose>
C18.452.394 <troubles du métabolisme glucidique>
C18.452.394.750 <diabète>
C18.452.394.750.124 <diabète de type 1>
C18.452.394.750.124.960 <Wolfram, syndrome>
C18.452.394.750.149 <diabète de type 2>
C18.452.394.750.774 <état prédiabétique>
C18.452.394.937 <glycosurie>
C18.654 <troubles nutrition>
C23 <troubles liés environnement>

FIGURE III.3 – Extrait de l'arborescence C (domaine 'Maladie') de MeSH

La relation associative dans MeSH, exprimée par l'expression "*Voir aussi*", relie entre deux concepts proches d'une façon libre et non clairement définie.

- **Descripteur** : Un descripteur est constitué d'un ou de plusieurs concepts de significations proches et porte le nom d'un de ces concepts, dit préféré. Les autres concepts, dits subordonnés, présentent une relation sémantique avec le concept préféré, soit une relation hiérarchique (générique ou spécifique), soit une relation associative (associé). Pour un descripteur donné,

les termes d'un concept sont synonymes entre eux, mais ne sont pas synonymes des termes d'un autre concept, car ils héritent de la relation sémantique entre ces concepts. Enfin, tous les termes d'un descripteur sont des équivalents documentaires ou termes d'entrée et renvoient au descripteur dans le cadre de l'indexation et de la recherche documentaire. Les descripteurs MeSH sont répartis en 16 catégories recouvrant la biologie, la médecine et les domaines connexes (*cf.* le tableau III.1).

TABLEAU III.1 – Les différentes catégories ou domaines du MeSH

[A]	Anatomie
[B]	Organisme
[C]	Maladies
[D]	Produits chimiques et pharmaceutiques
[E]	Techniques analytiques, diagnostiques, thérapeutiques et équipements
[F]	Psychiatrie et psychologie
[G]	Phénomènes et processus
[H]	Disciplines et professions
[I]	Anthropologie, enseignement, sociologie, et phénomènes sociaux
[J]	Technologie, industrie et agriculture
[K]	Sciences humaines
[L]	Sciences de l'information
[M]	Individus
[N]	Santé
[V]	Caractéristiques d'une publication
[Z]	Lieux géographiques

Chaque catégorie de descripteurs est structurée en arborescence hiérarchique pouvant comprendre jusqu'à onze niveaux de hiérarchie. Chaque descripteur est représenté par un code alphanumérique, la lettre indiquant la catégorie et la séquence numérique précisant la localisation dans la hiérarchie (*cf.* la figure III.3).
Certains descripteurs ont plusieurs localisations, au sein de la même catégorie ou de catégories différentes, et plusieurs codes alphanumériques représentant chacun une localisation. Par exemple, le descripteur "*Pain*" appartient à plusieurs hiérarchies : *C10.597.617*, *C23.888.592.612*, *C23.888.646*, *F02.830.816.444* et *G11.561.600.810.444*.

– **Qualificatif** : Les qualificatifs sont des entrées MeSH qui peuvent être utilisées seuls ou associés à un descripteur pour décrire un aspect particulier. Par exemple, le sens du descripteur "*diabète de type ii*" accompagné du qualificatif "*thérapeutique*" (*diabète de type ii/thérapeutique*) évoque le traitement du diabète non insulinodépendant. Alors que seul, le descripteur "*diabète de type ii*" fait référence à la maladie en général.

3.2.2 Nomenclature SNOMED

La SNOMED est une nomenclature pluri-axiale couvrant tous les champs de la médecine et de la dentisterie humaines, ainsi que de la médecine vétérinaire. SNOMED-CT [8] représente la dernière version de la nomenclature mais seule la version SNOMED 3.5 (appelée également SNOMED International) a été traduite en français. La SNOMED 3.5 comporte 11 axes (*cf.* le tableau III.2).

TABLEAU III.2 – Les onze axes de la SNOMED

Axe	**Nom de l'axe**
A	Agents physiques
C	Produits chimiques
D	Diagnostics
F	Fonctions
G	Qualificatifs
J	Métiers
L	Organismes vivants
M	Morphologie
P	Procedure
S	Contexte Social
T	Topographie

Dans chaque axe, les concepts sont représentés par une série de termes au sein de laquelle on peut distinguer une formulation préférée et des synonymes de diverses natures syntaxiques. La version française comporte 97 485 concepts désignés par 144 796 termes.
Par ailleurs, chaque axe représente une hiérarchie simple de concepts qui peuvent représenter une combinaison de concepts. Par exemple, le concept "*pemphigoïde bulleux*" (D0-10431) est la combinaison des concepts "*peau, SAI ; cutané*" (T-01000), "*acantholyse*" (M-51551) et "*ampoule, SAI*" (M36-760).

La recherche en informatique médicale a montré que la SNOMED est la terminologie la plus adaptée à l'indexation des informations du dossier patient (Pereira *et al.*, 2009). Cependant, elle contient des éléments non pertinents à l'indexation. Ce sont les éléments de l'axe G contenant les qualificatifs et termes de relations qui n'ont pas de sens lorsqu'ils ne sont pas reliés aux autres termes SNOMED, par exemple : "*sans*", "*disponible*", "*diagnostic provisoire*", etc.

8. SNOMED Clinical Terms

3.2.3 Ontologie de gènes - GO

L'ontologie de gènes (GO - *Gene Ontology*) est une ressource terminologique destinée à structurer la description des gènes et des produits géniques dans le cadre d'une ontologie commune à toutes les espèces. Ce projet, qui s'inscrit dans la démarche plus large d'Open Biomedical Ontologies (OBO) regroupant d'autres projets bio-informatiques dans le domaine biomédical, poursuit trois objectifs :

- gérer et enrichir son vocabulaire contrôlé décrivant les gènes et leurs produits,
- gérer les *annotations*, c'est-à-dire les informations rattachées aux gènes et à leurs produits,
- fournir les outils permettant d'accéder aux informations structurées dans le cadre du projet.

La base GO est conçue comme un graphe orienté acyclique, chaque terme étant en relation avec un ou plusieurs termes du même domaine, et parfois d'autres domaines. Le vocabulaire GO est construit pour n'être pas dépendant des espèces considérées, avec des termes applicables à la fois aux organismes multicellulaires et unicellulaires, aux eucaryotes et aux procaryotes.

Dans le cadre de GO, les propriétés des produits géniques sont décrites selon trois axes :

- les composants cellulaires auxquels ils s'appliquent, qu'il s'agisse du milieu intracellulaire ou de l'environnement extracellulaire,
- les fonctions moléculaires réalisées, par exemple structurelles ou catalytiques pour une protéine
- les processus biologiques, c'est-à-dire les transformations moléculaires nécessaires au fonctionnement d'entités biologiques intégrées.

Chaque terme GO est défini par l'ontologie du projet à travers :

- un nom de terme, qui peut être un mot unique ou une suite de mots,
- un identifiant unique alphanumérique,
- une définition avec ses sources citées,
- un espace de nom indiquant le domaine auquel il appartient.

Un terme GO peut également avoir des attributs supplémentaires facultatifs (*cf.* le tableau III.3) :

– des synonymes, qui peuvent être classés comme exactement équivalents au nom de terme, plus large, plus restrictif ou en rapport avec lui,
– des références à des concepts équivalents dans d'autres bases de données,
– des commentaires sur la signification ou l'utilisation du terme.

TABLEAU III.3 – Description d'un terme dans GO

id :	GO :0000016
name :	lactase activity
namespace :	molecular_function
def :	"Catalysis of the reaction : lactose + H2O = D-glucose + D-galactose." [EC :3.2.1.108]
synonym :	"lactase-phlorizin hydrolase activity" BROAD [EC :3.2.1.108]
synonym :	"lactose galactohydrolase activity" EXACT [EC :3.2.1.108]
xref :	EC :3.2.1.108
xref :	MetaCyc :LACTASE-RXN
xref :	Reactome :20536
is_a :	GO :0004553 ! hydrolase activity, hydrolyzing O-glycosyl compounds

3.2.4 Méta-thésaurus UMLS

Développé par la NLM, l'UMLS (Unified Medical Language System) est un système qui réunit trois bases de connaissances : le métathesaurus, le réseau sémantique, et le SPECIALIST Lexicon. Ils peuvent être utilisés séparément ou ensemble. Chacune de ces bases constitue une unité dans l'UMLS que nous présentons dans ce qui suit :

– Le **métathesaurus** constitue la base unifiée des concepts issus de plus que 150 terminologies biomédicales (dont MeSH, CIM-10 et SNOMED) en 17 langues. UMLS regroupe les différents termes synonymes (issus des différentes terminologies) sous un même concept (*cf.* la figure III.4). Par exemple, les termes "*Addison Disease*" du MeSH, "*Addison's disease*" de la SNOMED CT, "*Bronzed disease*" de la SNOMED 3.5 sont regroupés sous le même concept "*Addison's disease*" (*cf.* la figure III.4). Chaque concept possède un identifiant unique, un terme préféré, un ou plusieurs types sémantiques (les types sémantiques sont définis dans le réseau sémantique de l'UMLS) et une ou plusieurs définitions. Le métathesaurus contient également des relations entre les concepts (par exemple des relations de hiérarchie, proximité, etc...).
  La présence du français dans le méta-thésaurus est faible et ne présente que moins de 2% des concepts (Delbecque et Zweigenbaum., 2005). Les versions françaises de la CIM-10 et du SNOMED 3.5 n'y sont pas encore introduites ; par contre la version française du MeSH y est, et elle est régulièrement mise à jour.

– Le **réseau sémantique** a pour objectif de fournir une catégorisation cohérente de tous les concepts représentés dans le méta-thésaurus UMLS et de fournir un ensemble de relations utiles entre ces concepts.

FIGURE III.4 – Regroupement des termes synonymes dans l'UMLS

Toutes les informations sur les concepts spécifiques se trouvent dans le méta-thésaurus. Le réseau fournit des informations sur l'ensemble des types sémantiques, ou catégories, qui peuvent être affectés à ces concepts, et il définit l'ensemble des relations qui peuvent exister entre eux. Il contient 133 types sémantiques et 54 relations. La figure III.5 illustre un exemple de type sémantique relié à d'autres types sémantiques.

– Le **SPECIALIST Lexicon** contient les informations syntaxiques, morphologiques et orthographiques nécessaires au traitement automatique de la langue anglaise. Chacune de ces entrées possède une forme de base (lemme), une catégorie syntaxique, un identifiant unique et éventuellement des variantes orthographiques. Il est généralement utilisé pour des tâches de traitement automatique de la langue.

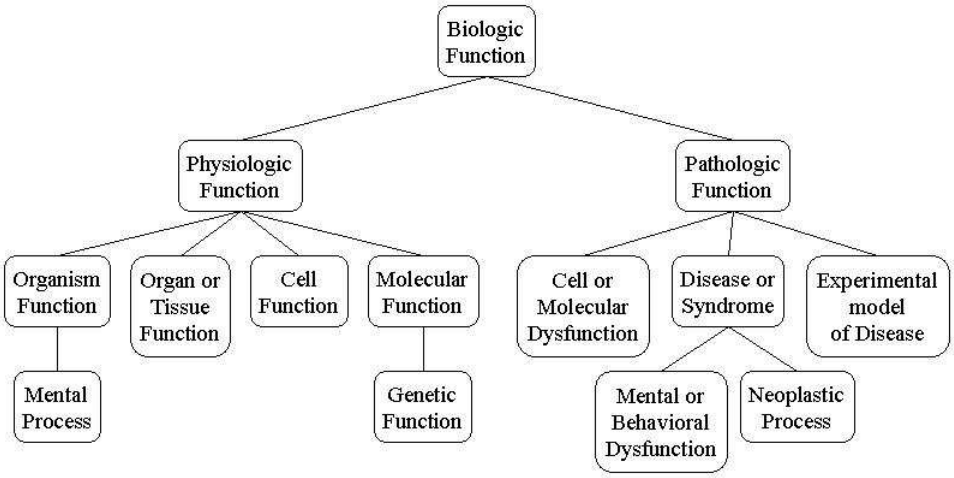


FIGURE III.5 – Le réseau sémantique *Biologic Function*

Nous présentons à ce niveau les différentes approches d'extraction des concepts biomédicaux ainsi que les outils d'extraction des concepts qui sont prédéfinis dans des ressources termino-ontologiques du domaine biomédical que nous avons abordées dans les sections précédentes.

4 Extraction des concepts biomédicaux

L'extraction des concepts est un processus de reconnaissance des expressions pertinentes dans le document afin de mettre en évidence les sujets les plus significatifs du document. Ceci est une des tâches les plus difficiles qui représentent de grands défis pour les chercheurs dans le domaine de l'extraction d'information (Cowie et Lehnert, 1996; Gaizauskas *et al.*, 2000; Krauthammer et Nenadic, 2004). Malgré la grande disponibilité des ressources termino-ontologiques qui ont été développées, plusieurs travaux de recherche ont rapporté qu'il existe un nombre de termes qui ne sont pas pertinents pour le document même s'ils sont définis dans la ressource (Gaizauskas *et al.*, 2000; Hirschman *et al.*, 2002; Tuason *et al.*, 2004). En général, les termes extraits du document doivent être vérifiés et confirmés par un ou plusieurs experts du domaine en raison de l'ambiguïté des termes. Par exemple, le gène "bride of sevenless" a pour acronyme "boss" dans FlyBase[9] ou la protéine "yotiao" dans Uniprot[10] a le nom d'un plat chinois.

9. `http://flybase.org/`
10. `http://www.uniprot.org/uniprot/`

4.1 Principales approches d'extraction de concepts

Nous présentons dans les sections suivantes les différentes approches d'extraction automatique des concepts, qui peuvent être subdivisées en quatre catégories : (1) *approche basée sur les règles linguistiques*, (2) *approche basée sur l'apprentissage automatique*, (3) *approche basée sur la recherche dans un dictionnaire*, et (4) *approche basées sur des mesures statiques.*

4.1.1 Approche basée sur des règles linguistiques

Les méthodes d'extraction des concepts basées sur des règles linguistiques consistent à définir des règles particulières pour décrire les entités nommées, les termes techniques qui désignent les concepts d'un domaine particulier. En général, les règles sont définies (manuellement) par les linguistes et les experts du domaine en se basant sur les caractéristiques de la langue comme les caractéristiques de l'orthographe, du lexique ou les caractéristiques morpho-syntaxiques. De plus, les listes des affixes, des suffixes ou des acronymes peuvent être utilisées pour enrichir les règles de base. Par exemple, une méthodologie d'identification de concepts basée sur les règles grammaticales a été proposée par (Ananiadou, 1994). Il s'agit d'une morphologie à quatre niveaux (non-negative compounding, class I affixation, class II affixation, native compounding) permettant de gérer les patrons de formations des termes simples et complexes. Leur système utilise donc une grammaire d'unification et un lexique des termes composés avec des instances des affixes spécifiques, des radicaux, et des formes composées des caractères grecs et latins. Ils ont distingué les deux notions *termes* et *mots* : en général, les termes sont constitués des mots et désignent les concepts dans un domaine particulier. Par exemple, les règles pour former un terme composé à partir des mots et à partir des suffixes peuvent être définies comme dans le tableau III.4.

Règles des termes composés	**Règles de la suffixation**
terme $\longrightarrow$ terme + mot	terme $\longrightarrow$ mot + terme_suffixe
terme $\longrightarrow$ terme + terme	terme $\longrightarrow$ terme + terme_suffixe
terme $\longrightarrow$ mot + terme	mot $\longrightarrow$ mot + mot_suffixe
mot $\longrightarrow$ mot + mot.	terme $\longrightarrow$ terme + mot_suffixe

TABLEAU III.4 – Règles de formation des termes composés

Inspirés de cette approche, (Gaizauskas *et al.*, 2003) ont proposé une grammaire terminologique non-contextuelle pour identifier les noms des protéines. Trois étapes principales sont à distinguer : (1) *analyse morphologique*, (2) *recherche lexicale*, et (3) *analyse grammaticale terminologique*. L'analyse morphologique a pour but de reconnaître les affixes biomédicales (*–ase*, *–in*...) qui

indiquent les noms de protéines ou d'enzymes. La recherche lexicale permet de reconnaître les termes biomédicaux composés de plusieurs mots qui sont définis dans les bases lexicales biomédicales comme CATH (Orengo *et al.*, 1997) et SCOP (Andreeva *et al.*, 2004). L'analyse grammaticale terminologique a pour but d'analyser les catégories grammaticales (nom, verbe, adjectif, adverbe, ...), les caractéristiques morphologiques et lexicales des termes composés et de les combiner en une unité multi-token unique ou en un terme composé de plusieurs mots.

D'autres systèmes d'extraction à base de règles simples utilisent des caractéristiques orthographiques et lexicales pour reconnaître les noms de protéines (Fukuda *et al.*, 1998; Hou, 2003; Narayanaswamy *et al.*, 2003). Par exemple, le système décrit dans (Fukuda *et al.*, 1998), appelé PROPER (PROtein Propernoun phrase Extracting Rules), utilise les caractéristiques de la description des noms propres dans les documents médicaux et biologiques, et ne nécessite aucun dictionnaire de termes spécifiques. Les auteurs ont observé que les mots caractéristiques contenant des majuscules, des chiffres numériques et de symboles spéciaux sont souvent utilisés pour décrire les noms de protéines. Ces mots fournissent une quantité importante d'informations et peuvent être considérés comme le noyau de noms de protéines, appelés "core terms" ou "terme de base". Par ailleurs, les mots-clés (par exemple, "protéine", "enzyme", "kinase"...) qui décrivent la fonction et les caractéristiques des termes techniques sont mentionnés comme "feature terms" ou "terme caractéristique". Par exemple, dans le terme "récepteur EGF", "EGF" est un terme de base et "récepteur" est un terme caractéristique. Les termes de base sont extraits en utilisant des caractéristiques orthographiques et lexicales (par exemple, lettres majuscules, chiffres numériques, et/ou symboles spéciaux, sauf ceux de plus de 9 caractères et qui se composent d'un tiret ('-') suivi d'un mot en minuscule).

L'avantage des méthodes d'extraction de concepts basées sur des règles est que les règles sont capables de reconnaître les termes en fonction des caractéristiques orthographiques et lexicales d'une langue spécifique. Ceci peut être facilement fait en utilisant des expressions régulières. Toutefois, ces approches sont connues pour être extrêmement coûteuses pour le développement, et nécessitent les connaissances linguistiques des développeurs. Une autre lacune de ces méthodes porte sur son application à d'autres entités, ce qui est généralement difficile et demande de définir de nouvelles règles appropriées.

4.1.2 Approche basée sur l'apprentissage automatique

Les approches basées sur l'apprentissage automatique utilisent des corpus manuellement annotés pour entraîner les classificateurs qui considèrent plusieurs caractéristiques des instances textuelles pour associer les termes tech-

niques à des classes prédéfinies. Les modèles de Markov cachés (HMM - Hidden Markov Models) ont été utilisés pour extraire les noms des gènes et des protéines (Collier *et al.*, 2000; Morgan *et al.*, 2003; Shen *et al.*, 2003). Chaque HMM, qui est essentiellement un automate à n états, se compose essentiellement d'un ensemble de paramètres (*cf.* la figure III.6) :

- x : ensemble des états,
- y : observations possibles,
- a : probabilités de transition d'un état à l'autre,
- b : probabilités d'émission.

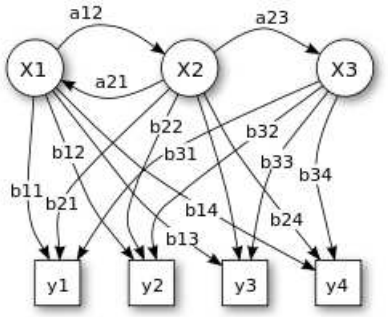


FIGURE III.6 – Les paramètres d'un modèle de Markov caché

Notons que l'adjectif "caché" employé pour caractériser le modèle traduit le fait que l'émission d'une donnée à partir d'un état est aléatoire.

(Collier *et al.*, 2000) ont utilisé le modèle de Markov caché en exploitant les caractéristiques orthographiques (par exemple, "termes composés des lettres majuscules et des chiffres", "termes commencés par une lettre majuscule", etc.) pour identifier les noms de gènes et de protéines parmi les dix classes prédéfinies (par exemple, "PROTEIN", "DNA", "RNA", "VIRUS"...). Plus spécifiquement, ils ont implémenté un modèle HMM d'ordre 2 en considérant l'état de la classe précédente par rapport à la classe courante. En intégrant les caractéristiques orthographiques au modèle HMM par une combinaison linéaire, ils ont observé une dégradation en terme de F-mesure pour les classes "PROTEIN", "DNA" et "RNA" par rapport au modèle HMM sans ces caractéristiques.

Les machines à vecteurs de support (SVM) constituent une méthode de classification binaire par apprentissage supervisé ; elle fut introduite par (Cortes et Vapnik, 1995). Pour deux classes d'exemples donnés, le but de SVM est de trouver un classificateur linéaire optimal, appelé *hyperplan*, qui sépare les données et maximise la distance entre ces deux classes. La figure III.7 illustre les vecteurs de support possibles qui séparent les deux groupes de données (points noirs *vs.* points blancs). Pour garantir la sécurité lors de la classification d'un nouvel exemple, on doit trouver un hyperplan dont la marge est maximale. Cela

signifie que la distance entre l'hyperplan et les exemples doit être la plus grande possible. Dans ce cas, l'hyperplan optimal devient un *hyperplan de marge maximale* ou *séparateur à vaste marge*.

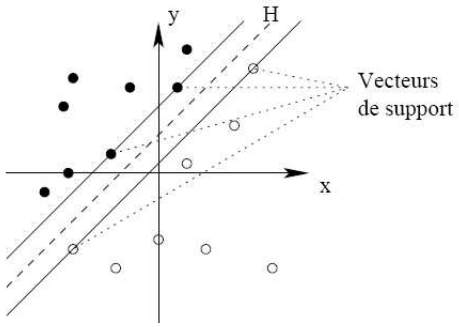


FIGURE III.7 – Illustration des vecteurs de support

Pour classifier plusieurs classes, il existe des SVM multi-classe qui sont une extension des SVM binaires. Par exemple, la méthode la plus intuitive pour la gestion des classificateurs multi-classes consiste à construire autant de classificateurs SVM que de classes (Platt *et al.*, 2000; Duda *et al.*, 2001).

Le travail de (Kazama *et al.*, 2002) consiste à entraîner les SVM multi-classes sur un corpus manuellement annoté, appelé GENIA, pour la tâche de reconnaissance d'entités nommées (notamment les protéines, les ADS, et les types de cellules). Plus précisément, leur méthode vise à prédire les balises composites B-I-O [11] indiquant une entité nommée en se basant sur les éléments comme les catégories grammaticales, les préfixes, les suffixes, les états d'un modèle Markov caché... Ils ont rapporté une valeur de 50% en terme de F-mesure. Ils ont conclu que la considération des classes précédemment identifiées ainsi que les suffixes sont utiles tandis que d'autres caractéristiques comme la catégorie grammaticale ou les préfixes n'ont pas d'influence positive sur les performances de l'identification des entités nommées.

(Yang et Chute, 1994a) ont proposé une méthode d'extraction de concepts est basée sur la technique LLSF (Linear Least Squares Fit [12]) pour identifier les concepts à partir du texte. Cette méthode consiste à apprendre les catégories sémantiques (concepts) manuellement annotés par les indexeurs humains dans un corpus de documents d'apprentissage. Deux matrices sont générées à partir du corpus d'apprentissage : la première matrice, dénotée A, correspond aux mots dans chaque document et la deuxième, dénotée B, correspond aux catégories sémantiques associées à chaque document (*cf.* la figure III.8). La technique LLSF ici a été définie comme le problème de réduire des erreurs de

11. B : Begin ; I : Inside ; O : Outside
12. La méthode des moindres carrés a été indépendamment élaborée par Legendre en 1805 et Gauss en 1809

l'identification des catégories sémantiques. Cela revient à trouver une matrice $F_{l\times n}$ qui minimise la somme des carrés comme suit :

$$\sum_{i=1}^{m} \|\vec{e_i}\|_2^2 = \sum_{i=1}^{m} \|F\vec{a_i}^T - \vec{b_i}^T\|_2^2 = \|FA^T - B^T\|_F^2 \qquad \text{(III.1)}$$

où

- $A_{m\times n}$ et $B_{m\times l}$ sont les deux matrices d'apprentissage
- A^T et B^T sont des matrices transposées
- $\vec{a_i}^T$ et $\vec{b_i}^T$ représentent le document i et l'ensemble de catégories sémantiques correspondant.
- $\vec{e_i} = F\vec{a_i}^T - \vec{b_i}^T$ représente l'erreur de la matrice F lors de la transformation de $\vec{a_i}$ en $\vec{b_i}$

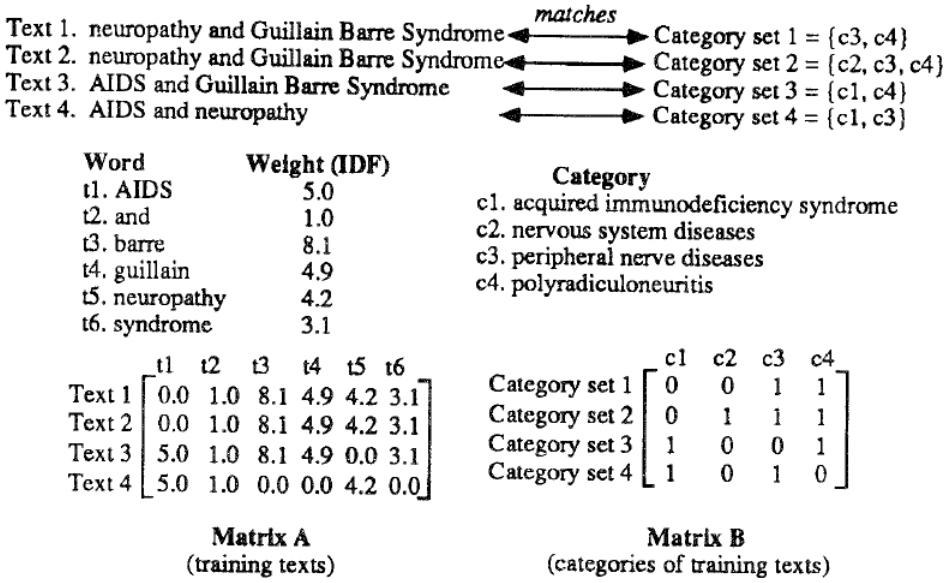


FIGURE III.8 – Exemples de la représentation des matrices LLSF

Cependant, les méthodes basées sur l'apprentissage supervisé sont confrontés à une grande difficulté comme l'insuffisance de données (data sparseness) car les données d'entraînement ne sont pas toujours disponibles et ne sont pas facilement mises à jour.

4.1.3 Approche basée sur la recherche dans un dictionnaire

Un des premiers travaux d'extraction des concepts à partir de textes biomédicaux en exploitant des ressources terminologiques externes, est l'algorithme d'identification de concepts mis en œuvre dans le système SAPHIRE (Hersh et Greenes, 1990). Leur algorithme essaie de chercher tous les synonymes de chaque mot dans un texte et associe toutes les combinaisons possibles des mots synonymes afin de générer un terme complexe qui est par la suite comparé aux entrées d'un dictionnaire des termes qui désignent les concepts dans la ressource. Inspiré de cette idée, MetaMap est un programme développé à la NLM pour associer aux textes biomédicaux les concepts du Méta-thésaurus UMLS (Aronson *et al.*, 2004b). Pour chaque terme complexe (groupe de mots consécutifs), les variantes sont générées en utilisant les connaissances dans la base lexicale SPECIALIST Lexicon de l'UMLS et une base de données supplémentaire de synonymes. Une variante d'un terme est composée de ses acronymes, ses abréviations, ses synonymes, les combinaisons significatives de ces derniers, les variantes flexionnelles et l'orthographe (Aronson, 1996). Les termes candidats dénotant des concepts UMLS sont récupérés lorsqu'elles contiennent au moins une des variantes générées. MetaMap est la principale composante de l'outil de MTI (Aronson *et al.*, 2004b), qui intègre plusieurs méthodes d'extraction de concepts ("PubMed related citations", "Restriction to MeSH") pour l'indexation des citations dans MEDLINE. Il retourne d'abord plusieurs concepts UMLS candidats et puis les concepts MeSH sont filtrés grâce aux associations entre UMLS et MeSH.

Les approches d'extraction de concepts basées sur la recherche dans un dictionnaire utilisent des ressources terminologiques pour comparer les instances textuelles à des entrées des concepts dans le dictionnaire qui enregistre tous les entrées possibles des concepts. La recherche dans le dictionnaire est basée sur la comparaison *exacte* ou *approximative* des chaînes de caractères entre les fragments du document et les entrées du dictionnaire. Plusieurs travaux ont rapporté que la méthode d'extraction basée sur la recherche exacte est simple, naïve et inefficace (Hirschman *et al.*, 2002), (Tuason *et al.*, 2004).

Les auteurs dans (Krauthammer *et al.*, 2000) ont proposé une méthode basée sur la recherche approximative des concepts représentés par des entrées dans un dictionnaire pour identifier les noms des gènes et des protéines ainsi que leurs variations. Dans leur approche, les dictionnaires de gènes et de protéines ainsi que le texte cible sont encodés en utilisant le codage de nucléotides (A, C, G, T) ; puis les techniques d'alignement d'ADN et de protéines dans les bases de données d'ADN et de protéines sont appliquées sur le texte converti pour identifier les chaînes de caractères (A, C, G, T) qui sont similaires aux gènes et aux protéines dans la base de données.

(Zhou *et al.*, 2006b) ont proposé une approche d'extraction des concepts basée sur la recherche approximative des entrées dans un dictionnaire en tenant compte des variations (lexicales, morphologiques) des termes. L'idée principale de leur approche est de capturer les mots significatifs au lieu de tous les mots d'un concept donné. MaxMatcher définit une mesure de similarité qui permet de faire une recherche approximative afin de trouver les termes qui représentent potentiellement les concepts candidats. Étant donné un terme t qui est constitué par un ensemble de mots simples $t = \{w_1, w_2, ..., w_m\}$ dans le texte, la similarité entre chaque mot constituent et le concept c, dénoté par un ensemble de termes $c = \{t_1, t_2, ..., t_n\}$, a été définie dans (Zhou *et al.*, 2006b) comme suit :

$$I(w_i, c) = max\{I(w_i, t_j) | j \leq n\} \tag{III.2}$$

où

- w_i est le mot constituent du terme t,
- t_j est le terme d'entrée j du concept c,
- n est le nombre de termes d'entrée du concept c,
- $I(w_i, t_j)$ est la similarité entre le mot w_i et le terme t_j, calculée comme suit :

$$I(w_i, t_j) = \begin{cases} \frac{1}{N(w_i) \times \sum_{w_{jk} \in t_j} 1/N(w_{jk})} & \text{if } w_i \in t_j \\ 0 & \text{otherwise} \end{cases} \tag{III.3}$$

où $N(x)$ est le nombre de concepts dont le terme contient le mot x. L'idée de base de cette mesure ressemble à la mesure IDF en RI : plus le mot apparaît dans plusieurs concepts, moins il est important dans la description du concept. La somme dans le dénominateur indique l'importance de chaque mot constituant le terme désignant le concept c.

Dans une étude comparative sur le corpus GENIA, leur méthode d'extraction approximative atteint une précision de 71.60% et un rappel de 75.18%.

4.1.4 Approche basée sur des mesures statistiques

Les approches basées sur les mesures statistiques ont été proposées pour identifier des termes techniques désignant les concepts. Par exemple, les auteurs dans (Frantzi *et al.*, 2000) ont proposé une méthode d'extraction des termes techniques, appelée *C/NC-value*, pour identifier automatiquement des termes dans plusieurs domaines (e.g., arts, médecine). Cette méthode a été utilisée plus tard pour reconnaître les termes désignant les concepts biomédicaux de la littérature (Hliaoutakis *et al.*, 2009). La méthode *C/NC-value* combine les informations statistiques et linguistiques pour extraire des termes et des sous-termes désignant les concepts spécifiques : d'abord, un ensemble de **règles de**

filtrage sont appliquées pour identifier les expressions ou chaînes de caractères candidates :

- $N + N$
- $(Adj|N) + N$
- $((Adj|N) + |((Adj|N) * (NP)?)(Adj|N)*)N$

où N est un nom, Adj est un adjectif et P est une préposition.

Ensuite, les mesures *C/NC-value* sont définies pour déterminer si une chaîne de caractères a est un terme ou un sous-terme significatif ou non. La partie C_{value}, qui indique l'importance du terme candidat dans le texte, est calculée comme suit :

$$C_{value}(a) = \begin{cases} log_2|a| \times f(a) & \text{si a n'est pas un sous-terme} \\ log_2|a| \times \left(f(a) - \frac{\sum_{b \in T_a} f(b)}{P(T_a)}\right) & \text{sinon} \end{cases} \quad \text{(III.4)}$$

où
- $|a|$ est la longueur du terme a,
- $f(x)$ est la fréquence du terme x dans le corpus,
- T_a désigne l'ensemble de termes contenant a,
- $P(a)$ désigne le nombre de termes contenant a.

La partie NC, qui prend en compte le contexte du terme candidat a et qui affecte un poids à chaque mot composant de a (appelé mot contextuel), est calculée comme suit :

$$NC_{value}(a) = 0.8 \times C_{value}(a) + 0.2 \times \sum_{b \in C_a} f_a(b) \times \frac{t(b)}{n} \quad \text{(III.5)}$$

où
- b est le mot contextuel (nom, verbe ou adjectif) du terme candidat a,
- C_a est le contexte du terme candidat a (ensemble de mots contextuels uniques de a),
- $t(b)$ est le nombre de termes contenant le mot contextuel b,
- n est le nombre total des termes considérés,
- $f_a(b)$ est le nombre de co-occurrences du mot b qui apparaît comme un mot contextuel du terme candidat a, c-à-d le nombre de fois le mot b apparaît ensemble avec a.

(Ruch, 2006) a introduit une méthode différente d'extraction de concepts basée sur la RI. Chaque concept MeSH (ses synonymes et sa description) est indexé comme un document unique. Étant donné un texte, qui devient la requête

du système de RI (le classificateur), une liste de concepts MeSH est extraite à partir de ce texte en se basant sur un modèle vectoriel. Afin d'améliorer la précision, ils ont intégré les règles d'appariement (expressions régulières) en utilisant les radicaux (stems) et les groupes nominaux (NP - noun phrases). Ils ont évalué leur méthode d'extraction en utilisant deux terminologies (MeSH, GO) séparément. Ils ont conclu que la combinaison de ces approches donne des résultats supérieurs à ceux obtenus avec chacune des méthodes séparément. De plus, le thésaurus MeSH est plus adéquat que l'ontologie des gènes GO pour extraire les concepts à partir des documents biomédicaux.

D'autres approches combinent plusieurs méthodes symboliques et statistiques pour identifier les concepts. Par exemple, l'outil MTI (Medical Text Indexer) (Aronson *et al.*, 2004b) intègre plusieurs méthodes d'extraction de concepts pour l'indexation des documents dans la base de données MEDLINE : d'abord, chaque méthode d'extraction de concepts (e.g., méthode symbolique de MetaMap (Aronson, 2001a), méthode statistique des K plus proches voisins (Kim *et al.*, 2001), tri-gramme) est appliquée pour extraire une liste de concepts issus du méta-thésaurus UMLS. Ensuite, les listes de concepts candidats générées par chaque méthode d'extraction sont fusionnées grâce aux associations de l'UMLS vers MeSH pour dégager à la fin la liste des concepts MeSH, qui sont finalement utilisés pour représenter la sémantique du document.

4.2 Principaux outils d'extraction de concepts

Nous présentons dans cette section les outils d'extraction de concepts **accessibles en ligne** uniquement ou **téléchargeables** comme un outil open source, à savoir PubMed ATM [13], MetaMap (Aronson, 2001a), MTI (Medical Text Indexer) (Aronson *et al.*, 2004b), MaxMatcher (Zhou *et al.*, 2006b).

4.2.1 PubMed ATM

PubMed ATM est un service implémenté dans le portail de PubMed visant à associer un morceau de texte (par exemple, la requête de l'utilisateur) à des termes ou concepts dans les différentes tables et les index associés dans l'ordre suivant [14] : (1) table des termes désignant les **concepts MeSH** ainsi que les informations supplémentaires comme *qualificatifs*, *types de publication*, *substances ...* ; (2) table des **Journaux**, (3) table des **Auteurs**. Étant donnée une requête, PubMed essaie de localiser les groupes de mots les plus longs qui sont sauvegardés dans les tables de concepts. Lorsqu'un terme désignant un

13. http://www.ncbi.nlm.nih.gov/pubmed
14. http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.How_PubMed_works_aut

concept est trouvé, le processus de recherche du terme candidat est terminé. Ensuite, les termes retrouvés sont groupés par des expressions booléennes pour reformuler une requête booléenne. Si aucun terme n'est trouvé dans les tables, les mots sont combinés par l'opérateur "AND" pour rechercher des documents contenant tous les mots. Le tableau III.5 illustre les résultats obtenus pour la requête "*mad cow disease*" en utilisant le service ATM de PubMed.

TABLEAU III.5 – Illustration des concepts extraits par ATM

"encephalopathy, bovine spongiform"[MeSH Terms]
OR ("encephalopathy"[All Fields] AND "bovine"[All Fields]
AND "spongiform"[All Fields])
OR "bovine spongiform encephalopathy"[All Fields]
OR ("mad"[All Fields] AND "cow"[All Fields]
AND "disease"[All Fields])
OR "mad cow disease"[All Fields]

La stratégie d'extraction des concepts du service ATM de PubMed utilise l'approche de la recherche exacte des termes dans la base de données. Il est capable de retrouver facilement les termes synonymes ainsi que les variants d'un terme donné. Par contre, les problèmes suivants peuvent être observés lors de l'extraction des concepts :

– PubMed ATM essaie plusieurs combinaisons possibles des mots pour formuler une nouvelle requête booléenne qui la rend finalement plus compliquée et plus difficile à interpréter par l'utilisateur. Par exemple, pour la requête "neck and back pain causes", les termes suivants sont identifiés : ("neck"[MeSH Terms] OR "neck"[All Fields]) AND ("back pain"[MeSH Terms] OR ("back"[All Fields] AND "pain"[All Fields]) OR "back pain"[All Fields] OR ("back"[All Fields] AND "pain"[All Fields]) OR "and back pain"[All Fields]) AND ("prevention and control"[Subheading] OR ("prevention"[All Fields] AND "control"[All Fields]) OR "prevention and control"[All Fields] OR "prevention"[All Fields]) tandis que les concepts de la requête sont : "neck pain", "back pain" et "cause".

– Lorsque la requête contient un terme simple, par exemple "Parkinson", PubMed ATM n'arrive pas à l'identifier comme un concept même si dans un contexte biomédical, il s'agit bien du concept "Parkinson disease" dans MeSH.

4.2.2 MetaMap

MetaMap est un programme permettant d'extraire les concepts UMLS à partir des textes biomédicaux (Aronson, 2001a). Ceci est le composant principal de l'outil MTI (Aronson *et al.*, 2004b) qui est utilisé pour indexer les documents dans la base MEDLINE. Le processus d'extraction est comme suit :

1. Identifier des groupes nominaux dans le texte en utilisant l'analyseur grammatical de Xerox (Cutting *et al.*, 1992),
2. Générer des variantes (synonymes, acronymes, ...) pour chaque groupe nominal en utilisant la ressource SPECIALIST Lexicon de l'UMLS,
3. Sélectionner des concepts candidats : tous les concepts ayant au moins un mot qui se trouve dans une des variantes sont récupérés,
4. Évaluer les concepts : les concepts candidats sont comparés avec le texte original en utilisant les quatre mesures suivantes : *centralité*, *variation*, *couverture* et *cohérence* (Aronson, 2001b). Les deux dernières prennent en compte la fréquence d'apparition des concepts dans le texte. Les concepts candidats sont finalement ordonnés en fonction du score final.
5. Construire les mappings : pour chaque groupe nominal, les concepts sont affectés en fonction du score de similarité.

MetaMap a des avantages et des inconvénients rapportés par plusieurs chercheurs dans le domaine (Hliaoutakis *et al.*, 2009; Trieschnigg, 2010). Deux principaux inconvénients qui influcent les performances de MetaMap sont liés au fait que la sélection des concepts candidats est basée sur les mots simples, ce qui pose le problème de **sur-génération** (over-generation) des variantes liées aux concepts non-pertients retournés. Par exemple, étant donné le groupe nominal "ocular complications", MetaMap l'associe à trois concepts "Ocular", "Complications" et "Complications Specific to Antepartum or Postpartum" car ils partagent au moins un mot en commun.

Le deuxième inconvénient concerne la comparaison stricte entre chaque groupe nominal et les entrées dans le métathésaurus. Cela entraîne le problème de **sous-génération** des variants pertinents. Par exemple, pour l'expression "gyrb and p53 protein", MetaMap n'arrive pas à identifier "gyrb" comme une protéine car celui-ci est enregistré comme "gyrb protein" dans le thésaurus MeSH ou dans l'UMLS.

Un autre inconvénient de MetaMap concerne le coût important en terme de temps de traitement parce que cet outil regroupe un ensemble de méthodes linguistiques sophistiquées comme l'analyse grammaticale, la génération des variantes, la recherche dans l'ensemble du Métathésaurus, ainsi que le calcul de plusieurs mesures statistiques.

4.2.3 MTI (Medical Text Indexer)

MTI a été présenté dans (Aronson *et al.*, 2004b) comme un outil d'assistance à l'indexation des citations d'articles de journaux dans MEDLINE. Plus précisément, MTI suggère les concepts les plus similaires pour chaque document aux indexeurs humains à la NLM. MTI est essentiellement basé sur trois algorithmes d'extraction de concepts suivants : MetaMap (Aronson, 2001a), "Related-PubMed citations" (PRC) (Wilbur, 2003) et "Restrict to MeSH" (Bodenreider *et al.*, 1998). Le premier composant sert à localiser les concepts candidats potentiels tandis que la deuxième sert à pondérer les mots selon un schéma TF-IDF, appelé selon les auteurs *poids local* et *poids global* des mots, en favorisant ceux qui apparaissent dans les documents voisins ou similaires à un document particulier. Le dernier composant a pour objectif de transformer les concepts issus du Métathésaurus en concepts MeSH pour mettre en évidence les sujets sémantiques du document.

MTI est utilisé dans la tâche d'indexation des articles de MEDLINE quotidiennement à la NLM. Du fait que MTI est principalement basé sur MetaMap comme un outil pour récupérer les concepts "similaires" à une instance textuelle (i.e., document ou requête), la liste de concepts candidats peut contenir des concepts non-pertinents vis-à-vis du texte (Aronson *et al.*, 2004b). Pour pallier ce problème, MTI applique trois différents niveaux de filtrage pour éliminer les concepts non-pertinents :

- **Filtrage strict** : consiste à supprimer tous les concepts qui ne sont pas identifiés ni par MetaMap ni par "PubMed Related Citations". Ceci permet de réduire un certain nombre de concepts candidats non-pertinents mais probablement certains qui sont pertinents afin de retenir seulement les meilleurs concepts au début de la liste. Par exemple, pour la requête "Find articles about the role of **NEIL1** in repair of **DNA**. ", le filtrage strict ne donne que "DNA" comme concept MeSH candidat. Dans cette requête, **NEIL1** est le nom d'un gène et aurait dû être également identifié. Cependant, il est probable que la troisième composante de MTI, i.e., "Restrict to MeSH" l'a supprimé si MetaMap ou PRC l'avait identifié correctement comme un nom de gène dans l'UMLS.

- **Filtrage moyen** : a pour objectif d'augmenter le rappel en assouplissant les conditions de sélection des concepts candidats : dix heuristiques sont impliquées dans la sélection des concepts candidats. Par exemple : les concepts identifiés par la méthode PRC n'ayant aucun terme qui se trouve dans un des concepts identifiés par MetaMap sont enlevés ; les termes désignant les concepts généraux par MetaMap sont supprimés si un terme désignant un concept plus spécifique est trouvé par PRC.

La spécificité d'un terme est déterminée grâce à l'architecture polyhiérarchique du MeSH où chaque concept est annoté par un numéro d'arbre où il se trouve ou grâce à la longueur du terme, etc.

– **Filtrage de base** : par défaut, MTI utilise le filtrage basique (base filtering) qui implique quatre fonctions suivantes : (1) *addition*, et (2) *suppression* des termes désignant les concepts MeSH ou les qualificatifs en se basant sur les résultats de chacune des deux méthodes MetaMap et PRC, (3) *promotion* des concepts identifiés par les deux méthodes et (4) *substitution* des qualificatifs pour certains concepts identifiés comme candidats. En général, le filtrage de base renvoie plus de concepts candidats et par conséquent plus de concepts dans la liste finale. Bien que le rappel puisse augmenter, il y en a plus de concepts non-pertinents mélangés avec ceux qui sont pertinents. Par exemple, pour la même requête précédente, le filtrage de base donne : *DNA ; Digitoxigenin ; Microsomes, Liver ; Digitoxin ; Hydroxylation ; Cytochrome P-450 Enzyme System ; Biotransformation ; Pregnenolone Carbonitrile ; Hydroxytestosterones ; Bile ; Liver ; Troleandomycin ; Drug Interactions ; Glucuronates ; Hydrolysis ; Kinetics ; Chromatography, Thin Layer ; Oxidation-Reduction ; Digoxin ; Phenobarbital ; Chromatography, Gel ; Testosterone ; Xenobiotics ; .... ; Wound Healing ; Sex Characteristics ; Half-Life ; Mesocricetus ; Macaca fascicularis ; Tritium ; Aging ; beta-Glucosidase ; Time Factors ; Mice, Inbred C3H ; Delivery of Health Care ; Analysis of Variance ;*

4.2.4 MaxMatcher

MaxMatcher est un outil d'extraction de concepts basé sur la **recherche** dans un dictionnaire des chaînes de caractères ou des termes composés d'un ou de plusieurs mots stockés dans un dictionnaire de termes désignant les concepts. La recherche d'une chaîne de caractères dans un dictionnaire de concepts peut se faire de manière *exacte* et/ou *approximative*. MaxMatcher utilise MeSH, et UMLS comme ressources terminologiques pour identifier les concepts. Du fait que le thésaurus MeSH est maintenu par une organisation (la NLM), ceci ne contient pas de termes ambigus désignant les concepts. Par conséquent, MaxMatcher adopte la recherche exacte pour localiser les concepts si MeSH est utilisé. Par contre, pour l'UMLS, un terme peut désigner plusieurs concepts, ce qui représente l'ambiguïté dans l'UMLS. De plus, un mot simple peut désigner un concept même s'il n'est pas explicitement défini dans la ressource terminoontologique. Par exemple, le mot "gyrb" est le nom d'une protéine qui est enregistré en tant que concept "gyrb protein" dans l'UMLS.

Étant donné un texte, MaxMatcher le découpe en phrases et puis localise la chaîne la plus longue qui correspond à une entrée dans le dictionnaire. Pour la

recherche approximative, la chaîne peut être plus courte que l'entrée du concept. Par exemple, le mot "gyrb" peut être désigner le concept "gyrb protein". Cependant, si un concept est constitué de deux sous-concepts, MaxMatcher retourne deux sous-concepts candidats plutôt que de retourner le concept plus spécifique. Par exemple, l'expression "**Ablation of liver tumor** by **injection** of **hypertonic saline**." contient trois concepts mis en gras dont l'identifiant dans l'UMLS est respectivement : *C2004650*, *C0021485* et *C0036085*. Cependant, MaxMatcher retourne quatre concepts suivants : **Ablation** (C0547070, T169) of **liver tumor** (C0023903, T191) by injection (C0021485, T061) of **hypertonic saline** (C0036085, T121, T197).

5 Indexation mono-terminologique *vs.* multi-terminologique de documents biomédicaux

La spécificité d'un modèle d'indexation en RI biomédicale réside essentiellement dans la méthode sous-jacente d'extraction de concepts (*cf.* la section 4). La différence entre les méthodes d'indexation biomédicale est essentiellement axée sur le choix de la méthode d'extraction de concepts ainsi que du référentiel de ressources terminologiques utilisées, plus précisément sur leur nombre et la stratégie de les intégrer dans un processus de RI : une terminologie, qui donne lieu à une **indexation mono-terminologique** ou plusieurs terminologies donnant lieu à une **indexation multi-terminologique**.

Cette section dresse un bref historique de l'indexation au sens de la RI biomédicale puis présente des éléments sur le principe d'indexation mono-terminologique *vs.* multi-terminologique.

5.1 Historique de l'indexation en RI biomédicale

Dans les années 1960s, pour accéder à l'information biomédicale dans la base MEDLINE, les utilisateurs devaient passer par une période d'entraînement à la bibliothèque nationale de la santé (National Library of Medicine - NLM). Pour chercher de l'information biomédicale, l'utilisateur devait remplir un formulaire et l'envoyer par courrier à la NLM, avec un temps de réponse de deux à trois semaines pour obtenir les résultats. Dans les années 1970s, aux États-Unis, l'accès à MEDLINE a été établi directement via le réseau Internet. Cependant, pour accéder à l'information, l'utilisateur devait attendre encore deux ou trois jours pour recevoir les résultats car la recherche d'information dans MEDLINE n'était assurée que par un ou des intermédiaires qui s'habituaient à l'utilisation des services d'accès à l'information de la NLM.

Dans les années 1980s, lorsque les bases de données en ligne étaient disponibles et accessibles, les premiers utilisateurs sur Internet apparaissaient. Et puis dans les années 1990s, avec l'apparition de la toile (Berners-Lee et Cailliau, 1990) et des premiers moteurs de recherche en ligne (Jansen et Spink, 2006), a connu une véritable explosion de l'information publiée sur Internet. De plus, les matériels informatiques, notamment la puissance et les performances des serveurs et des ordinateurs clients, ont été améliorés en termes de vitesse et de capacité de stockage. Les interfaces graphiques des moteurs de recherche ont facilité l'utilisation des ordinateurs ainsi que des moteurs de recherche et des services Web. De nos jours, l'accès à l'information via un moteur de recherche se fait par Internet avec un temps de réponse de quelques milli-secondes.

Avec l'accroissement régulier de publications dans le domaine biomédical, il est de plus en plus difficile de retrouver des informations pertinentes en réponse à la requête de l'utilisateur. Les moteurs de recherche d'information ou annuaires généralistes, présentent actuellement la majorité des outils de recherche sur Internet. En avril 2004, une étude de l'institut Harris, aux États-Unis, indique que plus d'un américain sur deux recherchent régulièrement de l'information médicale sur Internet, et utilisent plus souvent (65%) un moteur ou un annuaire généraliste plutôt qu'un outil spécialisé [15]. Cependant, ces moteurs de recherche classiques n'abordent pas les problématiques de l'information biomédicale comme la présence des synonymes, des acronymes, ou des termes ambigus dans les textes biomédicaux (Stokes *et al.*, 2009).

Il est connu que la grande majorité des revues rapportant les travaux de recherche les plus significatifs en biomédecine sont sélectionnées pour l'indexation dans MEDLINE après un examen minutieux effectué par un comité de sélection à la NLM. La sélection des articles à indexer est basée sur plusieurs critères comme par exemple la portée de la revue, la couverture ainsi que la qualité de son contenu scientifique. Un trait distinctif de MEDLINE est que les documents sont indexés par des termes MeSH dénotant des concepts du domaine. À partir de 1997, MEDLINE a été mise en ligne et accessible à tous les utilisateurs sur Internet via PubMed (Thirion *et al.*, 2009).

Les utilisateurs ainsi que les professionnels de santé ont besoin d'accéder à des ressources d'information biomédicale. Leur tâche de recherche se produit quotidiennement. Bien que les moteurs de recherche comme PubMed et Google scholar soient les plus populaires dans la recherche de l'information dans la littérature, ceux-ci ne fournissent pas de support explicite pour les requêtes plus ciblées comme la recherche des gènes, des protéines ou des maladies particulières. Dans le contexte de la recherche biomédicale et des sciences de la vie, il y a un besoin primordial de développer les systèmes de RI efficaces et performants pour aider les scientifiques à retrouver de bonnes informations.

15. `http://infodoc.inserm.fr/asso/2-selectionner-information/1-limites-outils.html`

5.2 Synthèse des travaux d'indexation des documents biomédicaux

L'accroissement des articles publiés dans MEDLINE a conduit donc au développement d'outils d'indexation (semi)-automatique tel que MTI (Medical Text Indexer) (Aronson, 2001a). L'indexation (semi)-automatique a été plus largement utilisée dans le domaine et vue comme ***une recommandation automatique*** de concepts ou de descripteurs sémantiques (Kim *et al.*, 2001; Pouliquen, 2002; Gaudinat *et al.*, 2002), ou de couples de descripteurs/qualificatifs (Névéol *et al.*, 2006). Récemment, les travaux de (Ruch, 2006; Trieschnigg *et al.*, 2009) montrent que la recommandation ou l'affectation des descripteurs MeSH dans les documents peut être considérée comme la ***catégorisation textuelle*** dans le sens que le classificateur décide lui-même si un descripteur est potentiellement associé à chaque document. Concernant particulièrement l'indexation des ressources francophones, les travaux d'indexation intègrent une ou plusieurs terminologies médicales (MeSH, ICD-10, CCAP, TUV, ...) en associant des descripteurs MeSH aux documents dans le catalogue CiSMEF (Névéol *et al.*, 2006; Pereira *et al.*, 2008). Ces travaux d'indexation constituent la base fondamentale des moteurs de recherche d'information biomédicale.

Dans ce qui suit, nous synthétisons les travaux d'indexation des documents biomédicaux en se basant sur les méthodes d'extraction de concepts qui peuvent être issus d'une ou de plusieurs terminologies.

5.2.1 Indexation mono-terminologique

Il existe une variété de systèmes d'indexation mono-terminologique des documents biomédicaux comme NOMINDEX (Pouliquen, 2002), MeSHMap (Ruch *et al.*, 2003), MAIF (Névéol *et al.*, 2006). Ces systèmes, qui sont basés sur les méthodes d'extraction de concepts (*cf.* la section 4) issus d'une seule terminologie notamment le thésaurus MeSH, fournissent une indexation contrôlée des ressources documentaires biomédicaux en anglais, en français ou les deux à la fois.

NOMINDEX (Pouliquen, 2002). Le système NOMINDEX a pour but d'indexer les documents biomédicaux par des concepts issus d'une ressource termino-ontologique. Dans un premier temps, les mots du document à indexer sont mis en correspondance avec les termes issus de l'ADM (Aide au Diagnostic Médical) (Lenoir *et al.*, 1981). Par exemple, l'expression "Néphrite glomérulaire lupique" sera indexée par les deux mots "Lupus" et "Glomérulonéphrite" (car le terme "Néphrite glomérulaire" est synonyme de "Glomérulonéphrite"). Ensuite, les termes identifiés sont rattachés à leurs équivalents MeSH, ainsi qu'à leur

identifiant unique dans l'UMLS. La figure III.9 présente l'architecture générale de leur système. Les termes désignant les concepts identifiés sont intégrés dans un modèle vectoriel (Salton, 1989) afin d'améliorer la représentation textuelle de documents. Plus précisément, la mesure *Cosinus* est utilisée pour calculer la similarité entre chaque document D et la requête Q :

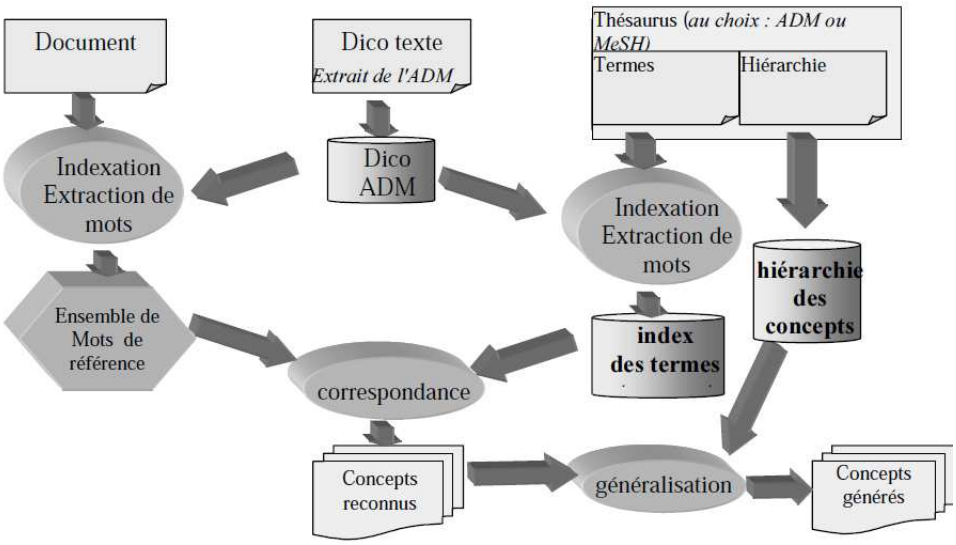


FIGURE III.9 – Architecture générale du système NOMINDEX (Pouliquen, 2002)

$$RSV(D,Q) = Cosine(D,Q) = \frac{\sum_{c \in D \cap Q} TFIDF_{c,D} \times TFIDF_{c,Q}}{\sqrt{\sum_{c \in D} TFIDF_{c,D}^2} \times \sqrt{\sum_{c \in Q} TFIDF_{c,Q}^2}} \quad \text{(III.6)}$$

où c est un concept commun dans la requête et le document, $TFIDF_{c,D}$ (resp. $TFIDF_{c,Q}$) est le poids TFIDF du concept c dans le document D (resp. dans la requête Q). Le poids du concept dans le document ou la requête X est calculé comme suit :

$$TFIDF_{c,X} = TF_{c,X} \times log_2 \frac{N}{DF_c} + 1 \quad \text{(III.7)}$$

avec $TF_{c,X}$: la fréquence d'apparition du concept dans le document ou la requête X et DF_c : le nombre de documents du corpus contenant le concept.

MeSHMap (Ruch *et al.*, 2003; Ruch, 2006). MeSHMap est un système de recherche d'information biomédicale basée sur une méthode d'extraction de

concepts MeSH en utilisant des mesures statistiques. Plus spécifiquement, il s'agit d'une méthode d'extraction basée sur la recherche d'information : les concepts sont considérés comme un ensemble de documents à indexer tandis que chaque document est considéré comme une requête à laquelle sont associés des "documents MeSH". L'extraction de concepts est constituée de deux étapes principales :

- l'extraction des composants MeSH en utilisant des expressions régulières permettant de reconnaître certaines variations, comme l'insertion d'un caractère spécial comme "-" (e.g., le terme "insulino-dépendant" sera reconnu comme "insulinodépendant") ainsi que l'insertion d'un mot (e.g., "maladies très rares" sera reconnu pour le terme "maladies rares"). Cette étape permet d'extraire les concepts avec une précision élevée puisque seuls les termes MeSH effectivement présents dans le texte peuvent être extraits.

- l'extraction des racines des composants sur l'ensemble du texte en utilisant le modèle vectoriel (Salton *et al.*, 1983) pour attribuer un score à chacun des termes du thésaurus MeSH. Cette étape vise à privilégier le rappel, puisque tous les termes MeSH contenant au moins l'une des racines identifiées sont considérés comme candidats à l'indexation.

(Ruch, 2006) n'a pas intégré les concepts ainsi identifiés dans un processus de RI. En utilisant la méthode MeSHMap, le travail de (Trieschnigg *et al.*, 2009) consiste à extraire les concepts à partir d'un ensemble de requêtes issues de TREC Genomics et combiner les scores obtenus d'une représentation textuelle et d'une représentation conceptuelle. Ce dernier a conclu que la méthode d'extraction de MeSHMap n'est pas performante car cette approche génère plusieurs termes non reliés aux sujets de la requête.

MAIF (Névéol *et al.*, 2006). Dans le cadre de l'indexation des ressources francophones, le système MAIF (MeSH Automatic Indexing for French) est dédié à l'indexation des des articles en texte intégral (Névéol *et al.*, 2006) (*cf.* la figure III.10). Étant donné un URL (adresse de la ressource), MAIF télécharge les documents et les indexe par des termes MeSH en français et éventuellement par des paires termes MeSH/qualificatifs. Le texte intégral de chaque document est traité par une **approche TAL** (Traitement Automatique du Langage) et le titre par une **approche k-PPV** (*k Plus Proche Voisins*).

- **L'approche TAL** consiste à identifier les termes MeSH et/ou qualificatifs en utilisant les transducteurs définis dans le logiciel INTEX (Silberztein, 1999). Les termes identifiés sont issus d'un dictionnaire des termes extraits du thésaurus MeSH. Les transducteurs sont des ***patrons d'extraction*** permettant de rendre compte de la grande variabilité de certains mots-clés (e.g., <adulte d'âge moyen>, <centre de rééducation

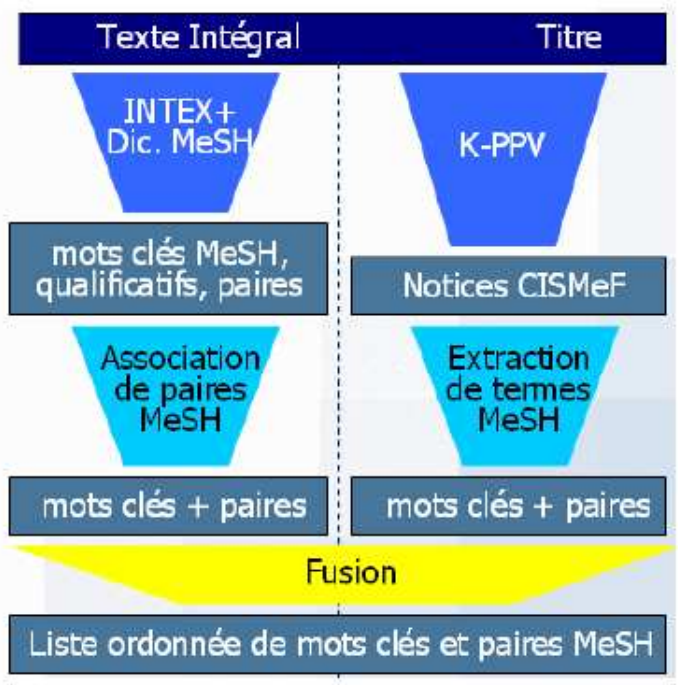


FIGURE III.10 – Architecture générale du système MAIF

et de réadaptation>, ...). Les transducteurs complètent la couverture du mot-clé par les entrées de dictionnaire :

- adulte d'âge moyen,adulte âge moyen.***N+MeSH :ms***
- adultes d'âge moyen,adulte âge moyen.***N+MeSH :mp***

– **L'approche k-PPV** est un algorithme issu de la reconnaissance des formes qui a été adapté à de nombreux autres domaines, y compris la classification de documents où il s'est révélé efficace (Lam *et al.*, 1999; Yang et Chute, 1994b). Cette méthode consiste à identifier les ressources (documents) les plus proches au titre de la ressource (document) à indexer. Les documents proches doivent contenir au moins un mot pertinent commun avec le titre du document courant. Pour chaque titre extrait, un score de similarité et de classement est calculé afin de ne retenir que les k premiers documents. Ce score a été calculé par deux méthodes différentes :

– La première méthode est fondée sur le décompte des mots non-

grammaticaux communs entre termes biomédicaux. Par exemple, "Le diabète et les maladies rénales" et "Diabète et grossesse" ont un mot non-grammatical commun : "diabète". Le mot grammatical "et", présent dans les deux titres, n'intervient pas dans le décompte. Le score de similarité entre les deux titres sera donc de 1.

– La deuxième méthode est fondée sur la distance de Levenshtein, ou distance d'édition (Levenshtein, 1966) entre les titres. La distance d'édition entre deux chaînes de caractères correspond au nombre minimum de transformations élémentaires (suppression, insertion et substitution) à réaliser pour passer d'une chaîne à l'autre.

Comme le travail de (Ruch, 2006), (Névéol *et al.*, 2006) n'a pas évalué l'impact d'intégrer les concepts termino-ontologiques extraits à partir de textes qui sont utilisés dans un processus de RI sémantique.

5.2.2 Indexation multi-terminologique

À notre connaissance, l'indexation multi-terminologique a été premièrement abordée par l'outil MetaMap (Aronson, 2001a) et puis l'outil MTI (Aronson *et al.*, 2004b) dans le cadre de l'indexation des articles de MEDLINE. Ces outils utilisent l'UMLS comme un ensemble de différentes terminologies pour suggérer les concepts aux indexeurs humains. Par exemple, avec MetaMap, on peut spécifier les terminologies à partir desquelles les concepts sont extraits (*cf.* la figure III.11). L'outil MTI ayant pour composante principale MetaMap extrait les concepts issus de plusieurs terminologies dans l'UMLS et à la fin il ne retient que les concepts MeSH pour indexer les articles dans MEDLINE (*cf.* la figure III.12).

Dans le cadre de la recherche d'information dans le portail de santé de CISMeF, l'outil F-MTI (French Multi-Terminology Indexer) s'est inspiré de l'outil MAIF afin d'intégrer plusieurs terminologies d'indexation (Pereira *et al.*, 2009). L'outil F-MTI a été conçu en particulier pour indexer les dossiers médicaux en utilisant plusieurs terminologies médicales à savoir la CIM-10, la CCAM, le thésaurus MeSH, la terminologie interne de la société Vidal ainsi que la nomenclature SNOMED. Plus récemment, F-MTI a été étendu dans le cadre de la thèse de (Sakji, 2010) pour la recherche d'information et l'indexation automatique des médicaments à l'aide de plusieurs terminologies de santé. (Sakji, 2010) a proposé une approche d'indexation automatique, par la classification ATC (Anatomique Thérapeutique et Chimique), pour les ressources du Portail d'Information sur les Médicaments (PIM), conçu dans le cadre du projet européen PSIP. Cette indexation a pour but d'améliorer l'indexation des médicaments afin de fournir à l'utilisateur une information plus fine et détaillée. Leur approche d'indexation automatique par la classification ATC, repose sur

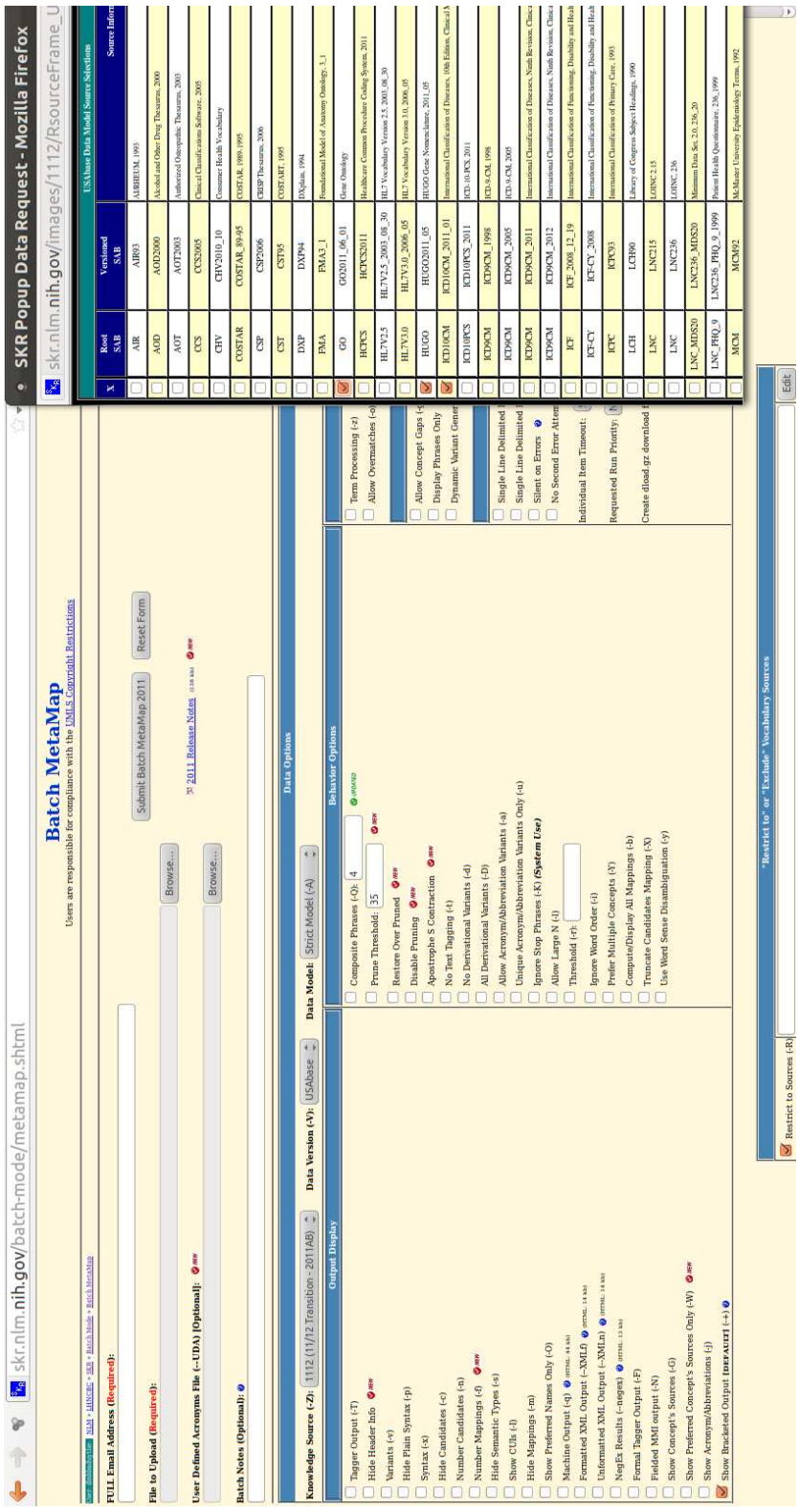


FIGURE III.11 – Extraction des concepts pour une indexation multi-terminologique avec MetaMap

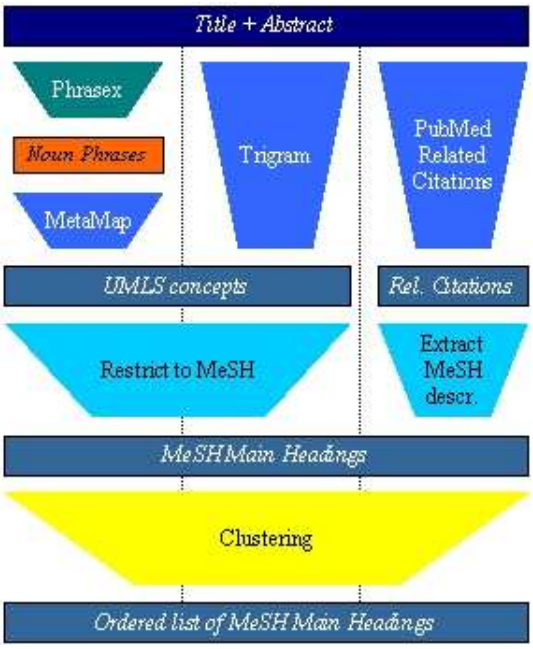


FIGURE III.12 – Les composantes principales de MTI (Medical Text Indexer)

les trois étapes séquentielles suivantes :

1. recherche du code ATC au niveau du titre de la ressource ;
2. recherche du nom commercial (NC) de la substance au niveau du titre de la ressource ; si c'est le cas, le code ATC, associé au nom commercial, est assigné à la ressource. Les associations ATC-NC sont sauvegardées dans une table reliant le nom commercial des médicaments et le code ATC correspondant ;
3. recherche du code ATC selon l'indexation de la ressource (indexation par les descripteurs et/ou les concepts chimiques supplémentaires du thésaurus MeSH).

(Avillach *et al.*, 2007) ont proposé une méthode d'indexation multi-terminologique pour indexer les résumés de sortie de l'hôpital des patients. Il s'agit d'une méthode d'indexation basée sur une méthode d'extraction de concepts en utilisant les connaissances symboliques biomédicales sous forme d'ontologies et des connaissances statistiques extraites à partir d'un domaine d'application. Les termes extraits sont ordonnés pour mettre en évidence leur importance dans le document. L'importance d'un terme est déterminée par le nombre de relations qu'il partage avec d'autres termes désignant les concepts. Les relations entre concepts peuvent être exploitées à partir du méta-thésaurus UMLS et des relations co-occurrences entre les concepts issus d'une ou de plusieurs terminologies (*cf.* la figure III.13). Ils ont conclu que l'utilisation de multiple terminologies en exploitant les relations sémantiques dans les termi-

nologies résulte en une meilleure précision de l'indexation.

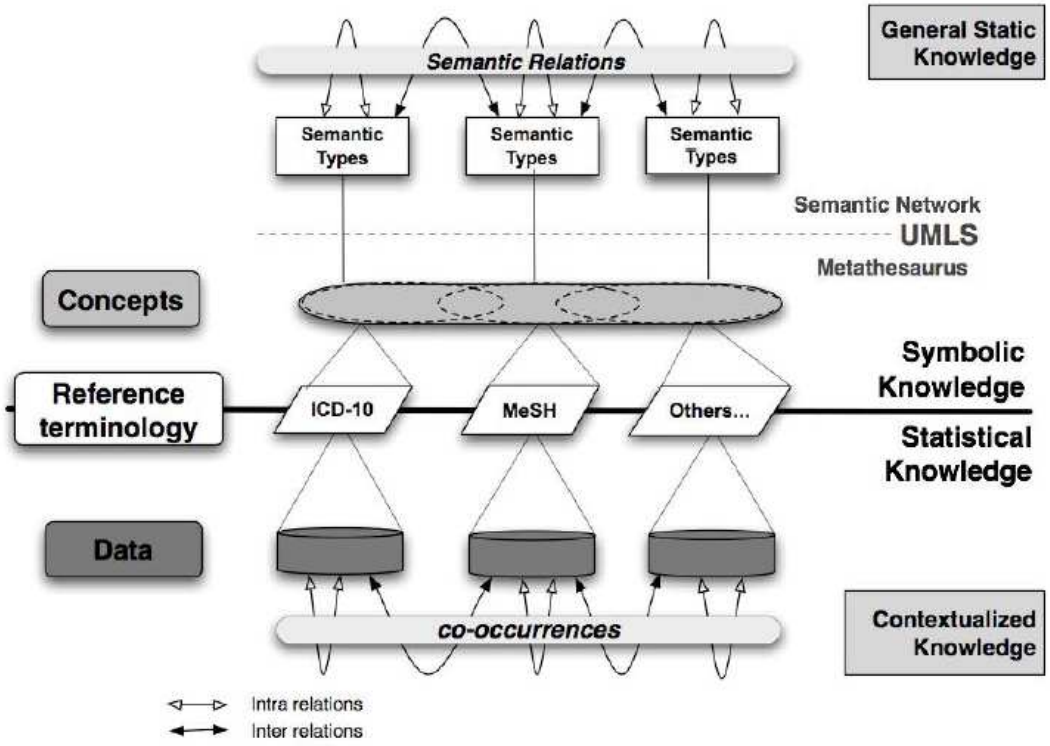


FIGURE III.13 – Les relations inter et intra entre les concepts issus de différentes terminologies (Avillach *et al.*, 2007)

Dans le cadre de la recherche d'information des documents biomédicaux de la littérature, le système Peregrine, développé par le groupe de recherche Biosemantics au Centre de Recherche Médicale à l'université ErasmusMC, a été utilisé pour identifier les concepts issus de l'UMLS (Trieschnigg, 2010). Le système Peregrine a pour but d'identifier à l'origine les noms des gènes et de protéines qui sont définis dans les ressources terminologiques. Leur méthode d'identification des termes repose sur une recherche dans un dictionnaire en combinaison avec un nombre d'étapes de traitement supplémentaires. Ces étapes concernent essentiellement la définition des règles d'indexation pour supprimer les termes incorrects et ambigus, l'intégration des règles linguistiques pour générer les variantes lexicales ainsi que des méthodes de désambiguïsation. Dans la thèse de (Trieschnigg, 2010), une méthode de classification de la requête, intitulée KNN, a été proposée pour identifier les concepts dans le méta-thésaurus UMLS. La probabilité que le concept c dans un document D soit pertinent à la requête Q est donnée par :

$$P(c, Q) = \sum_{D \in \Psi} P(D) \left(P(c|\phi_D) \prod_{i=1..n} P(q_i|\theta_D) \right) \tag{III.8}$$

où

- Ψ est les k premiers documents retournés par le système de RI vis-à-vis de la requête ;
- $P(D)$ est la probabilité *priori* d'observer le document D dans l'ensemble Ψ ;
- $P(Q|\theta_D)$ est la probabilité d'observer la requête à partir du document D (vraisemblance de la requête) ;
- $P(c|\theta_D)$ est la probabilité d'observer le concept c à partir du document D (vraisemblance du concept).

Dans un autre contexte de la fouille de données ou de l'exploration de données (*data mining*), les auteurs dans (Yu *et al.*, 2010) ont proposé une approche "multi-vue" pour identifier les noms des gènes dans les textes biomédicaux. Dans leur approche d'identification des gènes, plusieurs terminologies ont été intégrées comme GO, MeSH, eVOC, OMIM, LDDB, KO, MPO, SNOMED CT, et UniprotKB. Ils ont évalué l'impact de l'utilisation de plusieurs terminologies sur les performances de la catégorisation des noms des gènes en utilisant plusieurs approches de catégorisation comme la *fusion des matrices* (Sonnenburg *et al.*, 2006), le *regroupement des ensembles* (Aerts *et al.*, 2006; Tranchevent *et al.*, 2008). Ils ont conclu que l'approche multi-terminologique est prometteuse pour améliorer les performances de l'identification des gènes dans les textes biomédicaux.

6 Techniques et modèles d'appariement document-requête en RI biomédicale

La plupart des approches de RI biomédicale reposent sur des modèles de RI conceptuelle ou sémantique qui intègrent une ou plusieurs terminologies biomédicales pour améliorer les performances de la RI, notamment l'efficacité du système de RI en termes de *précision* et *rappel*. Plusieurs travaux sont basés sur une amélioration de la représentation de la requête via un processus de reformulation de requêtes (Zhou *et al.*, 2007a; Stokes *et al.*, 2009; Trieschnigg, 2010). D'autres travaux ont pour objectif d'améliorer à la fois la représentation des documents et de la requête via un processus d'extraction des concepts à partir des documents et de la requête pour améliorer leur représentation (Zhou *et al.*, 2006c; Le *et al.*, 2007; Gobeill *et al.*, 2009). D'autres modèles d'appariement, appelés modèles PICO [16], sont basés sur les patrons de besoins d'experts pour évaluer les requêtes cliniques (Schardt *et al.*, 2007; Zhao *et al.*, 2010).

16. Patients, Intervention, Comparison, Outcome

6.1 Reformulation de requêtes

Nous allons détailler dans ce qui suit les différentes approches de reformulation de requêtes qui ont été proposées dans le contexte de la RI biomédicale. Nous pouvons distinguer deux grandes catégories : (1) *reformulation conceptuelle de requêtes* basée sur des terminologies biomédicales et (2) *reformulation de requêtes par pseudo-réinjection de pertinence*. En général, ces deux approches sont souvent combinées pour améliorer les performances de la RI (Srinivasan, 1996; Fujita, 2004; Zhou *et al.*, 2007a).

6.1.1 Reformulation conceptuelle de requêtes

Les premiers travaux en RI biomédicale visant à construire les requêtes basées sur le thésaurus MeSH ont été effectuées vers la fin des années 1980s (Elkin *et al.*, 1988) et pendant les années 1990s (Aronson *et al.*, 1994; Hersh *et al.*, 1994; Yang et Chute, 1994a; Srinivasan, 1996; Aronson *et al.*, 1997). Depuis, plusieurs travaux ont exploité les concepts issus d'une ou de plusieurs ressources terminologiques pour reformuler la requête (Hersh *et al.*, 2000; Aronson *et al.*, 2004a; Seki *et al.*, 2004; Nakov *et al.*, 2004; Stokes *et al.*, 2009; Trieschnigg, 2010).

Certains travaux ont montré une amélioration de la reformulation conceptuelle de la requête basée sur les terminologies biomédicales (Srinivasan, 1996; Aronson *et al.*, 1997; Zhou *et al.*, 2007a; Stokes *et al.*, 2009) tandis que d'autres travaux n'ont signalé aucune amélioration des performances de RI en appliquant une expansion conceptuelle de la requête (Hersh *et al.*, 2000; Aronson *et al.*, 2004a; Seki *et al.*, 2004; Nakov *et al.*, 2004). En fait, les performances de la RI dépendent notamment de la façon dont les concepts sont exploités pour modifier ou reformuler la requête.

Par exemple, l'approche de RI conceptuelle proposée par (Aronson *et al.*, 1994) est basée sur une méthode d'extraction de concepts UMLS à partir du texte. Plus précisément, leur méthode d'extraction de concepts est basée sur deux techniques : (a) l'analyse syntaxique des phrases du texte pour identifier les groupes nominaux en utilisant l'outil Xerox POS tagger (Cutting *et al.*, 1992) et (b) l'association des groupes nominaux ou expressions dans le texte aux concepts issus d'un large thésaurus intitulé l'UMLS. Ils ont appliqué leur méthode d'extraction de concepts sur un corpus des documents biomédicaux en utilisant le modèle vectoriel implémenté dans le système SMART (Salton, 1991). Les concepts issus de l'UMLS sont extraits à partir de chaque document dans une collection de **3000 documents** extraits de la base MEDLINE et chaque requête de l'utilisateur parmi 150 requêtes. Ils ont **remplacé les groupes nominaux identifiés à partir de chaque document et dans**

la requête par les termes préférés désignant les concepts identifiés. Par exemple, si le texte (i.e., document, requête) contient l'expression "low blood pressure", celle-ci est sera remplacée par le terme "Hypotension". Avec cette méthode, ils ont obtenu une amélioration de +4 % en terme de précision moyenne par rapport à la méthode de RI sans la modification des documents et des requêtes. Cependant, il n'est pas évident de vérifier l'efficacité de leur méthode sur les collections des documents biomédicaux en réalité car elles sont en général beaucoup plus volumineuses (jusqu'à des millions de documents).

(Yang et Chute, 1994a) ont proposé une méthode de RI conceptuelle pour résoudre le problème de la différence du vocabulaire entre les requêtes et les documents pertinents. Leur méthode d'extraction de concepts est basée sur la technique LLSF (Linear Least Squares Fit [17]) pour identifier les concepts à partir des documents. Chaque document (plus précisément une citation de MEDLINE) est composé essentiellement d'un titre et d'un résumé et des termes MeSH manuellement ajoutés par les indexeurs humains. Ils ont divisé la collection de documents en deux sous ensembles : l'un est pour l'apprentissage et l'autre pour l'évaluation. Les requêtes ont été divisées en sorte que 88 % des requêtes test soient couvertes dans l'ensemble de requêtes d'apprentissage. Chaque requête d'apprentissage a été **annotée manuellement** par des catégories sémantiques. Chaque requête a été étendue par les catégories sémantiques (en l'occurrence termes désignant les concepts MeSH) extraites par l'approche LLSF.

(Aronson *et al.*, 1997) ont utilisé MetaMap (Aronson *et al.*, 1994) pour extraire les concepts UMLS à partir de la requête. Plus précisément, les groupes nominaux et concepts extraits sont ajoutés à la requête originale. L'indexation et la recherche se fait par un système de RI basée sur un modèle probabiliste intitulé INQUERY (Callan *et al.*, 1992). Les mots simples de la requête, les groupes nominaux ainsi que les termes désignant les concepts identifiés à partir de la requête sont re-pondérés et combinés linéairement en utilisant l'opérateur $\#SUM$ et $\#PHRASE$ (*cf.* l'exemple dans le tableau III.6). Ils ont évalué leur approche d'expansion conceptuelle sur une collection de 2334 citations MEDLINE et 75 requêtes (Haynes *et al.*, 1990). Ils ont conclu que l'expansion conceptuelle de la requête est efficace en dépassant une base d'évaluation de référence.

(Hersh *et al.*, 2000) ont identifié manuellement les concepts issus de l'UMLS à partir de chaque requête dans la collection OHSUMED (Robertson, 2002). Les requêtes sont ensuite reformulées par une expansion des termes désignant les concepts identifiés. Les termes désignant les concepts UMLS ajoutés dans la requête originale de l'utilisateur peuvent être : (1) les synonymes, (2) les

17. La méthode des moindres carrés a été indépendamment élaborée par Legendre en 1805 et Gauss en 1809

Requête originale	**Requête reformulée**
is there evidence to support the use of inhaled steroids in COPD when the patient is on intravenous steroids	#q34 = #*WSUM*(1 2 #*SUM*(is there evidence to support the use of inhaled steroids in COPD when the patient is on intravenous steroids) 1 #*SUM*(#*PHRASE*(use) #*PHRASE*(inhaled steroids) #*PHRASE*(copd) #*PHRASE*(patient) #*PHRASE*(intravenous steroids)) 5 #*SUM*(#*SUM*(Steroids) #*SUM*(Obstructive Lung Diseases) #*SUM*(Patients) #*SUM*(utilization) #*SUM*(Supports) #*SUM*(Inhaled) #*SUM*(IV))) ;

TABLEAU III.6 – Exemple de l'expansion conceptuelle de la requête dans (Aronson *et al.*, 1997)

termes issus des relations hiérarchiques (e.g., parent direct, enfant direct), (3) les termes reliés aux concepts ainsi que (4) les termes issus de la définition du concept. La recherche est basée sur les mots simples dans un modèle vectoriel implémenté dans le système SMART (Salton, 1991). Ils ont observé une dégradation des performances de la RI, notamment la précision et le rappel, lorsque les requêtes sont reformulées par des termes désignant les concepts. Ils ont conclu que l'expansion conceptuelle de requêtes n'est pas efficace par rapport aux performances obtenues par le modèle vectoriel classique.

(Hersh *et al.*, 2003) ont utilisé les différentes ressources terminologiques séparément (LocusLink, FlyBase, Gene Ontology, SwissProt, RefSeq...) pour reformuler la requête originale. Les noms des gènes sont identifiés automatiquement par les programmes en Perl et Python [18]. D'autre part, les noms des gènes issus de LocusLink et FlyBase sont identifiés manuellement. Ils ont exploité au total 13 sources d'évicence issues des ressources terminologiques pour étendre la requête, e.g., résumés des gènes (LocusLink), fonctions des protéines (SwissProt), liens aux maladies (SwissProt), séquences de mRNA (RefSeq)... Ils ont conclu que la reformulation par l'expansion conceptuelle de la requête

18. Ils n'ont pas indiqué clairement comment leur méthode d'identification des noms des gènes a été implémentée

en utilisant les base de données des gènes et de protéines n'apporte aucune utilité pour améliorer la MAP.

Le travail de (Zhou *et al.*, 2006a) consiste à représenter les requêtes (et les documents) par les signatures thématiques (*topic signatures*). Une signature thématique est définie comme un **couple de concepts** qui sont syntaxiquement et sémantiquement reliés. En s'inspirant du modèle de langue (Zhai et Lafferty, 2001a,b), ils ont proposé un modèle de reformulation conceptuelle de requêtes comme suit :

$$p_f(w|q) = \sum_{k:q\cap t_k\neq\phi} p_s(w|t_k)\frac{c(t_k,F)}{\sum_{i:q\cap t_i\neq\phi} c(t_k,F)} \tag{III.9}$$

où

- $p_s(w|t_k)$ est la probabilité que le mot w soit lié à la signature thématique t_k, calculée comme suit :

$$p_s(w|t_k) = \begin{cases} 0 & \text{si} \quad w \notin t_k \\ 1/|t_k| & \text{sinon} \end{cases} \tag{III.10}$$

- $c(t_k, F)$ est la fréquence de la signature thématique dans les premiers documents retournés F (appelés feedback documents) qui sont utilisés pour la reformulation.

(Lu *et al.*, 2009) ont proposé une méthode d'expansion de requête en utilisant les termes de MeSH qui sont automatiquement identifiés et associés à la requête de l'utilisateur via le service d'extraction de concepts de PubMed, appelé ATM (Automatic Term Mapping). Ce dernier associe à la requête une liste de termes pré-indexés dans les tables de traduction comme "MeSH", "Journal" et "Auteur". Leurs résultats expérimentaux ont montré que l'expansion de requête en utilisant MeSH dans PubMed permet d'améliorer en général le rappel des résultats de la RI. Pourtant, les résultats en termes de précision ne changent pas ou ne sont pas significatifs.

Durant les ateliers d'évaluation TREC Genomics, les participants ont montré l'utilité de l'expansion conceptuelle de la requête. Dans le cadre de TREC Genomics 2004, la méthode de (Fujita, 2004) a donné les meilleurs résultats en utilisant le schéma de pondération BM25 (avec la fréquence inverse de documents standard) entre le document d et la requête q comme suit :

$$w(d,q) = \sum_{t\in q}(k_4 + log\frac{N}{df(t)})\frac{(k_1+1)freq(d,t)}{k_1((1-b)+b\frac{dl_d}{avgdl})+freq(d,t)} \tag{III.11}$$

où

- N est le nombre total de documents dans la collection,
- $df(t)$ est le nombre de documents contenant le terme t,
- $freq(d, t)$ est la fréquence du terme t dans le document d,
- k_1, k_4 et b sont des paramètres du modèle.

D'une part, la requête est étendue par les termes issus des ressources termino-ontologiques. Plus précisément, ils ont indexé chaque enregistrement dans LocusLink et MeSH comme un document. Les **termes issus du premier document** qui est le plus similaire à la requête **sont utilisés pour l'expansion de la requête**. D'autre part, la requête originale est reformulée par la technique de reformulation PRF. La meilleure précision moyenne MAP obtenue est de 0.4075. De plus, en appliquant le modèle de langue avec la méthode de lissage Dirichlet-Prior, les résultats ont été améliorés jusqu'à 0.4264 en terme de MAP.

(Büttcher *et al.*, 2004), dont les meilleurs résultats (MAP=0.3867) ont été classés en deuxième dans TREC Genomics 2004 (Hersh *et al.*, 2004), ont utilisé une méthode similaire à celle de (Fujita, 2004), c-à-d l'expansion conceptuelle de requêtes par les différentes ressources comme AcroMed (Acr, 2004), EuGenes (euG, 2004), LocusLink (LL04, 2004) en combinaison avec la reformulation de la requête PRF. D'autres groupes (Aronson *et al.*, 2004a; Seki *et al.*, 2004; Nakov *et al.*, 2004), qui ont essayé l'association des vocabulaires contrôlés à la requête, ont obtenu des résultats moyens. La plupart des participants dans TREC ont utilisé une variété des approches d'expansion de la requête basée sur plusieurs ressources termino-ontologiques mais en général sans comparer à une référence de base (Hersh et Voorhees, 2009). De plus, les expérimentations ne sont pas exhaustives, ce qui rendent difficile à évaluer l'utilité des techniques d'expansion de requêtes ainsi que les ressources termino-ontologiques utilisées.

Dans le cadre de TREC Genomics 2005, la méthode de (Huang *et al.*, 2005) a donné les meilleurs résultats en se basant sur les configurations similaires à des méthodes utilisées dans TREC Genomics 2004. En effet, ces auteurs ont utilisé le schéma de pondération Okapi BM25 comme base de référence. De plus, leur approche consiste à enlever tous les mots vides standard et une liste additionnelle de mots fonctionnels du domaine biomédical ("gene", "role", "impact", "biological", "disease", "process" ...), à normaliser les mots notamment les noms de gènes par la définition de deux notions *point d'arrêt* et *remplaçant*. Un point d'arrêt (e.g., trait d'union '-', position entre deux lettres qui ont différentes casses sauf pour la première et deuxième position ou entre une lettre et un chiffre) définit la position à laquelle une chaîne de caractères ou un mot peut être divisée en deux parties séparées par un espace. Un remplaçant est une sous-chaîne dans une chaîne de caractères qui peut être remplacée par une autre sous-chaîne en sorte que la chaîne après avoir été remplacée représente toujours la même signification que celle d'origine. **L'expansion automatique**

de la requête par des synonymes, homonymes et acronymes extraits à partir de la requête en utilisant des ressources termino-ontologiques comme AcroMed (Acr, 2004) et Locus Link (LL04, 2004) a donné les meilleurs résultats (MAP=0.2888) pour une configuration automatique tandis que l'expansion manuelle a donné encore des meilleurs résultats (MAP=0.3020). Dans une configuration automatique, l'expansion de la requête consiste à extraire des variantes de noms de gènes figurés dans la requête originale et en y ajouter pour enrichir la sémantique de la requête.

Cependant, l'expansion de la requête doit être implémentée "avec soin" (Hersh *et al.*, 2007). Un grand nombre de critères (variables, paramètres) ont été pris en compte : le choix des ressources termino-ontologiques, la stratégie d'extraction des concepts, la méthode d'extraction des termes désignant les concepts à partir la requête, la stratégie d'intégration de ces concepts dans la requête, le modèle de recherche, etc. Étant donné la complexité et la variété des approches d'expansion conceptuelle de la requête, il n'est pas évident de retirer les bonnes conclusions sur l'utilité des approches d'expansion conceptuelle (Trieschnigg, 2010).

6.1.2 Reformulation de requêtes par pseudo-réinjection de pertinence

La reformulation par la méthode pseudo-réinjection de pertinence (PRF) a été développée la première fois par (Croft et Harper, 1979). Depuis, cette technique a été largement étudiée pour améliorer les ordonnancements des documents en particulier dans le cadre de TREC (Robertson *et al.*, 1994, 1998; Baeza-Yates et Ribeiro-Neto, 1999). Concernant l'application de la méthode PRF pour la RI biomédicale plusieurs travaux ont exploité la technique de reformulation PRF pour améliorer les performances de RI (Srinivasan, 1996; Fujita, 2004; Zhou *et al.*, 2007a; Jiang et Zhai, 2007). Par exemple, (Zhou *et al.*, 2007a) ont utilisé une version modifiée de la technique de PRF décrite dans (Baeza-Yates et Ribeiro-Neto, 1999) pour reformuler la requête. Leur méthode consiste à étendre la requête originale par les termes désignant les concepts MeSH trouvés dans les 15 premiers paragraphes retournés par le système de RI au lieu d'utiliser les termes généraux qui sont extraits du contenu des premiers paragraphes. Ensuite, les k premiers concepts les plus significatifs pour la requête sont retenus pour la reformulation. La similarité entre chaque concept c et la requête q est définie par :

$$sim(q,c) = \prod_{k_i \in q} \left(\delta + \frac{log(f(c,k_i) \times IDF_c}{logn} \right)^{IDF_i} \qquad \text{(III.12)}$$

où δ est une constante pour éviter les poids zéro (e.g., $\delta = 0.1$) ; n est le nombre des premiers paragraphes retournés. La fonction $f(c, k_i)$ mesure la corrélation entre le concept c et le terme k_i :

$$\sum_{j=1}^{n} pf_{i,j} \times pf_{c,j} \tag{III.13}$$

où $pf_{i,j}$ est la fréquence du terme k_i dans le paragraphe j et $pf_{c,j}$ est la fréquence du concept c dans le paragraphes j. Les fréquences inverses de documents IDF sont calculées comme suit :

$$\begin{aligned} IDF_i &= max\left(1, \frac{log_{10}(N/np_i)}{5}\right) \\ IDF_c &= max\left(1, \frac{log_{10}(N/np_c)}{5}\right) \end{aligned} \tag{III.14}$$

où N est le nombre de paragraphes dans la collection, np_i est le nombre de paragraphes contenant le terme k_i et np_c est le nombre de paragraphes contenant le concept c.

(Jiang et Zhai, 2007) ont appliqué la méthode PRF proposée dans (Zhai et Lafferty, 2001a) sur les deux collections TREC Genomics 2004 et 2005. Les meilleurs résultats MAP qu'ils ont obtenus sont 0.3357 et 0.2969 respectivement sur ces deux collections. En comparant aux meilleurs résultats officiels dans TREC, les résultats obtenus sur TREC Genomics 2004 sont moins performants que les meilleurs résultats dans TREC Genomics 2004 (Fujita, 2004) (0.3357 *vs.* 0.4075). Par contre, les performances en terme de MAP de leur méthode dépassent légèrement les meilleurs résultats dans TREC Genomics 2005 (Huang *et al.*, 2005) (0.2969 *vs.* 0.2888). Bien que la méthode de (Jiang et Zhai, 2007) soit basée sur plusieurs règles d'indexation (heuristiques) pour séparer les mots (tokenisation) et d'autres expressions régulières manuellement définies pour supprimer les caractères inutiles dans le texte, leurs meilleurs résultats basés sur le **modèle de langue** et la méthode de **reformulation PRF** ne montrent qu'une amélioration légère sur la collection TREC Genomics 2005 par rapport aux résultats de TREC Genomics 2005 tandis que sur la collection TREC Genomics 2004, leurs résultats montrent une dégradation significative des performances de RI par rapport aux meilleurs résultats de TREC Genomics 2004.

6.2 Expansion conceptuelle de documents

Les techniques d'expansion conceptuelle de documents ont été abordées dans la littérature dans le contexte de la RI pour résoudre le problème de dé-

faut d'appariement (souvent abordé comme "*term mismatch*" en anglais) entre les documents et les requêtes. L'expansion documentaire a pour but d'améliorer ou d'enrichir la sémantique du document par l'élargissement du contenu textuel par les concepts les plus représentatifs. Cette technique a été exploitée dans la recherche de discours (speech retrieval) (Singhal et Pereira, 1999). Dans le domaine de la RI biomédicale, la plupart des méthodes d'expansion documentaire sont basées sur une méthode d'extraction de concepts spécifique à partir du texte en exploitant des ressources comme le thésaurus MeSH (Medical Subject Headings thésaurus) (Gobeill *et al.*, 2009) ou le méta-thésaurus UMLS (Unified Medical Language Système) (Trieschnigg *et al.*, 2006; Le *et al.*, 2007) dans le contexte biomédical. Pour la plupart des travaux reliés à l'expansion conceptuelle, les méthodes d'expansion sont appliquées à la fois pour les documents et pour les requêtes :

(Trieschnigg *et al.*, 2006) ont ajouté aux documents et à la requête les concepts issus du méta-thésaurus UMLS (cf. l'exemple dans la figure III.14). La recherche d'information a été effectuée en exploitant les concepts identifiés dans les documents et dans la requête dans un modèle de langue uni-gramme avec la méthode de lissage de Jelinek-Mercer. Les résultats obtenus sur la collection TREC Genomics 2004 ont montré que l'annotation conceptuelle a conduit à la dégradation des performances de RI en comparaison à une base d'évaluation comparative basée sur les mots simples. Sur la collection TREC Genomics 2006, les performances en terme de MAP de la RI basée sur les concepts sont équivalentes aux performances de la RI classique basée sur un modèle de langue des mots simples.

(Le *et al.*, 2007) ont évalué l'impact de l'expansion documentaire et de la requête dans le domaine biomédical en exploitant les concepts ainsi que leurs relations sémantiques dans l'UMLS. Étant donné un texte (document ou requête), son contenu sera étendu avec des concepts UMLS identifiés par l'outil MetaMap. L'évaluation de leur approche, effectuée sur la collection ImageCLEF 2005, a montré une amélioration significative en termes de MAP par rapport à une base d'évaluation de RI basée sur les mots simples. À noter que la plupart des collections ImageCLEF de 2003 jusqu'à présent portent sur la tâche de recherche des images biomédicales. De ce fait, le contenu du document est essentiellement la fusion de plusieurs légendes des images.

(Gobeill *et al.*, 2009) ont combiné l'expansion de documents et l'expansion de requêtes pour évaluer leur approche d'indexation et de recherche d'information conceptuelle basée sur les concepts MeSH. Pour chaque concept MeSH, ses synonymes et sa description sont indexés comme un document unique dans une structure d'index. Étant donné un texte (document ou requête), celui-ci est associé à de "meilleurs concepts" issus du thésaurus MeSH en utilisant une approche d'extraction de concepts basée sur la RI (similaire à la méthode de (Ruch, 2006)). Enfin, les termes identifiés dénotant des concepts MeSH sont

utilisés pour étendre le document et la requête.

```
<article pmid="10901333">
<section id="title">
  <sentence id="0" start="22" end="116
    measurement
    low
    level
    arsenic
    exposure
    a
    comparison
    water
    toenail
    concentration
  <concept id="238690" tokens="3,4"/>
  <concept id="43047" tokens="7"/>
  <concept id="681563" tokens="7"/>
  <concept id="222007" tokens="8"/>
  <concept id="86045" tokens="9"/>
  </sentence>
</section>
<section ...>
   <sentence ...>...</sentence>
   ...
</section>
...
</article>
```

```
<topic id="160">
<section id="genesymbols">
  <sentence id="0">
  prnp
  <concept id="2008214" tokens="0"/>
  </sentence>
</section>
<section id="additional">
  <sentence id="0">
    mad
    cow
    disease
    <concept id="85209" tokens="0,1,2"/>
  </sentence>
</section>
<section id="question">
  <sentence id="0">
    what
    be
    role
    prnp
    mad
    cow
    disease
    <concept id="35820" tokens="2"/>
    <concept id="85209" tokens="4,5,6"/>
  </sentence>
</section>
</topic>
```

FIGURE III.14 – Exemple de document et requête annotés par des concepts UMLS

6.3 Appariement basé sur l'identification de patrons de besoins cliniques (modèle PICO)

Le modèle PICO (Patient Intervention Control Outcome) a pour objectif de définir les quatre éléments importants d'une question clinique permettant une recherche dans la littérature scientifique :

- **Patient** correspond à la description du patient : sexe, co-morbidité, race, âge, pathologie (SCORAP),
- **Intervention** définit une intervention appliquée,
- **Control** correspond à une autre intervention permettant la comparaison ou le contrôle,
- **Outcome** correspond aux résultats de l'expérience.

Le modèle PICO permet aux professionnels de santé d'enchaîner les différentes parties d'une question clinique concernant les patients et ainsi facilite la recherche d'information en identifiant les concepts clés de la requête en langage naturel (Snowball, 1997; Villanueva *et al.*, 2001). En effet, cette méthode a fait

récemment l'objet de quelques travaux en RI biomédicale (Schardt *et al.*, 2007; Boudin *et al.*, 2010).

Le travail de (Schardt *et al.*, 2007) a pour objectif d'exploiter un modèle PICO pour améliorer la recherche d'information dans PubMed en réponse à des **questions cliniques** (voir l'exemple dans le tablau III.7). Ils ont donc défini deux instances du modèle PICO comme suit :

- le premier modèle demande à l'utilisateur de remplir manuellement les éléments basiques du modèles PICO (problème du patient, intervention, comparaison et résultats), ainsi que son âge et son sexe. D'autres éléments peuvent être sélectionnés manuellement par l'utilisateur comme le type de publications (essai clinique, méta-analyse, revue ou guide de bonnes pratiques),

- Comme le premier modèle, le deuxième modèle PICO demande à l'utilisateur de remplir manuellement les éléments basiques en intégrant d'autres éléments comme le type de question (thérapie, diagnostic, étiologie, pronostic, recherche spécifique ou recherche sensitive).

Ils ont comparé par la suite la précision des résultats correspondant à trois questions cliniques des deux modèles PICO précédents avec la précision obtenue par le moteur de recherche PubMed dans un modèle booléen [19]. Ils ont observé une amélioration au niveau de la précision des modèles PICO par rapport à la précision obtenue par PubMed. D'une part, ils n'ont pas décrit comment les éléments PICO ont été intégrés ensemble lors de l'appariement, d'autre part il est difficile de conclure l'intérêt ainsi que l'avantage du modèle PICO en raison du nombre limité de (trois) questions testées. L'inconvénient majeur de leur méthode porte sur le fait que l'utilisateur doit saisir manuellement tous les éléments du modèle PICO. De plus, la recherche booléenne de PubMed ne supporte pas l'ordonnancement selon la notion de pertinence, ce qui présente la cause principale de la dégradation des performances de la RI.

Clinical Question : For a 54-year-old woman with periodontal disease, how effective is the therapeutic use of doxcyline decrease gum bleeding and recession compared to no treatment ?	
P	54-year-old (Age) woman (Sex) with periodontal disease (Pathology)
I	Doxcyline
C	No treatment
O	Decrease gum bleeding and recession

TABLEAU III.7 – Les informations extraites du modèle PICO

19. `http://www.ncbi.nlm.nih.gov/books/NBK3827/`

Contrairement au travail décrit dans (Schardt *et al.*, 2007), les auteurs dans (Boudin *et al.*, 2010) ont proposé une approche automatique d'identification des éléments PICO à partir des documents et des requêtes dans une collection de documents biomédicaux. Les termes qui correspondent aux éléments *P, I, O* de la requêtes sont pondérés grâce à une extension du modèle de langue :

$$p_1(t|M_Q) = \gamma \times \frac{count(t,Q)}{\|Q\|} \times \left(1 + \sum_{E \in P,I,O} w_{Q,E} \times \delta(Q_E, t)\right) \tag{III.15}$$

où

- $w_{Q,E}$ est le poids de l'élément E dans la requête Q, dénoté Q_E,
- $\delta(Q_E, t)$ est une fonction binaire :

$$\delta(Q_E, t) = \begin{cases} 1 & \text{si } t \in Q_E \\ 0 & sinon \end{cases} \tag{III.16}$$

- γ est un facteur de normalisation,
- $count(t, Q)$ est la fréquence du terme t dans Q et $\|Q\|$ est la longueur de la requête.

Le score du document D vis-à-vis de la requête Q est donné par une combinaison linéaire d'interpolation des poids comme suit :

$$score(Q, D) = score(Q_{all}, D) + \sum_{E \in P,I,O} w_{Q,E} \times score(Q_E, D) \tag{III.17}$$

où $score(Q_X, D)$ est le score correspondant à une partie de la requête, c-à-d, P, I, O ou le reste, qui est calculé par la somme des poids des termes figurant dans la requête.

Afin de tenir compte de l'importance des termes dans chaque élément PICO dans le document, (Boudin *et al.*, 2010) ont proposé d'étendre le modèle de représentation du document D comme suit :

$$p_2(t|M_D) = \gamma \times \left(p(t|M_{D_{all}}) + \sum_{E \in P,I,O} w_{D,E} \times p(t|M_{D_E})\right) \tag{III.18}$$

où

- γ est un facteur de normalisation,

- $p(t|M_{D_X})$ est la probabilité que le terme soit généré par le modèle de langue correspondant aux éléments E du document en utilisant la fonction de Dirichlet.

Enfin, le score final du document D vis-à-vis de la requête Q est recalculé par le produit des probabilités p_1 et p_2 comme suit :

$$score(D, Q) = \sum_{t \in Q} p_1(t|M_Q) \times p_2(t|M_D) \tag{III.19}$$

7 Évaluation de recherche d'information biomédicale

Il existe à ce jour deux campagnes d'évaluation en RI proposant des tâches dédiées à l'évaluation de la RI biomédicale : CLEF et TREC. Les sections suivantes présentent des éléments descriptifs de ces tâches.

7.1 Campagne d'évaluation CLEF

La campagne d'évaluation CLEF (Conference and Labs of the Evaluation Forum, ayant été connue en tant que Cross-Language Evaluation Forum) a pour objectif de promouvoir la recherche, l'innovation, et le développement des systèmes de RI. Plus spécifiquement, CLEF fournit un cadre général pour les objectifs suivants :

- la RI multi-lingue et multi-modale ;
- l'investigation de l'utilisation des données non-structurées, semi-structurées, structurées pour la RI ;
- la création des collections pour un cadre d'évaluation ;
- l'exploration de nouvelles méthodologies d'évaluation en RI ;
- la discussion des résultats expérimentaux, la comparaison des approches de RI, l'échange des idées innovantes en RI.

Les tâches et les objectifs de CLEF ne cessent d'évoluer afin de couvrir différentes tâches de la recherche d'images. Parmi plusieurs pistes d'évaluation dans CLEF, la piste ImageCLEF, qui apparaît pour la première fois en 2003, a pour objectif principal d'évaluer des techniques de recherche d'images, que ce soit par contexte ou par contenu. Par exemple, en 2011, les trois sous-tâches ont été définies dans ImageCLEF 2011 : *classification de modalités*, *recherche d'images* et *recherche des cas de patients* (Kalpathy-Cramer *et al.*, 2011).

<?xml version="1.0" encoding="UTF-8"?>
<article filename="10.1007_s12178-007-9001-4.xml" doi="10.1007/s12178-007-9001-4" url=""><fulltext>Anterior impingement is a common problem in dancers occurring primarily secondary to the repetitive forced ankle dorsiflexion inherent in ballet. Symptoms generally occur progressively and may respond to conservative treatment including addressing biomechanical faults that contribute to the problem. As impingement progresses, movements essential to ballet may become impossible and arthroscopic ankle surgery is often effective for both diagnosis and treatment, allowing athletes to return to dance. Introduction Injuries in classical ballet are common and often challenging to treat for a number of reasons. Many practitioners do not understand the physical demands of the sport or the terminology describing the common mechanisms of over use. Classical ballet dancers subject themselves to repetitive loads that require progressive training over hundreds of hours both increasing the risk of overuse injury and complicating rehabilitation that involves any rest from dance. The ability to dance en pointe (on the tips of the toes) for instance, requires progressive development of the kinetic chain from the back to the toes, any disruption of which may result in overuse injury anywhere in the chain. Foot and ankle injuries that are more common in classical ballet are both anterior and posterior ankle impingement, flexor hallucis longus tendonitis, and stress fractures at the base of the second metatarsal and fibula [1]. This article will address the diagnosis and treatment of anterior ankle impingement including when performers should return to dance. Terminology Successful treatment of injuries in classical ballet starts with an understanding of the basic positions and common movements which can lead to overuse injury. The five basic positions involve maximum turnout of the hips to achieve the foot position. Inadequate hip turnout results in excessive knee or ankle external rotation and rolling in of the ankle to achieve the desired foot position (Fig.   1). Rolling in excessively stretches the medial ankle and compresses the anterior lateral ankle which can contribute to anterior impingement. Pli é (Fig.   2) is the common dance movement contributing to anterior impingement and the “ rolled in ” position of the foot exaggerates the lateral compressive forces worsening the problem.
Fig.   1 Right foot demonstrates excessive pronation of ankle in attempt to exaggerate turnout Fig.   2 Pli é resulting in compressive force to anterior ankle Pathophysiology Anterior impingement in athletes is either secondary to hypertrophied soft tissue interposed in the anterior ankle joint or proliferation of osteophytes (Fig.   3) that limit the open space between the anterior lip of the tibia and the dorsal talar neck. Different theories have been proposed to account for the pathologic changes. Because impingement often occurs in athletes subject to forced plantar flexion (soccer players kicking or

FIGURE III.15 – Exemple d'un document lié à un cas de patients dans ImageCLEF 2011

TABLEAU III.8 – Exemples de requêtes dans la tâche de recherche des cas de patients dans ImageCLEF 2011

31	Female patient, 25 years old, with fatigue and a swallowing disorder (dysphagia worsening during a meal). The frontal chest X-ray shows opacity with clear contours in contact with the right heart border. Right hilar structures are visible through the mass. The lateral X-ray confirms the presence of a mass in the anterior mediastinum. On CT images, the mass has a relatively homogeneous tissue density.
32	Pain and incapacity to move after an accident. Slight deformation can be seen in the x-ray.
33	Accident on a public street, man, 38 years old, x-ray of the thorax shows displacement of mediastinum to the right. Fractures dominantly on the left.

La tâche de recherche des cas de patients (*Case-based Retrieval*), qui a été introduite pour la première fois en 2009, a pour but de rechercher des cas de patients qui répondent potentiellement à un essai clinique. Dans cette tâche, chaque document (unité de la recherche) est constitué du texte intégral y compris les légendes des photos biomédicales (*cf.* la figure III.15). Les requêtes sont constituées d'une description de cas de patients, avec la démographie des patients, des symptômes et des résultats de tests, y compris des études d'imagerie, sont fournies. Le tableau III.8 montre quelques exemples de requêtes dans dans la tâche de recherche des cas de patients dans ImageCLEF 2011 (Kalpathy-Cramer *et al.*, 2011). Quelques statistiques sur les collections ImageCLEF (Case-based Retrieval) sont présentées dans le tableau III.9.

ImageCLEF / Caract.	2009	2010	2011
Nb. de documents	70,000+	77,506	55,634
Taille moyenne	334 tokens	335 tokens	3,078 tokens
Nb. de requêtes	5	14	10

TABLEAU III.9 – Statistiques des collections ImageCLEF (Case-Based IR)

7.2 Campagne d'évaluation de TREC

La campagne d'évaluation TREC, qui a eu lieu pour la première fois en 1992, fournit des cadres d'évaluation des approches de recherche d'information ainsi qu'un forum pour comparer les résultats obtenus par les différents groupes de recherche. TREC est organisé comme un événement annuel qui sollicite les différentes équipes de recherche dans le monde entier à participer à plusieurs tâches de RI comme la recherche *ad-hoc* des documents, des systèmes question-réponse, ou des systèmes de recherche cross-lingue, etc. Parmi les différentes pistes de recherche, nous nous intéressons particulièrement à deux pistes suivantes : (1) **TREC Genomics** pour la RI de la littérature biomédicale et (2) **TRECMed** pour la RI des dossiers médicaux de patients.

7.2.1 TREC Genomics pour la RI de la littérature biomédicale

La piste TREC Genomics, qui a duré de 2003 à 2007, est devenue une des pistes de recherche importantes dans le domaine biomédical, notamment les documents dans la base bibliographique de MEDLINE[20]. Il s'agit de la base de données bibliographiques de premier ordre, développée et maintenue par la NLM[21]. MEDLINE contient plus de 21 millions de références d'articles en science de la vie, notamment de la biomédecine. Les documents de MEDLINE ont été extraits pour développer des collections tests dédiées à l'évaluation des approches de RI dans TREC Genomics.

Un des principaux objectifs de TREC Genomics porte sur la recherche *ad hoc* (Hersh et Bhupatiraju, 2003; Hersh *et al.*, 2004, 2005). Au début de TREC Genomics en 2003, une sous-collection de MEDLINE qui couvre les données entre 2002 et 2003 a été extraite pour construire une collection d'évaluation à partir de 525,938 enregistrements de MEDLINE. Chaque enregistrement (appelé MEDLINE record) contient plusieurs champs importants pour les expérimentations en RI comme : .PMID (PubMed Unique IDentifier), .TI (TItle), .AB (ABstract), MH (MeSH Headings), ...

20. `http://www.nlm.nih.gov/pubs/factsheets/medline.html`
21. US National Library of Medicine

Le tableau III.10 présente un exemple d'un document utilisé dans TREC Genomics sous le format de la NLM. Une description détaillée des champs d'un enregistrement de MEDLINE est décrite sur le site[22] de la NLM. Les participants doivent télécharger la collection, et puis l'indexer en utilisant un système d'indexation et de recherche d'information particulier pour obtenir les résultats. La tâche de RI *ad-hoc* a été évoluée en 2004 avec une collection test qui contient plus de documents extraits de MEDLINE (4,591,008 enregistrements de MEDLINE publiés entre 1994 et 2003). Les requêtes ont été construites en se basant sur les besoins d'informations des scientifiques dans le domaine biomédical. En 2005, TREC Genomics a réutilisé la même collection de documents créée en 2004 mais avec un ensemble de 50 nouvelles requêtes. Ces dernières ont été générées en se basant sur les requêtes génériques. Une requête générique permet de grouper les besoins d'informations ayant les mêmes sujets comme *gène*, *protéine* ou *maladie*. Par exemple, une instance de la requête générique "Find articles describing the role of a [**gene**] involved in a given [**disease**]" peut être comme suit :

"Find articles describing the role of **DRD4** involved in **alcoholism**"

Dans cette requête, **"DRD4"** est le nom d'une gène et **"alcoholism"** est le nom d'une maladie.

L'évaluation dans TREC Genomics, comme dans l'ensemble des pistes de TREC, est basée sur le "paradigme de Cranfield" qui mesure les performances d'un système de RI en se basant sur les quantités de documents pertinents retrouvés, en particulier les mesures de *rappel* et de *précision*. Sur le plan opérationnel, le rappel et la précision sont calculés en utilisant une collection de test des documents, les requêtes et les jugements de pertinence. Dans la plupart des pistes de TREC, les deux sont combinés en une seule mesure de performance, appelée *précision moyenne* (MAP - Mean Average Precision), qui mesure la précision après que chaque document pertinent est récupéré pour une requête donnée. La mesure MAP est en général calculée sur l'ensemble des requêtes.

Le tableau III.11 donne un aperçu sur les différentes tâches de RI *ad hoc* au fil des années de TREC Genomics (2003-2007). Chaque collection possède des caractéristiques différentes en termes de documents et de requêtes. Par exemple, la collection TREC Genomics 2003 porte uniquement sur la recherche des articles contenant les noms de gène d'un espèce particulier comme "Homo sapiens", d'où la requête ne contient que le nom de gène et ses variants associés manuellement en utilisant LocusLink (Hersh et Bhupatiraju, 2003). Les collections TREC Genomics 2004 et 2005 contiennent une quantité volumineuse de documents, 4,5 millions d'enregistrements issus de MEDLINE. La différence entre la collection TREC Genomics 2004 et 2005 porte sur la création de re-

22. `http://www.nlm.nih.gov/bsd/mms/medlineelements.html`

TABLEAU III.10 – Un exemple de documents utilisés dans TREC Genomics 2004-2005

PMID- 11096424
TI - Prenatal radiation-induced limb defects mediated by Trp53-dependent apoptosis in mice.
PG - 673-9
AB - We reported previously that in utero radiation-induced apoptosis in the predigital regions of embryonic limb buds was responsible for digital defects in mice. To investigate the possible involvement of the Trp53 gene, the present study was conducted using embryonic C57BL/6J mice with different Trp53 status. **...** These findings suggest that the wild-type Trp53 gene may be an intrinsic genetic susceptibility factor that is responsible for certain congenital defects induced by prenatal irradiation.
FAU - Wang, B
AU - Wang B
FAU - Ohyama, H
AU - Ohyama H
FAU - Haginoya, K
AU - Haginoya K
...
MH - Abnormalities, Radiation-Induced/*genetics/pathology
MH - Animals
MH - Apoptosis/*radiation effects
MH - Dose-Response Relationship, Radiation
...
MH - Pregnancy
MH - *Prenatal Exposure Delayed Effects
MH - Radiation Tolerance/genetics
MH - Tumor Suppressor Protein p53/deficiency/*genetics/metabolism
EDAT- 2000/11/30 11 :00
MHDA- 2001/02/28 10 :01
CRDT- 2000/11/30 11 :00
PST - ppublish
SO - Radiat Res. 2000 Dec ;154(6) :673-9.

quêtes test : les requêtes de la première représentent de vrais besoins d'informations des vrais biologistes ou professionnels de santé (Hersh *et al.*, 2004) avec l'utilisation des acronymes de noms de gènes sans leur forme complète tandis que celles en 2005 sont regroupés en plusieurs types de requêtes génériques avec l'utilisation des acronymes associés manuellement par leur forme complète. A partir de 2006, les tâches portent sur l'évaluation des performances des systèmes de question-réponse : étant donnée une question, le système doit renvoyer les réponses les plus pertinentes en terme de **paragraphes** ou **un segment de texte** dans les documents.

Année	Documents	Requêtes
2003	525.938 enregistrements de MEDLINE (4/2002-4/2003)	Recherche des enregistrements MEDLINE qui concernent les gènes et les protéines d'un organisme. Ex : "Homo sapiens : activating transcription factor 2 (ATF-2 ; HB16 ; CREB2 ; ...).
2004	**4.591.008** enregistrements MEDLINE (1994-2003)	50 requêtes construites en se basant sur les besoins d'informations de vrais biologistes, avec l'utilisation des **acronymes sans leur forme complète**. Ex : "Find articles about function of **FancD2**"
2005	**4.591.008** enregistrements MEDLINE (1994-2003)	50 requêtes construites en se basant sur les exemplaires prédéfinis, avec **l'utilisation des acronymes et leur forme complète**. Ex : "*Provide information on the role of the gene* **gamma-aminobutyric acid receptors** (GABABRs) in the process of inhibitory synaptic transmission"
2006	162.259 articles de journaux sous le format HTML publiés via Highwire Press	28 requêtes sous forme de questions basées sur les exemplaires prédéfinis. Ex : "What is the role of PrnP in mad cow disease ?"
2007	162.259 articles de journaux sous le format HTML publiés via Highwire Press	36 requêtes sous forme de questions basées sur 14 entitées. Ex : "Which [PATHWAYS] are mediated by CD44 ?"

TABLEAU III.11 – Résumé des tâches de RI dans le cadre de TREC Genomics

7.2.2 TRECMed pour la RI biomédicale des comptes-rendus médicaux de patients

Introduite pour la première fois dans TREC en 2011, la piste TRECMed a pour objectif de promouvoir la recherche et le développement des méthodologies de RI pour la récupération des documents liés aux dossiers médicaux de patients. TRECMed est donc un grand effort de la communauté de la RI biomédicale en fournissant un large cadre d'évaluation des dossiers médicaux de patients. Ces derniers contiennent des données obtenues des vrais patients dans les hôpitaux et donc nécessitent un processus de protection de données

personnelles du patient afin de garantir la confidentialité des données.

La tâche principale dans TRECMed est liée à la recherche *ad-hoc* des patients ou groupes de patients qui correspondent au besoin de l'utilisateur (exprimé par une ou plusieurs requêtes). La collection est fournie aux participants via l'accord avec l'université de Pittsburgh. Chaque compte-rendu (CR) dans la collection possède un identifiant unique, appelé *report checksum*. La plupart des CR sont liés à une consultation du patient. Une table d'association entre les CR et les consultations des patients (*mapping table*). Pour simplifier, nous pouvons appeler l'ensemble des CR correspondant à un patient un dossier médical du patient (DMP). Le nombre de CR d'un DMP peut varier entre 1 et 415. La figure III.16 illustre un compte-rendu du patient enregistré sous le format XML. Le tableau III.12 présente quelques requêtes exemples dans TRECMed 2011. Pour chaque requête, le système de RI doit retourner tous les DMP pertinents vis-à-vis de la requête. L'évaluation des performances de la RI est basée sur la capacité du système de RI de renvoyer les meilleurs DMP pour un ensemble de 35 requêtes au total.

<top>
<num> Number : 101
<desc>
Patients with hearing loss
</top>

<top>
<num> Number : 102
<desc>
Patients with complicated GERD who receive endoscopy
</top>

<top>
<num> Number : 103
<desc>
Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis
</top>

TABLEAU III.12 – Exemples de requêtes dans TRECMed 2011

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<report>
<checksum>20070224SP-jRp/Zo6oBtIc-489-832553573</checksum>
<subtype/>
<type>SP</type>
<chief_complaint/>
<admit_diagnosis/>
<discharge_diagnosis/>
<year>2007</year>
<downlaod_time>2009-10-05</downlaod_time>
<update_time/>
<deid>v.6.22.08.0</deid>
<report_text>[Report de-identified (Safe-harbor compliant) by De-ID v.6.22.08.0]


PATIENT HISTORY:
No clinical history is given.
PRE-OP DIAGNOSIS: Failed left total shoulder.
POST-OP DIAGNOSIS: Same.
PROCEDURE: Revision left total shoulder.
tw
_____________________________________________________________
FINAL DIAGNOSIS:
IMPLANTS REMOVAL AND TISSUE, LEFT SHOULDER,  EXCISION
A.          SCAR WITH SYOVIAL/GANGLION CYSTS AND NO EVIDENCE OF ACUTE
INFLAMMATION.
B.     ORTHOPEDIC HARDWARE (gross examination only).
**INITIALS
**INITIALS
     Pathologist:  **NAME[WWW M. M. XXX], M.D.
     Fellow/Chief Resident:  **NAME[VVV M. UUU], M.D.
     ** Report Electronically Signed Out **
     By Pathologist:  **NAME[WWW M. M. XXX], M.D.
     **DATE[Feb 26 2007] 11:12
My signature is attestation that I have personally reviewed the submitted
material(s) and the final diagnosis reflects that evaluation.
_____________________________________________________________
GROSS DESCRIPTION:
The specimen is received fresh, labeled with the patient's name, initials DG
and "removed implants left shoulder".  It consists of a silver metallic
humeral head, polyethylene implant and 2.0 x 1.0 x 1.0 cm aggregate of red-tan
tissue. The tissue is entirely submitted in cassette A.
tw
**INITIALS
MICROSCOPIC:
Microscopic examination substantiates the above diagnosis.
caa

The following statement applies to all immunohistochemistry, Insitu
Hybridization Assays (ISH &amp;amp; FISH), Molecular Anatomic Pathology, and
Immunofluorescent Testing:
The testing was developed and its performance characteristics determined by
```

FIGURE III.16 – Exemple d'un compte-rendu du DMP dans TRECMed 2011

8 Conclusion

Ce chapitre donne un aperçu sur l'état-de-l'art de la RI biomédicale. Nous avons donc présenté une typologie de l'information biomédicale, y compris la littérature biomédicale et les dossiers médicaux personnels. La littérature biomédicale a fait l'objet de nouveaux travaux de recherche en RI biomédicale, en particulier dans le cadre de TREC Genomics 2003-2007. Les comptes-rendus médicaux ont été mis à disposition pour la première fois dans le cadre de la campagne d'évaluation TRECMed 2011. Les cas de patients ont fait l'objet de la RI *ad hoc* avec la participation des groupes de recherche dans le cadre de la piste ImageCLEF depuis 2009.

Nous avons donc décrit les principales ressources termino-ontologiques les plus utilisées dans le domaine. Les approches ainsi que les outils d'extraction de concepts sont présentés dans le contexte de l'extraction d'information de manière générale. Ces méthodes peuvent être appliquées pour extraire les concepts à partir des documents ou de la requête de l'utilisateur. Quelques travaux, notamment ceux qui ont obtenu les meilleurs résultats dans TREC Genomics, ont rapporté l'utilité d'intégrer les concepts dans l'appariement de la requête, d'autres travaux n'ont pas trouvé l'intérêt de la détection des concepts de la requête. La conclusion peut être retirée est que les concepts extraits doivent être issus des "bonnes" ressources termino-ontologiques et doivent être intégrés de "manière adéquate" dans un processus de RI afin d'améliorer les performances de la RI. En général, les concepts extraits à partir du document ou de la requête sont d'abord vérifiés par un ou plusieurs experts du domaine ou des indexeurs humains afin de garantir la pertinence des concepts extraits à la description sémantique du contenu textuel.

Dans le cadre de la RI biomédicale, nous avons présenté en catégorisant les différentes approches de RI biomédicales, à savoir l'approche basée sur l'expansion conceptuelle en utilisant les termes extraits à partir des ressources terminologiques, l'approche basée sur la reformulation de la requête PRF ainsi que l'approche basée sur l'expansion documentaire automatique ou manuelle. En particulier, nous avons présenté les différentes approches de la RI biomédicale : *l'expansion conceptuelle de la requête*, *l'expansion de la requête par la méthode PRF* ainsi que *l'expansion documentaire*. Ces méthodes peuvent être combinées ensemble pour générer une meilleure représentation sémantique de l'information par rapport à la représentation classique par des sacs de mots simples. De plus, les nouveaux modèles d'appariement basés sur les patrons de besoins cliniques, appelés modèle PICO, peuvent être utiles afin de mieux cerner les besoins d'information spécifiques des professionnels de la santé.

Partie II

Proposition et Évaluation des Modèles d'Indexation et d'Accès à l'Information Biomédicale

Introduction

L'état-de-l'art présenté dans la première partie a pour but de donner un aperçu sur le domaine de la RI biomédicale. En particulier, nous avons présenté des méthodes et techniques d'indexation biomédicale en se basant sur une ou plusieurs ressources termino-ontologiques. Nous avons catégorisé ces travaux en deux principales approches : **indexation mono-terminologique** *vs.* **indexation multi-terminologique**. Pour les méthodes et techniques de recherche d'information biomédicale, nous avons présenté en particulier les techniques de reformulation de la requête (conceptuelle et/ou PRF) ainsi que les techniques d'expansion documentaire pour une indexation conceptuelle/sémantique.

Les défis majeurs en indexation et recherche d'information biomédicale sont résumés comme suit :

1. La plupart des approches d'indexation biomédicale ont pour but d'assister les indexeurs humains à sélectionner les termes d'index à partir des documents biomédicaux. À ce niveau, le "passage à échelle" devient un grand défi pour les méthodes d'extraction automatique des concepts sur les collections volumineuses, e.g., les collections des dizaines milliers de documents jusqu'à des millions de documents. Peu de travaux jusqu'à présent se focalisent sur l'évaluation de l'impact de l'indexation conceptuelle des documents sur une collection volumineuse. À noter que le travail de (Zhou *et al.*, 2006c) a pour but d'extraire les concepts à partir d'un ensemble d'une quarantaine de milliers de documents issus de TREC Genomics 2004 et de les indexer avec les concepts MeSH ou UMLS.

2. Les modèles d'appariement dans le domaine de la RI biomédicale s'orientent en général vers la représentation sémantique de la requête via les méthodes d'extraction des concepts automatique ou manuelle en utilisant des ressources termino-ontologiques. Le défi majeur à ce niveau consiste à proposer un nouveau modèle de représentation conceptuel/sémantique basé sur la granularité, notamment les concepts, qui est "le plus adéquat" à la tâche de recherche permettant de répondre à des besoins d'information des utilisateurs du domaine.

3. Les techniques d'extraction des concepts largement adoptées dans le domaine médical sont basées, en partie ou totalement, sur des processus d'analyse du langage naturel, ce qui peut s'avérer coûteux et peu généraliste en raison de la nécessité d'entraînement des méthodes sous-jacentes. De plus, peu de techniques exploitent la spécificité du langage médical, en terme de granularité et de poly-représentation.

4. Les approches d'indexation multi-terminologiques reposent sur l'utilisation directe du méta-thésaurus UMLS. Cependant, son exploitation directe, peut poser le problème du défaut d'appariement pour deux raisons : la première raison concerne l'usage du principe d'alignement de concepts (terminologies) proposé de fait dans le méta-thésaurus UMLS qui suppose des relations de correspondance pré-établies à partir du seul concept reconnu dans la langue identifiée. De plus, l'UMLS intègre une centaine de terminologies parmi lesquelles il y a des terminologies qui ne sont probablement pas pertinentes pour encoder l'information dans le texte (source du "bruit"). La seconde raison concerne la disponibilité de plus en plus accrue de ressources termino-ontologiques non intégrées dans UMLS. C'est de manière générale le cas de ressources non standardisées comme le thésaurus Vidal ou les terminologies privées comme le thésaurus TUV de la société Vidal (Pereira *et al.*, 2008; Darmoni *et al.*, 2009; Sakji, 2010).

Dans cette partie, nous présentons nos contributions dans le domaine de la RI biomédicale pour répondre à ces défis. Nos travaux se déclinent en trois volets principaux :

1. Un modèle d'indexation sémantique de documents biomédicaux. Ce modèle est basé sur une extraction de concepts pertinents à un document et/ou à la requête en tenant compte du granule termino-ontologique en termes de composition et ordre des termes préférentiels associés. Les entrées des concepts reconnus, dans le cas de la terminologie MeSH, sont de plus, désambiguïsés en tenant compte du sous-domaine d'appartenance. Les concepts ainsi identifiés sont alignés dynamiquement grâce à des techniques de fusion qui considèrent leur importance dans la représentativité du texte. L'intégration de plusieurs terminologies a pour objectif d'augmenter la précision et le rappel de l'extraction de concepts. Pour cela, nous utilisons différents modèles de vote visant à fusionner les concepts candidats termino-ontologiques extraits à partir de textes. Les meilleurs concepts candidats sont exploités pour étendre les documents et/ou la requête en utilisant les termes préférés désignant les concepts.

2. Des techniques d'expansion combinée de documents et requêtes pour une recherche d'information basée sur le contexte de données. Plus spécifiquement, nous étudions l'impact d'utiliser les mots-clés issus des concepts extraits pour étendre les documents et/ou la requête. Nous considérons les ressources termino-ontologiques comme le **contexte global** car les concepts sont définis de manière globale indépendamment du contenu du document ou de la requête courante. De plus, notre objectif est également d'étudier les performances des méthodes d'extraction de termes ou mots-clés pertinents à partir des premiers documents retournés afin de reformuler la requête. Nous appelons ces documents le **contexte**

local de la requête car ce contexte change en fonction de la sémantique de chaque requête. Nous étudions donc l'effet de combiner les contextes globaux ou locaux ensemble pour améliorer des performances de la RI biomédicale.

3. Enfin, nous présentons notre plateforme de RI biomédicale, intitulée BioSIR, visant à supporter les tâches d'indexation et de recherche d'information des documents biomédicaux. BioSIR permet d'indexer les documents avec les concepts issus d'une ou de plusieurs terminologies et d'étendre les documents et/ou les requêtes avec les concepts les plus importants pour améliorer les performances de la RI.

Chapitre IV

Résolution de l'ambiguïté des termes MeSH orientée domaine et son impact sur un processus de RI

Sommaire

1 Introduction **124**

2 Problématique et motivations **125**

3 RI basée sur les domaines des termes MeSH **128**

3.1 Indexation sémantique basée sur les domaines MeSH 130

3.1.1 Algorithme de désambiguïsation 1 (Left-to-Right TSD) 131

3.1.2 Algorithme de désambiguïsation 2 (Cluster-based TSD) 134

3.2 Appariement sémantique document-requête 136

4 Évaluation expérimentale **139**

4.1 Cadre d'évaluation 139

4.2 Résultats expérimentaux 141

4.3 Discussion 145

5 Conclusion **145**

"Everything that can be counted does not necessarily count ; everything that counts cannot necessarily be counted."
–Albert Einstein

Publication liée à ce travail

- **NLDB 2010 :** Duy Dinh, Lynda Tamine. *Sense-based biomedical indexing and retrieval*. Dans : the 15th International Conference on Applications of Natural Language to Information Systems, NLDB 2010, Juin 23-25, 2010, Cardiff University, Cardiff, Wales, UK, p. 24–35.

1 Introduction

Dans le chapitre III, nous avons présenté un état-de-l'art sur les méthodes d'indexation et de recherche d'information conceptuelle/sémantique dans le domaine biomédical. Ces méthodes exploitent donc les sources d'évidence dans les ressources termino-ontologiques comme la synonymie, les relations hiérarchiques, etc. pour améliorer les performances de la RI. Il est largement reconnu, que la terminologie MeSH est la plus exploitée pour l'indexation de documents de la littérature scientifique médicale. Ceci peut être attesté par les travaux d'indexation des articles ou résumés d'articles de la base MEDLINE à la NLM (Aronson *et al.*, 2004b). Chaque résumé peut être indexé par une dizaine de termes désignant les concepts MeSH (Névéol *et al.*, 2009).

Nous présentons dans ce chapitre une étude préliminaire effectuée dans le cadre de nos travaux de thèse, sur l'analyse structurelle de la terminologie MeSH et l'exploitation des résultats de cette analyse pour en mesurer l'impact sur un processus de RI médicale exploitant la seule terminologie MeSH. Notre objectif est d'étudier dans quelle mesure la prise en compte du domaine ou branche de domaine associée à une entrée de la terminologie permet d'améliorer le sens le plus "adéquat" du contenu documentaire. Plus précisément, nous exploitons la structure poly-hiérarchique du thésaurus MeSH pour "désambiguïser" les termes ayant plusieurs domaines dans les documents et les requêtes. Un terme MeSH est dit "ambigu" s'il est défini par plusieurs domaines dans la structure poly-hiérarchique.

Nous proposons ici deux méthodes de désambiguïsation des termes MeSH basées principalement sur la corrélation des termes voisins relativement à des sens candidats du même domaine. Ces méthodes de désambiguïsation sont évaluées sur la collection TREC OHSUMED constituée de résumés d'articles de journaux biomédicaux.

La suite de ce chapitre est organisée comme suit : la section 2 présente les problématiques et les objectifs principaux de notre contribution. La section 3 décrit nos méthodes d'identification des domaines MeSH (désambiguïsation) et le processus d'indexation des documents biomédicaux basée sur les domaines des termes désignant les concepts biomédicaux. Une évaluation expérimentale est présentée et discutée dans la section 4. La section 5 conclut ce chapitre.

2 Problématique et motivations

L'objectif principal de notre étude est d'améliorer les performances de la RI conceptuelle basée sur la ressource MeSH. Nous évaluons l'impact de l'intégration des termes MeSH ainsi que leurs domaines associés sur les performances de la RI. Nous rappelons que les concepts MeSH sont organisés en architecture poly-hiérarchique dans le sens où chaque concept est représenté par un nœud, qui appartient à une ou plusieurs hiérarchies dont chacune est identifiée par un numéro d'arbre. Chaque sous-arbre correspond à un des seize domaines du MeSH, e.g., A - Anatomie, B - Organismes, C - Maladies, etc. Chaque concept est dénoté par un terme préféré et/ou plusieurs termes non-préférés. Les termes préférés sont souvent retenus pour l'indexation (par les indexeurs humains) tandis que les termes non-préférés sont utilisés de manière libre (dans les documents et les requêtes). Chaque terme MeSH désigne un seul concept uniquement tandis qu'un concept peut être désigné par plusieurs termes (relation terme-concept de type 1-N).

En considérant ces caractéristiques du thésaurus MeSH, qui a été largement utilisé pour l'indexation conceptuelle/sémantique des documents biomédicaux, nous révélons ici deux types d'ambiguïté dans les documents biomédicaux :

1. Un mot a plusieurs sens (relation mot-sens de type 1-N), par exemple, le mot "cold" peut signifier "temperature" ou "disease".
2. Un terme est une entrée à un seul concept uniquement (relation terme-concept de type 1-1). Chaque concept est lié à plusieurs domaines ou sous-domaines, par exemple, dans le thésaurus MeSH, le terme "Necrotic DNA Degradation" peut appartenir à trois sous-domaines différents, en l'occurrence " Pathological Conditions, Signs and Symptoms" [C23], "Cell Physiological Phenomena [G04]" et "Genetic Phenomena" [G05].

Le premier type d'ambiguïté a été l'objet de nombreux travaux de recherche comme mentionnés dans la section 3.2 du chapitre II, dans un domaine général, qui a été abordé dans une tâche désambiguïsation basée sur les (sous-)domaines (Gliozzo *et al.*, 2004; Buitelaar *et al.*, 2007) ou dans un domaine spécifique comme le domaine biomédical (Schiemann *et al.*, 2008).

Nous nous intéressons en particulier au deuxième type d'ambiguïté où un terme (simple ou complexe) peut être associé à plusieurs domaines. À notre connaissance, aucun travail jusqu'à présent exploite cette caractéristique dans un processus de RI biomédicale. Cependant, notre approche trouve essence dans les travaux de désambiguïsation orientés domaines, présentés dans le chapitre III. Cette approche globale préconise la prise en compte du domaine contextuel dans la sélection du sens approprié du terme, appliqué plus particulièrement en utilisant WordNet, où un domaine est apparenté à un sens candidat. Dans notre cas, le domaine traduit la vue du concept selon ses différentes déclinaisons dans un contexte médical.

À titre illustratif, le concept "Pain" appartient à quatre branches de trois (sous-)domaines (*cf.* la figure IV.1) dont les concepts les plus génériques sont : *Nervous System Disease* (C10) ; *Pathological Conditions, Signs and Symptoms* (C23) ; *Psychological Phenomena and Processes* (F02) ; *Musculoskeletal and Neural Physiological Phenomena* (G11). Un autre exemple concernant le concept dénoté par le terme complexe "Sleep Initiation Dysfunctions" dans MeSH peut être interprété comme une "maladie du système nerveux" [C10.886.425.800.8005] ou "désordre mental" [F03.870.400.800.800].

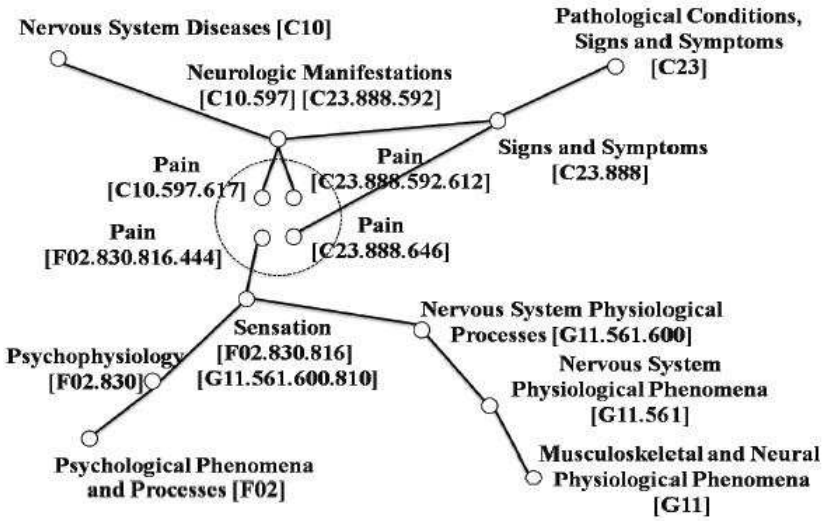


FIGURE IV.1 – Visualisation du concept "Pain" dans l'architecture polyhiérarchique de la terminologie MeSH

Plus généralement, le thésaurus MeSH (version 2009) est composé de 25186 concepts dont chacun est désigné par un terme préféré (main heading) et/ou un ou plusieurs termes non-préférés. Le tableau IV.1 présente en ordre croissant le nombre de (sous-)domaines partagés par les concepts MeSH. L'analyse de la répartition des concepts polyhiérarchiques de MESH (*cf.* le tableau IV.1) montre que plus de 50% de concepts MeSH sont associés à plus d'un domaine, ce qui nous a motivé d'avantage dans la réalisation de cette étude. Par exemple, 6593 concepts sont liés à 2 (sous-)domaines dans l'arborescence du MeSH.

TABLEAU IV.1 – Nombre de (sous-)domaines partagés par les concepts MeSH

Nb. (sous-)domaines	Nb. concepts	Pourcentage
1	12477	49.54%
2	6593	26.18%
3	3291	13.07%
4	1471	5.84%
5	652	2.59%
6	343	1.36%
7	157	0.62%
8	74	0.29%
9	55	0.22%
10	39	0.15%
11	11	0.04%
12	10	0.04%
13	5	0.02%
14	4	0.02%
17	1	0.004%
19	1	0.004%

De plus, les concepts issus du thésaurus MeSH ont des niveaux de spécificité différents, c-à-d que les concepts peuvent être génériques ou spécifiques. En RI biomédicale, les concepts plus spécifiques pourraient être intéressants car les informations exprimées via ces concepts sont plus ciblées que des informations génériques. Par exemple, la phrase "*Le patient souffre d'un* ***cancer de la thyroïde***." est plus informative que la phrase "*Le patient souffre d'un* ***cancer***.". Nous pouvons quantifier la spécificité des concepts grâce à leur profondeur dans la hiérarchie de concepts. La figure IV.2 indique la variation de la spécificité des concepts MeSH : les concepts les plus génériques se trouvent à la racine tandis que les concepts spécifiques se trouvent au niveau des nœuds feuilles.

Nos contributions associées à cette analyse exploratoire concernent :

1. Deux méthodes de désambiguïsation des termes désignant les concepts dans différents domaines en exploitant la structure poly-hiérarchique du MeSH. Plus spécifiquement, chaque terme est assigné par le sens le plus adéquat selon son contexte d'utilisation. Nos approches de désambiguïsation sont basées sur la densité des concepts dans cette structure poly-hiérarchique.
2. L'intégration de nos méthodes de désambiguïsation dans un processus de RI sémantique basée sur le sens précis des termes. L'appariement document-requête est basé sur les critères classiques (fréquence de termes, longueur du document, ...) et en plus sur la proximité sémantique entre le document et la requête.

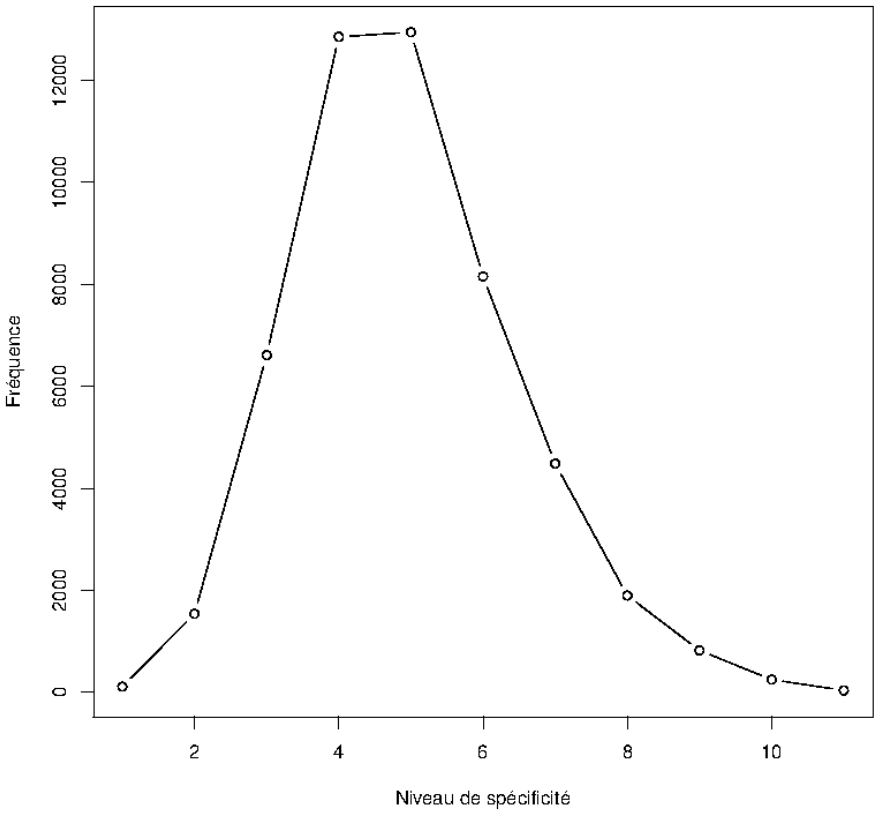


FIGURE IV.2 – Variation de la spécificité des concepts MeSH

3 RI basée sur les domaines des termes MeSH

Dans le cadre de la RI biomédicale, notre processus de RI sémantique est composé de deux étapes principales détaillées dans les sections qui suivent :

1. *Indexation sémantique basée sur les domaines MeSH :* l'objectif de l'indexation sémantique est d'assigner les concepts aux documents de la collection. Dans un premier temps, une analyse lexicale est appliquée aux documents qui sont annotés par des étiquettes grammaticales (e.g., nom, verbe, adjectif ...). Nous avons utilisé TreeTagger (Schmid, 1994) comme l'outil d'analyse lexicale, qui a été développé à l'Institut Computational Linguistics à l'université de Stuttgart. TreeTagger supporte plusieurs langues différentes comme l'anglais, le français... La sortie de l'analyse lexicale est un ensemble de tokens/mots sous forme canonique (lemme) avec leur catégorie grammaticale (*cf.* l'exemple dans le tableau IV.2).

   Ensuite, le texte est découpé en phrases où les termes désignant les concepts seront identifiés. Pour chaque phrase, nous utilisons un algorithme de recherche des groupes nominaux qui contiennent la séquence de mots la plus longue trouvée dans le dictionnaire de termes. Les termes

Entrée	**Sortie**		
A continuous surveillance by public health authorities will be critical to monitor the appearance of new influenza variants.	A	DT	a
	continuous	JJ	continuous
	surveillance	NN	surveillance
	by	IN	by
	public	JJ	public
	health	NN	health
	authorities	NNS	authority
	will	MD	will
	be	VB	be
	critical	JJ	critical
	to	TO	to
	monitor	VV	monitor
	the	DT	the
	appearance	NN	appearance
	of	IN	of
	new	JJ	new
	influenza	NN	influenza
	variants	NNS	variant
	.	SENT	.

TABLEAU IV.2 – Exemple d'analyse lexicale en utilisant TreeTagger

identifiés servent d'*index sémantique* tandis que les tokens qui ne sont pas identifiés dans la phrase servent d'*index classique* qui est basé sur les mots simples. Nous considérons ici deux types de descripteurs : termes issus du MeSH *vs.* mots du vocabulaire libre.

Une fois que les termes issus du MeSH sont identifiés, nous procédons une étape de désambiguïsation pour déterminer le domaine le plus adéquat de chaque terme ambigu en exploitant la structure poly-hiérarchique du thésaurus MeSH.

2. *Appariement sémantique document-requête :* lors de la recherche, la requête est traitée de la même manière que les documents. Les documents étant indexés par au moins un mot simple ou terme MeSH identifié dans la requête sont mis en correspondance avec la requête pour déterminer son degré de pertinence. L'appariement document-requête est réalisé selon un schéma de pondération sémantique qui combine les deux représentations : l'une est basée sur l'appariement des termes issus du MeSH avec leur domaine détecté (représentation sémantique) et l'autre est basée sur l'appariement des mots simples du vocabulaire (représentation classique).

3.1 Indexation sémantique basée sur les domaines MeSH

Notre approche de désambiguïsation consiste à sélectionner le domaine le plus adéquat d'un terme qui désigne un concept particulier dans le document. Les définitions et notations suivantes sont utilisées dans la description de nos méthodes de désambiguïsation :

Définition 1 : Un **mot** est une chaîne de caractères alphanumériques dont chacun est séparé de l'autre par un espace, e.g., "cancer", "prostate", "maladie", "rare"... sont considérés comme des mots simples.

Définition 2 : Un **terme** composé d'un ou plusieurs mots détermine une unité linguistique dans un vocabulaire contrôlé, e.g., "cancer du prostate", "maladie rare", ... sont des termes composés des mots. On peut également les appeler *groupes nominaux*.

Définition 3 : Un **concept** représente une classe sémantique d'un objet et peut être désigné par un ou plusieurs termes synonymes dont un terme est préféré et le reste correspond aux termes non-préférés du concept. Par exemple, dans la terminologie MeSH, "Neoplasms" est le terme préféré qui désigne le concept **cancer** tandis que les termes comme "cancer", "tumours" sont des termes non-préférés de ce concept.

Définition 4 : Le **sens** d'un concept est représenté par un nœud, indiqué par le numéro d'arbre dans une architecture polyhiérarchique. L'ensemble de sens d'un terme qui désigne un concept c sachant que c appartient à plusieurs domaines, est désigné par *syn(c)*. Par exemple, syn("Cold Temperature") = { G01.906.595.272, G16.500.275.063.725.710.300, G16.500.750.775.710.300, N06.230.300.100.725.154, N06.230.300.100.725.710.300 }.

Définition 5 : La relation **is-a** relie les concepts d'une même hiérarchie. Cette relation permet d'identifier les concepts les plus génériques ou les plus spécifiques liés à un sujet particulier.

Nous détaillons dans ce qui suit les deux méthodes de désambiguïsation proposées dans le cadre de la RI biomédicale (Dinh et Tamine, 2010a).

3.1.1 Algorithme de désambiguïsation 1 (Left-to-Right TSD)

La première méthode de désambiguïsation, appelée TSD1 [1], concerne la sélection du sens le plus adéquat de chaque terme ambigu, calculé de proche en proche en partant de gauche à droite du discours. Notre algorithme de désambiguïsation est basé sur les hypothèses suivantes :

- l'unicité du sens d'un terme qui désigne un concept spécifique dans le document (Gale *et al.*, 1992),
- la corrélation des sens des termes voisins : les sens associés à des termes voisins sur une fenêtre sont sémantiquement proches les uns des autres,
- la priorité du sens est définie par la position des termes désignant les concepts : le concept qui se trouve le plus à gauche détermine le sens global de la suite du discours. Ce principe est inspiré par la notion de *chaîne sémantique* du discours (Morris et Hirst, 1991). Cette dernière a été définie comme une séquence de mots $w_1, w_2, ..., w_i, w_{i+1}, ..., w_n$ qui sont reliés sémantiquement dans un texte en sorte que le mot w_i soit relié au mot w_{i+1} par une relation lexico-sémantique.

En se basant sur ces hypothèses, d'abord nous calculons la *similarité sémantique* entre le terme identifié le plus à gauche et son voisin le plus proche. La sélection du sens le plus adéquat pour chaque terme est déterminée selon la valeur de la similarité entre chaque couple de sens. Le troisième terme identifié est désambiguïsé en se basant sur le sens détecté du deuxième terme qui a été désambiguïsé et ainsi de suite. En se basant sur l'hypothèse d'unicité du sens (Gale *et al.*, 1992), une fois que le terme est désambiguïsé, son sens sera propagé pour toutes ses occurrences dans le document.

Nous illustrons le principe de notre méthode de désambiguïsation par l'algorithme 1. Les étapes principales sont les suivantes :

- **Découper le texte en phrases :** cette étape permet d'identifier les phrases où les termes désignant les concepts peuvent être identifiés.
- **Identifier les termes désignant les concepts :** cette étape permet d'identifier un ensemble de termes ou chaînes de caractères les plus longues qui correspondent aux entrées des concepts dans la ressource termino-ontologique.
- **Désambiguïser de proche en proche avec propagation de sens :** en considérant la liste de n termes qui désignent les concepts dans le

1. TSD signifie *Term Sense Disambiguation*

Algorithme 1 – Algorithme de désambiguïsation de proche en proche

Entrées : Document D
Sorties : Étiquettes correspondant aux sens des termes désambiguïsés S

$L_n \leftarrow ()$ {Initialiser la liste de concepts extraits du document}
{Découper le texte en phrases}
$Phrases \leftarrow extrairePhrases(D)$
Pour $p \in Phrases$ **faire**
  {Extraire les concepts dans chaque phrase}
  $L_p \leftarrow extraireConcepts(p)$ {Liste de concepts dans p}
  $L_n \leftarrow L_n \bigcup L_p$
Fin Pour
Pour $p \in Phrases$ **faire**
  {Désambiguïser de proche en proche avec propagation de sens}
  Si $S[1,p]$ est ambigu **alors**
    $S[1,p] \leftarrow argmax \sum_{s_1 \in syn(L_p[1], s_2 \in syn(L_p[2])} sim(s_1, s_2)$
    Si $max(sim(s_1, s_2)) = 0$ **alors**
      { s_1 n'a pas de similarité avec s_2, alors on calcule sa similarité avec tous les autres concepts dans le document}
      $S[1,p] \leftarrow argmax \sum_{s_1 \in syn(L_p[1], s_i \in syn(L_p[i])} sim(s_1, s_i)$
    Fin Si
    $propager(S[1,p], L_n)$
  Fin Si
  Pour $k = 2; k \leq n; k++$ **faire**
    $S[k,p] \leftarrow argmax \sum_{s_k \in syn(L_p[k])} sim(s_{k-1}, s_k)$
    $propager(S[k,p], L_n)$
  Fin Pour
Fin Pour
Retourner S

document, $L_n = \{c_1, c_2, ..., c_n\}$, nous proposons la formule suivante pour identifier le sens le plus adéquat du terme qui désigne le concept c_k :

$$\begin{cases} (s_1, s_2) \propto \arg\max\left(\sum_{s_1 \in syn(c_1), s_2 \in syn(c_2)} sim(s_1, s_2)\right) & \text{si } k \leq 2 \\ s_k \propto \arg\max\left(\sum_{s \in syn(c_k)} sim(s_{k-1}, s)\right) & \text{si } k > 2 \end{cases} \quad \text{(IV.1)}$$

où

s_k : le sens du terme désignant le concept c_k,
$syn(c_k)$: l'ensemble de sens du terme désignant le concept c_k,
$sim(s_u, s_v)$: similarité basée sur les hiérarchies de s_u and s_v.

La similarité entre deux sens de deux termes est calculée en utilisant la similarité de graphes des hiérarchies de concepts associés selon la formule de (Leacock et Chodorow, 1998) :

$$sim(s_1, s_2) = \max \left\{ -\log \frac{length(s_1, s_2)}{2 * D} \right\} \quad \text{(IV.2)}$$

où $length(s_1, s_2)$ est le chemin le plus court entre s_1 and s_2, and D est le niveau le plus profond de la hiérarchie.

La figure IV.3 illustre les résultats de cet algorithme appliqué sur un fragment de texte :

- les rectangles représentent la liste des différents domaines de chaque concept. Chaque rectangle contient une liste de (sous-)domaines ou hiérarchies à laquelle un concept peut être associé. Par exemple, le concept "headaches" est associé à trois hiérarchies identifiées par les numéros d'arbres : "C23.888.592.612.441", "C23.888.646.487", et "C10.597.617.470".

- les flèches continues associées par des numérotations indiquent les étapes pour calculer le sens le plus adéquat pour chaque terme désignant un concept (mis en gras dans le texte d'entrée). Les flèches sont numérotées de gauche à droite dans l'ordre du discours. Dans cet exemple, le concept "Pain" a une influence sur le sens d'autres termes dans le texte. La décision du domaine associé à chaque terme est faite en fonction de la similarité entre les concepts voisins et le concept le plus dominant (le plus à gauche, c-à-d l'endroit où le discours commence).

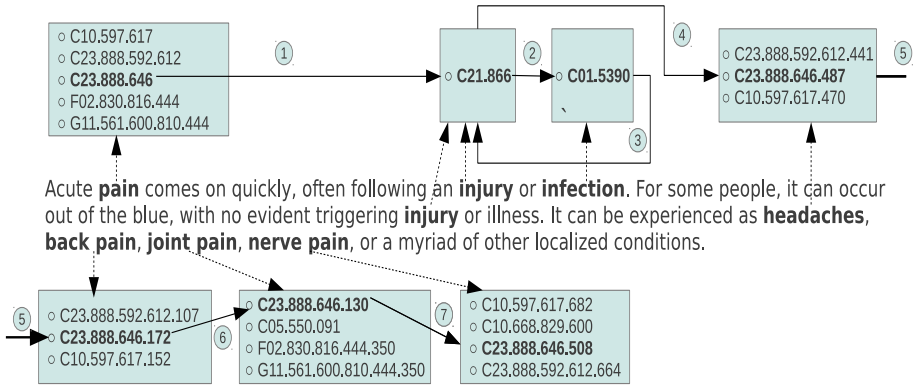


FIGURE IV.3 – Illustration de la première méthode de désambiguïsation

3.1.2 Algorithme de désambiguïsation 2 (Cluster-based TSD)

La deuxième méthode de désambiguïsation, appelée *Cluster-based TSD*, concerne les groupes de termes ou clusters de termes qui désignent les concepts dans le document. Chaque cluster correspond à un (sous-)domaine de concepts. Par exemple, le concept "**Acne Keloid**" appartient à deux sous-domaines **A** (*Anatomy*) et **C** (*Diseases*) dans le thésaurus MeSH car il est associé aux branches suivantes : *A10.165.450.300.425.125 ; C17.300.200.425.125 ; C17.800.030.030 ; C17.800.271.125.125 ; C17.800.329.500.261.*

Nous supposons que les concepts dans le document ont une similarité en termes de domaines entre eux. Cette similarité une fois déterminée permettrait de traduire de manière adéquate la spécificité des concepts dans le document. Par exemple, si le document contient les concepts "Acne Keloid" et " Facial Dermatoses", nous déterminons un domaine (une hiérarchie) pour chaque concept qui maximise leur similarité.

Nous considérons deux hypothèses suivantes :

- l'unicité du sens d'un terme qui désigne un concept spécifique dans le document (Gale *et al.*, 1992), et
- le sens d'un terme dépend du sens d'autres termes qui sont reliés hiérarchiquement quelle que soit leur position dans le document.

En se basant sur ces hypothèses, notre algorithme de désambiguïsation fonctionne comme suit : d'abord, les termes qui désignent les concepts dans le document sont identifiés. Ceux-ci sont regroupés dans les différents groupes dont chacun correspond à une hiérarchie ou un domaine. Du fait qu'un terme peut désigner un concept qui peut correspondre à plusieurs domaines, il est possible qu'un terme soit classé dans les différents groupes de concepts, d'où l'ambiguïté à résoudre. Les sens qui maximisent la similarité sémantique entre les termes dans chaque groupe sont retenus pour les termes identifiés dans le document. Une fois que le terme est désambiguïsé, son sens est propagé dans tout le document en adoptant l'hypothèse d'unicité de sens (Gale *et al.*, 1992).

Nous illustrons le principe de notre méthode de désambiguïsation par l'algorithme 2. Les étapes principales sont les suivantes :

- **Extraire les concepts dans le document :** cette étape permet d'extraire les termes désignant les concepts candidats dans le document.
- **Détecter les clusters de concepts :** les concepts candidats peuvent appartenir à plusieurs clusters.

Algorithme 2 – Algorithme de désambiguïsation par cluster de concepts

Entrées : Document D
Sorties : Étiquettes correspondant aux sens des termes désambiguïsés S

{Extraire les concepts dans le document}
$L_n \leftarrow extraireConcepts(D)$ {n : nombre de concepts extraits}
{Détecter les clusters de concepts}
$C \leftarrow extraireCluster(L_n)$
{Distribuer les concepts dans les clusters}
Pour $i = 1; i \leq |C|; i++$ **faire**
  {Calculer la similarité entre chaque concept et d'autres concepts dans le même cluster}
  $Calculer_Similarité(C[i])$
Fin Pour
Pour $i = 1; i \leq n; i++$ **faire**
  {Choisir le sens (domaine) le plus adéquat pour chaque concept}
  $maxScore \leftarrow -1.0$
  $sens \leftarrow NULL$
  Pour $j = 1; j \leq |C|; j++$ **faire**
    Pour $k = 1; k \leq |C[j]|; k++$ **faire**
      Si $L_n[i] == C[j,k]$ **alors**
        {il s'agit du concept k dans le cluster j}
        Si $maxScore < score(C[j,k])$ **alors**
          $maxScore \leftarrow score(C[j,k])$
          $sens \leftarrow sens(C[j,k])$
        Fin Si
      Fin Si
    Fin Pour
  Fin Pour
  $S[i] \leftarrow sens$
Fin Pour
Retourner S

- **Distribuer les concepts candidats dans les clusters :** les concepts candidats sont distribués dans chaque cluster correspondant pour créer un "pool" de concepts. Chaque cluster peut contenir les concepts issus de plusieurs branches que partagent le même domaine.

- **Choisir le sens (domaine) le plus adéquat pour chaque concept :** Soit $K = \{k_1, k_2, ..., k_{16}\}$ l'ensemble de 16 domaines du MeSH, où k_i est le nom d'un domaine particulier, $L_n = \{c_1, c_2, ..., c_n\}$ la liste de termes identifiés, le sens du terme qui désigne le concept c_i est calculé en se basant sur la similarité entre les couples de sens des termes dans chaque

groupe. Formellement :

$$s_i \propto \arg\max\Big(\sum_{c_i,c_j \in k_u, i \neq j} \; \sum_{s_a \in syn(c_i), s_b \in syn(c_j)} sim(s_a, s_b)\Big) \qquad \text{(IV.3)}$$

où

- k_u est un des 16 domaines dans le thésaurus,
- $sim(s_a, s_b)$ est la similarité sémantique entre les deux sens liés aux termes désignant les concepts c_a et c_b.

La différence entre la première et la deuxième méthode de désambiguïsation porte sur le fait que la méthode "Left-to-Right TSD" se focalise sur le terme désignant un concept biomédical qui se trouve le plus à gauche tandis que la méthode "Cluster-based TSD" se focalise sur la similarité en termes de sens entre les concepts dans le même domaine.

La figure IV.4 présente les résultats de la désambiguïsation où le texte d'entrée est le même que celui dans l'exemple illustratif dans la section 3.1.1 : nous avons trois clusters de concepts correspondant à trois sous-domaines C, F et G. Le groupe C contient 7 concepts identifiés, les groupes F et G contiennent 2 concepts. Les étiquettes mises en gras correspondent aux domaines de concepts qui maximisent la similarité entre les concepts dans chaque groupe.

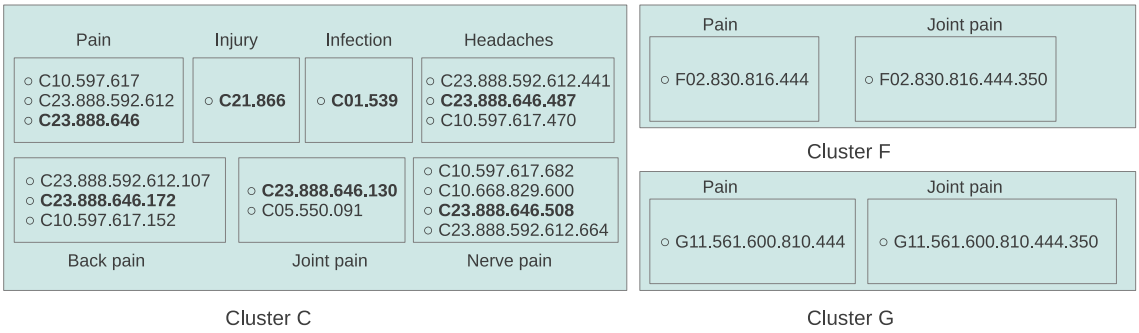


FIGURE IV.4 – Illustration de la deuxième méthode de désambiguïsation

3.2 Appariement sémantique document-requête

Notre objectif ici est de définir un schéma d'appariement sémantique qui intègre le sens le plus adéquat de chaque terme MeSH identifié à partir des documents et de la requête. Ce schéma sémantique exploite un index sémantique qui contient à la fois les termes désignant les concepts identifiés en utilisant une des deux approches de désambiguïsation précédentes et les mots simples qui ne correspondent à aucune entrée du MeSH [2].

2. Une entrée peut être un terme préféré ou non-préféré désignant un concept.

La similarité entre les documents et la requête est donc calculée selon un schéma sémantique basé sur le sens des termes identifiés et désambiguïsés dans la requête ainsi que dans les documents de la collection. Nous distinguons les étapes suivantes :

Étape 1 : Représentation des documents. Étant donné le document initial D_i qui contient à la fois des termes désignant les concepts du vocabulaire contrôlé et des mots simples du vocabulaire libre, D_i peut être représenté formellement comme suit :

$$\begin{aligned} D_i^s &= \{d_{1i}^s, d_{2i}^s, ..., d_{mi}^s\} \\ D_i^w &= \{d_{1i}^w, d_{2i}^w, ..., d_{ni}^w\} \end{aligned} \tag{IV.4}$$

où

- D_i^s est l'ensemble des termes MeSH identifiés et désambiguïsés,
- D_i^w est l'ensemble des mots simples,
- m et n sont respectivement le nombre de concepts identifiés et le nombre de mots simples du document D_i,
- d_{ji}^s est le *j-ième* concept identifié et d_{ji}^w est le *j-ième* mot du document.

Par exemple, le document correspondant au morceau de texte dans la figure IV.3 peut être représenté comme deux ensembles de mots simples et termes désignant les concepts dans la figure IV.5. Chaque terme MeSH identifié est donc annoté par le numéro d'arbre (étiquette) correspondant au sens identifié (*cf.* la figure IV.6). Cette annotation sémantique permet d'interpréter le domaine du concept ainsi que son niveau de la pronfondeur dans le thésaurus.

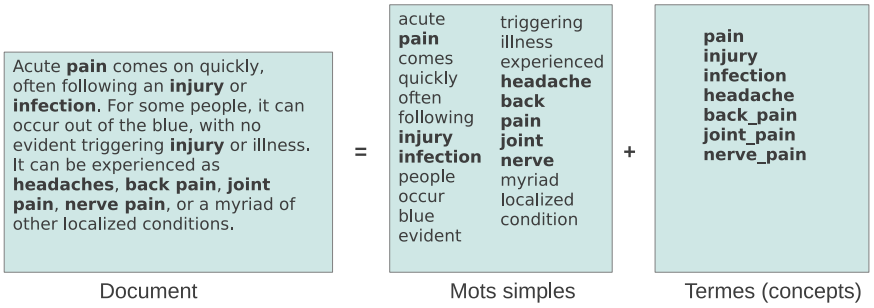


FIGURE IV.5 – Représentation du document par deux ensembles de mots simples et termes désignant les concepts biomédicaux

Acute <C23.888.646>**pain**</C23.888.646> comes on quickly, often following an <C21.866>**injury**</C21.866> or <C01.5390>**infection**</C01.5390>. For some people, it can occur out of the blue, with no evident triggering <C21.866>**injury**</C21.866> or illness. It can be experienced as <C23.888.646.487>**headaches**</C23.888.646.487>, <C23.888.646.172>**back pain**</C23.888.646.172>, <C23.888.646.130>**joint pain**</C23.888.646.130>, <C23.888.646.508>**nerve pain**</C23.888.646.508>, or a myriad of other localized conditions.

FIGURE IV.6 – Illustration d'un fragment de texte annoté avec les étiquettes correspondant aux sens des termes ambigus

Étape 2 : Représentation de la requête. Les requêtes sont traitées de la même manière que les documents. Par conséquent, la requête Q peut être représentée formellement comme suit :

$$\begin{array}{rcl} Q^s & = & \{q_1^s, q_2^s, ..., q_u^s\} \\ Q^w & = & \{q_1^w, q_2^w, ..., q_v^w\} \end{array} \quad \text{(IV.5)}$$

où Q^s, Q^w sont respectivement l'ensemble de concepts et de mots simples, u et v sont respectivement le nombre de concepts et mots de Q, q_k^s est le *k-ième* concept et q_k^w est le *k-ième* mot de Q.

Étape 3 : Calcul de la pertinence. La mesure de pertinence du document D_i vis-à-vis de la requête Q considère dans notre cas deux principaux facteurs : (1) l'adéquation du sens des concepts dans le document et de la requête, (2) la spécificité des concepts dans le document. Formellement, la similarité sémantique entre la requête Q et le document D_i, dénotée $RSV(Q, D_i)$, est donnée par :

$$RSV(Q, D_i) = RSV(Q^w, D_i^w) + RSV(Q^s, D_i^s) \quad \text{(IV.6)}$$

où

- $RSV(Q^w, D_i^w)$ est la mesure *TF-IDF* classique basée sur les mots simples
- $RSV(Q^s, D_i^s)$ est la mesure de pertinence basée sur le sens du terme désignant un concept du document vis-à-vis de la requête, calculée comme suit :

$$\begin{array}{rcl} RSV(Q^w, D_i^w) & = & \sum_{q_k^w \in Q^w} TF_i(q_k^w) * IDF(q_k^w) \\ RSV(Q^s, D_i^s) & = & \sum_{q_k^s \in Q^s} (1 + h(q_k^s)) * TF_i(q_k^s) * IDF(q_k^s) \end{array} \quad \text{(IV.7)}$$

où

- TF_i est la fréquence normalisée du mot q_k^w ou du concept q_k^s dans le document D_i,
- IDF est la fréquence inverse de documents de q_k^w ou q_k^s dans la collection,
- $h(q_k^s)$ est la spécificité du terme q_k^s associée à son propre sens dans la requête, calculée comme suit :

$$h(q_k^s) = \frac{niveau(q_k^s)}{MaxDepth} \quad \text{(IV.8)}$$

où $niveau(q_k^s)$ correspond au niveau de la profondeur de q_k^s et $MaxDepth$ est la profondeur maximale des concepts dans le thésaurus.

4 Évaluation expérimentale

Notre évaluation expérimentale consiste à mesurer l'impact de l'utilisation des sens liés aux domaines de concepts dans la ressource terminologique MeSH sur l'efficacité d'un processus de RI sémantique. Nous décrivons dans ce qui suit le cadre d'évaluation et présentons, puis discutons les résultats obtenus.

4.1 Cadre d'évaluation

– ***Collection test*** *:* Nous utilisons la collection OHSUMED, proposée dans le cadre de la tâche TREC9-Filtering en 2000, qui est constituée des titres et/ou des résumés de 270 journaux médicaux publiés entre 1987-1991 (Hersh *et al.*, 1994). Un document contient six champs : *titre (.T), résumé (.W), concepts indexés de MeSH (.M), auteur (.A), source (.S), and publication (.P)*. Le tableau IV.3 illustre un document dans la collection OHSUMED en utilisant le format de TREC (similaire au format XML). Selon ce format, chaque document est représenté par la balise <DOC>. Celui-ci est identifié par l'élément DOCNO. Les autres éléments correspondent au contenu du document (par exemple TITLE, ABSTRACT, etc.). Quelques caractéristiques statistiques de la collection sont données dans le tableau IV.4.
Nous avons testé notre approche de RI basée sur les méthodes de désambiguïsation sur un ensemble de 48 requêtes de la collection OHSUMED. Chacune est fournie avec un ensemble de documents jugés pertinents par un groupe de médecins. Le champ *titre* indique la *description du patient* (patient description) et le champ *description* annonce *le besoin en information* (information request).

  Le tableau IV.5 présente quelques exemples de requêtes issues de la collection OHSUMED. Chaque requête est identifiée par un numéro unique.

– ***Mesures d'évaluation*** *:* Nous utilisons les mesures P@5, P@10 qui sont respectivement la précision moyenne aux 5, 10 premiers documents retournés et MAP (*Mean Average Precision*) sur l'ensemble de 48 requêtes. Pour chaque requête, les 1000 premiers documents sont renvoyés par le système et les précisions moyennes (P@5, P@10, MAP) sont calculées pour mesurer la performance de la RI.

TABLEAU IV.3 – Exemples de documents de la collection OHSUMED

<DOC>
<DOCNO>88000001</DOCNO>
<TITLE>The binding of acetaldehyde to the active site of ribonuclease : alterations in catalytic activity and effects of phosphate.</TITLE>
<ABSTRACT>Ribonuclease A was reacted with [1-13C,1,2-14C]acetaldehyde and sodium cyanoborohydride in the presence or absence of 0.2 M phosphate. After several hours of incubation at 4 degrees C (pH 7.4) stable acetaldehyde-RNase adducts were formed, and the extent of their formation was similar regardless of the presence of phosphat e. Although the total amount of covalent binding was comparable in the absence or presence of phosphate, this active site ligand prevented the inhibition of enzymatic activity seen in its absence. This protective action of phosphate diminished with progressive ethylation of RNase, indicating that the reversible association of phosph ate with the active site lysyl residue was overcome by the irreversible process of reductive ethylation. Modified RNase was analysed using 13C proton decoupled NMR spe ctroscopy. Peaks arising from the covalent binding of enriched acetaldehyde to free amino groups in the absence of phosphate were as follows : NH2-terminal alpha amino group, 47.3 ppm ; bulk ethylation at epsilon amino groups of nonessential lysyl residues, 43.0 ppm ; and the epsilon amino group of lysine-41 at the active site, 47.4 pp m. In the spectrum of RNase ethylated in the presence of phosphate, the peak at 47.4 ppm was absent. When RNase was selectively premethylated in the presence of phosph ate, to block all but the active site lysyl residues and then ethylated in its absence, the signal at 43.0 ppm was greatly diminished, and that arising from the active site lysyl residue at 47.4 ppm was enhanced. These results indicate that phosphate specifically protected the active site lysine from reaction with acetaldehyde, and that modification of this lysine by acetaldehyde adduct formation resulted in inhibition of catalytic activity.</ABSTRACT>
</DOC>

TABLEAU IV.4 – Description statistique de la collection test

Nombre de documents	293.856
Longueur moyenne du document	100
Longueur moyenne de la requête	6 (TITRE) 12 (TITRE+DESC)
Nombre moyen de concepts/requête	1,50 (TITRE) 3,33 (TITRE+DESC)
Nombre de documents jugés pertinents/requête	50

TABLEAU IV.5 – Exemples de requêtes de la collection OHSUMED

<top>
<num> Number : OHSU1
<title> 60 year old menopausal woman without hormone replacement therapy
<desc> Description :
Are there adverse effects on lipids when progesterone is given with estrogen replacement therapy
</top>

<top>
<num> Number : OHSU2
<title> 60 yo male with disseminated intravascular coagulation
<desc> Description :
pathophysiology and treatment of disseminated intravascular coagulation
</top>

4.2 Résultats expérimentaux

Pour évaluer l'utilité de nos méthodes de désambiguïsation ainsi que leur impact sur les performances de notre approche d'indexation et de recherche d'information sémantique, nous avons réalisé deux séries d'expérimentations :

- la première est basée sur l'indexation classique du texte contenu dans les balises *titre* et *résumé* d'articles de MEDLINE en utilisant la configuration standard sous la plateforme Terrier [3] (version 2.2.1) avec le schéma de pondération de référence OKAPI BM25 (Robertson *et al.*, 1998). Cette configuration est utilisée comme la base d'évaluation comparative (baseline), dénotée *BM25*.

- la seconde série d'expérimentations concerne notre méthode d'indexation sémantique qui se décline en trois scénarios :
  1. le premier est basé sur la sélection naïve du premier sens du terme désignant un concept spécifique trouvé dans le thésaurus, dénotée *TSD-0*,
  2. le second est basé sur notre première méthode de désambiguïsation décrite dans la section 3.1.1, dénotée *TSD-1*,
  3. le troisième est basé sur notre seconde méthode de désambiguïsation décrite dans la section 3.1.2, dénotée *TSD-2*.

3. `http://ir.dcs.gla.ac.uk/terrier/`

Nous utilisons à la fois les termes qui représentent les entrées MeSH (*concepts* ou *main headings*), et les mots simples du vocabulaire libre qui ne font pas partie des entrées de ce thésaurus. Dans l'approche d'indexation classique, les documents et les requêtes sont indexés en utilisant la plateforme Terrier. L'indexation comprend une séquence d'étapes consécutives de traitements : suppression de mots vides, identification de termes désignant les concepts dans le thésaurus MeSH, racinisation des termes (mots simples ou termes complexes).

Dans notre approche basée sur le sens identifié pour chaque terme désignant un concept spécifique, les documents et les requêtes sont d'abord désambiguïsés et indexés avec les sens appropriés des termes du MeSH. Puis, le schéma de pondération est appliqué à chaque terme (mot simple ou entrée du MeSH) dans la requête en utilisant la formule IV.7.

Le tableau IV.6 présente les performances de la recherche basée sur la base d'évaluation comparative et sur nos méthodes de désambiguïsation pour les requêtes courtes et pour les requêtes longues. Nous avons obtenu les résultats suivants : les deux méthodes de TSD permettent d'améliorer les performances de la RI par rapport au schéma de pondération classique BM25, quelle que soit la longueur de la requête. Concernant les performances des requêtes courtes, les taux d'amélioration en terme de la précision moyenne MAP sont de +5.16% et +4.77% pour la méthode TSD-1 et TSD-2 respectivement. Concernant les performances des requêtes longues, les taux d'amélioration de la MAP sont de +17.35% et +17.06% pour la méthode TSD-1 et TSD-2 respectivement.

Les t-tests ($p \leq 0.05$, dénoté *) montrent que notre approche de RI sémantique basée sur nos méthodes de désambiguïsation est statistiquement significative par rapport à la baseline. Cela montre l'intérêt de la prise en compte de la sémantique dans le processus de la RI. Ce processus intègre donc l'identification et la désamnbiguïsation des termes ambigus permettant de détecter les meilleurs domaines MeSH ainsi que la spécificité des concepts dans le document et dans la requête. En plus, les résultats montrent que la sélection naïve du sens des termes (*TSD-0*) n'améliore pas les performances de la recherche (e.g., la précision P@10 de la méthode *TSD-0* se détériore). En revanche, une affectation correcte de sens à chaque terme désignant un concept spécifique dans le document permet d'améliorer les performances de la RI.

Les figures IV.7 et IV.8 présentent les courbes de *précision* − *rappel* à 11 points de rappel de 0.0 à 1.0 avec un pas de 0.1. Nous observons que les méthodes de désambiguïsation TSD-1 et TSD-2 donnent toujours une meilleure précision que la baseline ainsi que la méthode de désambiguïsation naïve *TSD-0*. En effet, les courbes correspondant aux performances des méthodes TSD-1 et TSD-2 sont plus élevées que les performances obtenues par la baseline ainsi que la méthode TSD-0 lorsque les requêtes contiennent plus de concepts, notamment pour les requêtes longues.

TABLEAU IV.6 – Les performances de RI (P@5, P@10 et MAP) obtenues sur la collection OHSUMED

(a) Performances des requêtes basées sur le champ *titre*

Mesure	BM25	TSD-0	TSD-1	TSD-2
P@5 (%)	0.1750	0.1792 (+2.40)	0.1917 (+9.54)	0.1792 (+2.40)
P@10 (%)	0.1854	0.1771 (-4.48)	0.1875 (+1.13)	0.1875 (+1.13)
MAP (%)	0.1027	0.1034 (+0.68)	0.1080 (+5.16)	0.1076 (+4.77)

(b) Performances des requêtes basées sur les champs *titre* et *description*

Mesure	BM25	TSD-0	TSD-1	TSD-2
P@5 (%)	0.5042	0.5083 (+0.81)	0.5083 (+0.81)	0.5125 (+1.65)
P@10 (%)	0.4563	0.4606 (+0.94)	0.4833 (+5.92)	0.4812 (+5.46)
MAP (%)	0.2421	0.2545 (+5.12)	0.2841* (+17.35)	0.2834* (+17.06)

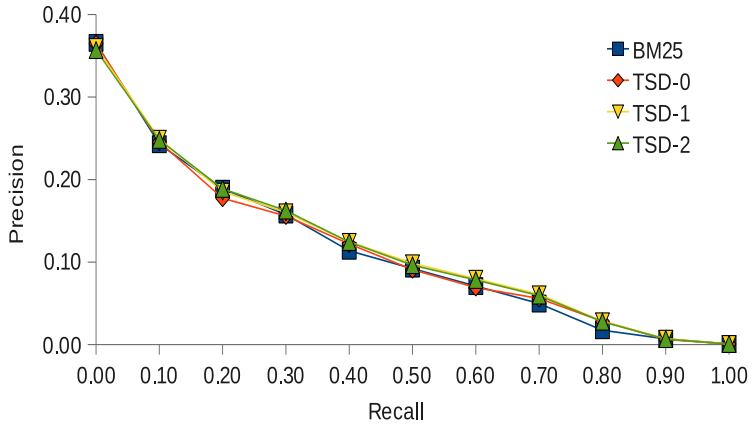


FIGURE IV.7 – Courbe rappel-précision des requêtes basées sur *titre*

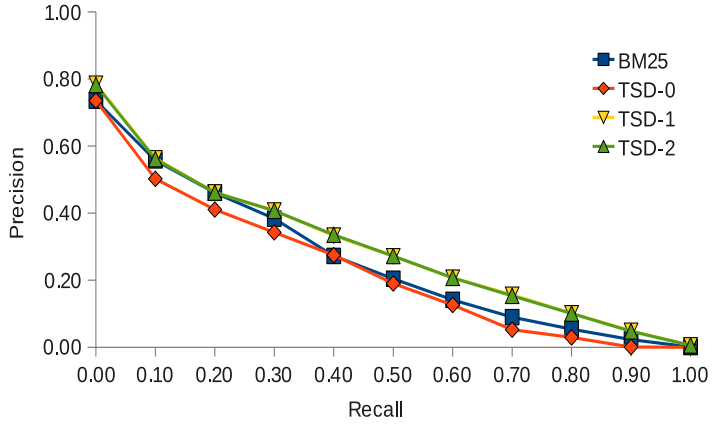


FIGURE IV.8 – Rappel-précision des requêtes basées sur *titre* et *description*

Nous avons également mené une analyse plus fine au niveau de la requête pour évaluer l'impact de la spécificité en fonction de la longueur de la requête et le nombre de concepts utilisés dans la requête. Comme présenté dans la figure IV.9, pour chaque requête, nous calculons la spécificité moyenne qui est essentiellement la moyenne des niveaux de la pronfondeur des concepts dans la requête. Nous obtenons ainsi des valeurs réelles entre 2 et 6, c-à-d que les concepts ont un niveau de spécificité entre 2 et 6. Les requêtes sont ensuite regroupées en fonction de la spécificité approximative dans l'ensemble {2, 3, 4, 5, 6}.

Pour chaque groupe de requêtes, nous calculons la longueur moyenne (en termes de mots-clés ou tokens), le nombre moyen de termes désignant les concepts identifiés, et le taux d'accroissement en moyenne correspondant à chaque groupe de requêtes. Nous nous apercevons que plus la requête est spécifique, c'est-à-dire que si les concepts observés dans la requête sont plus spécifiques, le nombre de concepts dans la requête est moindre. Cela pourrait s'expliquer par le fait que quelques-uns des concepts les plus spécifiques couvrent suffisamment le besoin en information de l'utilisateur alors qu'il nécessite un nombre plus élevé de concepts qui sont moins spécifiques afin de mieux exprimer son besoin en information.

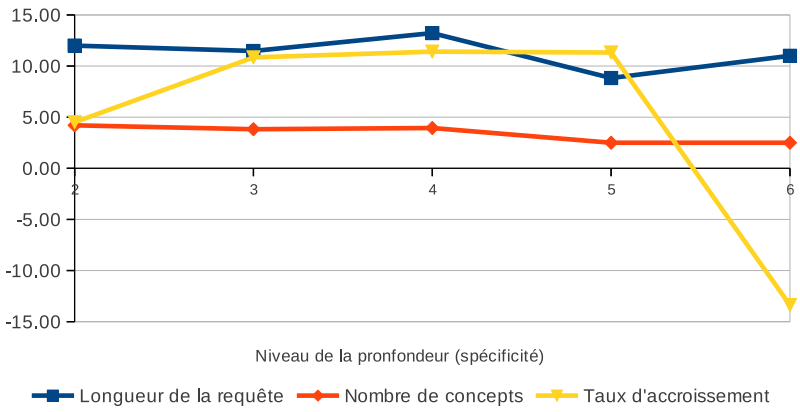


FIGURE IV.9 – Analyse de résultats en fonction de la spécificité de la requête

Dans la plupart des cas, notre approche privilégie les documents contenant les concepts à un niveau de spécificité élevé et montre une amélioration par rapport à la baseline. Toutefois, si la requête est longue mais le nombre de concepts ayant un degré de spécificité élevé est moindre, notre approche tend à retourner les premiers documents contenant ces concepts. Cela peut être la cause de la dégradation des performances lorsque quelques-uns des concepts les plus spécifiques ont un impact important sur d'autres termes dans la requête.

4.3 Discussion

L'objectif principal de nos méthodes de désambiguïsation est d'identifier le sens le plus adéquat de chaque terme issus du thésaurus MeSH qui est donc lié à des domaines ou sous-domaines dans cette architecture poly-hiérarchique. De plus, nous avons exploité les informations liées aux sens des termes identifiés dans un processus de RI comme la spécificité du concept dans un contexte du document et de la requête. L'évaluation des résultats obtenus a montré l'utilité de tenir compte de la sémantique du document et de la requête ainsi que la spécificité des concepts identifiés dans le document et dans la requête. La sélection naïve du sens d'un terme (la méthode TSD-0) n'est pas utile pour la RI. Par contre, l'identification du domaine le plus adéquat du terme (les méthodes TSD-1 et TSD-2) permet d'améliorer les performances de la RI. Nous observons que nos méthodes TSD-1 et TSD-2 donnent de meilleurs résultats par rapport à la méthode TSD-0 ainsi qu'à la base d'évaluation BM25. Les performances de la RI obtenues par nos méthodes TSD-1 et TSD-2 sont légèrement différentes pour les requêtes longues. Cela peut s'expliquer par le fait que pour les requêtes longues, le nombre de concepts identifiés est souvent plus grand que celui dans les requêtes courtes et donc il est probable que les méthodes de désambiguïsation ont tendance à donner les mêmes résultats.

5 Conclusion

Dans ce chapitre, nous avons proposé et évalué une approche d'indexation conceptuelle/sémantique basée sur le domaine des termes désignant les concepts biomédicaux. Le sens du terme est reflété par son propre domaine dans l'architecture poly-hiérarchique du thésaurus MeSH. Notre approche de RI sémantique s'appuie essentiellement sur nos deux méthodes de désambiguïsation de termes ambigus en termes de domaines biomédicaux définis dans le MeSH. Nos méthodes de désambiguïsation sont principalement basées sur le contexte local du document et de la requête où apparaissent les concepts. Notre modèle d'indexation et d'appariement conceptuel/sémantique proposé prend en compte l'adéquation du domaine des termes ambigus ainsi que leur spécificité dans le document et dans la requête.

L'évaluation de la méthode d'indexation et d'appariement conceptuelle sur le corpus standard OHSUMED montre une amélioration par rapport à une base d'évaluation de référence basée sur un schéma de pondération probabiliste. Cependant, notre approche de RI conceptuelle est basée sur une combinaison simple des représentations textuelle et conceptuelle, ce qui peut causer des problèmes lorsque la requête est longue et que le nombre de concepts ayant un degré de spécificité élevée dans la requête est moindre, notre approche tend à re-

tourner les premiers documents contenant ces concepts. Cela peut être la cause de la dégradation de la performance lorsque quelques-uns des concepts les plus spécifiques ont un impact important sur d'autres termes dans la requête. Il faudrait peut être explorer une meilleure combinaison ou une autre meilleure façon pour améliorer la sémantique du document et/ou de la requête. Dans nos futurs travaux, nous envisageons d'investir la précision ainsi que les performances de la méthode d'extraction de concepts. Cela nous permet de mieux reconnaître les concepts pertinents dont nous exploitons les informations conceptuelles dans un processus d'indexation et de recherche d'information conceptuelle/sémantique.

Chapitre V

Extraction de concepts biomédicaux : approche basée sur la pertinence et la corrélation des contextes documentaires et terminologiques

Sommaire

1 Introduction **149**

2 Problématiques et motivations **150**

3 Extraction de concepts biomédicaux basée sur la combinaison du score thématique et du score de corrélation d'ordre de mots . . . **153**
- 3.1 Représentation des concepts de la terminologie 153
- 3.2 Calcul du score de pertinence des concepts candidats 155
  - 3.2.1 Calcul du score thématique des concepts candidats . . 155
  - 3.2.2 Calcul du score de la corrélation d'ordre de mots . . . 156
- 3.3 Illustration de l'extraction des concepts par des exemples concrets 161

4 Évaluation expérimentale **163**
- 4.1 Cadre d'évaluation de TREC Genomics 163
  - 4.1.1 Description de la collection de documents 163
  - 4.1.2 Description de l'ensemble de requêtes 164
  - 4.1.3 Ressources termino-ontologiques des concepts biomédicaux 166
  - 4.1.4 Acronymes et variants des noms de gènes 167
- 4.2 Scénarios d'évaluation 167
  - 4.2.1 Évaluation de l'efficacité des méthodes d'extraction de concepts sur les documents 167
  - 4.2.2 Évaluation de l'efficacité des méthodes d'extraction de concepts sur les requêtes 169
- 4.3 Mesures d'évaluation des performances de la RI 171
- 4.4 Résultats expérimentaux 172
  - 4.4.1 Évaluation de notre méthode d'extraction de concepts sur les documents 172

4.4.1.1 Impact du nombre de termes préférés utilisés pour l'expansion de documents 172
4.4.1.2 Paramètres du modèle de reformulation de requêtes 173
4.4.1.3 Évaluation de l'efficacité de notre méthode d'extraction de concepts via l'expansion de documents et de requêtes 174
4.4.2 Résultats expérimentaux des méthodes d'extraction de concepts appliquées sur les requêtes 175
4.4.2.1 Évaluation de l'efficacité de la RI basée sur des méthodes d'extraction de concepts via l'expansion conceptuelle de requêtes 176
4.4.2.2 Évaluation de l'efficacité de la RI basée sur la combinaison de l'expansion conceptuelle et la reformulation de requêtes basée sur la technique PRF 178
4.5 Évaluation comparative 180
5 Conclusion **184**

"It's hardware that makes a machine fast. It's software that makes a fast machine slow." –*Craig Bruce*

Publications liées à ce travail

- **ECIR 2011 :** Duy Dinh, Lynda Tamine. *Combining global and local semantic contexts for improving biomedical information retrieval.* Dans : the 33rd European Conference on Information Retrieval, ECIR 2011, Avril 18-21, 2011, Dublin, Ireland, p. 375–386 ;
- **SAC 2011 :** Duy Dinh, Lynda Tamine. *Biomedical concept extraction based on combining the content-based and word order similarities.* Dans : the 26th ACM Symposium on Applied Computing, SAC 2011, Mars 21-25, 2011, Taichung, Taiwan, p. 1159–1163 ;

1 Introduction

Dans le chapitre IV, nous avons identifié les termes désignant les concepts candidats en utilisant une méthode d'extraction simple qui est basée sur la recherche exacte des termes dans le thésaurus MeSH. Cependant, la recherche exacte peut causer la dégradation des performances des résultats d'extraction (Krauthammer et Nenadic, 2004; Zhou *et al.*, 2006b). Dans ce chapitre, nous introduisons une nouvelle méthode d'extraction approximative des concepts biomédicaux pour la RI biomédicale. Notre méthode d'extraction des concepts biomédicaux peut être intégrée dans n'importe quelle approche de RI biomédicale.

Les techniques de recherche d'information actuelles qui sont basées sur des mots-clés fournissent des capacités limitées pour capturer la sémantique associée à des besoins d'information des utilisateurs. Cette limitation peut être due à la représentation du contenu textuel par des sacs de mots isolés sans prendre en compte leur contexte d'utilisation. Pour résoudre les problèmes liés à la limitation des modèles de RI basés sur des sacs de mots, les modèles de RI conceptuels ou sémantiques ont été proposés dans les années quatre-vingts du dernier siècle (Croft, 1986). Ces modèles utilisent les informations sur les concepts (e.g., synonymes, acronymes, abréviations, etc.) dans les ressources termino-ontologiques comme les ontologies, les thésaurus ou les dictionnaires pour améliorer la représentation textuelle des documents/requêtes. L'hypothèse de base des approches de RI conceptuelle est que les mots-clés indépendants ne sont pas capables de capturer le contenu sémantique des documents ainsi

que des besoins d'information. Par conséquent, une des solutions qui semblent appropriées est d'atteindre le niveau conceptuel du contenu textuel via une représentation conceptuelle par des sources de connaissances.

Plusieurs sources d'évidence peuvent être exploitées pour améliorer la représentation du document et de la requête telles que les connaissances sur la tâche, le problème à résoudre, l'intention de l'utilisateur, ou le domaine, etc. Nous nous intéressons en particulier à l'utilisation de ce dernier pour extraire les concepts à partir du texte. Plus précisément, nous nous intéressons à la représentation conceptuelle du contenu textuel pour la recherche d'information biomédicale. Dans le domaine biomédical, il existe plusieurs ressources comme MeSH [1], UMLS [2], CIM [3], ... Ces ressources termino-ontologiques contiennent en général des concepts qui sont désignés par des termes médicaux ainsi que les relations sémantiques entre eux.

Pour atteindre la représentation conceptuelle du contenu textuel et exploiter les informations sémantiques, il faut d'abord identifier les concepts à partir du texte. Il existe de nombreux travaux en traitement du langage naturel traitant de l'identification ou l'extraction automatique des concepts à partir du contenu textuel. Les travaux les plus récents dans le domaine de la biomédecine sont (Gaizauskas *et al.*, 2000; Humphreys *et al.*, 2000; Aronson *et al.*, 2004b; Hliaoutakis *et al.*, 2009; Ruch, 2006; Zhou *et al.*, 2006b; Sohn *et al.*, 2008). Ces travaux ont été menés dans le cadre de la tâche de catégorisation multi-étiquettes, où le classificateur décide si un concept donné est significatif pour un texte spécifique ou non.

Ce chapitre est organisé comme suit : nous présentons d'abord les problématiques de recherche et nos motivations dans la section 2. Ensuite, nous détaillons notre méthode d'extraction de concepts biomédicaux pour la RI conceptuelle/sémantique dans la section 3. Les résultats expérimentaux obtenus sont présentés et discutés dans la section 4. Enfin, la section 5 conclut le chapitre.

2 Problématiques et motivations

Notre objectif principal ici est donc d'exploiter les ressources termino-ontologiques pour identifier les meilleurs concepts dans les textes biomédicaux. Notre approche d'extraction de concepts est inspirée de la RI et des caractéristiques du langage naturel pour estimer la pertinence des concepts candidats vis-à-vis d'un texte donné. Notre hypothèse de base derrière la pertinence des

1. Medical Subject Headings
2. Unified Medical Language System
3. Classification Internationale des Maladies

concepts porte sur le fait qu'un concept, qui est désigné par un ou plusieurs termes synonymes, est un "bon candidat" pour représenter la sémantique du texte si les deux conditions suivantes sont satisfaites :

– Plus le concept, vu comme un ensemble de termes comprenant son terme d'entrée préféré et ses termes d'entrée non préférés, partage des mots ou termes avec le texte considéré, plus il est représentatif de ce contenu. Cette condition permet le "rapprochement approximatif" document-concept et répond de fait à la question de la variation lexicale posée dans le domaine. Nous illustrons le contenu d'un concept par l'exemple dans le tableau V.1 : le concept "**Minisatellite Repeat**" dont l'identifiant unique est "C0242827" est désigné par le terme préféré "Minisatellite Repeat" et plusieurs termes non-préférés (e.g., "Variable Tandem Repeats", "Simple Repetitive Sequence', etc.). Nous supposons que tout document qui contient une de ces différentes variantes lexicales est candidat à être pertinent vis-à-vis de la requête qui contient un de ces termes d'entrée.

CUI : **C0242827**	
MH : **Minisatellite Repeats**	
ENTRY : Variable Tandem Repeats	ENTRY : Locus VNTR
ENTRY : Minisatellites	ENTRY : VNTR Locus
ENTRY : Simple Repetitive Sequence	ENTRY : Loci VNTR
ENTRY : VNTR Loci	ENTRY : Regions VNTR
ENTRY : Variable Number of Tandem Repeats	ENTRY : VNTR Regions
ENTRY : VNTR Region	ENTRY : Region VNTR
ENTRY : VNTR Sequences	ENTRY : Repeat Variable Tandem
ENTRY : VNTR	ENTRY : Tandem Repeat Variable
ENTRY : Minisatellite Repeat	ENTRY : Variable Tandem Repeat
ENTRY : Repeat Minisatellite	ENTRY : Repeats Variable Tandem
ENTRY : Repeats Minisatellite	ENTRY : Tandem Repeats Variable
ENTRY : Repetitive Sequences Simple	ENTRY : Sequence VNTR
ENTRY : Sequences Simple Repetitive	ENTRY : VNTR Sequence
ENTRY : Simple Repetitive Sequences	ENTRY : Sequences VNTR
ENTRY : Repetitive Sequence Simple	ENTRY : Minisatellite
ENTRY : Sequence Simple Repetitive	

TABLEAU V.1 – Le concept "Minisatellite repeats" (C0242827), issu du thésaurus MeSH, vu comme un document qui est constitué par ses termes d'entrée

– Plus la structure ou la formation des termes terminologiques repérés dans les textes est corrélée (en termes d'ordre d'apparition des mots) à celle des entrées terminologiques, plus le concept s'apparie avec le document associé. Cette condition est liée en effet, à la granularité du langage médical lié d'une part au sens précis de chaque terme, dans chaque contexte lié aux termes voisins, à la branche terminologique, à la terminologie utilisée etc. Nous illustrons ce principe d'extraction par l'exemple présenté dans

le tableau V.2 : dans cet exemple, les concepts comme "**variable number of tandem repeat**", "**dopamine receptor D4**" sont constitués de plusieurs mots dans un ordre défini. De plus, ils sont accompagnés par des acronymes qui sont souvent utilisés dans les documents biomédicaux. Ces acronymes peuvent être utiles pour formuler la requête lors la recherche d'information. Les mots simples ou termes composés qui sont soulignés correspondent à un terme d'entrée défini dans la terminologie. En tenant compte de la corrélation en termes d'ordre des mots, les concepts candidats peuvent être identifiés de manière précise ou approximative.

"*The* ***polymorphism*** *of* ***variable number of tandem repeat*** *(VNTR) in* ***dopamine receptor D4*** *(DRD4)* ***gene*** *exon III has been linked to various neuro-psychiatric* ***conditions*** *with disinhibition/impulsivity as one of the core features. This study examined the modulatory* ***effects*** *of long-allele* ***variant*** *of DRD4 VNTR on the regional neural activity as well as inter-regional neural interactions in a young female* ***population***."

TABLEAU V.2 – Illustration de la corrélation en termes d'ordre de mots entre le texte et les entrées des concepts.

Sur la base de ces motivations, nous proposons une approche approximative d'extraction de concepts à partir du texte libre basée sur l'appariement thématique et la corrélation des contextes textuels et terminologiques. D'un point de vue pratique, la terminologie est représentée comme une collection de concepts ; de plus l'extraction de concepts terminologiques à partir d'un texte revient à sélectionner, parmi les concepts de la terminologie, ceux qui sont pertinents au contenu textuel. La pertinence est calculée selon une combinaison de scores thématiques et score de corrélation d'ordre de mot.

Notre contribution présentée ici porte sur les principaux points suivants :

- la proposition d'une nouvelle méthode d'extraction de concepts terminoontologiques à partir du texte libre. L'originalité de notre méthode comparativement aux travaux liés de l'état de l'art réside dans les éléments suivants :
  - aucune analyse morphologique des contenus textuels n'est reprise,
  - l'exploitation de la similarité thématique (liée aux thèmes) et structurelle (liée à la structure ou la formation des termes, notamment l'orde de mots constituant des termes) entre le texte et les concepts dans la terminologie pour couvrir au mieux les granules d'information associés à un terme d'entrée d'un concept particulier dans la terminologie.

- l'évaluation expérimentale et comparative des méthodes d'extraction de concepts qui sont largement adoptées en RI.

3 Extraction de concepts biomédicaux basée sur la combinaison du score thématique et du score de corrélation d'ordre de mots

Notre méthode d'extraction de concepts termino-ontologiques est essentiellement basée sur une approche de recherche des concepts pertinents qui requiert :

- la représentation des concepts définis dans la ressource termino-ontologique qui peut être considérée comme un corpus de concepts. En effet, comme le travail de (Ruch, 2006), chaque concept peut être vu comme un document constitué des mots issus des termes d'entrée du concept.

- le calcul du score de pertinence des concepts qui est donc basée sur la similarité thématique et structurelle entre le texte vu comme une requête à laquelle les concepts candidats, vus comme documents, peuvent être associés.

3.1 Représentation des concepts de la terminologie

Chaque concept dans la ressource termino-ontologique est désigné par un ensemble de termes d'entrée préférés ou non-préférés. Le terme préféré est utilisé comme la forme standard pour désigner le concept tandis que le terme non-préféré est considéré comme une variante du terme préféré. Les termes préférés sont souvent utilisés pour indexer les textes biomédicaux et vus comme les sujets du texte (Névéol *et al.*, 2009). Le concept C est donc représenté par un ensemble de termes d'entrée :

$$C = \{E_1, E_2, ..., E_m\} \tag{V.1}$$

où

- E_i est un terme d'entrée (préféré ou non-préféré du concept),
- m est le nombre de termes d'entrée du concept.

Le concept C est considéré comme un document constitué des mots simples issus des termes d'entrée du concept. Formellement, le "document" représentant le concept C, dénoté C^D, peut être considéré comme le vecteur des mots simples :

$$C^D =< w_1, w_2, ..., w_N > \tag{V.2}$$

où w_j est le poids du j-ième mot observé dans le concept et N est le nombre de mots uniques dans la terminologie.

Le tableau V.3 illustre deux concepts MeSH représentés comme deux documents différents qui sont identifiés grâce à la balise “DOCNO”. Ce dernier correspond en effet à l’identifiant du concept dans le thésaurus MeSH. Chaque document est constitué des mots simples issus des termes d’entrée préférés ou non-préférés. Les balises “MH” et “ENTRY” correspondent aux termes préférés, appelé également *termes vedettes* ou *Main Headings* en anglais, et aux termes non-préférés (e.g., synonymes, abréviations, acronymes, ou variantes lexicales, etc.). Les documents représentant les concepts MeSH sont indexés par les mots-clés les plus significatifs.

<DOC>
<DOCNO>**C0039986**</DOCNO>
<MH>Thoracic Surgery</MH>
<ENTRY>Surgery Thoracic</ENTRY>
<ENTRY>Surgery Heart</ENTRY>
<ENTRY>Cardiac Surgery</ENTRY>
<ENTRY>Heart Surgery</ENTRY>
<ENTRY>Surgery Cardiac</ENTRY>
</DOC>

<DOC>
<DOCNO>**C0273115**</DOCNO>
<MH>Lung Injury</MH>
<ENTRY>Lung Injuries</ENTRY>
<ENTRY>Pulmonary Injury</ENTRY>
<ENTRY>Injuries Lung</ENTRY>
<ENTRY>Injury Lung</ENTRY>
<ENTRY>Injuries Pulmonary</ENTRY>
<ENTRY>Pulmonary Injuries</ENTRY>
<ENTRY>Injury Pulmonary</ENTRY>
<ENTRY>Chronic Lung Injury</ENTRY>
<ENTRY>Chronic Lung Injuries</ENTRY>
<ENTRY>Lung Injuries Chronic</ENTRY>
<ENTRY>Lung Injury Chronic</ENTRY>
</DOC>

TABLEAU V.3 – Exemples des concepts du thésaurus MeSH vus comme les documents d’une collection de concepts

3.2 Calcul du score de pertinence des concepts candidats

Notre objectif ici est de calculer un score de pertinence pour les concepts candidats qui sont susceptibles d'être pertinents à un texte donné. Comme nous avons mentionné précédemment, les concepts candidats sont ordonnés en fonction de la combinaison de la similarité thématique et structurelle en se basant sur la corrélation d'ordre de mots constituant les concepts. De ce fait, le poids du concept C vis-à-vis du texte Θ (document ou requête), dénoté $Rel(C, \Theta)$, peut être donné par :

$$Rel(C, \Theta) = (1 + Sim(C, \Theta)) \times (1 + \rho(C, \Theta)) \tag{V.3}$$

où

- $Sim(C, \Theta)$ correspond à la similarité thématique entre le concept C et le texte Θ,
- $\rho(C, \Theta)$ correspond à la similarité structurelle qui modélise la corrélation en termes d'ordre de mots entre le texte et chaque entrée d'un concept particulier dans la ressource termino-ontologique. Pour ce faire, nous pouvons utiliser la mesure de Spearman en statistiques pour modéliser la corrélation des rangs liés à deux objets (e.g., deux textes).

Nous détaillons le calcul de ces deux mesures de similarité dans les sections suivantes.

3.2.1 Calcul du score thématique des concepts candidats

Le score thématique $Sim(C, \Theta)$ a pour but de récupérer et d'ordonner dans un premier temps les premiers concepts qui sont les plus similaires à un texte Θ (e.g., le document ou la requête). Étant donné un texte, le système de RI, qui est essentiellement basé sur la mesure Cosinus, retourne une liste de concepts qui sont thématiquement similaire au texte original. Formellement, la similarité entre Θ et C, dénotée $Sim(C, \Theta)$, est donnée par :

$$Sim(C, \Theta) = \frac{\sum_{j=1}^{N_c} w_j * d_j}{\sqrt{\sum_{j=1}^{N_c} w_j^2} * \sqrt{\sum_{j=1}^{N_c} d_j^2}} \tag{V.4}$$

où

- N_c est le nombre total de concepts dans la terminologie,
- w_j et d_j sont le poids du mot w_j dans le document représentant le concept C et dans le texte Θ. Ce poids, dénoté p_j, est calculé en utilisant le schéma

de pondération BM25 (Robertson et Walker, 1994) :

$$p_j = tf_j * \frac{log\frac{N-n_j+0.5}{n_j+0.5}}{k_1 * ((1-b) + b * \frac{len}{avgl}) + tf_j} \quad \text{(V.5)}$$

où

- tf_j est la fréquence du mot w_j dans le texte (C, Θ),
- N est le nombre total de textes (concepts ou documents) dans la collection,
- n_j est le nombre de textes (concepts, documents) contenant le mot w_j,
- len est la longueur du textes, i.e. le nombre de mots distincts,
- $avgl$ est la longueur moyenne des textes.

3.2.2 Calcul du score de la corrélation d'ordre de mots

Dans cette étape, les concepts extraits à l'étape précédente grâce au score de similarité thématique Cosinus sont réordonnés en fonction de la corrélation en termes d'ordre de mots entre un morceau de texte dans le document et les termes d'entrée désignant les concepts dans l'ontologie. L'idée ici est donc de tenir compte de la formation des termes, c-à-d l'ordre d'apparition des mots constituants, désignant les concepts ayant une corrélation élevée avec la formation du texte, c-à-d l'odre d'apparition des mots dans le texte. Du fait que les concepts se trouvent en général au niveau de la phrase, le texte Θ est d'abord découpé en phrases. Le découpage du texte est réalisé grâce à la reconnaissance des séparateurs de phrases (e.g., '.', ';', '?', '!'). De plus, nous utilisons une heuristique pour vérifier si le symbole "." est à la fin d'une phrase ou non. Cette heuristique porte sur la vérification des abréviations (e.g., "U.S.A"), des noms de gènes ou de protéines (e.g., "E. coli") ou des chiffres (e.g., 1.0). Formellement, le texte Θ peut être représenté comme un ensemble de phrases :

$$\{p_1, p_2, ..., p_q\} \quad \text{(V.6)}$$

où p_i est une phrase dans le texte et q est le nombre de phrases identifiées.

Afin d'éviter les erreurs d'identification et de réduire le nombre de calcul à effectuer, une fenêtre W délimitée par le premier et le dernier mot de chaque terme d'entrée est capturée (*cf.* l'exemple dans la figure V.1). Pour chaque concept candidat, nous calculons la corrélation d'ordre de mots entre chaque terme d'entrée E et chaque fenêtre W, dénotée $\rho(E, W)$. Pour mesurer cette corrélation, nous représentons chaque instance textuelle (c-à-d la phrase ou le terme d'entrée) par une liste ordonnée en fonction de la position des mots. De ce fait, nous pouvons appliquer la mesure de Spearman en statistiques pour calculer la corrélation entre les deux rangs. Si $\rho = -1$, la corrélation entre

les deux rangs est nulle. Si $\rho = 1$, les deux rangs sont parfaitement corrélés. Dans les autres cas, la valeur de ρ est comprise entre -1 et 1 en fonction de la corrélation entre les deux rangs.

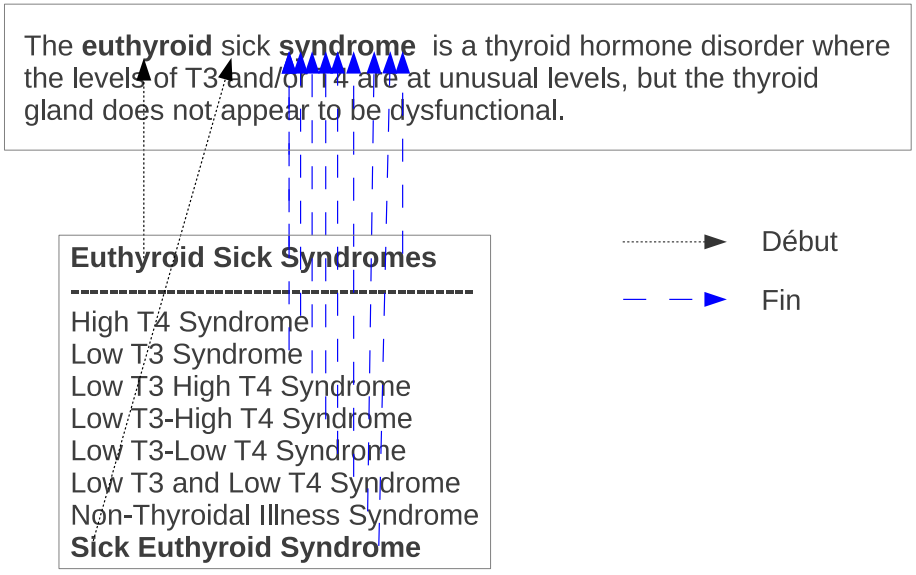


FIGURE V.1 – Une fenêtre délimitée par le premier et le dernier mot de chaque terme d'entrée du concept

L'algorithme 3 illustre de manière générale les principales étapes de notre méthode d'extraction de concepts basée sur la mesure de corrélation d'ordre de mots. La corrélation de l'ordre de mots entre la fenêtre W délimitée par le premier et le dernier mot d'un terme d'entrée E et ce dernier est donc calculée par :

$$\rho(W, E) = 1 - \frac{6 * \sum_i^L [r(w_i, W) - r(w_i, E)]^2}{L * (L^2 - 1)} \tag{V.7}$$

où

- $r(w_i, W)$ représente l'ordre ou la position relative du mot-clé w_i dans la fenêtre du fragment de texte délimitée par le premier et le dernier mot qui constituent le terme d'entrée E du concept,
- $r(w_i, E)$ est l'ordre du mot-clé w_i constituant l'entrée E,
- L est le nombre de mots dans les vecteurs représentant le fragment de texte et l'entrée du concept. Les vecteurs de termes sont normalisés en sorte qu'ils possèdent la même taille.

Enfin, la similarité structurelle entre le concept candidat et le texte donné est retenue comme la corrélation maximale entre chacun de ses termes d'entrée et le texte :

$$\rho(c, \Theta) = max_{E \in Entries(c)} \rho(E, \Theta) \tag{V.8}$$

où

- $Entries(c)$ désigne l'ensemble de termes d'entrée du concept c.

- $\rho(E, \Theta) = max_{p \in \Theta \wedge W \in p} \rho(W, E)$, où p est une phrase contenant la fenêtre W. Ici, la corrélation entre un terme d'entrée et le texte est la **corrélation maximale** entre le terme d'entrée et la fenêtre correspondante.

Le tableau V.4 illustre le calcul de la corrélation entre la fenêtre du fragment de texte et le terme d'entrée du concept. Dans cet exemple illustratif, le terme "*Colorectal cancer*" est défini dans le thésaurus MeSH comme un terme *non-préféré* du concept "**Colorectal Neoplasms**" (*terme-préféré*). Nous pouvons ignorer le dernier car celui-ci a une corrélation faible avec le morceau de texte (un seul mot commun entre la fenêtre et le terme). Pour le terme non-préféré ayant deux mots communs entre celui-ci et le texte, la fenêtre "**colorectal, lung and prostate cancer**", déterminée par le premier et le dernier mot de ce terme, est utilisée pour estimer la corrélation d'ordre de mots entre le concept et le morceau de texte.

TABLEAU V.4 – Calcul de la corrélation de positions des mots entre un terme d'entrée du concept désigné par le terme "Colorectal cancer" et la fenêtre d'un fragment de texte délimitée par le premier et le dernier mot du terme d'entrée.

Terme d'entrée E	**colorectal**	**cancer**		
Vecteur v_E	33573	592		
Fenêtre W	**colorectal**	**lung**	**prostate**	**cancer**
Vecteur v_W	33573	4874	13927	592
après normalisation				
Vecteur normalisé v_W	33573	4874	13927	592
Vecteur normalisé v_E	33573	592	4874	13927
conversion des positions relatives en rangs				
$A = r(w_i, v_W)$	0	1	2	3
$B = r(w_i, v_E)$	0	3	1	2
$A[i] - B[i]$	0	-2	1	1
corrélation des rangs entre A et B				
$\rho(A, B)$	$1 - 6 \times \frac{0+4+1+1)}{4 \times (4^2-1)} = 1 - 6 \times \frac{6}{4 \times 15} = 0.4$			
normalisation par la taille de la fenêtre				
$score(W, E) = \rho(A, B)/\|W\| = 0.4/4 = 0.1$				

Algorithme 3 – Extraction de concepts basée sur la corrélation des positions de mots en utilisant la mesure de Spearman

Entrées : Texte Θ, Terminologie T
Sorties : Liste des concepts L_c ordonnés en fonction du score thématique et structurelle

{Extraire les concepts candidats en utilisant d'abord la similarité thématique}
$L_c \leftarrow extraireConcepts(\Theta, T)$
{Calculer la corrélation d'ordre de mots (positions des mots-clés) en utilisant la mesure de Spearman}
{Découper le texte en phrases}
$Phrases \leftarrow extrairePhrases(\Theta)$
Pour $p \in Phrases$ **faire**
  {Extraire et normaliser la liste de mots-clés de p}
  $L_p \leftarrow normaliser(extraireMots(p))$
  {Calculer la corrélation entre chaque phrase et chaque concept candidat}
  Pour $c \in L_c$ **faire**
    {Extraire tous les termes d'entrée (variantes) du concept c}
    $V \leftarrow extraireVariantes(c)$
    Pour $v \in V$ **faire**
      $L_v \leftarrow normaliser(extraireMots(v))$
      Si $L_p \cap L_E \neq NULL$ **alors**
        {Appliquer l'algorithme de Spearman pour mesurer la corrélation entre p et v}
        $\rho(p, E) \leftarrow corelSpearman(L_p, L_v)$ {*cf.* algo 4}
        {Retenir la corrélation maximale entre le texte Θ et le concept c}
        Si $\rho(\Theta, c) < \rho(p, E)$ **alors**
          $\rho(\Theta, c) \leftarrow \rho(p, E)$
        Fin Si
      Fin Si
    Fin Pour
    {Combiner la similarité thématique et la similarité structurelle}
    $score(\Theta, c) = (1 + Sim(\Theta, c)) \times (1 + \rho(\Theta, c))$
  Fin Pour
Fin Pour
{Trier les concepts en fonction du score combiné}
$trier(L_c)$
Retourner L_c

Algorithme 4 – Calculer la corrélation de Spearman

Entrées : Phrase p, Terme d'entrée E d'un concept
Sorties : Corrélation de Spearman $-1 \leq \rho(p, E) \leq 1$

{Construire les vecteurs de mots (ids)}
$v_p \leftarrow vecteur(p)$; $v_E \leftarrow vecteur(E)$
$min \leftarrow \|v_p\| < \|v_E\| ? \|v_p\| : \|v_E\|$
{capturer la fenêtre W}
$first \leftarrow v_p(v_E[0])$; $last \leftarrow v_p(v_E[\|v_E\| - 1), first)$
Si $first < 0 \vee last < 0$ **alors**
  Retourner -1
Fin Si
Si $first == last$ **alors**
  Retourner 0 {terme simple constitué d'un mot}
Fin Si
$v_W \leftarrow extraireFenêtre(v_p, first, last)$
{Normaliser la fenêtre par rapport à v_E}
Si $\|v_W\| < \|v_E\|$ **alors**
  $k = 0$
  Tant que $k < \|v_E\|$ **faire**
    Si $W.indexOf(v_E[k]) < 0$ **alors**
      $ajouter(v_W, v_E[k])$
    Fin Si
    $k \leftarrow k + 1$
  Fin Tant que
Fin Si
$supprimerDoublons(v_W)$
{normaliser le vecteur v_E du terme d'entrée E}
$k = 0$
Tant que $k < \|v_W\|$ **faire**
  Si $v_t(v_W[k]) < 0$ **alors**
    $ajouter(v_E, v_W[k])$
  Fin Si
  $k \leftarrow k + 1$
Fin Tant que
{convertir les deux vecteurs en rangs des positions}
$i \leftarrow 0$
Tant que $i < \|v_W\|$ **faire**
  $A[i] \leftarrow i; B[i] \leftarrow v_E(v_W[i])$
  $i \leftarrow i + 1$
Fin Tant que
{calculer la corrélation de Spearman entre deux rangs}
$\rho(A, B) \leftarrow cor.test(A, B, method =$ "$spearman$")
{normaliser le score final par la taille de la fenêtre}
Si $\rho(A, B) < 1$ **alors**
  $\rho(A, B) \leftarrow \rho(A, B)/\|v_W\|$
Fin Si
Retourner $\rho(A, B)$

3.3 Illustration de l'extraction des concepts par des exemples concrets

Dans le tableau V.5, nous illustrons les résultats notre méthode d'extraction de concepts. La colonne de gauche du tableau correspond aux concepts extraits dans la première étape via la recherche des concepts similaires au texte (méthode basée sur la RI en utilisant la mesure Cosinus). La colonne de droite correspond aux concepts extraits par l'algorithme d'extraction basé sur la corrélation des positions de mots constituant le texte et les concepts MeSH.

Comme nous le voyons dans le tableau V.5, parmi les 20 premiers concepts extraits en utilisant l'algorithme basé sur la mesure thématique Cosinus, les quatre premiers concepts, le huitième et le dix-huitième (soulignés) correspondent aux sujets du texte illustratif. Les autres concepts sont en général reliés au concept "cancer". Par exemple, le concept "BRCA1" est une protéine liée au cancer du sein. Le concept "Infectious skin diseases" parmi des sujets sémantiques du morceau de texte est classé en dix-huitième. Nous remarquons que parmi les premiers concepts retournés, les concepts les plus liés aux sujets extraits à partir du texte illustratif en utilisant la méthode Cosinus sont éparpillés et intercalés par d'autres concepts qui sont plus ou moins reliés aux sujets principaux du texte. Il est pourtant difficile d'estimer la pertinence de ces concepts ainsi identifiés comme candidats.

En revanche, en utilisant l'algorithme d'extraction de concepts basé sur la corrélation de Spearman, les concepts les plus reliés aux sujets du texte sont renvoyés en premier (avec un score positif) car ils partagent en même temps un nombre maximum de mots avec chaque entrée du concept et plus particulièrement les mots constituant une phrase du texte apparaissent dans le même un ordre d'apparition que ceux dans les termes d'entrée du concept. Nous soulignons que ces deux sources d'évidence permettent d'avoir une **estimation approximative** de la similarité entre chaque terme d'entrée d'un concept particulier et chaque phrase, ou plus précisément la fenêtre correspondante. L'exemple présenté dans le tableau V.5 montre que les concepts comme "Infectious skin diseases" (*Skin diseases, Infectious*), "Cancer of the skin" (*Skin Neoplasms*), "Prostate cancer" (*Prostate Neoplasms*) sont mieux ordonnés que les concepts "Colorectal cancer", "Breast cancer", etc. parce que les premiers sont observés "clairement" dans le texte tandis que mots figurant dans les derniers sont intercalés par d'autres mots. La normalisation par la taille de la fenêtre dans l'algorithme 4 vise à estimer de manière approximative la probabilité que le concept soit observé dans le texte.

Cette technique est donc capable de **distinguer** de manière efficace **les concepts pertinents de ceux qui ne le sont pas**. En effet, les concepts non-pertinents ou moins pertinents par rapport au sémantique du fragment de

"The ***cancer of the skin*** *represents the most commonly diagnosed* ***cancer****, surpassing* ***breasts****,* ***colorectal****,* ***lung*** *and* ***prostate cancer****.* An **infectious skin disease** can have implications for **cancer** treatment..."

Cosinus	**Corrélation de Spearman**
0|C0037286|Skin Neoplasms|10.55	0|C0037278|Skin Diseases, Infectious|1.00
1|C0033578|Prostatic Neoplasms|9.10	1|C0033578|Prostatic Neoplasms|1.00
2|C0024121|Lung Neoplasms|8.52	2|C0037286|Skin Neoplasms|1.00
3|C0009404|Colorectal Neoplasms|8.47	3|C0024121|Lung Neoplasms|0.17
4|C0949634|Antineoplastic Protocols|8.29	4|C0009404|Colorectal Neoplasms|0.10
5|C0009405|Colorectal Neoplasms, Hereditary Nonpolyposis|7.97	5|C1458155|Breast Neoplasms|0.09
6|C0085183|Neoplasms, Second Primary|7.81	6|C0027651|Neoplasms|0.00
7|C0037278|Skin Diseases, Infectious|7.10	7|C0540004|Nuclear Receptor Coactivator-3|-1.00
8|C0259275|BRCA1 Protein|7.09	8|C0029897|Otorhinolaryngologic N.|-1.00
9|C0149925|Small Cell Lung Carcinoma|6.75	9|C0027646|Neoplasm Staging|-1.00
10|C0006827|Cancer Care Facilities|6.70	10|C0376659|Cancer Vaccines|-1.00
11|C2607925|Early Detection of Cancer|6.68	11|C0085118|Oncology Service, Hospital|-1.00
12|C0016978|Gallbladder Neoplasms|6.59	12|C0294028|BRCA2 Protein|-1.00
13|C0022374|Jejunal Neoplasms|6.57	13|C0242596|Neoplasm, Residual|-1.00
14|C0020876|Ileal Neoplasms|6.57	14|C0032019|Pituitary Neoplasms|-1.00
15|C0031347|Pharyngeal Neoplasms|6.55	15|C0027658|Neoplasms, Germ Cell and Embryonal|-1.00
16|C0029295|Oropharyngeal Neoplasms|6.55	16|C0013207|Drug Screening Assays, Antitumor|-1.00
17|C1458155|Breast Neoplasms|6.53	17|C0040136|Thyroid Neoplasms|-1.00
18|C0002455|American Cancer Society|6.50	18|C0206710|Neoplasms, Basal Cell|-1.00
19|C0023055|Laryngeal Neoplasms|6.45	19|C0036095|Salivary Gland Neoplasms|-1.00
...	...

TABLEAU V.5 – Résultats de l'extraction de concepts basée sur la corrélation de Spearman en comparaison avec ceux obtenus par la méthode basée sur la mesure Cosinus en RI.

Les termes mis en gras sont prédéfinis comme *termes MeSH*. Les termes soulignés sont des termes préférés qui sont identifiés à partir du fragment de texte. Par exemple, "Neoplasms" est le terme préféré et "cancer" est le terme non-préféré du concept dont l'identifiant unique est *C0027651*.

texte possèdent une corrélation faible en terme d'ordre des mots. Ils sont en général affectés par un score de corrélation non positif et sont par conséquent classés après les concepts ayant une plus grande corrélation d'ordre des mots.

4 Évaluation expérimentale

Dans cette section, nous décrivons les différentes expérimentations réalisées dans le but de valider **notre approche de RI basée sur notre méthode d'extraction de concepts biomédicaux appliquée sur les documents ainsi que sur les requêtes**. L'évaluation porte particulièrement sur les performances de la RI obtenues par :

- les **différents modèles de RI de l'état-de-l'art**, à savoir le modèle BM25 (Robertson et Walker, 1994) pour pondérer les mots-clés de la requête et le modèle Bo1 (Amati, 2003) pour l'expansion de requêtes (*cf.* la section 2.4 du chapitre II),

- l'**expansion conceptuelle de requêtes** basée sur notre méthode d'*extraction de concepts* biomédicaux.

Nous présentons le cadre d'évaluation de la RI biomédicale défini par TREC, intitulé TREC Genomics *Ad Hoc* Retrieval, ainsi que les ressources expérimentales utilisées dans la section 4.1, et puis nous définissons les différents scénarios d'évaluation constituant notre protocole d'évaluation de notre approche de RI biomédicale dans la section 4.2. Les résultats expérimentaux sont présentés dans la section 4.4. Dans la section 4.5, nous comparons nos meilleurs résultats à ceux qui ont été obtenus par les participants de TREC Genomics.

4.1 Cadre d'évaluation de TREC Genomics

4.1.1 Description de la collection de documents

Nous utilisons le corpus de TREC Genomics 2004 pour les raisons suivantes :

- elle contient une quantité volumineuse de documents (plus de 4.6 millions de documents),
- elle représente les vrais besoins d'informations des professionnels de santé,
- la plupart des requêtes utilisent des acronymes sans leur forme complète, ce qui représente un grand défi pour pallier le problème de défaut d'appariement (mismatch) entre la requête et les documents car en général les

documents (articles scientifiques) contiennent à la fois la forme complète des noms de gènes associés par leur(s) acronyme(s). Donc, l'expansion de requêtes avec la forme complète des noms de gènes peut avoir un effet sur les performances de la RI.

Le tableau V.6 présente les statistiques de la collection TREC Genomics 2004. Il s'agit d'un sous-ensemble de la base bibliographique, intitulée MEDLINE[4], des résumés d'articles de journaux biomédicaux entre **1994** et **2003**.

Nombre de documents	4.6 millions
Nombre de documents jugés	42255
Longueur moyenne du document	202
Nombre de requêtes	50
Longueur moyenne de la requête	17
Nombre de documents pertinents par requête	75

TABLEAU V.6 – Statistiques de la collection TREC Genomics 2004

Le nombre de documents était environ **4.6 millions d'enregistrements** représentant approximativement un tiers de la taille de MEDLINE jusqu'en 2004. La taille de la collection occupe **9.5 Giga octets** au total. Il existe parmi ces enregistrements 1209243 (soit 26.3 %) qui n'ont pas de résumés. Chaque enregistrement contient essentiellement les quatre champs les plus importants pour une recherche basée sur les mots-clés :

- **PMID** : identifiant de l'enregistrement dans PubMED,
- **TI** : titre de l'article,
- **AB** : le résumé de l'article,
- **MH** : un ensemble de termes préférés issus du MeSH ajoutés manuellement ou semi-automatiquement par les indexeurs humains.

Nous utilisons tous ces quatre champs dans nos expérimentations pour indexer toute la collection de 4.6 millions de documents, y compris le champ MH pour optimiser les performances de la RI. En effet, l'utilisation des termes MeSH manuellement annotés par les indexeurs humains a montré une amélioration des performances de la RI (Srinivasan, 1996; Abdou et Savoy, 2008; Dinh et Tamine, 2011a).

4.1.2 Description de l'ensemble de requêtes

Les requêtes sont collectées à partir des vrais besoins d'informations des biologistes (*cf.* les exemples dans le tableau V.7). Un ensemble de 50 requêtes

4. `http://www.nlm.nih.gov/bsd/pmresources.html`

TABLEAU V.7 – Exemples de requêtes de TREC Genomics

<ID> 6
<TITLE> FancD2
<NEED> Find articles about function of FancD2
<CONTEXT> There are many genes involved in Fanconi Anemia and the downstream pathways of FancD2 in flies. The FancD2 is monoubiquitylated and there are 2 components of the FancD2 pathway. The researcher studies the FancD2 pathway in flies.

<ID> 36
<TITLE> RAB3A
<NEED> Background information on RAB3A
<CONTEXT> Further information about a gene is needed after it is identified through a gene expression profile. The genes are related to synaptic plasticity in learning and memory.

ont été créées et jugés par les utilisateurs. Chaque requête contient trois champs principaux :

- **ID** : identifiant de la requête,
- **TITLE** : besoin d'information bref ou requête courte,
- **NEED** : besoin d'information détaillé ou requête longue,
- **CONTEXT** : information supplémentaire sur la requête.

Dans nos expérimentations, nous avons exclu le champ "CONTEXT" car celui-ci contient des termes non-informatifs pour la requête comme "involved in", "studies", "learning". En effet, selon les meilleurs résultats obtenus par les participants de TREC Genomics 2004, les termes de la requête dans le champ "CONTEXT" doivent avoir un poids moins important que ceux dans les autres champs (Fujita, 2004). Cependant, l'estimation du poids des termes dans les champs différents demande un entraînement sur la collection pour déterminer les cœfficients adéquats. En général, les termes dans la requête courte sont les plus importants, alors que ceux dans la requête longue complètent la description du besoin d'information. En les fusionnant, les mots-clés importants réapparaissent plusieurs fois et de ce fait ils possèdent un poids plus important. Par exemple, dans la requête 36 (*cf.* le tableau V.7), l'acronyme "RAB3A" apparaît deux fois dans la requête fusionnée ; par conséquent, il est plus important que les autres termes comme "background" ou "information".

4.1.3 Ressources termino-ontologiques des concepts biomédicaux

Pour identifier les concepts biomédicaux, nous utilisons le thésaurus MeSH (version 2011) en supposant que l'évolution de la terminologie au fil des années reste toujours valide et n'a pas beaucoup d'influence sur l'identification des concepts dans les enregistrements de MEDLINE entre 1994 et 2003. Nous avons extrait les concepts MeSH (termes préférés et non-préférées) qui ont été intégrés dans l'UMLS. Il y a donc au total 25.585 concepts qui sont représentés par 136.384 entrées préférées ou non-préférées. La figure V.2 illustre la structure poly-hiérarchique du MeSH comprenant 16 domaines au total. Chaque domaine peut avoir un ou plusieurs sous domaines et ainsi de suite. Par exemple, le domaine des "Maladies [C]" possède 26 sous domaines notés de C01 à C26. Chaque niveau correspond à un concept biomédical particulier.

1. + **Anatomy [A]**
2. + **Organisms [B]**
3. – **Diseases [C]**
   - **Bacterial Infections and Mycoses [C01] +**
   - **Virus Diseases [C02] +**
   - **Parasitic Diseases [C03] +**
   - **Neoplasms [C04] +**
   - **Musculoskeletal Diseases [C05] +**
   - **Digestive System Diseases [C06] +**
   - **Stomatognathic Diseases [C07] +**
   - **Respiratory Tract Diseases [C08] +**
   - **Otorhinolaryngologic Diseases [C09] +**
   - **Nervous System Diseases [C10] +**
   - **Eye Diseases [C11] +**
   - **Male Urogenital Diseases [C12] +**
   - **Female Urogenital Diseases and Pregnancy Complications [C13] +**
   - **Cardiovascular Diseases [C14] +**
   - **Hemic and Lymphatic Diseases [C15] +**
   - **Congenital, Hereditary, and Neonatal Diseases and Abnormalities [C16] +**
   - **Skin and Connective Tissue Diseases [C17] +**
   - **Nutritional and Metabolic Diseases [C18] +**
   - **Endocrine System Diseases [C19] +**
   - **Immune System Diseases [C20] +**
   - **Disorders of Environmental Origin [C21] +**
   - **Animal Diseases [C22] +**
   - **Pathological Conditions, Signs and Symptoms [C23] +**
   - **Occupational Diseases [C24] +**
   - **Substance-Related Disorders [C25] +**
   - **Wounds and Injuries [C26] +**
4. + **Chemicals and Drugs [D]**
5. + **Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]**
6. + **Psychiatry and Psychology [F]**
7. + **Phenomena and Processes [G]**
8. + **Disciplines and Occupations [H]**
9. + **Anthropology, Education, Sociology and Social Phenomena [I]**
10. + **Technology, Industry, Agriculture [J]**
11. + **Humanities [K]**
12. + **Information Science [L]**
13. + **Named Groups [M]**
14. + **Health Care [N]**
15. + **Publication Characteristics [V]**
16. + **Geographicals [Z]**

FIGURE V.2 – Structure poly-hiérarchique du MeSH

4.1.4 Acronymes et variants des noms de gènes

Pour identifier les noms de gènes ainsi que leurs acronymes et variants, il existe plusieurs ressources ou bases de données de gènes comme GO [5] (Gene Ontology), Entrez Gene [6], HUGO [7] (Human Genome Organisation), OMIM [8] (Online Mendelian Inheritance in Man) ... Nous avons utilisé la dernière car elle a été conçue pour l'utilisation par des experts et des chercheurs du domaine de la génétique [9]. L'OMIM, qui a été intégrée dans l'UMLS version 2011, contient un nombre de 19.352 noms de gènes au total, avec un nombre de 43946 d'entrées (noms complets, acronymes, et variants). Chaque nom complet de gène (ex : EPIDERMAL GROWTH FACTOR) peut être désigné par un ou plusieurs acronymes (ex : EGF, URG). Nous avons donc extrait pour chaque nom de gène tous les acronymes ainsi que leurs variantes, puis nous les avons sauvegardés dans un dictionnaire de noms des gènes. Chaque gène est sauvegardé dans un enregistrement, appelé RECORD séparément. Finalement, nous avons indexé tous les enregistrements par leurs acronymes ou leurs variantes, ce qui nous permet de retrouver leur forme complète. Le tableau V.8 présente quelques noms de gènes dont le champ **MH** représente la forme complète du nom de gène et **ENTRY** représente leurs acronymes et variantes.

4.2 Scénarios d'évaluation

Dans cette section, nous décrivons les différents scénarios d'évaluation afin de valider notre approche de RI biomédicale basée sur les ressources termino-ontologiques, notamment les méthodes d'extraction de concepts. Nous comparons les performances de notre approche de RI biomédicale aux performances de la RI obtenue par les modèles de RI de l'état-de-l'art ainsi que les performances obtenues en se basant sur les différentes méthodes d'extraction de concepts. Nous évaluons d'abord l'efficacité de notre méthode d'extraction de concepts en l'appliquant sur les documents et puis sur l'ensemble de requêtes.

4.2.1 Évaluation de l'efficacité des méthodes d'extraction de concepts sur les documents

Dans ce scénario, nous évaluons l'intérêt d'utiliser notre méthode d'extraction de concepts basée sur la corrélation d'ordre de mots pour déterminer le

5. http://www.geneontology.org/
6. http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
7. http://www.genenames.org/
8. http://www.ncbi.nlm.nih.gov/omim
9. http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim

TABLEAU V.8 – Un extrait des noms de gènes stockés dans un dictionnaire

```
*NEWRECORD
CUI = C1825770
MH = KERATIN 83
ENTRY = KRTHB3
ENTRY = HB3
ENTRY = KRT83

*NEWRECORD
CUI = C1367721
MH = EPIDERMAL GROWTH FACTOR
ENTRY = EGF
ENTRY = URG

*NEWRECORD
CUI = C1849719
MH = POLYSYNDACTYLY WITH CARDIAC MALFORMATION

*NEWRECORD
CUI = C1418092
MH = OLFACTORY RECEPTOR, FAMILY 5, SUBFAMILY I, MEMBER 1
ENTRY = OR5I1
```

contexte documentaire qui est défini par les meilleurs concepts extraits du document. Pour ce faire, nous menons une série d'expérimentations à travers les cinq scénarios suivants :

- le premier scénario (dénoté $\mathbf{DE^m}$) concerne une stratégie d'intégrer le contexte documentaire via une **expansion conceptuelle manuelle** de documents comme suit : les concepts issus du thésaurus MeSH sont extraits manuellement ou semi-automatiquement à partir des documents par les indexeurs humains ; les termes préférés désignant les concepts sont utilisés pour étendre le document,

- le deuxième scénario (dénoté $\mathbf{DE^a}$) concerne l'expansion conceptuelle de document par des concepts MeSH extraits en utilisant notre méthode d'extraction de concepts. Ce scénario est similaire au premier mais l'extraction de concepts est complètement automatique,

- le troisième scénario (dénoté **QE**) concerne uniquement la reformulation de requêtes par la méthode de réinjection de pertinence PRF appliquée sur les documents originaux sans l'expansion de documents,

- le quatrième scénario (dénoté $\mathbf{QE + DE^m}$) concerne la combinaison des contextes documentaires et de la requête : lors de l'indexation, les

concepts sont extraits à partir des documents en utilisant MeSH, puis les documents sont étendus par les termes préférés désignant les concepts MeSH (comme le premier scénario). Ensuite, lors de la recherche, les meilleurs termes issus des premiers documents étendus renvoyés par le système de RI en réponse à la requête originale. Ces termes sont utilisés pour reformuler la requête afin d'améliorer les performances de la RI.

– le cinquième scénario (dénoté $\mathbf{QE + DE^a}$) est similaire au quatrième scénario mais le processus d'extraction de concepts est automatique en utilisant notre méthode d'extraction de concepts.

Il est à noter que les documents jugés pertinents par les experts sont générés en fusionnant les documents envoyés par chaque participant dans TREC pour générer un ensemble de 45255 documents dans TREC Genomics 2004. Du fait que la tâche d'extraction de concepts peut prendre du temps, nous décidons d'extraire les concepts à partir des documents qui ont été jugés. L'indexation dans ce cas est basée uniquement sur 45255 documents.

4.2.2 Évaluation de l'efficacité des méthodes d'extraction de concepts sur les requêtes

À ce niveau, nous évaluons l'efficacité de notre méthode d'extraction de concepts sur les performances de la RI. Nous définissons les scénarios d'évaluation suivants :

– le premier scénario concerne les résultats obtenus par le modèle **Best Match - BM25** ou plus connu sous le nom **Okapi BM25**, qui est un modèle probabiliste bien fondé sur la loi de la distribution de Poisson de la fréquence des mots dans un document et dans la collection (Robertson *et al.*, 1994). Nous utilisons le modèle BM25 comme la base de référence d'évaluation (**baseline**),

– le deuxième scénario vise à évaluer les performances de la RI conceptuelle en utilisant l'algorithme d'extraction des concepts qui se base uniquement sur la recherche des concepts similaires, dénotée **RI**. Ceci est basé uniquement sur le calcul d'un score thématique entre les concepts dans l'ontologie et le texte en utilisant le modèle BM25 (Robertson et Walker, 1994),

– le troisième scénario vise à évaluer l'impact de prise en compte de la corrélation de l'ordre des mots en utilisant la mesure de corrélation de Spearman. Ceci inclut le calcul d'un score thématique et d'un score de corrélation d'ordre des mots entre les entrées des concepts et le texte,

- le quatrième scénario utilise le service ATM de PubMed [10] pour extraire les termes préférés désignant les concepts MeSH (dénoté **Pubmed**),

- le cinquième scénario utilise l'outil MetaMap (Aronson, 2001a) pour extraire les termes préférés désignant les concepts UMLS (dénoté **MetaMap**),

- le sixième scénario utilise l'outil Medical Text Indexer (MTI) (Aronson *et al.*, 2004b) pour extraire les termes préférés désignant les concepts MeSH avec un filtrage de base, un filtrage moyen et un filtrage strict (dénoté **MTI-base**, **MTI-moyen** et **MTI-strict** respectivement),

- le dernier utilise l'outil MaxMatcher (MM) (Zhou *et al.*, 2006b) pour extraire les termes préférés désignant les concepts MeSH par une recherche exacte des concepts MeSH (dénoté **MM-exact**) *vs.* des termes désignant les concepts UMLS par une recherche approximative (dénoté **MM-app**).

De plus, dans notre méthode d'extraction basée sur la mesure de Spearman (le troisième scénario), afin de mettre en évidence les concepts ayant une corrélation élevée en terme d'ordre des mots, nous définissons les deux niveaux de filtrage suivants :

- **Filtrage approximatif** (dénoté **Spa**) : a pour but de sélectionner les concepts les plus significatifs et les plus similaires à un morceau de texte. Nous supposons que les concepts ayant une corrélation positive ($0 < \rho \leq 1$) sont plus significatifs que ceux ayant une corrélation négative ($-1 \leq \rho \leq 0$). Par conséquent, nous éliminons tous les concepts ayant un score de corrélation faible (négative) avec le morceau de texte.

- **Filtrage exact** (dénoté **Spe**) : le filtrage exact est plus strict que le filtrage approximatif. Nous supposons que les concepts désignés par les termes composés sont plus importants que les concepts qui sont constitués d'un mot unique. Par conséquent, nous ne retenons que les concepts désignant par les termes ayant une corrélation absolue ($\rho = 1$).

Pour tous les modes de filtrage, si le concept qui est constitué d'un mot simple, celui-ci est affecté par un score neutre ($\rho = 0$). En revanche, si le concept est un acronyme ou une abréviation qui désigne un nom d'un gène ou d'une protéine, celui-ci est considéré comme important et reçoit un score maximal ($\rho = 1$). La reconnaissance des acronymes et abréviations est effectuée grâce aux expressions régulières suivantes :

10. http://www.ncbi.nlm.nih.gov/pubmed

- un acronyme doit contenir des lettres (minuscules et/ou majuscules) et des chiffres ($[a-zA-Z]+[0-9]+$), e.g., "FancD2", "NEIL1",
- une ou plusieurs lettres majuscules suivies par le symbole '.', un espace blanc et une ou plusieurs lettres minuscules ($[A-Z]+[.][a-z]+$), e.g., "E. coli",
- toutes les consonnes ($[\hat{}aeiouAEIOU]*$), e.g., "TGFB", "PGRP".

Concernant l'évaluation de l'efficacité de l'extraction de concepts pour étendre les requêtes, nous avons indexé toute la collection de 4.6 millions de documents afin de pouvoir comparer nos résultats à ceux obtenus par les participants dans TREC Genomics 2004. Nous construisons un dictionnaire des entrées des concepts à partir du thésaurus MeSH (noté MH), un dictionnaire à partir des termes d'entrées des concepts dans l'ontologie OMIM (noté OM) et un dictionnaire contenant à la fois les concepts MeSH et OMIM (MH-OM) pour extraire les concepts à partir d'une instance textuelle, notamment la requête de l'utilisateur. Au total, nous évaluons cinq sous-scénarios suivants concernant notre méthode d'extraction des concepts : **RI-MH**, **Spa-MH**, **Spe-MH**, **Spe-OM** et **Spe-MH-OM**.

4.3 Mesures d'évaluation des performances de la RI

Dans ce cadre d'évaluation, nous avons utilisé trois mesures d'évaluation qui sont définies dans le cadre de la campagnes d'évaluation TREC :

- **précision à** X premiers documents (dénotée $P@X$), est donc la proportion des documents pertinents par rapport aux X premiers documents renvoyés par le SRI. Elle mesure la satisfaction de l'utilisateur concernant les X premiers documents pertinents. Dans notre cas, nous retenons les précisions pour les 5, 10 premiers documents retournés, dénotées respectivement $P@5$, $P@10$.

- **précision moyenne** (Mean Average Precision, dénotée MAP) correspond à la précision moyenne calculée sur tout l'ensemble des documents pertinents retournés. Elle mesure la capacité du modèle d'appariement ou d'un SRI de pouvoir sélectionner les documents pertinents, en réponse à un ensemble de requêtes.

- **rappel** mesure la capacité du système de RI à renvoyer les documents pertinents. Ceci correspond à la proportion des documents pertinents renvoyés par le système de RI par rapport à l'ensemble des documents pertinents possibles.

Ces mesures, obtenues sur un ensemble de requêtes, sont générées par l'outil *trec_eval*[11] qui est définitivement utilisé par la communauté TREC pour évaluer les performances de la RI.

4.4 Résultats expérimentaux

À ce niveau, notre objectif est d'étudier l'impact de l'extraction de concepts sur les performances de RI. Pour cela, notre démarche consiste à étendre les documents et/ou les requêtes avec les concepts extraits. Nous présentons d'abord les résultats expérimentaux obtenus en se basant sur notre méthode d'extraction de concepts appliquée sur les documents et puis sur l'ensemble de requêtes.

4.4.1 Évaluation de notre méthode d'extraction de concepts sur les documents

Tout d'abord, nous évaluons l'impact des termes préférés ajoutés au contenu des documents en faisant varier le nombre de termes préférés désignant les concepts extraits. Le nombre de mots-clés (tokens ou mots simples) extraits à partir des premiers documents (étendus) est également un paramètre à expérimenter. Ensuite, nous évaluons l'efficacité de notre méthode d'extraction de concepts en utilisant les meilleurs paramètres expérimentés pour l'expansion conceptuelle de documents et l'expansion de requêtes.

4.4.1.1 Impact du nombre de termes préférés utilisés pour l'expansion de documents

Pour évaluer l'utilité de l'expansion de documents, le nombre de concepts, dénoté N, a été expérimenté dans l'intervalle de 0 à 10 avec un pas de 1 et de 10 à 25, avec un pas de 5 à chaque itération. Le tableau V.9 montre les résultats MAP obtenus par notre méthode d'extraction de concepts qui sont utilisés pour l'expansion de documents. Selon les résultats, l'expansion de documents par notre méthode d'extraction ne permet pas d'améliorer les performances sur l'ensemble de requêtes. Ceci peut être expliqué par le fait que les mots-clés simples issus de plusieurs termes préférés désignant les concepts extraits ne contribuent pas à améliorer la sémantique du document car ils ont été traités comme un sac de mots indépendamment.

Par contre, en observant en détail les performances de chaque requête avec

11. `http://trec.nist.gov/trec_eval/`

les différentes valeurs de N, nous avons pu observé une amélioration pour quelques requêtes en fonction de la valeur de N. Si l'on peut déterminer une valeur dynamique de N qui optimise la MAP pour chaque requête sans causer la dégradation, on peut améliorer les résultats MAP jusqu'à 0.4352 (soit +6.23 % par rapport à la base d'évaluation BM25). De même, les résultats en termes de P@5 (resp. P@10) peuvent être mieux améliorés jusqu'à 0.7180 (resp. 0.6366) avec un taux d'amélioration de +14.33% (resp. +7.53%). Du fait que les performances de la RI changent en fonction de la requête et du nombre de concepts étendus aux documents, nous décidons de retenir $N = 5$ (la valeur optimale) pour évaluer les scénarios présentés dans la section suivante.

TABLEAU V.9 – Performances de la RI (P@5, P@10, MAP) pour $N = 0..25$

N	P@5	P@10	MAP
Baseline	**0.6280**	**0.5920**	0.4097
1	0.6120 (-2.55)	0.5862 (-0.98)	0.4067 (-0.73)
2	0.6120 (-2.55)	0.5800 (-2.03)	0.4072 (-0.61)
3	0.6160 (-1.91)	0.5720 (-3.38)	0.4046 (-1.24)
4	0.6160 (-1.91)	0.5720 (-3.38)	0.4032 (-1.59)
5	0.6080 (-3.18)	0.5780 (-2.36)	**0.4118** (+0.51)
6	0.6120 (-2.55)	0.5700 (-3.72)	0.3995 (-2.49)
7	0.6200 (-1.27)	0.5740 (-3.04)	0.3881 (-5.27)
8	0.6080 (-3.18)	0.5740 (-3.04)	0.3877 (-5.37)
9	0.5320 (-15.29)	0.5620 (-5.07)	0.3861 (-5.76)
10	0.5240 (-16.56)	**0.5920** (+0.00)	0.4031 (-1.61)
15	0.5200 (-17.20)	0.5840 (-1.35)	0.3999 (-2.39)
20	0.5360 (-14.65)	0.5660 (-4.39)	0.4032 (-1.59)
25	0.5240 (-16.56)	0.5660 (-4.39)	0.4007 (-2.20)
N optimal	0.7180 (+14.33)	0.6366 (+7.53)	0.4352 (+6.23)

4.4.1.2 Paramètres du modèle de reformulation de requêtes

À ce niveau, notre objectif est de trouver la meilleure configuration du modèle de reformulation de requêtes par la méthode PRF. Les deux paramètres suivants sont à expérimenter : le nombre de documents impliqués et le nombre de termes extraits pour étendre la requête originale. De ce fait, nous faisons varier ces deux paramètres dans l'intervalle de 5 à 25 avec un pas de 5 à chaque itération.

Le tableau V.10 montrent les résultats MAP obtenus sur l'ensemble de 50 requêtes en se basant sur l'index des documents originaux (basés uniquement sur les champs TITLE et ABSTRACT sans l'expansion de documents). La meilleure valeur MAP a été observée à 20 termes (mots-clés) extraits à partir de 20 premiers documents retournés. De ce fait, nous retenons ces valeurs pour valider notre approche de RI dans la section suivante.

TABLEAU V.10 – Les résultats MAP obtenus par la reformulation de requêtes par la méthode PRF

Nb. docs \ **Nb. termes**	5	10	15	**20**	25
5	0.4369	0.4347	0.4455	0.4422	0.4440
10	0.4204	0.4232	0.4286	0.4289	0.4332
15	0.4357	0.4407	0.4431	0.4463	0.4428
20	0.4373	0.4395	0.4454	**0.4475**	0.4467
25	0.4347	0.4403	0.4429	0.4448	0.4473

4.4.1.3 Évaluation de l'efficacité de notre méthode d'extraction de concepts via l'expansion de documents et de requêtes

À ce niveau, notre objectif est d'étudier l'impact de la combinaison de l'expansion de documents et de la reformulation de requêtes sur les performances de la RI biomédicale. Nous rappelons que l'expansion de documents peut être considérée comme la détermination du contexte documentaire via les concepts candidats extraits dans une vue globale de la ressource terminologique indépendamment de la requête et l'expansion de requêtes par la méthode PRF se fait en déterminant le contexte local de la requête, c-à-d les premiers documents (étendus) retournés vis-à-vis de la requête. Pour ce faire, nous évaluons l'efficacité de la RI en se basant sur les scénarios définis dans la section 4.2.1. Nous présentons et discutons par la suite les résultats expérimentaux.

Le tableau V.11 présente les performances de la RI obtenues sur l'ensemble de 50 requêtes. Selon les résultats, nous observons que l'expansion de documents par des concepts manuellement identifiés par les indexeurs humains a permis une amélioration faible au niveau des résultats MAP et P@5. Par contre sa P@10 est légèrement dégradée. Bien que notre méthode d'extraction de concepts pour l'expansion de documents ne montre aucune amélioration au niveau de la MAP et P@10, elle donne une amélioration légère au niveau de la P@5. La différence entre ces deux scénarios est que la première est effectuée avec l'intervention des indexeurs humains qui sélectionnent environ une dizaine de concepts pour chaque document (Hersh *et al.*, 2004) tandis que notre méthode automatique utilise par défaut N=5 concepts pour étendre le contenu du document.

La méthode de reformulation de requêtes PRF en utilisant le modèle Bo1 (Amati, 2003) permet d'améliorer mieux les résultats MAP avec un taux d'amélioration de +9.23 % par rapport à la base d'évaluation de référence. Concernant les scénarios d'évaluation basés sur la combinaison de l'expansion de documents et de la reformulation de requêtes, nous remarquons que les méthodes $QE + DE^m$ et $QE + DE^a$ donnent en général une meilleure performance que

TABLEAU V.11 – Performances de la RI $P@5, P@10, MAP$ (% taux d'amélioration) de la combinaison des contextes documentaires et de requêtes en comparaison à la base d'évaluation de référence.

	P@5	**P@10**	**MAP**
$BM25$ (baseline)	0.6280	**0.5920**	0.4097
DE^m	**0.6320** (+0.64)	0.5900 (-00.34)	0.4139 (+01.03) †††
DE^a	**0.6320** (+0.64)	0.5780 (-02.36)	0.4118 (+00.51)
QE	0.6200 (-1.27)	0.5720 (-03.38)	0.4475 (+09.23)
$QE + DE^m$	0.5680 (-9.55)	0.5320 (-10.14)††	**0.4567** (+11.47)
$QE + DE^a$ (notre méthode)	0.6280 (+0.00)	0.5860 (-01.01)	0.4532 (+10.62) †††

les t-tests : † significatif ($p < 0.05$), †† très significatif ($p < 0.01$), et ††† extrêmement significatif ($p < 0.001$).

la baseline ainsi que chaque composante séparément, c-à-d DE et QE. Par exemple, la méthode manuelle $QE+DE^m$ permet une amélioration de +11.47% en termes de MAP tandis que notre méthode automatique $QE + DE^a$ donne une amélioration de +10.62% en termes de MAP par rapport à la baseline. Par contre, les précisions P@5 et P@10 ne sont pas améliorées.

4.4.2 Résultats expérimentaux des méthodes d'extraction de concepts appliquées sur les requêtes

Nous présentons dans cette section les résultats obtenus par la base d'évaluation de référence **BM25** sur l'ensemble de 4.6 millions de documents, l'expansion conceptuelle de requêtes en utilisant notre méthode d'extraction de concepts. Par la suite, nous présentons les résultats obtenus par la combinaison de l'expansion conceptuelle de requêtes et l'expansion de requête basée sur la méthode PRF (Amati, 2003). Nous comparons nos meilleurs résultats obtenus à ceux qui sont obtenus par les participants de TREC Genomics 2004. Pour tous les résultats obtenus, nous utilisons les tests-t (Student t-tests) pour déterminer la significativité des résultats. Nous utilisons le symbole † pour indiquer la significativité des résultats pour tous les tests dont la valeur de p est inférieure ou égale à 0.05 et †† pour p inférieure ou égale à 0.01.

4.4.2.1 Évaluation de l'efficacité de la RI basée sur des méthodes d'extraction de concepts via l'expansion conceptuelle de requêtes

Le tableau V.12 présente les résultats obtenus par l'expansion conceptuelle de requêtes en utilisant plusieurs méthodes d'extraction de concepts. Selon les résultats obtenus, nous avons observé que : **la plupart des termes désignant les concepts issus du MeSH ou de l'UMLS identifiés par les différentes méthodes d'extraction existant ne donne aucune amélioration importante** ou plutôt **résulte en une dégradation significative**. Par exemple, l'utilisation des termes MeSH identifiés par Pubmed ATM (grâce à l'étiquette [MeSH fields]) a diminué la MAP jusqu'à -26.58% par rapport à la *baseline*. Du fait que MetaMap a pour but d'optimiser le rappel de l'extraction de concepts en générant un nombre maximum de variantes pour chaque groupe nominal, il retourne également des termes contenant des "mots bruyants" pour la requête. Par exemple, pour la requête "*Find articles about the function of mutY in humans*.", MetaMap a identifié les termes suivants : "*Finding ; Article ; physiological aspects ; Function ; Function Axis ; Mathematical Operator ; Homo sapiens ;"*. Parmi les termes retournés par MetaMap, les mots comme "finding", "article", "physiological", "aspects", "mathematical", "operator" ne sont pas utiles pour améliorer la sémantique de la requête.

Bien que MTI propose différents niveaux de filtrage des concepts, les résultats MAP sont toujours dégradés : pour le filtrage strict, MTI ne retient que les concepts UMLS identifiés par MetaMap et la méthode "PubMed Related Citations" et puis ces concepts sont filtrés afin de ne retenir que les concepts MeSH (plus précisément les termes préférés). Par conséquent, les acronymes ont été ignorés par MTI bien qu'ils soient utiles pour améliorer la requête. Pour le filtrage de base, MTI identifie plus de concepts pour la requête mais ceux-ci sont mélangés avec les concepts non-pertinents. Par conséquent, les résultats obtenus sont plus faibles que les autres méthodes d'extraction testées dans nos expérimentations. Le filtrage moyen, qui est, selon (Aronson *et al.*, 2004b), le compromis entre les deux derniers modes de filtrage, permet d'augmenter les précisions P@5 et P@10 et de conserver le rappel bien que la MAP soit légèrement dégradée. Les taux de différence en terme de MAP de ces trois scénarios sont respectivement de -1.37%, -96.46% et de -11.07% par rapport à la baseline.

Les deux méthodes d'extraction de concepts de MaxMatcher (**MM-app** *vs.* **MM-exact**) (intégrées dans notre outil cxtractor) sont plus stables que PubMed, MetaMap et MTI. En effet, bien que l'expansion de requêtes en utilisant les concepts extraits par MaxMatcher ne donne pas de résultats significatifs par rapport à la baseline, MaxMatcher permet d'améliorer légèrement la MAP (+0.67%) et le rappel (+1.41%) pour la méthode de recherche approximative dans l'UMLS. Les résultats montrent également que la recherche approximative

TABLEAU V.12 – Performances des méthodes d'extraction de concepts pour l'expansion conceptuelle de requêtes en comparaison avec les résultats de la base d'évaluation de référence BM25.

	MAP	**P@5**	**P@10**	**Rappel**
BM25	0.3732	0.5960	0.5700	0.6124
PubMed	0.2740†† (-26.58)	0.6000 (+0.67)	0.5460 (-6.43)	0.5177† (-15.46)
MetaMap	0.2246†† (-39.82)	0.4960†† (-16.78)	0.4720†† (-26.26)	0.5218†† (-14.79)
MTI-strict	0.3681 (-1.37)	0.5840 (-2.01)	0.5740 (+1.07)	0.6006 (-1.93)
MTI-base	0.0132†† (-96.46)	0.0320†† (-94.63)	0.0280†† (-145.23)	0.0394†† (-93.57)
MTI-moyen	0.3319† (-11.07)	0.6320 (+6.04)	0.5980 (+7.50)	0.6092 (-0.52)
MM-app	0.3757 (+0.67)	0.6040 (+1.34)	0.5920 (+5.89)	0.6254 (+2.12)
MM-exact	0.3648 (-2.25)	0.6160 (+3.36)	0.5860 (+4.29)	0.6161 (+0.60)
RI	0.0570†† (-84.73)	0.1480†† (-75.17)	0.1260†† (-118.97)	0.2407†† (-60.70)
Spa-MH	0.0420†† (-88.75)	0.1080†† (-81.88)	0.0920†† (-128.08)	0.1830†† (-70.12)
Spe-MH	0.3746 (+0.38)	0.5960 (+0.00)	0.5800 (+2.68)	0.6086 (-0.62)
Spe-OM	0.3803 (+1.90)	0.5960 (+0.00)	0.5720 (+0.54)	0.6190† (+1.08)
Spe-MH-OM	0.3838 (+2.84)	0.5880 (-1.34)	0.5700 (+0.00)	0.6344 (+3.59)
Max	0.3838	0.6320	0.5980	0.6344

Les chiffres entre parenthèses représentent la différence (%) par rapport à la baseline.

dans MaxMatcher dépasse les performances de la recherche exacte en général : la MAP, le rappel et la P@10 de la recherche approximative sont meilleurs que la recherche exacte même si les taux d'amélioration par rapport à la baseline sont très compétitifs. En particulier, les scénarios basés sur l'extraction de concepts de MaxMatcher permet d'améliorer la P@10 de +5.89% et +4.29% pour la recherche approximative et exacte respectivement.

Pour nos cinq scénarios basés sur notre méthode d'extraction des concepts issus des ressources comme MeSH et OMIM (**RI-MH**, **Spa-MH**, **Spe-MH**, **Spe-OM** et **Spe-MH-OM**), nous avons des observations suivantes : la méthode RI retourne les concepts similaires à la requête, qui partagent au moins un mot avec la dernière, mais qui contiennent également des "mots inutiles" pour la requête. Par exemple, pour la requête "2 Generating transgenic mice. Find protocols for generating transgenic mice.", les concepts suivants sont retournés : "Transgenes" ; "Radionuclide Generators", "Resources", "Cohort Effect", "Income", ..., "Computer Communication Networks", "Neurologic Manifestations" ... Il est évident que les termes comme "Computer Communication Networks", "Income" ou "Resources" ... sont des termes non-pertinents pour cette requête. Les performances en termes de MAP, P@5, et P@10 et rappel sont dégradées comme le cas de MTI-base. La méthode **Spa-MH** qui est essentiellement basée sur la méthode de RI pour collecter dans un premier temps les concepts les plus

similaires à la requête mais avec une contrainte, que les termes désignant les concepts doivent avoir une corrélation positive en terme d'ordre de mots par rapport à la requête. Malheureusement, les termes retournés contiennent les mots non-pertinents pour la requête, ce qui explique la dégradation totale des performances de la RI. Pour la méthode **Spe**, seulement les concepts désignés par les termes ayant une corrélation absolue en terme d'ordre de mots sont retenus. Nous avons remarqué que l'expansion conceptuelle de requêtes en utilisant les mots simples issus des termes préférés identifiés par notre meilleure méthode d'extraction de concepts, intitulée **Spe-MH-OM**, où les termes préférés ainsi que les acronymes et leurs variantes sont extraits, permet d'améliorer le rappel de +3.59% et la MAP de +2.84% tout en conservant les précisions P@5 et P@10 par rapport à la baseline.

4.4.2.2 Évaluation de l'efficacité de la RI basée sur la combinaison de l'expansion conceptuelle et la reformulation de requêtes basée sur la technique PRF

À ce niveau, nous présentons et discutons les résultats obtenus en combinant l'expansion conceptuelle de requêtes et la reformulation de requêtes basée sur la technique PRF. Dans le modèle de reformulation PRF, nous avons deux paramètres à expérimenter : le nombre de termes extraits (paramètre 1) à partir d'un certain nombre des premiers documents retournés (paramètre 2). Le nombre de documents impliqués dans la reformulation est parmi les valeurs suivantes : {10, 15, 20} et le nombre de termes extraits est parmi les valeurs de 10 à 50 avec un pas de 10 à chaque itération. Les figures V.3, V.4 et V.5 présentent les performances (MAP, P@5 et P@10 respectivement) obtenues sur la collection originale de TREC Genomics 2004.

Nous pouvons constater qu'en général, la méthode PRF ainsi que la combinaison de celle-ci et la méthode d'expansion conceptuelle de requêtes basée sur notre meilleure méthode d'extraction de concepts est plus performante que la baseline. Cela prouve que la technique de reformulation de requêtes est plus performante que la baseline. De plus, l'expansion conceptuelle de requêtes permet d'améliorer mieux les précisions P@5 et P@10 par rapport à la méthode PRF avec un taux d'amélioration moyen de +3.1% et de +1.16% respectivement. Par contre, les résultats MAP de ces deux scénarios sont légèrement différents. Nous concluons que l'expansion conceptuelle de requêtes par des termes préférés permet d'améliorer les précisions à X documents sans perturber les résultats MAP.

Le tableau V.13 présente les meilleures performances en termes de MAP, P@5 et P@10 pour quatre combinaisons binaires entre les deux techniques d'expansion. Les meilleurs résultats sont obtenus en utilisant **40 termes** extraits à

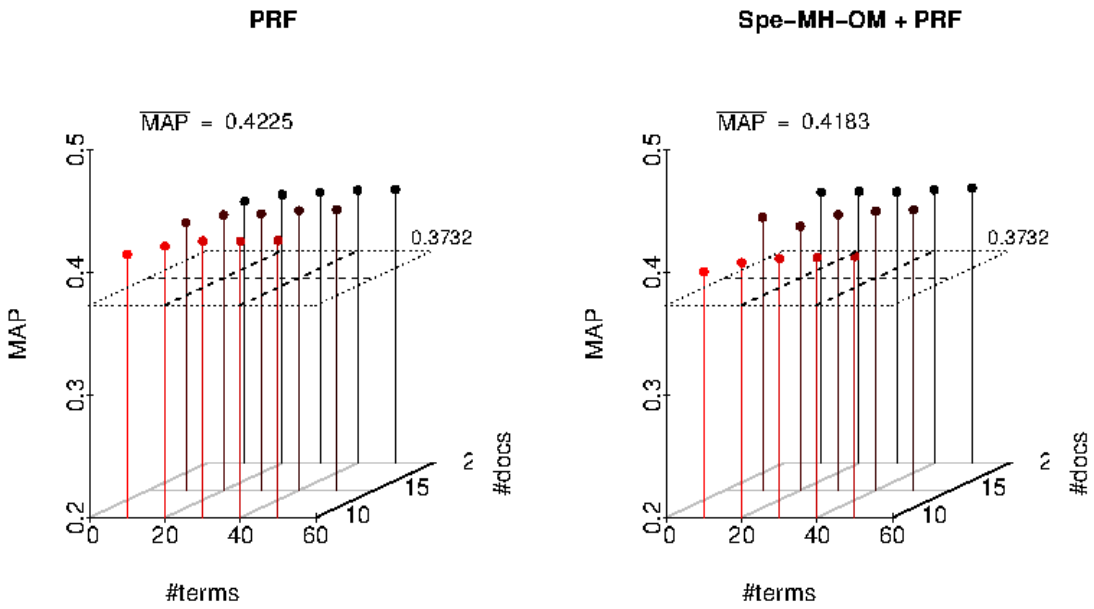


FIGURE V.3 – Résultats MAP de la méthode PRF en combinaison avec la méthode d'expansion conceptuelle

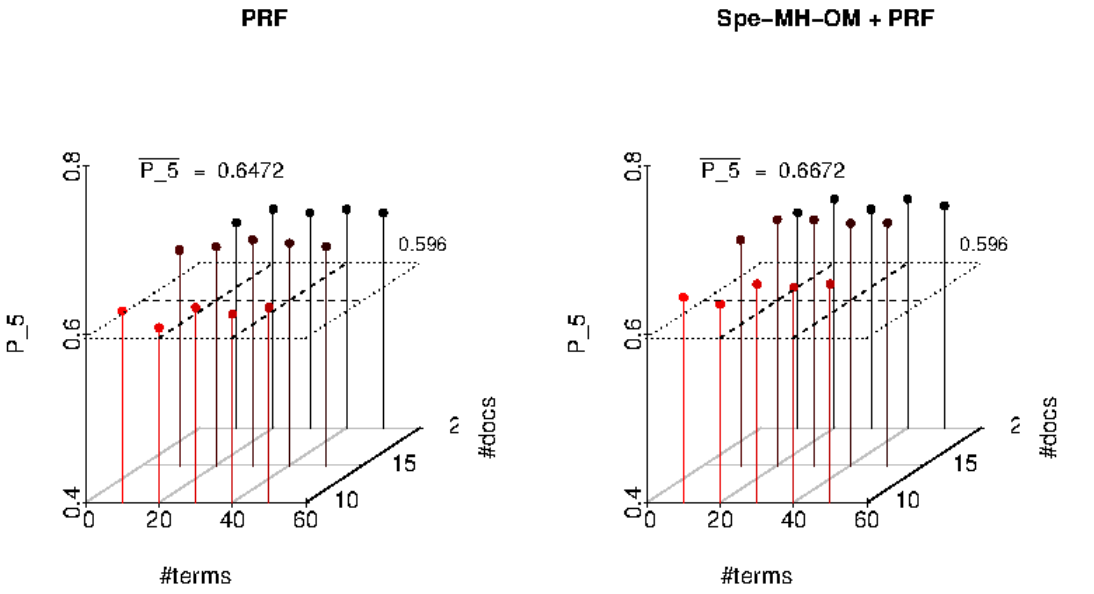


FIGURE V.4 – Résultats P@5 de la méthode PRF en combinaison avec la méthode d'expansion conceptuelle

partir de **15 premiers documents**. Le couple (0, 0) dans la colonne de gauche signifie que ni l'expansion conceptuelle de reququêtes ni la reformulation de requêtes basée sur la méthode PRF est utilisée. Le couple (1, 1) indique que ces deux techniques sont utilisées à la fois. Bien que les résultats en termes de MAP et P@10 obtenus par la combinaison de la méthode d'expansion conceptuelle et de la méthode de reformulation PRF soient légèrement différents de ceux

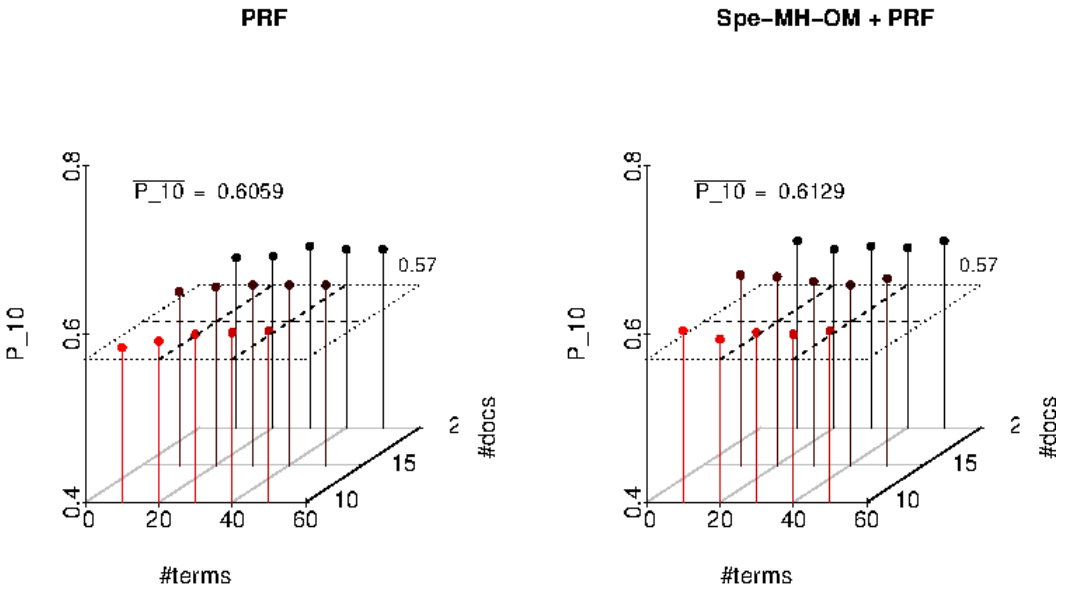


FIGURE V.5 – Résultats P@10 de la méthode PRF en combinaison avec la méthode d'expansion conceptuelle

qui sont obtenus uniquement par la méthode PRF, cette combinaison a permis d'augmenter la précision P@5 avec un taux d'amélioration de +15.44% par rapport à la baseline. La précision P@5 obtenue par cette combinaison dépasse également la précision P@5 obtenue par la méthode PRF uniquement avec un taux d'amélioration de +7.5%. Nous concluons que l'expansion conceptuelle de requêtes en utilisant les mots-clés appropriés issus des termes préférés désignant les concepts ainsi que la prise en compte des acronymes peuvent être efficaces pour augmenter le rappel tandis que la combinaison de la dernière avec la méthode d'expansion PRF permet d'améliorer la pertinence des premiers documents en conservant la MAP.

4.5 Évaluation comparative

Dans cette section, notre objectif est de comparer nos meilleurs résultats, qui sont obtenus par notre méthode d'expansion conceptuelle de requêtes exploitant à la base notre méthode d'extraction de concepts, aux résultats officiels obtenus par les participants de TREC Genomics 2004. Dans notre meilleure approche, les concepts sont extraits directement à partir de chaque requête en se basant sur notre meilleure méthode d'extraction de concepts présentée dans la section 3.2.2, les acronymes désignant les noms de gènes sont identifiés grâce aux expressions régulières et les termes correspondant à leur forme complète sont ajoutés à la requête. Par exemple, pour la requête "**6 FancD2. Find articles about function of FancD2.**", nous ajoutons les termes suivants : "**FANCONI ANEMIA, COMPLEMENTATION GROUP D2 ; FA4 ;**

TABLEAU V.13 – Performances de la RI (MAP, P@5, P@10) basée sur la reformulation PRF en combinaison avec l'expansion conceptuelle de requêtes.

A	**B**	**MAP**	**P@5**	**P@10**	
0	0	0.3732	0.596	0.570	**baseline**
0	1	0.4295†† (+15.09)	0.640†† (+7.38)	0.618†† (+8.42)	**PRF**
1	0	0.3838 (+2.84)	0.588 (-1.34)	0.570 (+0.00)	**Spe-MH-OM**
1	1	0.4280†† (+14.68)	0.688†† (+15.44)	0.620† (+8.77)	**Spe-MH-OM + PRF**

A : expansion conceptuelle, **B** : pseudo relevance feedback (PRF)

FANCD ; FANCD2 ;" qui correspondent à la forme complète et les variantes de l'acronyme "**FancD2**".

Le tableau V.14 présente les trois premiers meilleurs résultats obtenus par les participants dans TREC Genomics 2004 ainsi que les résultats du groupe qui a été classé en *médian*. Nous comparons deux de nos méthodes aux meilleurs résultats obtenus dans TREC : la première est automatique et la deuxième est manuelle. La méthode manuelle ressemble à la méthode automatique sauf le fait que dans la première, nous avons supprimé les mots suivants que nous trouvons inutiles pour la requête : "family", "members", "information", "provide", "find" et "article". Les résultats montrent que nos résultats sont meilleurs que les meilleurs résultats de TREC Genomics 2004. La MAP de notre méthode automatique (resp. manuelle) montre un taux d'accroissement de +5.03% (resp. +5.57%) par rapport à la MAP du premier groupe. Les résultats en terme de P@5 et P@10 sont également meilleurs que le premier groupe avec un taux d'accroissement de +6.17 % et +2.65% pour la méthode automatique et de +4.94% et +1.99% pour la méthode manuelle. De plus, nos méthodes permettent également d'optimiser le rappel avec un taux d'accroissement de +5.89% et +6.20% par rapport au rappel du premier groupe. Ces résultats montrent que les performances de la RI obtenues par notre approche d'extraction de concepts sont plus pertinentes que les meilleurs résultats de TREC Genomics 2004.

La figure V.6 dessine les courbes liées à la distribution de la densité par noyau (kernel density distribution) des valeurs MAP obtenues par les trois

TABLEAU V.14 – Comparaison aux résultats de TREC Genomics 2004

Rang officiel	**Run**	**MAP**	**P@5**	**P@10**	**Rappel**
1	pllsgen4a2	0.4075	0.6480	0.6040	0.6491
2	uwmtDg04tn	0.3867	0.6720	**0.6240**	0.6705
		(-5.10)	(+3.70)	(+3.31)	(+3.30)
3	pllsgen4a1	0.3689	0.6120	0.5700	0.5858
		(-4.60)	(-8.93)	(-8.65)	(-12.64)
médian	PDTNsmp4	0.2074	0.5120	0.4560	0.3697
		(-43.78)	(-16.34)	(-20.00)	(-36.88)
	Spe-MH-OM+PRF	**0.4280**	**0.6880**	0.6200	**0.6873**
	(automatique)	(+5.03)	(+6.17)	(+2.65)	(+5.89)
	Spe-MH-OM+PRF	*0.4302*	0.6800	0.6160	**0.6894**
	(manuel)	(+5.57)	(+4.94)	(+1.99)	(+6.20)

Les chiffres entre parenthèses correspondent aux taux d'accroissement par rapport aux meilleurs résultats de TREC (pllsgen4a2)

meilleurs résultats de TREC, les résultats classés en médian et nos résultats MAP obtenus par la meilleure méthode Spe-MH-OMIM (auto). La distribution de la densité par noyau est une méthode non-paramétrique visant à estimer la probabilité d'une fonction de densité d'une variable aléatoire (Sheather et Jones, 1991). Comme nous le voyons, nos résultats sont plus stables en terme de MAP que les meilleurs résultats de TREC Genomics 2004 car les valeurs les plus élevées de la MAP (à partir de 0.4) de notre approche sont distribuées au dessus des valeurs MAP de la dernière. Bien que nos résultats ne soient pas statistiquement significatifs par rapport aux meilleurs résultats, les performances de la RI de notre méthode ont montré une amélioration prometteuse en termes de la MAP et des précisons P@5, P@10 ainsi que du rappel.

Nous discutons à ce niveau les techniques utilisées par le premier groupe. Comme décrite dans (Fujita, 2004), la méthode *pllsgen4a2* a utilisé le modèle BM25 pour pondérer les termes ; les requête sont basées sur trois champs *TITLE*, *NEED* et *CONTEXT* ; les mots vides sont éliminés et les mots-clés sont normalisés par l'algorithme de Porter (Porter, 1997). Chaque requête est étendue d'abord par les termes issus de MeSH et LocusLink et ensuite étendue la deuxième fois par la méthode Pseudo Relevance Feedback (PRF). Notre méthode est différente de la meilleure méthode de TREC Genomics 2004 selon les points suivants :

– nous avons utilisé le thésaurus MeSH comme la ressource standard et en plus nous avons utilisé OMIM qui est une base d'acronymes des noms de gènes. LocusLink n'est plus à jour depuis 2005 et désormais remplacé

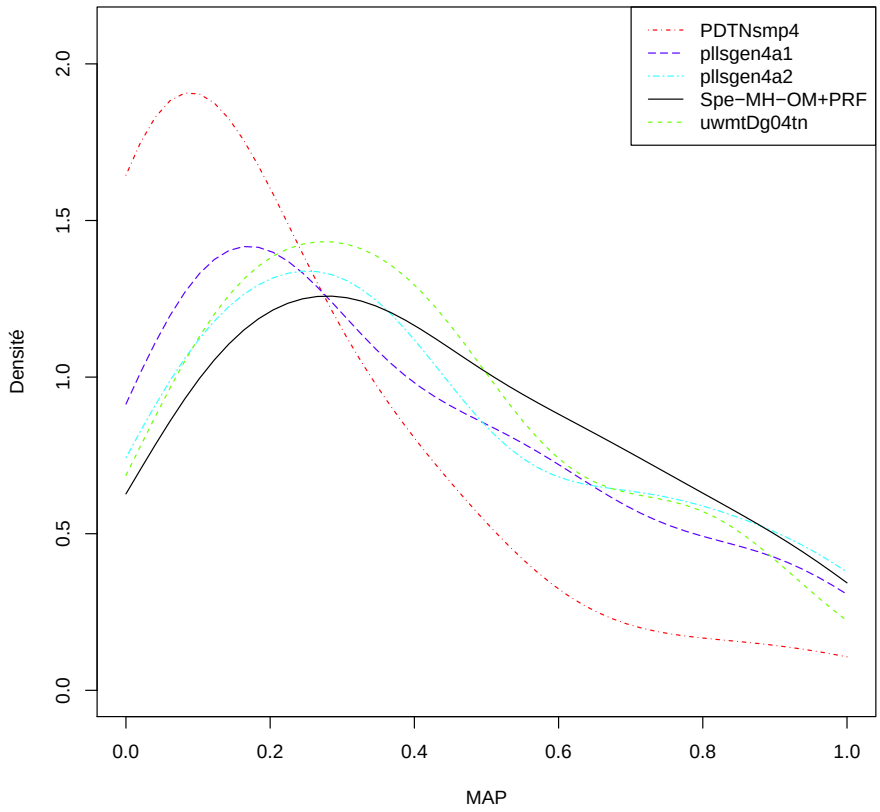


FIGURE V.6 – Distribution de la densité par noyau des résultats MAP en comparaison aux résultats officiels de TREC Genomics 2004

par l'Entrez Gene qui est essentiellement une méta base de données de gène, y compris OMIM, HUGO, GeneRIF, GO, etc. Nous avons utilisé OMIM car il s'agit d'un sous-ensemble de l'Entrez Gene qui permet de reconnaître les noms de gène de l'humanité.

– pour les requêtes, nous n'avons utilisé que les deux champs *TITLE* et *NEED* ; par conséquent, notre requête est plus courte tout en conservant la description du besoin d'information. Cela peut aider à récupérer plus de documents pertinents, car ces deux champs représentent de manière complète le besoin de l'utilisateur. De ce fait, nous pouvons optimiser le temps de réponse du système de RI, notamment le temps de recherche des mots-clés dans l'index.

– nous avons combiné l'expansion conceptuelle de requêtes et la reformulation de requête basée sur la méthode PRF en utilisant les statistiques de Bose-Einstein pour extraire les meilleurs termes issus des premiers documents renovoyés par le système de RI. Cette combinaison permet

d'optimiser mieux le rappel et les précisions ainsi que la MAP.

5 Conclusion

Nous avons présenté au cours de ce chapitre notre contribution portant sur les différentes méthodes d'extraction de concepts pour la RI biomédicale. Nous avons proposé une nouvelle méthode d'extraction de concepts qui est essentiellement basée sur la combinaison de deux mesures de similarité : thématique et structurelle. Pour cela, nous avons choisi la mesure de corrélation de Spearman permettant de traduire la similarité sémantique entre un morceau de texte (e.g., document ou requête) et un concept dans la ressource termino-ontologique. Dans le cadre la recherche d'information biomédicale, notamment dans TREC Genomics, notre algorithme d'extraction de concepts issus du thésaurus MeSH est capable d'identifier les acronymes et les abréviations qui sont définis dans l'ontologie de gènes OMIM.

Nous avons évalué les performances de la RI qui intègre l'expansion conceptuelle de requêtes en utilisant plusieurs outils d'extraction. Les résultats obtenus en utilisant les outils d'extraction de concepts existants montrent que les concepts ainsi identifiés ne permettent pas d'améliorer les performances de la RI. Concernant notre algorithme d'extraction de concepts, nous avons pu sélectionner les meilleurs concepts candidats issus du thésaurus MeSH et l'ontologie des gènes OMIM. L'expansion conceptuelle se fait par l'ajout des mots-clés issus des termes préférés désignant les concepts dans les ontologies. Si le terme identifié est un acronyme, sa forme complète (s'il en existe) est également ajoutée pour pallier au problème de défaut d'appariement entre la requête de l'utilisateur et les documents de la collection. Bien que nos résultats aient une amélioration légère en terme de rappel par rapport à la baseline, la combinaison de l'expansion conceptuelle basée sur notre meilleur algorithme d'extraction de concepts et la reformulation de requêtes basée sur la méthode PRF permet d'améliorer la précision P@5 ainsi que le rappel tout en conservant la MAP de la méthode PRF, qui est déjà statistiquement et significativement performante par rapport à la base d'évaluation de référence.

Chapitre VI

Indexation multi-terminologique pour la RI biomédicale

Sommaire

1 Introduction **186**

2 Problématiques et motivations **187**

3 Architecture générale de notre approche de RI multi-terminologique **189**

4 Indexation multi-terminologique basée sur des techniques de vote **191**
- 4.1 Extraction mono-terminologique de concepts 191
- 4.2 Extraction de concepts multi-terminologique 193

5 Appariement multi-terminologique basé sur la combinaison des contextes document et requête **198**
- 5.1 Expansion conceptuelle de documents 199
- 5.2 Combinaison de l'expansion documentaire et la reformulation de la requête par la méthode PRF 200

6 Évaluation expérimentale **202**
- 6.1 Objectifs d'évaluation 202
- 6.2 Cadre d'évaluation 202
  - 6.2.1 Collections de TREC Genomics 203
  - 6.2.2 Protocole d'évaluation 205
  - 6.2.3 Schémas d'appariement document-requête 207
  - 6.2.4 Modèles de reformulation de la requête par la méthode PRF 208
  - 6.2.5 Mesures d'évaluation des performances de la RI . . . 209
- 6.3 Résultats expérimentaux 210
  - 6.3.1 Entraînement des modèles de pondération et modèles de reformulation de requêtes 210
  - 6.3.2 Évaluation de l'efficacité de l'indexation mono-terminologique 211
  - 6.3.3 Évaluation de l'efficacité de l'indexation multi-terminologique 218
- 6.4 Discussion 221

7 Conclusion **223**

"It is unworthy of excellent men to lose hours like slaves in the labor of calculation which could be relegated to anyone else if machines were used."
–Gottfried Wilhelm von Leibnitz

Publications liées à ce travail

- **JWS 2012 :** Duy Dinh, Lynda Tamine. Towards a context sensitive biomedical information retrieval based on domain knowledge sources. *Journal of Web Semantics : Science, Services and Agents on the World Wide Web*, Elsevier, 12-13 :41-52

- **AIIM 2012 :** Duy Dinh, Lynda Tamine, Fatiha Boubekeur. Factors affecting the effectiveness of biomedical document indexing and retrieval based on terminologies. *Artificial Intelligence in Medicine, Elsevier, 2012.*

- **AIME 2011 :** Duy Dinh, Lynda Tamine. *Voting techniques for a multi-terminology based biomedical information retrieval.* Dans : the 13th Conference on Artificial Intelligence in Medicine, **AIME 2011**, Juillet 2-6, 2011, Bled, Slovenia, p. 184–193 ;

1 Introduction

Dans les chapitres IV et V, nous nous sommes concentrés sur les techniques de la RI conceptuelle en exploitant les ressources termino-ontologiques dans un schéma d'indexation et de recherche d'information mono-terminologique, c-à-d une seule terminologie est utilisée pour identifier les concepts à partir du texte (document et/ou requête) en utilisant une méthode d'extraction de concepts particulière. Dans ce chapitre, nous proposons une approche d'indexation et de recherche d'information multi-terminologique. Notre objectif ici est d'évaluer l'impact de l'intégration de plusieurs terminologies sur les performances de la RI biomédicale. Notre approche est inspirée par le principe de la poly-représentation de l'information en RI (Ingwersen, 1996). Selon ce principe, la valeur de l'information dans le document peut être augmentée en combinant plusieurs sources d'évidence. De cette manière, nous espérons obtenir une meilleure représentation du document via les concepts extraits à partir de documents en utilisant plusieurs terminologies (vues comme sources d'évidence). Nous considérons la tâche d'extraction de concepts prédéfinis dans plusieurs terminologies comme un problème de vote consistant à fusionner les listes de concepts identifiés. Pour chaque document, nous appliquons une méthode d'extraction de concepts basée sur la recherche approximative des concepts issus d'une terminologie. Étant donné que les concepts ainsi identifiés ne sont pas

pondérés, nous proposons de les pondérer en utilisant un schéma de pondération particulier (e.g., probabiliste (Robertson *et al.*, 1998)) en RI. Par la suite, les listes de concepts sont fusionnées en utilisant plusieurs techniques de fusion de données (Fox et Shaw, 1994) pour obtenir une meilleure représentation conceptuelle du document.

Ce chapitre est organisé comme suit. La section 2 présente les problématiques et les motivations de la RI conceptuelle basée sur les terminologies biomédicales. La section 3 présente l'architecture générale de notre approche de RI basée sur les terminologies. La section 4 présente les modèles de vote dédiés à la fusion des concepts extraits à partir de plusieurs terminologies. La section 5 décrit notre approche de RI basée sur les terminologies biomédicales pour améliorer les performances de la RI biomédicale. Nous évaluons notre approche de RI conceptuelle dans la section 6. La dernière section conclut le chapitre.

2 Problématiques et motivations

Il est bien connu à présent que de nombreuses terminologies médicales sont disponibles et sont en constante évolution (*cf.* le chapitre III, la section 3.2). Ces ressources termino-ontologiques peuvent être utilisées pour indexer les différents types de documents biomédicaux (Névéol *et al.*, 2006). L'objectif de l'indexation terminologique en se basant sur les terminologies est de faciliter l'accès à la littérature biomédicale en affectant une liste de concepts issus d'une ou de plusieurs terminologies biomédicales (Névéol *et al.*, 2006; Darmoni *et al.*, 2009). En pratique, les terminologies les plus utilisées pour l'indexation contrôlée sont : MeSH, SNOMED, ICD-10, GO, UMLS... Par exemple, les documents de la base MEDLINE sont indexés par les descripteurs issus du thésaurus MeSH qui sont sélectionnés par les indexeurs humains à la NLM (Aronson *et al.*, 2004b).

En RI biomédicale, bien que le thésaurus MeSH ait été largement utilisé pour indexer les documents biomédicaux, celui-ci jusqu'à présent ne couvre pas tous les termes biomédicaux dans tous les domaines de la médecine (Keizer *et al.*, 2000). Par exemple, l'utilisation de SNOMED comme le standard pour la codification des dossiers de patients représente un intérêt majeur dans les systèmes de santé (Cornet et de Keizer, 2008). De plus, les travaux les plus récents montrent que l'intégration d'autres terminologies comme SNOMED, ICD-10, CCAM, TUV, etc. dans le processus d'indexation contrôlée apporte une utilité certaine (Pereira *et al.*, 2008; Darmoni *et al.*, 2009). L'exploitation de plusieurs terminologies a été abordée dans la littérature par deux approches : la première vise à construire des associations entre terminologies tandis que la seconde se focalise sur l'extraction de concepts à partir de différentes terminologies pour

une meilleure représentation sémantique du document (Avillach *et al.*, 2007; Pereira *et al.*, 2008; Darmoni *et al.*, 2009).

C'est précisément, dans le cadre de la seconde approche que se situe la méthode d'indexation multi terminologique que nous proposons. Cette méthode est essentiellement basée sur la valeur ajoutée issue du principe de la polyreprésentation d'un document (Ingwersen, 1996). Ce principe a montré, notamment dans le domaine de la RI que l'association de différents descripteurs à un document, issus de différentes sources d'évidence, permet d'améliorer la "significativité" de son contenu. L'application de ce principe au contexte multi-terminologique nous motive à associer à un document un descripteur de concepts extraits de différentes terminologies de base.

Plus précisément, les contributions originales présentées dans ce chapitre portent sur :

- la proposition de techniques d'indexation basée sur des méthodes de fusion appliquées aux terminologies. Nous évaluons l'intérêt d'utiliser des algorithmes de vote afin de fusionner les concepts extraits à partir du texte libre. La fusion permet donc de générer une liste des concepts uniques en utilisant plusieurs terminologies. L'objectif est donc de fournir une liste de concepts les plus représentatifs qui peut être associée au texte biomédical.

- L'évaluation expérimentale permettant de mesurer l'impact réel de l'indexation multi-terminologique en RI. À notre connaissance cette étude n'a pas été réalisée à ce jour. Plus précisément, nous menons une série d'expérimentations visant à évaluer plusieurs facteurs qui influencent les performances de la RI biomédicales : (1) les connaissances sur les documents, c-à-d les concepts extraits qui sont prédéfinis dans une ou plusieurs ressources termino-ontologiques et (2) les connaissances sur la requête de l'utilisateur, c-à-d les termes extraits à partir des premiers documents retournés lors de la première phase de recherche. Nous pouvons distinguer les deux types de "contexte" : *contexte global vs. contexte local*. La notion "contexte" indique la source d'information d'où les termes les plus significatifs qui sont reliés à la requête peuvent être extraits. Le contexte global (e.g., thésaurus, ontologies, collection de documents ...) est déterminé indépendamment de la requête tandis que le contexte local (e.g., les k premiers documents retournés) est déterminé en fonction de la requête.

Nous considérons la tâche d'extraction de concepts prédéfinis dans plusieurs terminologies comme un problème de vote consistant à fusionner les listes de concepts identifiés. Pour chaque document, nous appliquons une méthode d'extraction de concepts basée sur la recherche approximative des concepts issus d'une terminologie (Zhou *et al.*, 2006b). Étant donné que les concepts ainsi

identifiés ne sont pas pondérés, nous proposons de les pondérer en utilisant un schéma de pondération particulier (e.g., probabiliste (Robertson *et al.*, 1998)) en RI. Par la suite, les listes de concepts sont fusionnées en utilisant plusieurs techniques de fusion de données (Fox et Shaw, 1994) pour obtenir une meilleure représentation conceptuelle du document.

3 Architecture générale de notre approche de RI multi-terminologique

La figure VI.1 présente l'architecture générale de notre approche de RI conceptuelle multi-terminologique. Nous y distinguons deux composantes principales : (1) *indexation multi-terminologique* et (2) *recherche d'information terminologique*. Nous détaillons dans ce qui suit ces deux composantes.

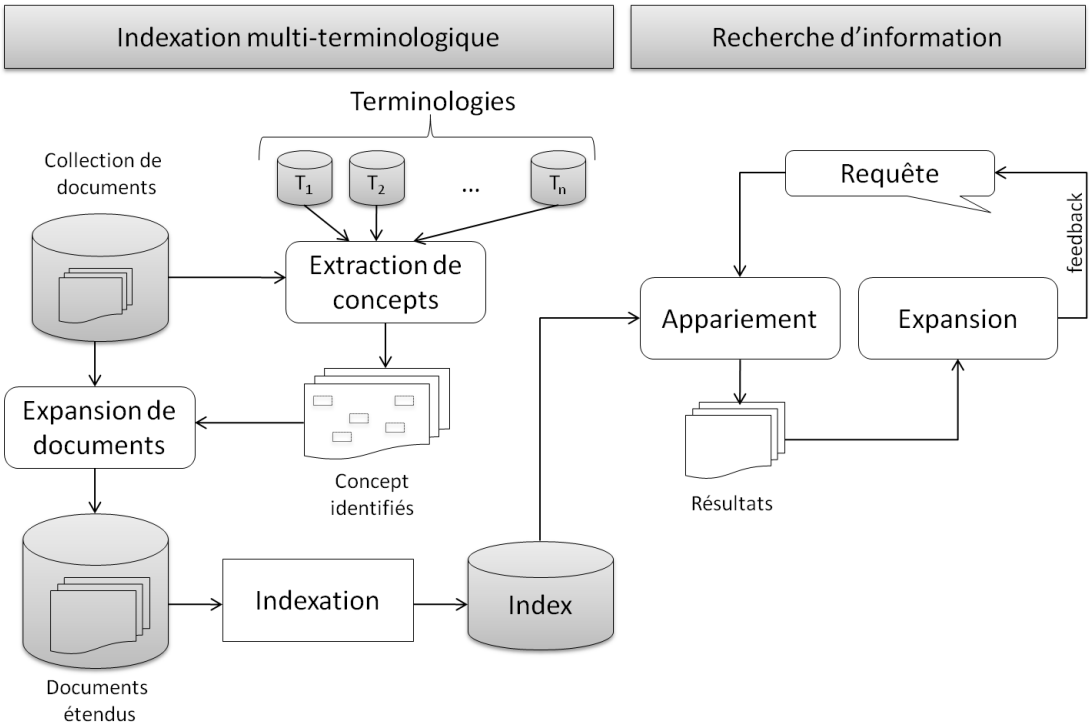


FIGURE VI.1 – Architecture générale du système de RI conceptuelle multi-terminologique

Indexation multi-terminologique

L'indexation multi-terminologique est essentiellement basée sur plusieurs terminologies biomédicales. Étant donnée une collection de documents et un

ensemble de terminologies, l'objectif est d'associer les meilleurs concepts issus des terminologies aux documents pour améliorer leur indexation. Chaque document est donc traité selon les étapes suivantes :

- *Découpage en phrases* : consiste à normaliser le texte, à le transformer en paragraphes qui sont découpés par la suite en phrases en utilisant les délimiteurs (‘.’, ‘;’, ‘?’, ‘!’, ‘\n’, ‘\t’, etc.).
- *Analyse grammaticale* : consiste à décomposer et à rechercher la nature et la fonction des mots (e.g., nom, verbe, adjectif, adverbe, ...).
- *Extraction de concepts* : consiste à localiser les termes candidats qui sont définis dans une terminologie spécifique et à les pondérer en utilisant un schéma de pondération particulier (e.g., TF-IDF). Les termes candidats doivent contenir un nombre maximum de (mots) constituants.
- *Fusion de concepts* : pour chaque terminologie, une liste de concepts est extraite à partir du document. L'idée de la fusion des concepts issus de plusieurs terminologies est d'obtenir une liste de concepts uniques qui sont classés par ordre décroissant en fonction du poids associé.
- *Expansion conceptuelle de documents* : les concepts ainsi identifiés sont intégrés de nouveau dans le document pour mettre en évidence les sujets sémantiques du document. Notre méthode d'expansion conceptuelle de documents s'inspire de la tâche quotidienne des indexeurs humains à la bibliothèque nationale de la santé aux États-Unis qui consiste à affecter les termes MeSH à chaque document dans MEDLINE.

Recherche d'information terminologique

Après avoir généré les structures d'index de la collection de documents étendus ou annotés par des concepts identifiés préalablement, nous pouvons traiter les requêtes de l'utilisateur lors de la recherche d'information. Cette procédure consiste à trouver les documents pertinents en réponse à la requête de l'utilisateur. Ce traitement correspond à une tâche de récupérer, d'extraire et de combiner les sources d'information dans les documents et dans la requête originale pour enrichir les résultats de la requête. La recherche se réalise en deux étapes consécutives : la première consiste à localiser les documents candidats pour la requête tandis que la deuxième permet de sélectionner les meilleurs termes dans les premiers documents retournés pour enrichir la requête via une deuxième recherche. Cette dernière étape a pour but de réordonner et d'optimiser le rappel de la recherche afin d'améliorer les performances de la RI.

4 Indexation multi-terminologique basée sur des techniques de vote

Les étapes de l'indexation multi-terminologique sont les suivantes : (1) extraction mono-terminologique de concepts à partir de documents, (2) fusion de concepts multi-terminologique et (3) indexation multi-terminologique de documents. Dans ce qui suit, nous allons détailler chacune de ces étapes.

4.1 Extraction mono-terminologique de concepts

L'algorithme d'extraction de concepts à partir d'une terminologie permet d'extraire les termes désignant les concepts spécifiques issus d'une terminologie. Afin de mettre en évidence les concepts identifiés, nous proposons de les pondérer en utilisant un schéma de pondération particulier. Le poids du concept c_j issu de la terminologie T_i et identifié à partir du document D, est calculé en se basant sur le schéma de pondération BM25 (Robertson *et al.*, 1994). Formellement, le poids du concept c_j dans le document vis-à-vis du document D est le poids maximal du terme d'entrée v_p du concept c_j dans le document D :

$$w_{ji}^{D} = max_{v_p \in c_j} score(v_p, D) \tag{VI.1}$$

où $score(v_p, D)$ est le poids du terme d'entrée v_p du concept c_j dans le document D, calculé comme suit :

$$score(v_p, D) = \sum_{k=1}^{\ell} tf(t_k) * \frac{log\frac{N-n_k+0.5}{n_k+0.5}}{k_1 * ((1-b) + b * \frac{dl}{avg\_dl}) + tf(t_k)} \tag{VI.2}$$

où

- v_p est la p-ième variante, appelée également *terme d'entrée* du concept c_j,
- t_k est le mot constituant (jeton simple) k du concept c_j dans la terminologie T_i ;
- ℓ est le nombre de mots constituant le concept c_j ;
- $tf(t_k)$ est la fréquence de t_k dans le document D ;
- N_{t_k} est le nombre de documents (concepts) contenant t_k ;
- N est le nombre total de documents (concepts) dans la collection (terminologie) ;
- dl est la longueur du document (concept) ;
- avg_dl est la longueur moyenne de documents (concepts) ;
- k_1, et b sont des paramètres expérimentaux ($k_1 = 1.2, b = 0.75$ par défaut).

Algorithme 5 – Extraction de concepts mono-terminologique

Entrées : Document D, Terminologie T
Sorties : Liste des concepts L

```
1: P ← extraire_paragraphes(D)
2: Pour p ∈ P faire
3:    S ← extraire_phrases(p)
4:    Pour s ∈ S faire
5:       analyse_grammaticale(s)
6:       w ← s.motDebut {1er mot de la phrase}
7:       t ← NULL {Terme trouvé}
8:       Tant que w ≠ NULL faire
9:          {vérifier si le mot courant est le début d'un terme}
10:         Si !debut(w) alors
11:            w ← w.suivant
12:            continue
13:            {chercher un terme commençant par w}
14:            t ← recherche(w, T)
15:            Si t ≠ NULL alors
16:               ponderer(t)
17:               ajouter(L, t)
18:               w ← t.MotFin.suivant
19:            Sinon
20:               w ← w.suivant
21:            Fin Si
22:         Fin Si
23:      Fin Tant que
24:   Fin Pour
25: Fin Pour
26: Retourner L
```

Nous illustrons ce principe d'extraction de concepts basé sur une mono-terminologie par l'algorithme 5. Dans l'algorithme 5, la méthode *analyse_grammaticale* permet de localiser les groupes nominaux et ainsi de réduire le nombre de recherches inutiles. La méthode *recherche*, qui est considéré comme le cœur de cet algorithme, vise à vérifier si un terme donné existe dans le dictionnaire des concepts ou non. Pour cela, nous normalisons tous les mots constituant le terme à rechercher ainsi que ceux constituant l'entrée d'un concept en utilisant l'algorithme de racinisation de (Porter, 1997).

Le tableau VI.1 présente trois enregistrements correspondant à trois concepts MeSH différents. Chaque concept est représenté par un identifiant unique (CUI), une entrée préférée (MH) et les entrées non-préférées (ENTRY). Toutes les entrées sont normalisées et indexées par la position "Début" qui enregistre le début de l'enregistrement et la position "Fin" qui indique la fin

de cet enregistrement dans le dictionnaire. Les entrées ayant la même forme normalisée qu'une des entrées précédentes sont ignorées.

Nous illustrons le fonctionnement de notre algorithme d'extraction par le texte biomédical suivant :

"Sertoli-Leydig cell tumor is a cancer that starts in the female ovaries. The cancer cells produce and release a male sex hormone, which may cause the women to develop male physical characteristics (virilization), including facial hair and a deep voice."

Cette méthode d'extraction de concepts à partir du texte fournit une liste de concepts comme illustré dans le tableau VI.2. La colonne de gauche présente l'annotation du texte original avec des concepts identifiés ; la colonne de droite présente la liste de concepts extraits à partir du texte. Pour activer le calcul du poids des concepts grâce à un schéma de pondération TF-IDF, tous les textes doivent être sauvegardés dans des documents différents avec un format spécifique, par exemple le format de TREC avec la délimitation des éléments textuels par des balises XML ou des nœuds :

DOC : élément qui contient tout le contenu d'un document
DOCNO : identifiant du document
TITLE : titre du document
ABSTRACT : résumé du document

Le tableau VI.3 illustre un document TREC dont l'identifiant est dénoté par la balise *DOCNO*. Étant donnée une collection de documents, cxtractor indexe les documents et génère les structures d'index de la collection. Les résultats de l'extraction de concepts mono-terminologiques issus du MeSH (colonne gauche) et de SNOMED (colonne droite) avec l'affectation de poids à chaque concept identifié sont présentés dans le tableau VI.4.

4.2 Extraction de concepts multi-terminologique

Après avoir extrait les concepts en utilisant chacune des terminologies séparément, notre objectif à ce niveau est de fusionner les concepts candidats pour générer une liste des concepts unique grâce à des associations entre les terminologies via les identifiants uniques dans l'UMLS.

Nous évaluons huit techniques de fusion de données pour agréger les votes de concepts identifiés à partir de chaque document en utilisant plusieurs terminologies. Les techniques de fusion (e.g., $CombMIN, CombMAX, CombSUM$, $ComMNZ$, etc.) ont été d'abord définies dans (Fox et Shaw, 1994) et de-

TABLEAU VI.1 – Exemples de concepts MeSH enregistrés dans un dictionnaire

Concept	**Terme normalisé**	**Début**	**Fin**
***NEWRECORD**			
UI = D008309			
CUI = C0024633			
MH = Mallory-Weiss Syndrome	mallori weiss syndrom	0	129
ENTRY = Mallory Weiss Syndrome			
ENTRY = Syndrome, Mallory-Weiss	syndrom mallori weiss	0	129
***NEWRECORD**			
UI = D008308			
CUI = C0024632			
MH = Mallophaga	mallophaga	142	383
ENTRY = Biting Lice	bite lice	142	383
ENTRY = Chewing Lice	chew lice	142	383
ENTRY = Bovicola	bovicola	142	383
ENTRY = Damalinia	damalinia	142	383
ENTRY = Amblycera	amblycera	142	383
ENTRY = Ischnocera	ischnocera	142	383
ENTRY = Lice, Biting	lice bite	142	383
ENTRY = Lice, Chewing	lice chew	142	383
ENTRY = Mallophagas			
ENTRY = Chewing Lices			
ENTRY = Lices, Chewing			
ENTRY = Biting Lices			
ENTRY = Lices, Biting			
ENTRY = Bovicolas			
ENTRY = Damalinias			
ENTRY = Amblyceras			
ENTRY = Ischnoceras			
...			
***NEWRECORD**			
UI = D000001			
CUI = C0000699			
MH = Calcimycin	calcimycin	3155759	130
ENTRY = Antibiotic A23187	antibiot a23187	3155759	130
ENTRY = A23187	a23187	3155759	130
ENTRY = A-23187	a23187	3155759	130
ENTRY = A 23187	a23187	3155759	130
ENTRY = A23187, Antibiotic	a23187 antibiot	3155759	130

TABLEAU VI.2 – Exemple de l'extraction de concepts mono-terminologique

Annotation	**Ordre|CUI|Terme préféré|Poids**
Sertoli-Leydig cell tumor (C0206723, Sertoli-Leydig Cell Tumor) is a cancer (C0027651, Neoplasms) that starts in the female (C0015780, Female) ovaries (C0029939, Ovary). The cancer (C0027651, Neoplasms) cells (C0007634, Cells) produce and release a male (C0024554, Male) sex hormone (C0036884, Gonadal Steroid Hormones), which may cause the women (C0043210, Women) to develop male (C0024554, Male) physical characteristics (virilization (C0042755, Virilism)), including facial hair (C0018494, Hair) and a deep voice (C0042939, Voice).	0|C0206723|Sertoli-Leydig Cell Tumor|0.00 1|C0027651|Neoplasms|0.00 2|C0015780|Female|0.00 3|C0029939|Ovary|0.00 4|C0007634|Cells|0.00 5|C0024554|Male|0.00 6|C0036884|Gonadal Steroid Hormones|0.00 7|C0043210|Women|0.00 8|C0042755|Virilism|0.00 9|C0018494|Hair|0.00 10|C0042939|Voice|0.00

TABLEAU VI.3 – Un documents biomédical sous le format TREC

<DOC>
<DOCNO>*11096424* **</DOCNO>**
<TITLE>- Prenatal radiation-induced limb defects mediated by Trp53-dependent apoptosis in mice.**</TITLE>**
<ABSTRACT>- We reported previously that in utero radiation-induced apoptosis in the predigital regions of embryonic limb buds was responsible for digital defects in mice. To investigate the possible involvement of the Trp53 gene, the present study was conducted using embryonic C57BL/6J mice with different Trp53 status. Susceptibility to radiation-induced apoptosis in the predigital regions and digital defects depended on both Trp53 status and the radiation dose ; i.e., Trp53 wild-type (Trp53(+/+)) mice appeared to be the most sensitive, Trp53 heterozygous (Trp53(+/-)) mice were intermediate, and Trp53 knockout (Trp53(-/-)) mice were the most resistant. These results indicate that induction of apoptosis and digital defects by prenatal irradiation in the later period of organogenesis are mediated by the Trp53 gene. These findings suggest that the wild-type Trp53 gene may be an intrinsic genetic susceptibility factor that is responsible for certain congenital defects induced by prenatal irradiation. Radiological Sciences, Chiba, Japan. **</ABSTRACT>**
</DOC>

viennent l'objet de plusieurs travaux de recherche dans le domaine de la RI (Lee, 1997; Montague et Aslam, 2001; Wu et McClean, 2006). Ces techniques peuvent se subdiviser en deux catégories : la première est basée sur les rangs des documents retournés (concepts dans le cas de notre travail) et la deuxième est basée sur les poids des documents retournés (ou concepts identifiés).

Soit $R(D, T_i)$ l'ensemble de concepts extraits à partir du document D en utilisant la terminologie T_i, alors la liste de concepts extraits à partir de D en utilisant plusieurs terminologies, notées $T = \{T_1, T_2, ..., T_m\}$, peut être définie comme : $R(D, T) = \cup_{i=1}^{m} R(D, T_i)$, où m est le nombre de terminologies utilisées pour l'extraction de concepts. Chaque concept possède un identifiant unique

Liste de concepts MeSH	**Liste de concepts SNOMED**
<DOCNO> *11096424* 0|C0282505|Limb Buds|1.8030 1|C0162638|Apoptosis|1.7039 2|C0000768|Congenital Abnormalities|1.6676 3|C0017337|Genes|0.9117 4|C0031084|Periodicity|0.9117 5|C0242290|Organogenesis|0.9117 6|C0036397|Science|0.9117 7|C0022341|Japan|0.9117 8|C0015385|Extremities|0.8912 9|C0851346|Radiation|0.7921 10|C0026809|Mice|0.7268	**<DOCNO>** *11096424* 0|C1314939|Involved|0.9117 1|C0332182|Periodic|0.9117 2|C0205147|Regional|0.8912 3|C0012655|Diathesis, NOS|0.8912 4|C0162638|Apoptosis|0.7921

TABLEAU VI.4 – Listes de concepts extraits à partir des documents

dans l'UMLS grâce auquel ils peuvent être fusionnés ensemble. Nous utilisons les notations suivantes pour déterminer le score d'un concept candidat dans le document D :

- $score(c_j, D)$ est le score combiné du concept c_j dans le document D. Ce score est obtenu par une combinaison de scores du concept en utilisant plusieurs terminologies ;
- $\|R(D, T_i)\|$ est le nombre total de concepts extraits à partir du document D en utilisant la terminologie T_i ;
- r_{ji}^D est le rang du concept c_j identifié comme concept candidat pour le document D dans la liste des concepts extraits en utilisant la terminologie T_i ;
- w_{ji}^D est le score du concept c_j dans le document D en utilisant la terminologie T_i ;
- $\|\{c_j \in R(D, T)\}\|$ est le nombre de terminologies contenant le concepts c_j.

Les techniques de vote auxquelles nous nous intéressons peuvent être regroupées en deux catégories en fonction de la source d'évidence utilisée, à savoir le **rang** et le **poids** des concepts candidats.

- **Fusion basée sur les rangs des concepts :** Le principe est de considérer les rangs des concepts comme un critère pour les ordonner : plus le rang du concept est élevé (c-à-d le concept est renvoyé vers la fin de la liste), moins il est significatif pour le document. Nous évaluons ici deux techniques de vote : la technique *CombRank* est utilisée pour calculer la somme des rangs du concept figurant dans plusieurs listes de concepts candidats. Formellement, le score du concept c_j vis-à-vis du document D

est donné par :

$$score(c_j, D) = \sum_{i=1}^{n} (\|R(D, T_i)\| - r_{ji}^D) \qquad [CombRank] \qquad \text{(VI.3)}$$

La technique $CombRCP$ calcule la somme des rangs inversés du concept dans plusieurs listes. Finalement, les concepts sont fusionnés dans une liste de concepts uniques selon leur nouveau rang en ordre croissant. Formellement :

$$score(c_j, D) = \sum_{i=1}^{n} 1/r_{ji}^D \qquad [CombRCP] \qquad \text{(VI.4)}$$

- **Fusion basée sur l'importance des concepts :** Le principe est de tenir compte de l'importance des concepts comme un critère pour les ordonner : plus le poids du concept est élevé, plus il est important pour le document. Nous évaluons ici six techniques de vote : la technique $CombSUM$ fusionne les concepts candidats selon la somme des poids de concepts identifiés. Les techniques $CombMIN, CombMAX, CombMED$ fonctionnent de la même manière en utilisant les fonctions d'agrégation min, max et $median$ respectivement. Formellement, le score du concept c_j vis-à-vis du document D , en appliquant ces fonctions d'agrégation, est donné par :

$$\begin{cases} score(c_j, D) = \sum_{i=1}^{n} w_{ji}^D & [CombSUM] \\ score(c_j, D) = \min\{w_{ji}^D, i = \overline{1..n}\} & [CombMIN] \\ score(c_j, D) = \max\{w_{ji}^D, i = \overline{1..n}\} & [CombMAX] \\ score(c_j, D) = median\{w_{ji}^D, i = \overline{1..n}\} & [CombMED] \end{cases} \qquad \text{(VI.5)}$$

où n est le nombre de concepts candidats identifiés à partir du document.

Les deux dernières techniques $CombANZ$ et $CombMNZ$ sont les variantes de $CombSUM$ où le nombre de terminologies est pris en compte. Dans $CombMNZ$, plus le nombre de terminologies d'où le concept est extrait est élevé, plus le concept candidat est important. La technique $CombANZ$ favorise les concepts issus d'une minorité de terminologies. Formellement, le score du concept c_j vis-à-vis du document D , en appliquant ces deux dernières fonctions, est donné par :

$$\begin{cases} score(c_j, D) = \sum_{i=1}^{n} w_{ji}^D \times \|\{c_j \in R(D, T)\}\| & [CombMNZ] \\ score(c_j, D) = \sum_{i=1}^{n} w_{ji}^D \div \|\{c_j \in R(D, T)\}\| & [CombANZ] \end{cases} \qquad \text{(VI.6)}$$

Reprenons le premier document de l'exemple dans le tableau VI.3, l'extraction de concepts multi-terminologique en utilisant les deux techniques

CombSUM et *CombRank* à partir de deux terminologies MeSH et SNOMED résulte en deux listes de concepts différentes (*cf.* le tableau VI.5). Comme nous le voyons, les rangs ainsi que les poids de chaque concept dans la liste combiné sont modifiés en fonction de chaque technique de vote. Par exemple, pour la technique *CombSUM*, le concept "Apoptosis" (C0162638) est classé en premier car son poids est important dans les deux listes de concepts identifiés en utilisant chaque terminologie séparément au début.

TABLEAU VI.5 – Résultats de l'extraction multi-terminologique de concepts

CombSUM	**CombRank**
<DOCNO> *11096424*	**<DOCNO>** *11096424*
0|C0162638|Apoptosis|2.4960	0|C0282505|Limb Buds|1.0000
1|C0282505|Limb Buds|1.8030	1|C1314939|Involved|1.0000
2|C0000768|Congenital Abnormalities|1.67	2|C0332182|Periodic|0.9800
3|C0017337|Genes|0.9117	3|C0162638|Apoptosis|0.9721
4|C0022341|Japan|0.9117	4|C0205147|Regional|0.9600
5|C0031084|Periodicity|0.9117	5|C0000768|Congenital Abnormalities|0.94
6|C0036397|Science|0.9117	6|C0012655|Diathesis, NOS|0.9400
7|C0242290|Organogenesis|0.9117	7|C0017337|Genes|0.9200
8|C0332182|Periodic|0.9117	8|C0031084|Periodicity|0.9000
9|C1314939|Involved|0.9117	9|C0242290|Organogenesis|0.8800
10|C0012655|Diathesis, NOS|0.8912	10|C0036397|Science|0.8600
11|C0015385|Extremities|0.8912	11|C0022341|Japan|0.8400
12|C0205147|Regional|0.8912	12|C0015385|Extremities|0.8200
13|C0851346|Radiation|0.7921	13|C0851346|Radiation|0.8000
14|C0026809|Mice|0.7268	14|C0026809|Mice|0.7800

5 Appariement multi-terminologique basé sur la combinaison des contextes document et requête

Nous présentons dans cette section deux principales utilisations des concepts extraits par notre méthode d'extraction décrite dans la section précédente, à savoir (1) *l'expansion conceptuelle de documents* et (2) *la combinaison de l'expansion conceptuelle de documents et de la reformulation de requêtes par la méthode PRF*.

5.1 Expansion conceptuelle de documents

Comme nous avons mentionné dans la section 4 du chapitre III, l'objectif principal de l'extraction de concepts consiste à identifier les meilleurs termes qui désignent les concepts biomédicaux. Ces derniers peuvent être utilisés pour étendre le contenu textuel du document en mettant en évidence les sujets sémantiques ou concepts du document. De plus, en ajoutant les termes préférés dans l'index, les résultats de la recherche peuvent être améliorés notamment pour les requêtes contenant les termes préférés désignant les concepts biomédicaux qui ne sont pas observés dans les documents originaux. La figure VI.2 illustre un exemple de défaut d'appariement (term mismatch) entre les documents et la requête de l'utilisateur. Dans cet exemple, le thème principal de la requête est "*laser therapy*". Le document $D1$ contient les mots "laser ablation" et le document $D2$ ne contient que le mot "laser". Bien que le premier document soit plus pertinent que le deuxième, la fonction d'appariement basée sur un modèle TF_IDF classique renvoie $D2$ en premier. Il s'agit dans ce cas d'un défaut d'appariement document-requête parce que le document $D1$ doit être retourné en premier.

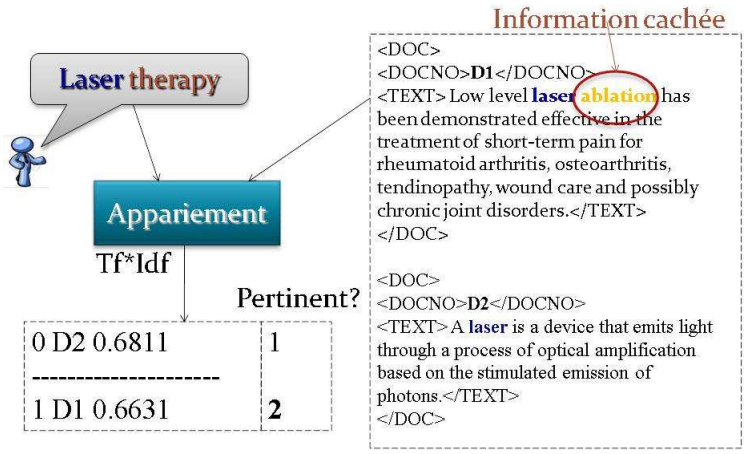


FIGURE VI.2 – Un exemple de défaut d'appariement document-requête

Pour pallier au défaut d'appariement dû à la synonymie, nous proposons de normaliser le contenu du document en ajoutant des termes préférés désignant les concepts biomédicaux. Ce mécanisme, appelé *expansion documentaire conceptuelle* ou *expansion conceptuelle de documents*, permettrait de mieux récupérer les documents ayant le(s) même(s) sujet(s) sémantique(s) que la requête de l'utilisateur lorsque les documents originaux ne contiennent pas de termes préférés utilisés dans la requête (*cf.* l'illustration dans la figure VI.3). Le document étendu, dénoté D^e, peut être considéré comme l'union de la représentation du document original D et la représentation conceptuelle du document via la liste

de concepts extraits, dénotée L_C :

$$Index(D^e) = Index(D) \cup Index(L_C) \tag{VI.7}$$

où D est l'ensemble de mots-clés du document original et L_C est l'ensemble des mots-clés issus de l'ensemble de termes préférés désignant les concepts identifiés, sachant que les mots-clés issus de L_C peuvent apparaître dans D.

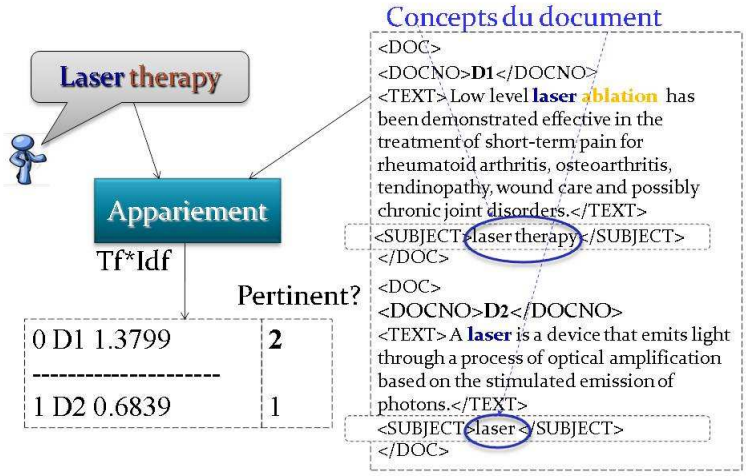


FIGURE VI.3 – Expansion documentaire en utilisation les termes préférés

5.2 Combinaison de l'expansion documentaire et la reformulation de la requête par la méthode PRF

Étant donné un ensemble de documents et une requête de l'utilisateur, nous définissons les notations suivantes :

- $DS = \{D_1, D_2, ..., D_n\}$ l'ensemble de documents de la collection,
- $DS^{(E)} = \{D_1^{(E)}, D_2^{(E)}, ..., D_n^{(E)}\}$ l'ensemble de documents étendus par les concepts identifiés à partir des documents originaux,
- $QS = \{D_1^{(Q)}, D_2^{(Q)}, ..., D_n^{(Q)}\}$ l'ensemble de documents pertinents vis-à-vis des requêtes testées,
- $QS^{(E)} = \{D_1^{(E,Q)}, D_2^{(E,Q)}, ..., D_n^{(E,Q)}\}$ l'ensemble de documents étendus pertinents vis-à-vis des requêtes testées.

On espère augmenter le rappel de la RI en récupérant plus de documents pertinents en réponse à un ensemble de requêtes. Cet ensemble peut être dénoté par : $DS^{(E)} \bigcap QS$. De manière générale, l'ensemble de documents pertinents

retournés par le système de RI ($DS \bigcap QS$) devrait être un sous-ensemble des documents étendus pertinents retournés par le même système de RI :

$$\{D, D \in DS \wedge D \in QS\} \subseteq \{D, D \in DS^{(E)} \wedge D \in QS\} \tag{VI.8}$$

L'expansion de la requête permet de retrouver plus de documents liés à la requête ($DS \bigcap QS^{(E)}$) grâce aux nouveaux termes extraits à partir des premiers documents retournés. Formellement :

$$\{D, D \in DS \wedge D \in QS\} \subseteq \{D, D \in DS \wedge D \in QS^{(E)}\} \tag{VI.9}$$

Nous proposons de combiner l'expansion documentaire et l'expansion de la requête pour résoudre le défaut d'appariement document-requête en élargissant l'espace de documents pertinents retournés (cf. figure VI.4). Formellement :

$$\begin{cases} \{D, D \in DS \wedge D \in QS\} \subseteq \{D, D \in DS^{(E)} \wedge D \in QS\} \subseteq \{D : D \in DS^{(E)} \wedge D \in QS^{(E)}\} \\ \{D, D \in DS \wedge D \in QS\} \subseteq \{D, D \in DS \wedge D \in QS^{(E)}\} \subseteq \{D, D \in DS^{(E)} \wedge QS^{(E)}\} \end{cases} \tag{VI.10}$$

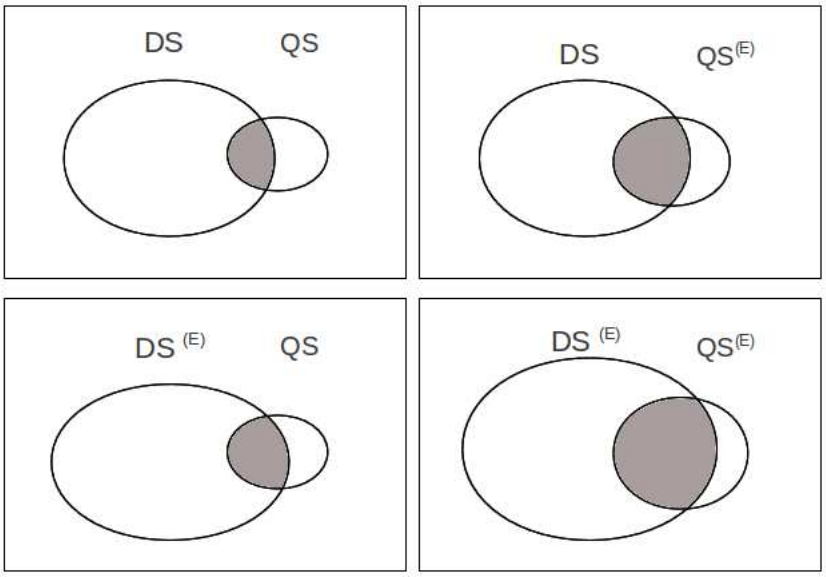


FIGURE VI.4 – Expansion documentaire *vs.* expansion de la requête

6 Évaluation expérimentale

Dans cette section, nous décrivons les différentes expérimentations réalisées dans le but de valider notre approche de RI basée sur plusieurs terminologies biomédicales. L'évaluation porte particulièrement sur les facteurs de base liés à l'expansion conceptuelle de documents et l'expansion de la requête, l'impact des modèles de vote pour fusionner les concepts issus de plusieurs terminologies sur les performances de la recherche.

6.1 Objectifs d'évaluation

Les principaux objectifs de nos expérimentations sont de :

- premièrement, montrer l'intérêt de la combinaison de l'expansion documentaire conceptuelle via les concepts biomédicaux extraits à partir de chaque document et l'expansion de la requête par des termes extraits à partir des premiers documents étendus retournés ;

- deuxièmement, étudier l'utilité de notre approche de RI conceptuelle basée sur l'utilisation d'une terminologie ou de plusieurs terminologies comparativement à l'approche de RI classique sans l'utilisation de terminologies.

Pour cela, nous considérons deux bases de référence d'évaluation suivantes :

- *Baseline_Classic* : concerne l'approche de RI classique basé sur les modèles de pondération de l'état de l'art,
- *Baseline_QE* : concerne l'approche de RI qui utilise les modèles d'expansion de la requête pour améliorer les performances de la RI.

6.2 Cadre d'évaluation

Nous définissions dans ce qui suit le cadre d'évaluation portant sur les collections de documents biomédicaux et les mesures d'évaluation des performances du système de RI.

TABLEAU VI.6 – Un exemple de citation ou document dans MEDLINE

PMID- 11096424
TI - Prenatal radiation-induced limb defects mediated by Trp53-dependent apoptosis in mice.
PG - 673-9
AB - We reported previously that in utero radiation-induced apoptosis in the predigital regions of embryonic limb buds was responsible for digital defects in mice. To investigate the possible involvement of the Trp53 gene, the present study was conducted using embryonic C57BL/6J mice with different Trp53 status. **...** These findings suggest that the wild-type Trp53 gene may be an intrinsic genetic susceptibility factor that is responsible for certain congenital defects induced by prenatal irradiation.
...
MH - Abnormalities, Radiation-Induced/*genetics/pathology
MH - Animals
MH - Apoptosis/*radiation effects
MH - Dose-Response Relationship, Radiation
...
MH - Pregnancy
MH - *Prenatal Exposure Delayed Effects
MH - Radiation Tolerance/genetics
MH - Tumor Suppressor Protein p53/deficiency/*genetics/metabolism

6.2.1 Collections de TREC Genomics

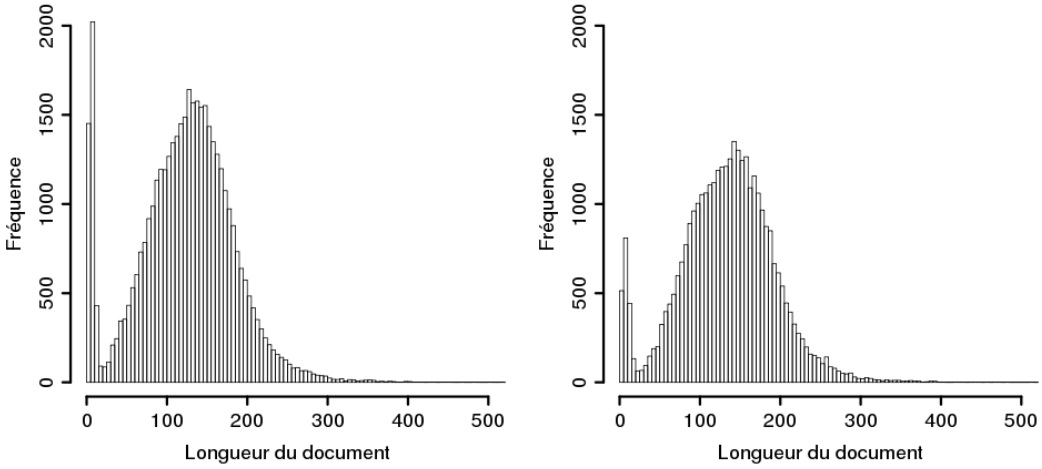
Un ensemble d'environ 4.6 millions de citations a été extrait à partir de la base MEDLINE pour créer des collections d'évaluation. Chaque document de la collection se compose principalement d'un titre et/ou un résumé qui constituent l'essentiel du contenu textuel du document. Le tableau VI.6 montre un exemple d'une citation de MEDLINE sous le format défini par la NLM [1]. Nous nous intéressons particulièrement aux trois champs suivants :

- **PMID** : représente l'identifiant ou le numéro unique d'une citation stockée dans MEDLINE. Chaque citation correspond donc à un article de journal qui a été sélectionné pour être indexé dans MEDLINE.
- **TI** : représente le titre de l'article original.
- **AB** : représente le résumé de l'article original.

Selon la méthode d'évaluation mise en œuvre par TREC en adaptant la méthodologie d'évaluation du paradigme de CRANFIELD, le jugement de la pertinence des documents résultats de chaque requête ne peut pas être effectué sur tous les documents de la collection. En effet, un ensemble de documents obtenus par la fusion des premiers documents résultats soumis par chaque groupe

1. `http://www.ncbi.nlm.nih.gov/`

est retenu pour le jugement. D'autre part, l'extraction de concepts à partir d'une collection volumineuse comme TREC Genomics, présente un coût élevé pour accomplir la tâche de RI conceptuelle. Par exemple, les indexeurs humains ont mis une dizaine voire une vingtaine d'années ou plus pour extraire et assigner les termes MeSH à chaque citation dans MEDLINE. Il est pratiquement impossible de mener une série d'expérimentations avec les moyens existants dans un environnement du laboratoire. Par conséquent, nous adoptons la démarche d'évaluation proposée par les auteurs dans (Zhou *et al.*, 2006c,a, 2007b), notre système prototype ne traite que les documents qui ont été jugés, c'est à dire 48.753 citations dans TREC Genomics 2004 et 41.018 citations dans TREC Genomics 2005. Nous utilisons principalement le texte du titre et/ou du résumé de chaque citation sans utiliser les termes MeSH manuellement ou semi-manuellement ajoutés par les indexeurs humains.



(a) TREC Genomics 2004 (b) TREC Genomics 2005

FIGURE VI.5 – Distribution de la longueur du document dans les collections TREC Genomics

La figure VI.5 présente la variation de la longueur du document dans TREC Genomics 2004 et 2005. La longueur moyenne d'un document dans TREC Genomics 2004 est de 122 tokens (jetons) et celle de TREC Genomics 2005 est de 134 tokens (jetons). D'après les statistiques sur les collections TREC Genomics 2004 et 2005, on peut considérer que ces deux collections ont à peu près les mêmes caractéristiques, ce qui nous permet par exemple d'entraîner une ou des fonctions d'appariement (ranking functions) sur l'une des deux collections et de les tester sur l'autre.

Les requêtes pour la tâche de RI *ad hoc* dans TREC Genomics ont été créées à partir de vrais besoins d'information des biologistes et médecins et ont été modifiées le moins possible pour générer un nombre raisonable de documents à

TREC Genomics / Caractéristique	2004	2005
Nombre de requêtes	50	49
Longueur moyenne	17	9
Documents jugés par requête	975	816
Documents pertinents par requête	166	94

TABLEAU VI.7 – Quelques statistiques sur les requêtes de TREC Genomics

juger pour chaque requête (Hersh *et al.*, 2004, 2005). Le tableau VI.7 fournit quelques statistiques sur les requêtes de TREC Genomics. On peut constater que les requêtes dans TREC Genomics 2004 sont en général plus longues que celles dans TREC Genomics 2005, ce qui traduit une plus grande spécificité à propos du besoin d'information de l'utilisateur dans TREC Genomics 2004. Par conséquent, le nombre moyen de documents pertinents pour chaque requête dans TREC Genomics 2004 est plus élevé que celui dans TREC Genomics 2005. Le nombre de documents jugés ainsi que le nombre de documents pertinents pour chaque requête ont une influence importante sur l'évaluation des performances d'un SRI.

6.2.2 Protocole d'évaluation

Pour atteindre les objectifs définis précédemment, nous menons une série d'expérimentions en utilisant l'outil IRToolkit [2] qui intègre plusieurs moteurs de recherche, y compris Terrier (Ounis *et al.*, 2005), Lemur [3] et Lucene [4]. L'objectif principal de IRToolkit est de faciliter les expérimentations en RI en offrant une possibilité de lancer l'indexation, la recherche d'information et l'évaluation en sélectionnant un moteur de recherche spécifique dans une plateforme homogène.

Nous définissons une série de scénarios d'expérimentations comme suit :

– le premier concerne l'**approche de RI classique** sans utiliser ni les sources de connaissances du domaine, ni les termes MeSH qui sont affectés manuellement par des indexeurs humains : lors de l'indexation, les mots vides sont enlevés du document, les mots ou jetons en général sont normalisés par la racinisation de Porter (Porter, 1997). Au moment de la recherche, les documents sont pondérés en utilisant un des trois modèles de pondération décrits dans les paragraphes qui suivent,

2. `http://sourceforge.net/projects/irtoolkit/files/`
3. `http://www.lemurproject.org/`
4. `http://lucene.apache.org/`

– la deuxième série d'expérimentations concerne une **approche de RI mono-terminologique** basée sur trois différentes terminologies : MeSH, SNOMED et GO. Ici, nous avons adopté MaxMatcher (MM) (Zhou *et al.*, 2006b), qui est essentiellement un extracteur de termes/concepts à partir de documents biomédicaux. Nous l'avons modifié afin de pouvoir mesurer l'importance ou le degrée de description des concepts extraits du document (cf. la formule VI.2, section 4.1). Cette modification a été implémentée dans le logiciel **cxtractor**[5]. Les termes extraits en utilisant chacune des terminologies sont utilisés pour l'expansion du document (DE). Enfin, nous appliquons l'expansion de la requête basée sur la méthode de reformulation de requêtes PRF sur la collection de documents étendus. Plus précisément, cette technique est abordée ici comme la combinaison de contextes documentaires et terminologiques via l'**expansion conceptuelle de documents** et de l'**expansion de requêtes** pour optimiser les performances de la RI,

– la troisième série d'expérimentations concerne notre approche de RI basée sur l'utilisation de plusieurs terminologies (MeSH, SNOMED, et GO), appelée approche de **RI multi-terminologique**. Comparativement à la deuxième expérimentation, les concepts prédéfinis dans plusieurs terminologies et extraits à partir de chaque document sont combinés en utilisant les différents modèles de vote pour obtenir une liste de concepts uniques. Cette liste de concepts va être utilisée pour l'expansion documentaire. Enfin, comme dans le scénario précédent, nous combinons QE et DE.

Du fait que nos approches de RI sont basées sur l'utilisation des terminologies via un processus d'indexation et de recherche d'information qui intègre l'expansion documentaire et l'expansion de la requête, nous avons besoin d'apprendre une fonction d'appariement intégrant trois paramètres principaux : (1) le nombre de termes extraits de (2) k premiers documents pour l'expansion de la requête et (3) le nombre de termes désignant des concepts du domaine pour l'expansion conceptuelle documentaire. Nous entraînons ces paramètres sur la collection TREC Genomics 2004 et appliquons les meilleures configurations sur la collection TREC Genomics 2005 pour démontrer l'efficacité de nos approches de RI basées sur les terminologies biomédicales.

Pour cela, nous expérimentons les trois différents modèles de pondération de termes et les trois différents modèles de reformulation de requêtes implémentés dans Terrier. Ces trois modèles représentent donc notre base de référence d'évaluation classique (Baseline_Classic).

5. `https://sourceforge.net/projects/cxtractor/files/`

6.2.3 Schémas d'appariement document-requête

Le modèle BM25 : dans ce modèle de pondération probabiliste BM25 Robertson *et al.* (1994), le poids du document D vis-à-vis de la requête Q est donné par :

$$RSV(D,Q) = \sum_{t \in Q} w^{(1)} * \frac{(k_1+1)*tfn}{K+tfn} * \frac{(k_3+1)*qtf}{k_3+qtf} \quad \text{(VI.11)}$$

où

- $w^{(1)}$ est la fréquence inverse du document calculée comme suit :

$$w^{(1)} = log_2 \frac{N - N_t + 0.5}{N_t + 0.5} \quad \text{(VI.12)}$$

- tfn est la fréquence normalisée du terme t du document D. Formellement :

$$tfn = \frac{tf_w}{(1+b) + b * \frac{dl}{avg\_dl}}, (0 \leq b \leq 1) \quad \text{(VI.13)}$$

- qtf est la fréquence du terme t dans la requête ; $K = k_1 * (1 - b * (1 - dl/avg\_dl))$

Le modèle In_expB2 : dans ce modèle, les termes sont pondérés par la fréquence documentaire inverse estimée par la probabilité de Bernoulli (dénotée In_expB2 [6]) et par la fréquence normalisée des termes dans le document (Amati, 2003). Formellement, le poids du document D vis-à-vis de la requête Q est donné par :

$$RSV(D,Q) = \sum_{t \in Q} \frac{qtf \times (tf+1) \times tfn_2}{N_t \times (tfn_2+1) \times ln2} \times \log_2 \frac{N+1}{N \times (1 - e^{\frac{-tf}{N}}) + 0.5} \quad \text{(VI.14)}$$

où

- t est le terme de la requête qui apparaît dans le document D,
- N_t est la fréquence de documents, c-à-d le nombre de document contenant t,
- N est la taille de la collection, c-à-d le nombre total de documents de la collection,
- qtf est la fréquence du terme t dans la requête,
- tf est la fréquence du terme t dans le document,
- tfn_2 est la fréquence normalisée du terme t dans le document, calculée par :

$$tfn_2 = \frac{tf}{ln2} \times log_2 \left[1 + c \times \frac{avg\_dl}{dl} \right] \quad \text{(VI.15)}$$

6. **In**verse **Exp**ected document frequencey - **B**ernoulli after-effect

où
- dl est la longueur du document,
- avg_dl est la longueur moyenne des documents
- c est un paramètre du modèle (c=1.0 par défaut)

Le modèle LGD : dans ce modèle, les termes sont pondérés en utilisant la loi log-logistique (connue aussi comme la distribution de Fisk en économie) (Clinchant et Gaussier, 2010). Formellement, le poids du document D vis-à-vis de la requête Q est donné par :

$$RSV(D,Q) = \sum_{t \in Q} qtf \times \left[log_2(\frac{N_t}{N} + tf_n) - log_2(\frac{N_t}{N}) \right] \quad \text{(VI.16)}$$

où
- t est le terme dans le document D,
- N_t est la fréquence de documents (le nombre de documents contenant le terme t),
- N est la taille de la collection,
- qtf est la fréquence du terme de la requête,
- tf_n est la fréquence normalisée du terme t dans le document, calculée par :

$$tf_n = tf \times log_2(1 + c \times \frac{avg\_dl}{dl}) \quad \text{(VI.17)}$$

où
- avg_dl est la longueur moyenne de documents dans la collection (nombre de jetons en moyenne d'un document),
- dl est la longueur du document et c est un paramètre du modèle.

6.2.4 Modèles de reformulation de la requête par la méthode PRF

La reformulation ou l'expansion de la requête (QE) vise à enrichir la requête originale de l'utilisateur afin de récupérer plus de documents pertinents ou de modifier les rangs des documents pour améliorer les performances de la RI. Deux catégories de QE peuvent être distinguées : l'une est basée sur un **contexte global** et l'autre est basée sur un **contexte local** de la requête.

Le contexte global de la requête est déterminé par un ensemble intégral de données (e.g., documents) ou sources de connaissance (dictionnaire, thésaurus, ontologies, etc.). Le contexte local de la requête est souvent défini par les premiers documents retournés pour la requête qui sont utilisés par la suite pour sélectionner les meilleures termes à ajouter dans la nouvelle requête. Cette méthode est souvent mentionnée comme une méthode de reformulation *pseudo-relevance feedback* (PRF) (*cf.* la section 2.4 du chapitre II). Dans ce processus

de reformulation, les termes dans les premiers documents sont pondérés par un modèle de pondération issu de la plateforme DFR (Amati, 2003).

Dans nos expérimentations, nous utilisons la technique de reformulation de requêtes basée sur la méthode PRF pour sélectionner les meilleurs termes dans les meilleurs documents retournés par le système. Plus spécifiquement, nous adoptons les statistiques de Bose-Einstein et Kullback-Leibler (Amati, 2003) pour sélectionner les meilleurs termes qui sont susceptibles d'être pertinents pour générer la nouvelle requête Q^e obtenue par l'expansion de la requête originale Q. Plus spécifiquement, nous évaluons les performances de la RI en utilisant les trois différents modèles de reformulation PRF, à savoir les modèles Bo1, Bo2 et KL (*cf.* la section 2.4.2 du chapitre II).

6.2.5 Mesures d'évaluation des performances de la RI

Dans ce cadre d'évaluation, nous avons utilisé trois mesures d'évaluation qui sont définies dans le cadre de la campagnes d'évaluation TREC :

- la **précision à** X premiers documents (dénotée $P@X$), est donc la proportion de documents pertinents par rapport aux X premiers documents renvoyés par le SRI. La précision à X mesure la satisfaction de l'utilisateur concernant les X premiers documents pertinents. Dans notre cas, nous retenons les précisions pour les 10 et 20 premiers documents retournés, dénotées respectivement $P@10$, $P@20$;

- la **précision moyenne** (Mean Average Precision, dénotée MAP) correspond à la précision moyenne calculée sur l'ensemble des documents pertinents retournés. La MAP mesure la capacité du modèle d'appariement ou d'un SRI à pouvoir sélectionner les documents pertinents, en réponse à un ensemble de requêtes ;

- le **rappel** mesure la capacité du système de RI à renvoyer les documents pertinents. Le rappel correspond à la proportion de documents pertinents renvoyés par le système de RI par rapport à l'ensemble de documents pertinents possibles.

Toutes ces mesures, obtenues sur un ensemble de requêtes, sont générées par l'outil *trec_eval*[7] qui est définitivement utilisé par la communauté TREC.

7. `http://trec.nist.gov/trec_eval/`

6.3 Résultats expérimentaux

Les résultats présentés dans ce qui suit sont liés aux trois séries d'expérimentations décrites dans la section 6.2.2. Tout d'abord, nous comparons les performances des résultats de recherche obtenus par des modèles d'expansion de la requête aux résultats obtenus par les trois modèles de pondération sans l'expansion documentaire ni l'expansion de la requête. Cette étude est considérée comme une étape d'entraînement pour retenir les meilleurs paramètres du système de RI. Elle est donc réalisée sur la collection TREC Genomics 2004. Nous étudions ensuite l'efficacité de notre approche de RI conceptuelle basée sur une mono-terminologie à l'aide de l'expansion documentaire (DE) sur la base de trois différentes terminologies sur une collection de test qui est en l'occurrence la collection TREC Genomics 2005. Par la suite, nous évaluons l'efficacité de la combinaison de l'expansion documentaire (DE) mono-terminologique et l'expansion de la requête (QE) basée des mesures statistiques de la sous-collection (méthode de (Rocchio, 1971)). Enfin, nous démontrons l'utilité d'utiliser des techniques de vote pour combiner des concepts issus de plusieurs terminologies afin de fournir une liste cohérente des concepts représentant la sémantique du document et nous évaluons l'impact de l'indexation et de recherche d'information multi-terminologique.

6.3.1 Entraînement des modèles de pondération et modèles de reformulation de requêtes

L'objectif de cette étape est donc de montrer les performances du **modèle de pondération** de termes et du modèle d'**expansion de requêtes** sur la collection de documents d'entraînement. Chaque modèle de pondération est combiné avec chacun des trois modèles d'expansion de requêtes pour lancer la recherche des documents en réponse à un ensemble de requêtes. Par la suite, nous retenons la meilleure configuration, notamment le modèle de pondération, le modèle d'expansion de requêtes ainsi que les meilleures valeurs des paramètres qui permettent d'optimiser la MAP pour établir notre base de référence d'évaluation la plus forte (strong baseline).

La figure VI.6 montre les résultats MAP obtenus sur la collection TREC Genomics 2004. Ici, nous avons faire varier les paramètres comme le *nombre de termes* sélectionnés (dénoté #terms) qui se trouvent dans un ensemble de k premiers documents retournés (dénoté #docs) pour entraîner le modèle de reformulation de requêtes. Nous pouvons constater que, parmi les trois modèles de pondération, le modèle In_expB2 (sans QE) résulte une valeur MAP de **0.4117** qui est légèrement meilleure que les deux autres modèles de pondération (0.3997 et 0.4018 pour le modèle BM25 et LGD respectivement). En utilisant l'expansion de requêtes sur toutes les requêtes, la plupart des modèles QE

dépassent chacun des modèles de pondération de termes (Baseline_Classic). Par exemple, la combinaison du modèle de pondération *LGD* et le modèle Bo1 pour QE (dénotée *LGD_Bo*1) donne la meilleure valeur MAP de **0.4637** en utilisant 40 termes extraits à partir de 15 premiers documents retournés, ce qui donne une amélioration jusqu'à +15.41% par rapport à la baseline 1 qui utilise le modèle de pondération LGD. La combinaison du modèle de LGD avec le modèle Bo2 (dénotée *LGD_Bo*2) génère une valeur MAP de 0.4464 ($\#terme = 20, \#docs = 10$) dans le meilleur des cas et une valeur MAP de 0.3937 (#termes = 5, #docs = 50) dans le pire des cas. C'est probablement parce que les 5 premiers termes extraits à partir de 50 premiers documents ne sont pas pertinents vis-à-vis de la requête.

La figure VI.7 illustre la distribution, par une estimation par noyau, d'un échantillon de 300 valeurs MAP obtenues par chaque modèle de pondération en combinaison avec chacun des trois modèles d'expansion de requêtes et vice-versa. La distribution de la densité par noyau est une méthode non-paramétrique visant à estimer la probabilité d'une fonction de densité d'une variable aléatoire (Sheather et Jones, 1991). Le modèle **LGD** dont la valeur moyenne de la MAP est de **0.4411** dépasse les deux autres modèles de pondération de termes. En ce qui concerne les modèles QE, bien que les trois modèles QE aient une performance très compétitive en termes de MAP, le modèle **Bo1** dont la moyenne de la MAP est de **0.4389** donne une meilleure performance que les deux autres modèles QE. Par conséquent, nous choisissons le modèle de pondération **LGD** en combinaison avec le modèle d'expansion **Bo1** comme notre base de référence d'évaluation en utilisant **40** termes extraits à partir des **15** premiers documents. Dans la section suivante, nous allons évaluer nos approches de RI basées sur une ou plusieurs terminologies biomédicales sur la collection TREC Genomics 2005.

6.3.2 Évaluation de l'efficacité de l'indexation mono-terminologique

Dans cette sous-section, nous considérons l'expansion des documents originaux, qui sont constitués des titres et/ou des résumés, par des termes préférés désignant les concepts extraits à l'aide de chacune des terminologies suivantes : MeSH, SNOMED, et GO. Notre méthode d'expansion conceptuelle documentaire est inspirée par le travail des indexeurs humains fournissant une douzaine de termes MeSH dans les citations MEDLINE (Névéol *et al.*, 2009). Dans cette expérimentation, nous faisons varier le nombre de termes préférés (de 5 à 50 avec un pas de 5 termes à chaque itération) définis dans chaque terminologie séparément pour enrichir le contenu du document. Les trois histogrammes dans la figure VI.8 révèlent la distribution des nombres de concepts MeSH, SNOMED, GO qui sont extraits à partir de chaque document dans l'ensemble de documents utilisés pour tester notre approche. Nous pouvons voir que la

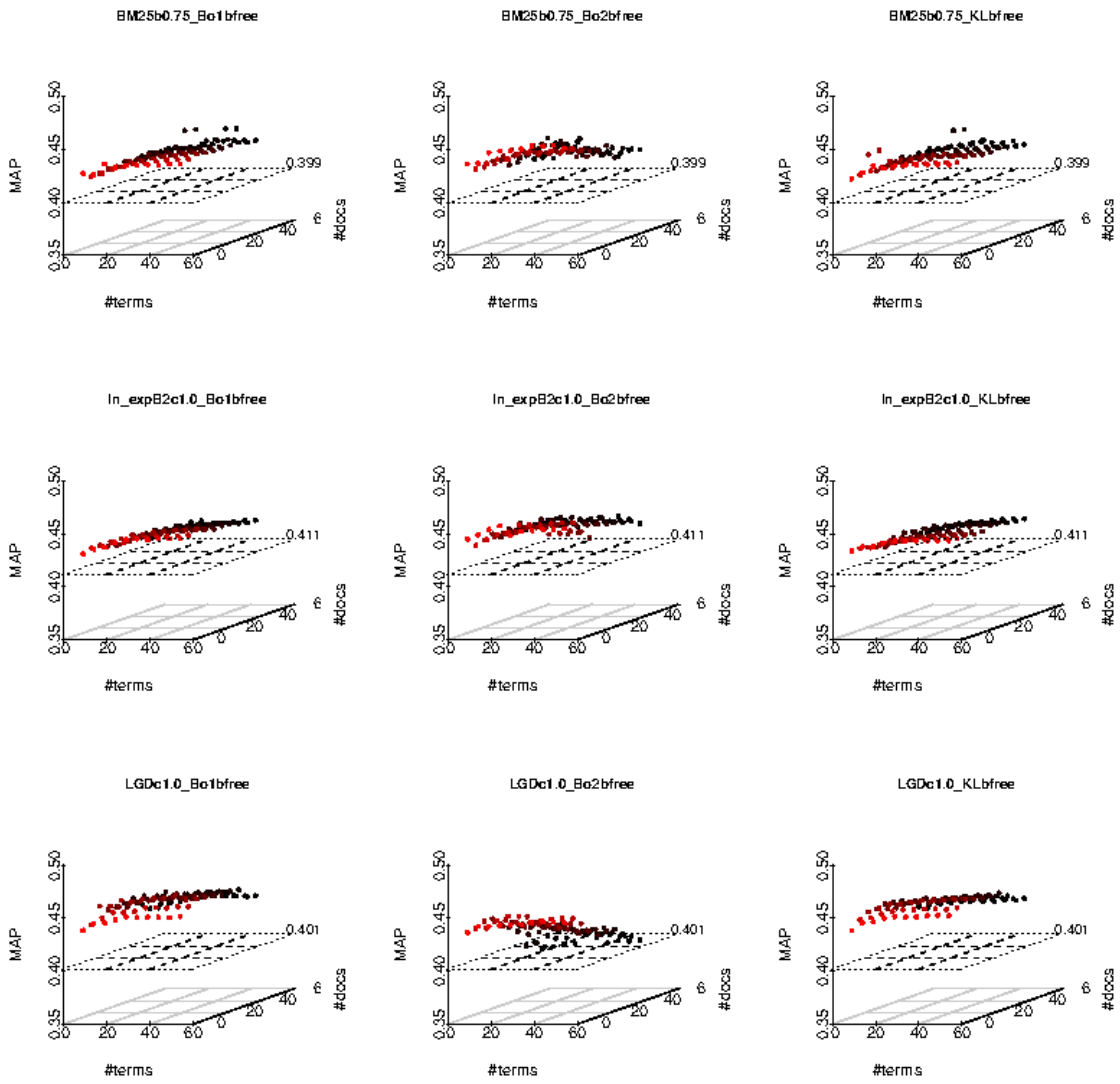


FIGURE VI.6 – Distribution de la MAP pour chaque modèle de pondération en combinaison avec chaque modèle d'expansion de la requête et vice-versa. Le nombre de termes (dénoté $\#terms$) extraits à partir des premiers documents (dénoté $\#docs = 5..50$) sont de 5 à 50 avec un pas de 5 termes/documents

moyenne des nombres de concepts MeSH est $\overline{N} = 19.27$. Celle-ci est supérieure à celle des nombres de concepts SNOMED ($\overline{N} = 10.77$) et GO ($\overline{N} = 4.45$). Cela prouve que les concepts les plus utilisés dans la littérature biomédicale sont les concepts issus du thésaurus MeSH mais les concepts issus d'autres terminologies peuvent être également utilisés.

Le tableau VI.8 présente les résultats en terme de MAP obtenus par notre approche de RI mono-terminologique sur la collection de test TREC Genomics 2005. D'après les résultats, l'utilisation des terminologies MeSH et SNOMED permet d'améliorer les résultats de RI en terme de MAP par rapport à la

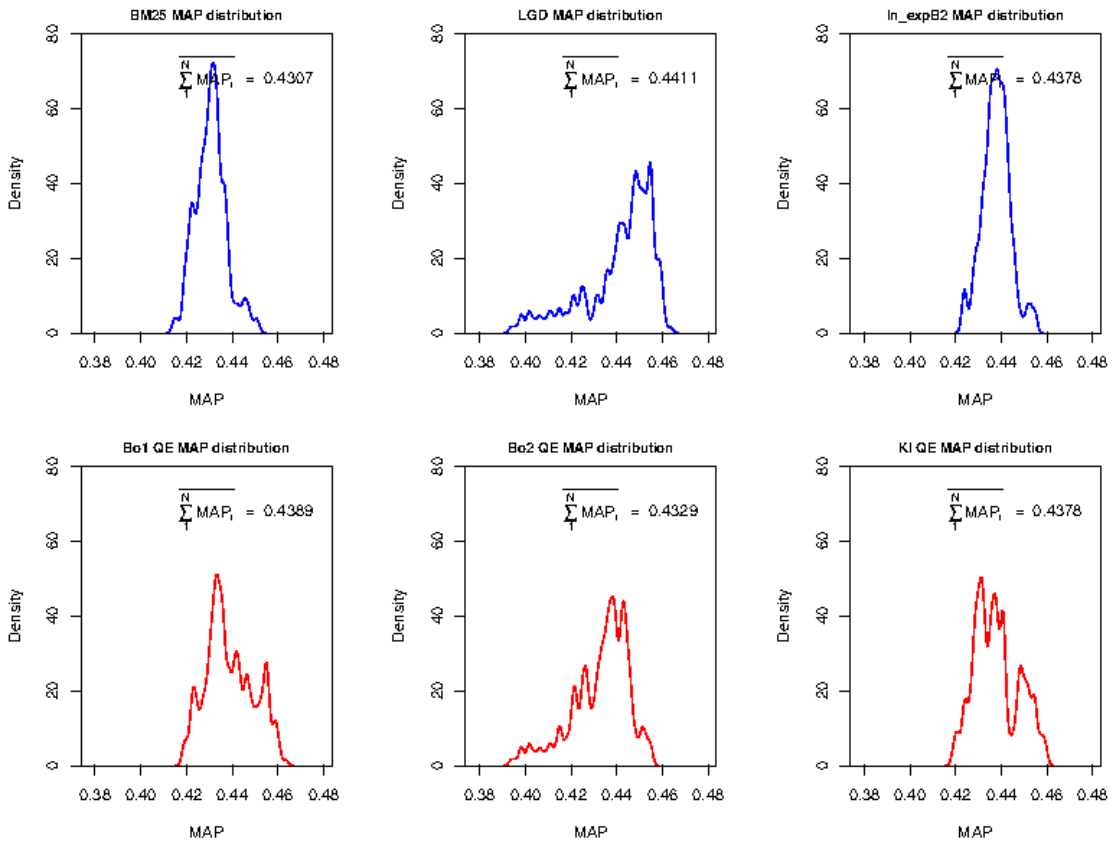


FIGURE VI.7 – Estimation par la méthode du noyau d'un échantillon de 300 (3x10x10) valeurs de MAP obtenues pour chaque modèle de pondération en combinaison avec une des **trois** modèles d'expansion de requêtes.
Le nombre de termes (dénoté $\#terms$) extraits à partir des premiers documents (dénoté $\#docs = 5..50$) sont de 5 à 50 avec un pas de 5 termes/documents.

base de référence d'évaluation (baseline 2 avec la combinaison du modèle de pondération LGD et du modèle de reformulation de requêtes en utilisant le modèle Bo1). Les taux d'amélioration sont variés entre +0.90% et +4.77% pour toutes les trois terminologies. En particulier, nous observons que l'expansion documentaire en utilisant **15 concepts** donne les meilleurs résultats et que ce nombre correspond au nombre de termes préférés (*main headings*) MeSH sélectionnés par les indexeurs humains à la NLM pour indexer les documents dans la base bibliographique MEDLINE.

Concernant les termes issus de l'ontologie de gènes (GO), nous observons que l'expansion documentaire avec 5 premiers concepts permet d'améliorer la MAP jusqu'à +1.99%. Cependant, à partir de 10 concepts, les performances diminuent parce qu'il est probable les concepts GO identifiés ne correspondent pas aux sujets sémantiques du document ou parce qu'ils ne permettent pas de traduire la sémantique du document qui est importante pour améliorer les

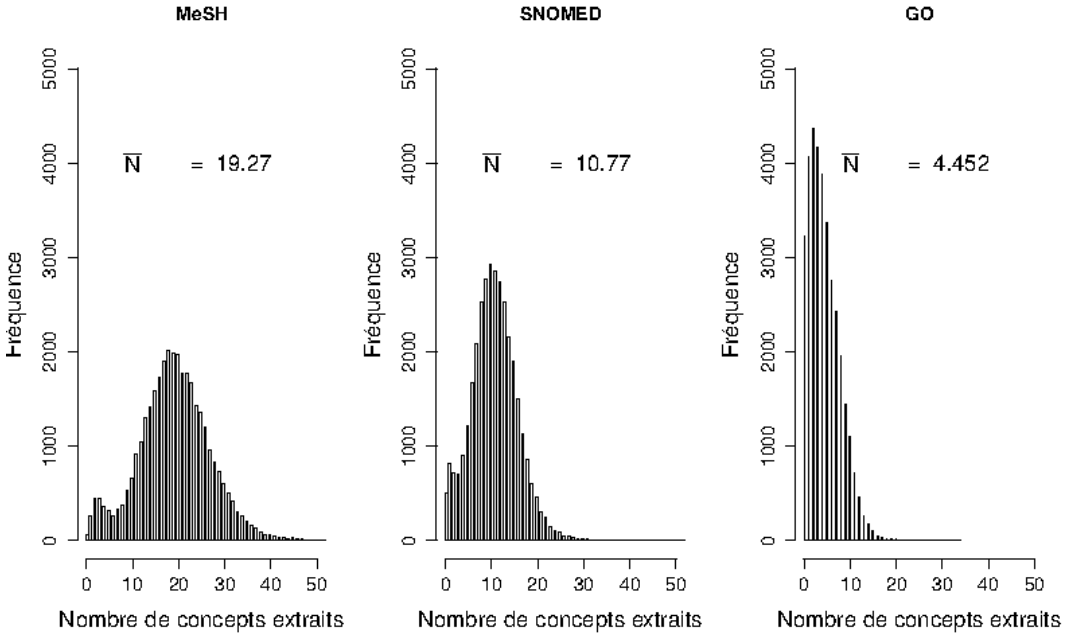


FIGURE VI.8 – Histogrammes des nombres de concepts MeSH, SNOMED, GO extraits à partir des documents

TABLEAU VI.8 – Résultats en terme de MAP de notre approche de RI basée sur une mono-terminologie sur la collection TREC Genomics 2005.

Terminology / N	**MeSH**	**SNOMED**	**GO**
Baseline_QE	0.2664		
5	0.2724 (+2.25)	0.2697 (+1.24)	**0.2717** (+1.99)
10	0.2763 (+3.72)	0.2688 (+0.90)	0.2622 (-1.58)
15	**0.2791**† (+4.77)	**0.2736** (+2.70)	0.2623 (-1.54)
20	0.2778 (+4.28)	0.2732 (+2.55)	0.2623 (-1.54)
25	0.2759 (+3.57)	0.2733 (+2.59)	0.2623 (-1.54)
30	0.2758 (+3.53)	0.2731 (+2.52)	0.2623 (-1.54)
35	0.2756 (+3.45)	0.2731 (+2.52)	0.2623 (-1.54)
40	0.2752 (+3.30)	0.2731 (+2.52)	0.2623 (-1.54)
45	0.2753 (+3.34)	0.2731 (+2.52)	0.2623 (-1.54)
50	0.2754 (+3.38)	0.2731 (+2.52)	0.2623 (-1.54)

N est le nombre de termes préférés désignant les concepts utilisés pour l'expansion conceptuelle documentaire - Les chiffres en gras (resp. en italique) indique la valeur MAP optimale obtenue pour une terminologie en faisant varier N (resp. pour une valeur donnée de N) - Les chiffres entre parenthèses indiquent les taux d'accroissement comparativement à la base de référence d'évaluation - Les taux d'accroissement significatifs sont dénotés par † pour la valeur de $p \leq 0.05$ des tests T calculés sous R (R Development Core Team, 2008)

résultats vis-à-vis de la requête. Pour démontrer la différence entre les concepts issus de chaque terminologie utilisée pour l'extraction de concepts, nous présentons dans la figure VI.9 la distribution des nombres de concepts extraits communs entre chacun des couples de terminologies. À titre d'exemple, le nombre moyen de concepts communs entre les concepts MeSH et SNOMED extraits est environ 2 ou 3 concepts en moyenne. Par contre, on observe rarement les concepts communs entre GO et les deux autres terminologies. Le tableau VI.9 révèle la différence entre les concepts GO extraits du document ayant l'identifiant **10022301** comparativement aux concepts MeSH et SNOMED. Parmi les concepts GO extraits de ce document, quatre sur dix (4/10) concepts GO sont retrouvés dans le thésaurus MeSH et/ou dans la nomenclature SNOMED. Le reste ne contient qu'un ou quelques mots qui constituent les concepts MeSH et/ou SNOMED, par exemple, le terme "DNA" désignant le concept C1235786 est un constituant du terme "DNA repair" du concept C0012899 issu du MeSH ou de la SNOMED.

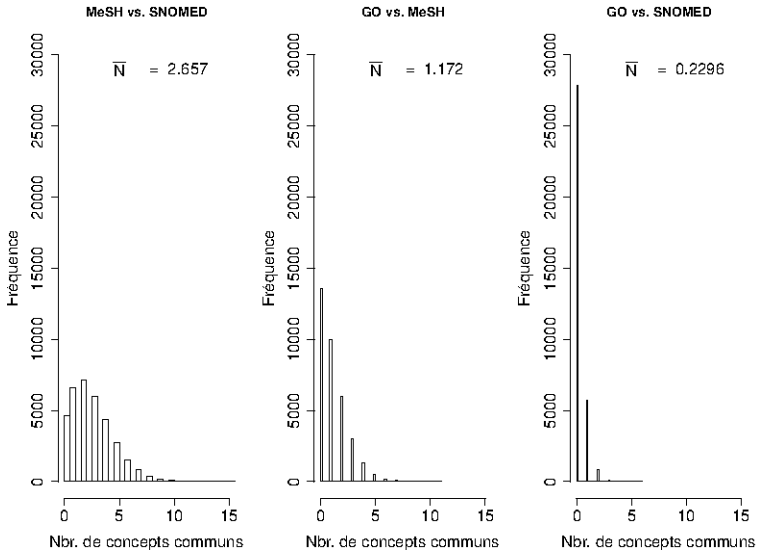


FIGURE VI.9 – Distribution des concepts communs entre les terminologies MeSH, SNOMED et GO extraits à partir des documents de la collection TREC Genomics 2005

À ce niveau, nous présentons les performances (MAP, P@10, P@20, Rappel) de l'indexation basée sur une mono-terminologie en les comparant aux performances de la base de référence d'évaluation (Baseline_QE). Le tableau VI.10 montre les résultats obtenus par la combinaison de l'expansion documentaire par des concepts issus de chacune des terminologies MeSH, SNOMED, GO et de l'expansion de requêtes comparativement aux résultats de la base de référence d'évaluation qui est basée uniquement sur l'expansion de requêtes. La figure VI.10 visualise les performances de la RI (P@10, Rappel et MAP) obtenues par chacune des terminologies ainsi que les performances de la base de référence d'évaluation. Les concepts MeSH, SNOMED et GO sont extraits automatiquement à partir de chaque document et sont utilisés par la suite pour

TABLEAU VI.9 – Listes des concepts extraits à partir du document 1002230

<DOC>
<DOCNO>10022301</DOCNO>
<TITLE>- Mechanisms of p53-induced apoptosis : in search of genes which are regulated during p53-mediated cell death.</TITLE>
<ABSTRACT>- The tumor suppressor gene p53 is a major player in the protection of cells from DNA damage. In the majority of human cancers, p53 is functionally inactivated–mostly by mutations but also by interaction with viral or cellular proteins. Wild-type p53 is involved in essential functions such as DNA repair, transcription, genomic stability, senescence, cell cycle control and apoptosis. It was shown to be a sequence-specific transcriptional activator, and this activity appears to be necessary to impose growth arrest. A major target gene which participates in p53-mediated growth arrest is p21/Waf1, an inhibitor of cyclin-dependent kinases. Whether or not transcriptional activation of target genes is required for p53-mediated apoptosis may depend on the cell type and external factors, and the mechanism of cell death induction is not clear yet. We have employed clones of the M1 myeloid leukemic cell line expressing a temperature-sensitive p53 mutant to study genes which are regulated during p53-induced apoptosis. Fontenay-aux-Roses, France.</ABSTRACT>
</DOC>

	MeSH	**SNOMED**	**GO**
0	C0243045 Cyclin-Dependent K.	C0162638 Apoptosis	**C1155872** regulation of cell cycle
1	C0162638 Apoptosis	C0001792 Old-age	**C2265405** senescence
3	C0919532 Genomic Instability	C0237477 Arrested	**C1325410** tumor suppressor
4	C0001811 Aging	C0205164 Major	C0012899 DNA repair
5	C0242613 Gene Targeting	C0012860 DNA damage	**C0040649** transcription
6	C0007587 Cell Death	C0012899 DNA repair	**C1621983** induction
7	C0162493 Transcriptional A.	C0205245 Functional	C0018270 growth
8	C0012860 DNA Damage	C0205132 Linear	**C1325786** DNA
9	C0012899 DNA Repair	C0205177 Active	C0007634 cell
10	C0525050 Cloning, Organism	C0308718 CONTROL	C1325816 protein
11	C0007586 Cell Cycle	C0027651 Neoplasm	
12	C0007600 Cell Line	C0086418 Homo sapiens	
13	C0039476 Temperature		
14	C0376706 Mechanics		
15	C0033684 Proteins		

Les concepts GO qui ne figurent pas dans le MeSH et la SNOMED sont mis en gras. Les concepts communs entre MeSH et SNOMED sont mis en italique.

étendre son contenu. De plus, pour avoir un aperçu sur l'efficacité de notre approche de RI terminologique, nous présentons également les résultats obtenus par l'utilisation des termes MeSH (main headings, subheadings) qui ont été sélectionnés par les indexeurs humains. Notons que ces termes ont été ajoutés à chaque document (citation) en se basant sur le contenu intégral des articles orignaux avant d'être stockés par le titre et/ou le résumé ainsi que la liste de termes MeSH associés dans la base MEDLINE. La principale différence entre les termes identifiés par notre méthode d'extraction de concepts et celle des termes MeSH manuellement sélectionnés concerne donc le contenu du document. Plus précisément, notre approche d'extraction de concepts est basée uniquement sur le titre et/ou le résumé des documents tandis que les indexeurs humains ont accès aux contenus intégraux des articles.

Nous observons que les termes MeSH manuellement ajoutés par les indexeurs humains permettent d'améliorer la MAP et le rappel en RI, par rapport à la base de référence d'évaluation. Les taux d'accroissement de la MAP (+7.13%) et du rappel (+2.91%) sont meilleurs que ceux qui sont obtenus par les scénarios automatiques. Par contre, concernant les précisions P@10 et P@20, notre approche de RI mono-terminologique permet d'optimiser les pré-

cisions P@10 avec un taux d'accroissement de +3.61% en utilisant SNOMED (*vs.* -1.54% pour MeSH manuel) et les précisions P@20 avec un taux d'accroissement de +2.95% en utilisant MeSH (*vs.* +2.08% pour MeSH manuel).

TABLEAU VI.10 – Comparaison des performances (MAP, P@10, P@20, Rappel) de l'indexation mono-terminologique aux performances de la base de référence d'évaluation (Baseline_QE)

	MAP	**P@10**	**P@20**	**Rappel**
Baseline_QE (*LGD+Bo1*)	0.2664	0.3959	0.3459	0.8833
MeSH (*auto*)	0.2791 (+4.77)	0.3980 (+0.53)	**0.3561** (+2.95)	0.8859 (+1.43)
MeSH (*manuel*)	**0.2854** (+7.13)	0.3898 (-1.54)	0.3531 (+2.08)	**0.9090** (+2.91)
SNOMED	0.2736 (+2.70)	**0.4102** (+3.61)	0.3439 (-0.58)	0.8842 (+0.10)
GO	0.2717 (+1.99)	0.3939 (-0.51)	0.3449 (-0.29)	0.8805 (-0.32)

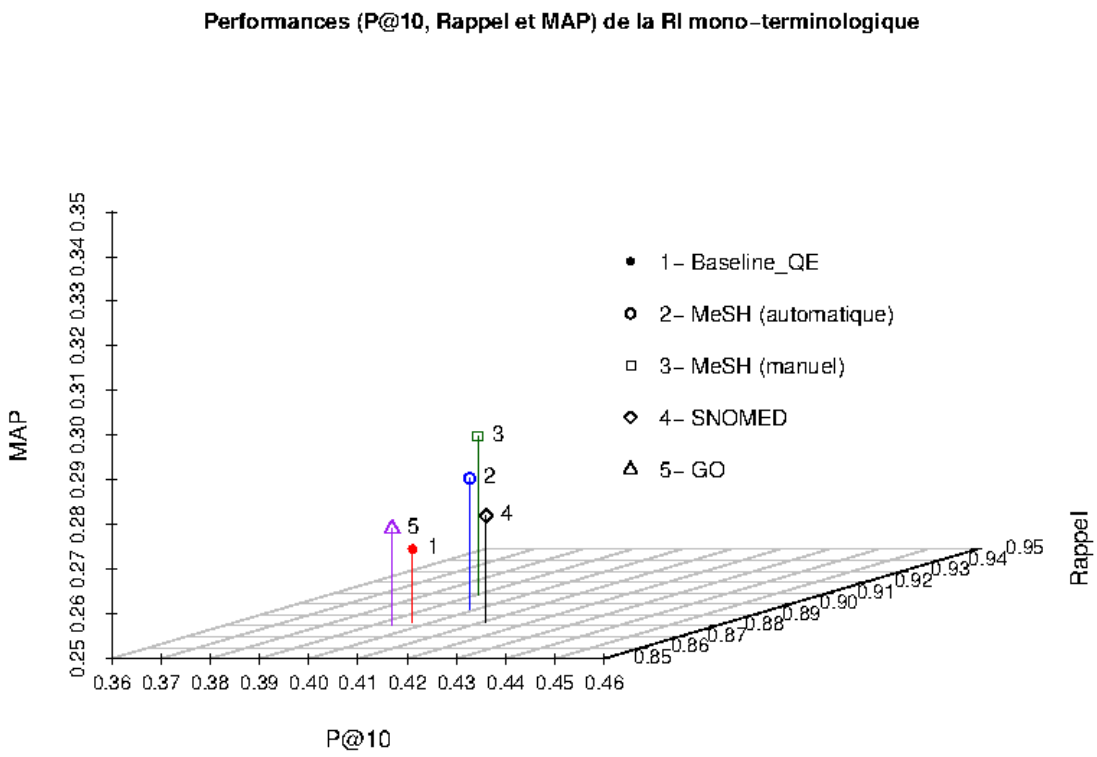


FIGURE VI.10 – Visualisation des performances (P@10, Rappel, MAP) de la RI mono-terminologique en 3D en comparaison aux performances de la base de référence d'évaluation sans utiliser aucune terminologie.

D'après les résultats obtenus et présentés dans le tableau VI.8, nous concluons que l'expansion documentaire en utilisant une quinzaine de concepts extraits à partir de chaque document permet d'améliorer les performances de la RI en terme de MAP. Cependant, les termes désignant les concepts extraits doivent être les plus significatifs dans la représentation du document. Cela signifie que la terminologie utilisée pour l'extraction de concepts doit être capable

de couvrir de manière adéquate les sujets sémantiques du document. MeSH et SNOMED sont susceptibles d'être les terminologies qui sont appropriées pour étendre le contenu textuel du document. Bien que les termes GO ne contribuent pas beaucoup dans l'amélioration de la MAP, nous voulons les utiliser dans notre approche de RI multi-terminologique pour évaluer l'influence des concepts issus de plusieurs terminologies sur les performances de la RI biomédicale.

6.3.3 Évaluation de l'efficacité de l'indexation multi-terminologique

Dans cette sous-section, nous évaluons l'intérêt de l'intégration de plusieurs terminologies dans un processus de RI biomédicale. Nous présentons les résultats obtenus par notre approche de RI multi-terminologique, puis nous comparons nos résultats à ceux obtenus par la base de référence d'évaluation ainsi que par notre approche de RI mono-terminologique. Ici, les concepts biomédicaux issus de trois terminologies (MeSH, SNOMED et GO) sont extraits à partir de chaque document. Nous utilisons la même méthode d'extraction de concepts qui a été utilisée pour extraire les concepts issus d'une mono-terminologie, c-à-d MaxMatcher++ avec la pondération des concepts par le schéma de pondération BM25 (Robertson *et al.*, 1994) implémenté dans l'outil d'extraction de concepts **cxtractor** [8]. Ensuite, les concepts ainsi extraits sont fusionnés ou combinés en utilisant les huit modèles de vote (décrits dans la section 4.2). Comme nous avons montré dans les expérimentations précédentes, la combinaison de l'expansion conceptuelle documentaire et de l'expansion de requêtes produit de meilleures performances de RI. De plus, l'utilisation de 40 termes extraits à partir des premiers documents résultats permet d'optimiser la MAP. Afin d'évaluer la qualité des ces termes, nous faisons varier le nombre de documents impliqués dans la sélection des termes, puis nous évaluons les performances des résultats obtenus par chaque modèle de vote pour l'extraction de concepts et l'indexation conceptuelle basée sur plusieurs terminologie. Le nombre de documents concernés pour l'expansion de requêtes est parmi les valeurs suivantes {10, 12, 15, 18, 20}.

Le tableau VI.11 montre les résultats en terme de MAP obtenus sur la collection TREC Genomics 2005 en utilisant les huit techniques de vote pour fusionner les concepts issus des trois terminologies MeSH, SNOMED et GO. Tout d'abord, nous comparons les résultats MAP de chaque modèle de vote à la base de référence d'évaluation (baseline_QE avec la combinaison du modèle de pondération de termes LGD et le modèle d'expansion de requêtes Bo1, dénotée LGD_Bo_1). Ensuite, nous montrons la valeur ajoutée de notre approche de RI multi-terminologique par rapport à l'approche de RI mono-terminologique.

8. `https://sourceforge.net/projects/cxtractor/files/`

TABLEAU VI.11 – Performances de notre approche de RI multi-terminologique sur la collection TREC Genomics 2005.

Modèle de vote \ **#docs**	**10**	**12**	**15**	**18**	**20**
Baseline_QE	0.2664				
CombANZ	0.2810 (+5.48)	0.2848 (+6.91)	0.2779 (+4.32)	0.2700 (+1.35)	0.2648 (-0.60)
CombMAX	0.2800 (+5.11)	**0.2859**† (+7.32)	0.2776 (+4.20)	0.2670 (+0.23)	0.2637 (-1.01)
CombMED	0.2812 (+5.56)	0.2849 (+6.94)	0.2777 (+4.24)	0.2675 (+0.41)	0.2650 (-0.53)
CombMIN	0.2812 (+5.56)	0.2823 (+5.97)	0.2776 (+4.20)	0.2678 (+0.53)	0.2651 (-0.49)
CombMNZ	0.2805 (+5.29)	0.2846 (+6.83)	0.2782 (+4.43)	0.2699 (+1.31)	0.2652 (-0.45)
CombRank	**0.2831** (+6.27)	0.2849 (+6.94)	0.2750 (+3.23)	0.2695 (+1.16)	0.2648 (-0.60)
CombRCP	0.2814† (+5.63)	0.2790 (+4.73)	0.2744 (+3.00)	0.2684 (+0.75)	0.2623 (-1.54)
CombSUM	0.2806 (+5.33)	0.2850† (+6.98)	**0.2785** (+4.54)	**0.2704** (+1.50)	**0.2666** (+0.08)

Les chiffres en gras indiquent les valeurs MAP optimales obtenue par le meilleur modèle de vote en utilsant 40 termes pour l'expasion de requêtes. Les chiffres soulignés indiquent les valeurs MAP optimales obtenues par chaque modèle de vote. Les chiffres entre parenthèses indiquent les taux d'accroissement de la MAP par rapport à la base de référence d'évaluation. Les taux d'accroissement significatifs ($p \leq 0.05$) sont dénotés par le symbole †

Comme nous le constatons, la plupart des modèles de vote donnent de meilleurs résultats que la base de référence d'évaluation lorsque le nombre de termes étendus à chaque requête est inférieur à 20. Le modèle $CombSUM$, qui sélectionne les concepts candidats en calculant la somme des poids des concepts extraits, donne globalement de meilleurs résultats MAP que d'autres modèles de vote. Cependant, la meilleure valeur de la MAP est obtenue par le modèle $CombMAX$ (MAP=0.2859, #terms=40, #docs=12). Ces résultats montrent que les concepts ayant un poids important dans le document sont susceptibles d'être pertinents pour représenter les sujets sémantiques du document. En effet, ces concepts sont pondérés par un schéma TF-IDF, en l'occurrence le modèle BM25, qui est un des modèles les plus appropriés pour modéliser la pertinence d'un document vis-à-vis d'une requête spécifique en RI.

Nous remarquons également que pour la même configuration, c-à-d les

mêmes paramètres (e.g., nombre de termes étendus à la requête, nombre de documents impliqués dans l'expansion de requêtes, etc.), les modèles de vote donnent des résultats compétitifs avec peu de différence en terme de MAP entre eux. Cependant, les résultats obtenus par chaque modèle de vote peuvent être statistiquement significatifs ou non par rapport à ceux obtenus par la base de référence d'évaluation. Afin d'évaluer le degré de signification de nos résultats, les tests T (tests de Student) deux à deux sont calculés en utilisant le langage de programmation R, qui est dédiée au traitement de données et l'analyse statistique (R Development Core Team, 2008). Les résultats les plus significatifs ($p - value \leq 0,05$) des tests T deux à deux entre les résultats obtenus par chaque modèle de vote et ceux obtenus par la base de référence d'évaluation sont : $CombSUM$ ($p = 0.0303, df = 49, t = 2.2311, M = 0.0115$), $CombMAX$ ($p = 0.0277, df = 49, t = 2.2697, M = 0.0112$) and $CombRCP$ ($p = 0.0144, df = 49, t = 2.5369, M = 0.0026$). Par conséquent, ces résultats montrent que notre approche de RI multi-terminologique est statistiquement significative par rapport à la base de référence d'évaluation.

TABLEAU VI.12 – Comparaison des performances (MAP, P@10, P@20, Rappel) de l'approche de RI **mono-** *vs.* **multi-**terminologique

	MAP	**P@10**	**P@20**	**Rappel**
MeSH (auto)	0.2791	0.3980	0.3561	0.8859
MeSH (manuel)	0.2854	0.3890	0.3531	**0.9090**
SNOMED	0.2736	0.4102	0.3439	0.8842
GO	0.2623	0.3837	0.3398	0.8805
CombMax (*#docs 12*)	**0.2859**	0.4061	0.3531	0.8887
CombSUM (*#docs 12*)	0.2850	0.4061	0.3551	0.8861
CombSUM (*#docs 15*)	0.2785	**0.4122**	**0.3633**	0.8905
CombRank (*#docs 15*)	0.2750	0.4082	0.3520	0.8953

Après avoir montré l'efficacité de notre approche de RI par rapport à la base de référence d'évaluation, en l'occurrence l'approche de RI qui est basée uniquement sur l'expansion de requêtes, notre objectif est de montrer l'intérêt de l'utilisation de plusieurs terminologies dans un processus de RI biomédicale. Nous comparons les performances de RI, notamment les mesures MAP, P@10, P@20 et le rappel, entre les résultats obtenus par notre approche de RI multi-terminologique et ceux obtenus par l'approche de RI mono-terminologique (*cf.* le tableau VI.12). Nous observons que les résultats MAP de la première sont légèrement différents de ceux de la deuxième. En particulier, la MAP de notre approche de RI multi-terminologique basée sur le modèle de vote CombMax dépasse les autres scénarios automatiques de RI mono-terminologique et est légèrement supérieure à la MAP obtenue par le scénario manuel de RI mono-

terminologique (0.2859 *vs.* 0.2854). Concernant les précisions P@10 et P@20, les résultats obtenus par le modèle CombSUM (avec 15 documents étendus qui sont impliqués dans l'expansion de requêtes) sont meilleurs que ceux obtenus par les scénarios (automatique ou manuel) de RI mono-terminologique. Concernant le rappel, l'expansion documentaire conceptuelle par des termes MeSH manuellement ajoutés dans chaque document permet d'optimiser le rappel qui est légèrement supérieur au rappel obtenu par le modèle CombRank (0.9090 pour le premier contre 0.8953 pour le deuxième). Il est également à noter que les termes MeSH ont été ajoutés par les indexeurs humains se basant sur les textes intégraux des articles tandis que notre approche de RI multi-terminologique est uniquement basée sur les titres et/ou résumés des articles pour extraire les termes désignant les concepts issus de plusieurs terminologies.

6.4 Discussion

Notre approche d'indexation dédiée à la RI biomédicale est essentiellement basée sur l'exploitation des ressources termino-ontologiques biomédicales ainsi que des caractéristiques statistiques de la sous-collection liées au contexte local de la requête, notamment les premiers documents retournés pour reformuler la requête de l'utilisateur via la combinaison de l'expansion documentaire conceptuelle et l'expansion de requêtes. À travers les expérimentations menées dans le cadre de la RI biomédicale sur les collections TREC Genomics, nous avons pu identifier les facteurs les plus pertinents ayant un impact significatif sur l'efficacité de la recherche biomédicale IR, à savoir (1) le choix du modèle de pondération des termes, (2) le modèle d'expansion de requêtes en utilisant un nombre approprié de termes extraits à partir d'un ensemble de meilleurs documents retournés par le modèle de pondération de termes et (3) l'expansion documentaire avec une quinzaine de termes désignant des concepts du domaine issus d'une terminologie unique ou de plusieurs terminologies.

Nous affirmons que l'approche de RI biomédicale basée sur la combinaison d'un modèle de pondération des termes approprié avec un modèle d'expansion de requête pertinent peut être une solution efficace pour améliorer les performances de la RI biomédicale. Par exemple, sur les collections de TREC Genomics, la combinaison du modèle de pondération des termes *LGD* (Clinchant et Gaussier, 2010) et du modèle d'expansion de requêtes *Bo*1 a montré une amélioration par rapport à la base de référence d'évaluation classique sans l'expansion de requêtes.

Sur la collection TREC Genomics 2004, nous avons identifié une meilleure configuration qui constitue notre base de référence d'évaluation solide pour évaluer notre approche de RI basée sur la ou les terminologies sur la collection TREC Genomics 2005. Nous avons introduit dans notre approche de RI un

nouveau facteur, qui est le nombre de termes désignant des concepts extraits à partir de chaque document. Ces concepts sont par la suite intégrés dans le processus de RI conceptuelle via l'expansion documentaire. Notre intuition de l'expansion documentaire est essentiellement basée sur l'hypothèse que les concepts pertinents extraits du contenu du document sont susceptibles de représenter les sujets sémantiques les plus descriptifs du document. Ces sujets sont représentés par les termes préférés désignant les concepts. Par conséquent, cela permettrait de résoudre le problème de la synonymie en RI biomédicale. En effet, les résultats obtenus par notre approche de RI mono-terminologie montrent que l'expansion documentaire conceptuelle en combinaison avec l'expansion de requêtes est utile pour améliorer les performances de la RI comparativement aux résultats obtenus par la reformulation de requêtes en utilisant la méthode PRF uniquement.

Nous avons évalué les différents modèles de vote pour combiner plusieurs concepts issus de multiples sources d'information du domaine biomédical. Sur les collections TREC Genomics, les résultats obtenus par notre approche de RI mono-terminologique (qui se base sur DE et QE) dépassent ceux qui sont obtenus par l'approche de RI classique (qui est basée uniquement sur QE). D'ailleurs, lorsque les techniques de vote sont utilisées pour sélectionner les meilleurs concepts, issus de plusieurs terminologies, extraits à partir de chaque document, les résultats de notre approche de RI multi-terminologique ont démontré l'utilité de prendre en compte les poids des concepts candidats extraits. En effet, comme indiqué dans la section 6.3.3, notre approche de RI multi-terminologique est systématiquement et significativement plus performant que la base de référence d'évaluation (Baseline_QE) et a des performances plus stables que l'approche de RI mono-terminologique en termes de MAP et de précision P@10 et P@20.

Il est également intéressant de noter que de nombreux travaux dans le domaine biomédical ont montré la valeur ajoutée de l'utilisation des termes MeSH qui sont manuellement ou semi-automatiquement associés à chaque citation de MEDLINE (Srinivasan, 1996; Aronson *et al.*, 2004b; Zhou *et al.*, 2006a; Abdou et Savoy, 2008). Dans le cadre de la RI *ad-hoc* du domaine biomédical, nous avons extrait automatiquement les concepts en utilisant une mono-terminologie ou plusieurs terminologies pour associer automatiquement aux documents. Ces concepts qui sont intégrés dans un processus de RI conceptuelle via la combinaison de l'expansion documentaire et l'expansion de requêtes ont montré l'efficacité de notre approche de RI terminologique pour améliorer les performances de la RI en comparaison avec les modèles de l'état-de-l'art en RI.

7 Conclusion

Nous avons présenté au cours de ce chapitre nos contributions portant sur la proposition d'une nouvelle approche de recherche d'information qui est basée sur l'utilisation des terminologies biomédicales. Notre approche de RI utilise deux méthodes d'extraction de concepts différentes : l'une consiste à extraire les concepts qui sont prédéfinis dans une terminologie tandis que l'autre vise à extraire les concepts issus de plusieurs terminologies via les huit modèles de vote. Nous utilisons ces concepts comme un moyen pour représenter les sujets sémantiques du document. Dans un contexte global (i.e., terminologies biomédicales), nous supposons que les concepts extraits permettent de normaliser les mots utilisés dans le document via les termes préférés désignant les concepts. De cette manière, les problèmes liés à la synonymie, à l'utilisation des abréviations ou des acronymes dans le document sont automatiquement résolus.

Concernant les requêtes de l'utilisateur, nous utilisons la technique de reformulation de requêtes via l'expansion des termes sélectionnés à partir des premiers documents étendus retournés par un modèle de pondération de termes de l'état-de-l'art. Les premiers documents retournés peuvent être considérés comme le contexte local de la requête car ils sont dépendants de chaque requête, c-à-d que ce contexte change en fonction de chaque requête. Nous avons combiné le contexte global du document et le contexte local de la requête dans un effort de récupérer plus de documents pertinents vis-à-vis de chaque requête pour améliorer les performances de RI. Les résultats obtenus sur la collection TREC Genomics 2005 montrent que notre approche de RI basée sur les contextes permet d'avoir une amélioration significative par rapport à la base de référence d'évaluation. D'une part, notre approche de RI mono-terminologique dépasse la base de référence d'évaluation ; d'autre part, en introduisant plusieurs terminologies, notre approche de RI multi-terminologique devient plus stable que les dernières. Nous concluons que l'indexation des documents biomédicaux par les termes désignant les concepts du domaine apporte les meilleurs résultats en termes de précision et rappel.

Chapitre VII

BioSIR - système prototype de RI biomédicale

Sommaire

1 Introduction **225**

2 Extraction de concepts **228**

3 Expansion conceptuelle de documents **231**

4 Évaluation de requêtes **235**

5 Outils d'évaluation **239**

1 Introduction

BioSIR (**Bio**medical **S**emantic **I**nformation **R**etrieval) : est notre premier système (version prototype) pour l'indexation et la recherche d'information biomédicale. Cette plateforme permet d'extraire des concepts biomédicaux qui sont définis dans les terminologies (e.g., MeSH, SNOMED, GO ou UMLS) à partir des textes biomédicaux, d'indexer les documents biomédicaux avec les concepts biomédicaux qui représentent les sujets sémantiques du document et/ou de la requête, de sélectionner de l'information biomédicale pertinente en réponse à une ou plusieurs requêtes de l'utilisateur ainsi que d'évaluer les performances des modèles de RI biomédicale.

L'interface graphique dans la figure VII.1 donne un aperçu sur les principaux modules de notre système : (1) *extraction de concepts* , (2) *indexation conceptuelle/sémantique*, (3) *recherche d'information conceptuelle/sémantique* et (4) *évaluation des résultats de la recherche d'information.*

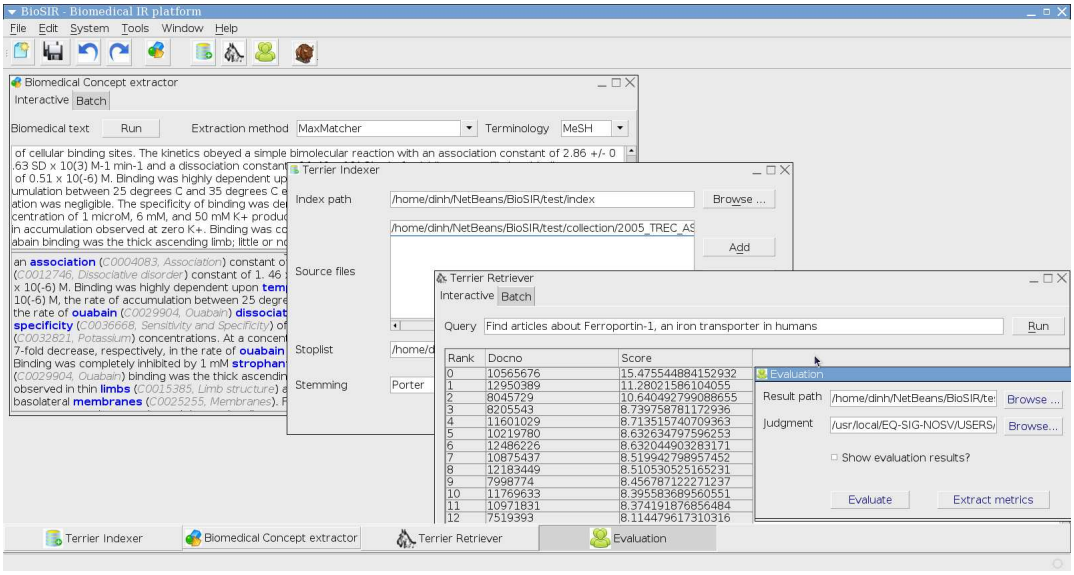


FIGURE VII.1 – Plateforme de la RI biomédicale BioSIR (version prototype)

Notre projet BioSIR est hébergé sur la plateforme OSIRIM (Open Services for Indexing and Research Information in Multimedia [1]) qui est un projet fédératif conduit par les équipes de recherche SAMOVA [2] et SIG [3], et principalement soutenu par le gouvernement Français, la région Midi-Pyrénées et le Centre National de la Recherche Scientifique (CNRS). Il s'agit d'un environnement homogène pour la recherche sur l'indexation et la recherche d'information dans des contenus multimédias. Cette architecture matérielle et logicielle (*cf.*

1. `http://osirim.irit.fr`
2. `http://www.irit.fr/-Equipe-SAMoVA-`
3. `http://www.irit.fr/-Equipe-SIG-`

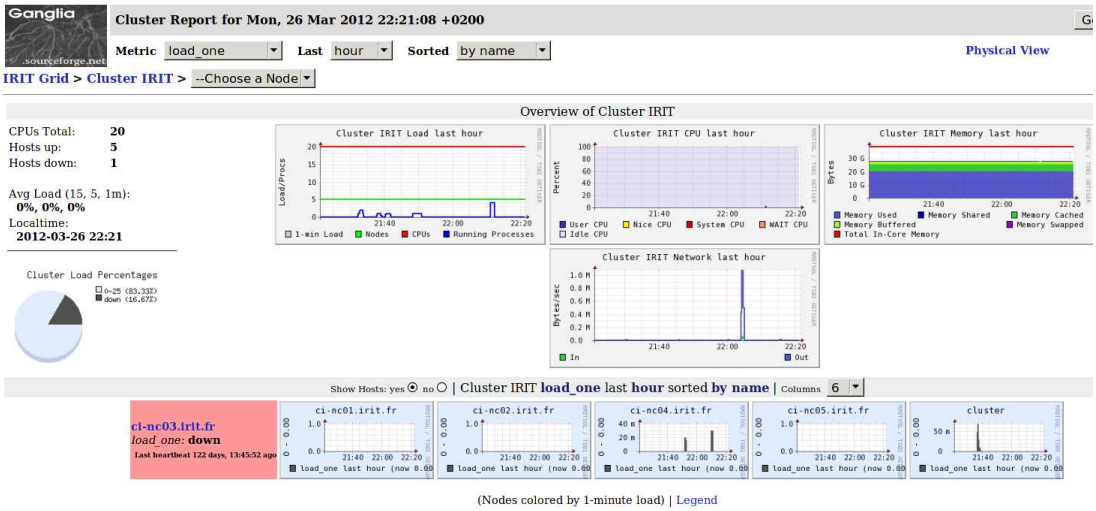


FIGURE VII.2 – Interface Ganglia : outil de visualisation des ressources en temps réels de la plateforme OSIRIM

la figure VII.2) permet de construire, tester et évaluer facilement des tâches les plus compliquées en RI comme l'apprentissage des modèles d'appariement ou des fonctions d'ordonnancement en RI ainsi que des tâches liées au traitement du langage naturel qui demandent un temps d'exécution important comme l'annotation des corpus, l'extraction de concepts, etc. Cette plateforme rassemble entre autres un ensemble de corpus, d'outils d'évaluation, de logiciels d'analyse, de moteurs de recherche dans le principal but d'offrir un espace de mutualisation où les chercheurs pourront échanger leurs connaissances dans le but de bénéficier de résultats obtenus dans d'autres laboratoires.

Dans ce contexte de partage et d'échange entre les chercheurs au laboratoire, notre plateforme d'indexation et de recherche d'information biomédicale BioSIR fournit un cadre général pour exploiter de manière efficace les ressources de la plateforme comme :

1. la puissance de calcul apportée par plusieurs nœuds de calcul (6 nœuds x 2 CPU x 2 cœurs de 2,8 GHz)
2. la capacité d'exécution de plusieurs tâches en parallèle
3. la capacité de stockage et de mise à disposition de données et collections
4. l'accès à des collections et corpus du domaine
5. des logiciels de la communauté RI et Indexation multimédia
6. l'hébergement de projets

L'interaction entre BioSIR et OSIRM s'effectue via des lignes de commande pour soumettre des scripts .pbs, appelés *jobs*. Un job représente une tâche

particulière qui est soumise à la plateforme OSIRIM. Le tableau VII.1 présente quelques lignes de commande utiles pour interagir avec OSIRIM.

$/opt/pbs/bin/qsub	soumettre un script .pbs
$/opt/pbs/bin/qstat	afficher les status des nœuds de calcul
$/opt/pbs/bin/qdel	supprimer un job dans la queue

TABLEAU VII.1 – Lignes de commande pour interagir avec OSIRIM

Il arrive souvent que l'utilisateur souhaite supprimer tous ses jobs soumis à OSIRIM. Le script Shell suivant récupère tous les jobs ayant été soumis par l'utilisateur et les arrête automatiquement.

```
#!/bin/bash
# Deletes all jobs submitted by current user
# Duy Dinh, IRIT - University of Toulouse, March. 2012
#
# Syntax: $sh deleteJobs.sh

arr=(); # value returned by the split function
# split a string with a set of delimiters
split()
{
        saveIFS=$IFS
        IFS="$2" # delimiter
        arr=($1) # convert string to arrays of strings delimited by a set of delimeters
        IFS=$saveIFS
}

# run qstat, get first (jobID) and third (user name) column
str=$(/opt/pbs/bin/qstat | awk 'BEGIN { OFS = ";"; ORS = "\n\n" } {print $1,$3}')
# convert to array of elements
elements=($str)
i=0
while [ $i -lt ${#elements[@]} ]
do
        s=${elements[$i]}
        split $s ';'
        if test ${arr[1]} = ${USER}
        then
                split ${arr[0]} '.'
                echo "Deleting job " ${arr[0]}
                /opt/pbs/bin/qdel ${arr[0]}
        fi
    (( i=i+1 ))
done
```

2 Extraction de concepts

Nous avons implémenté plusieurs méthodes d'extraction de concepts à partir des documents biomédicaux, à savoir les méthodes d'extraction basées sur un modèle de RI (e.g., TF_IDF, BM25, etc.) ainsi que la mesure statistique de Spearman permettant de modéliser la corrélation entre chaque concept candidat et un texte donné grâce aux positions des mots communs entre eux. Les algorithmes d'extraction de concepts sont implémentés dans notre logiciel Open Source cxtractor [4] qui est intégré dans notre plateforme BioSIR. cxtractor est un projet Open Source développé en Java visant à intégrer les algorithmes d'extraction de concepts de l'état-de-l'art dans une plateforme générique.

La figure VII.3 illustre les résultats de la méthode d'extraction de concepts proposée dans (Dinh et Tamine, 2011b). Les termes identifiés sont associés à leur concept représenté par un identifiant unique et sa forme préférée.

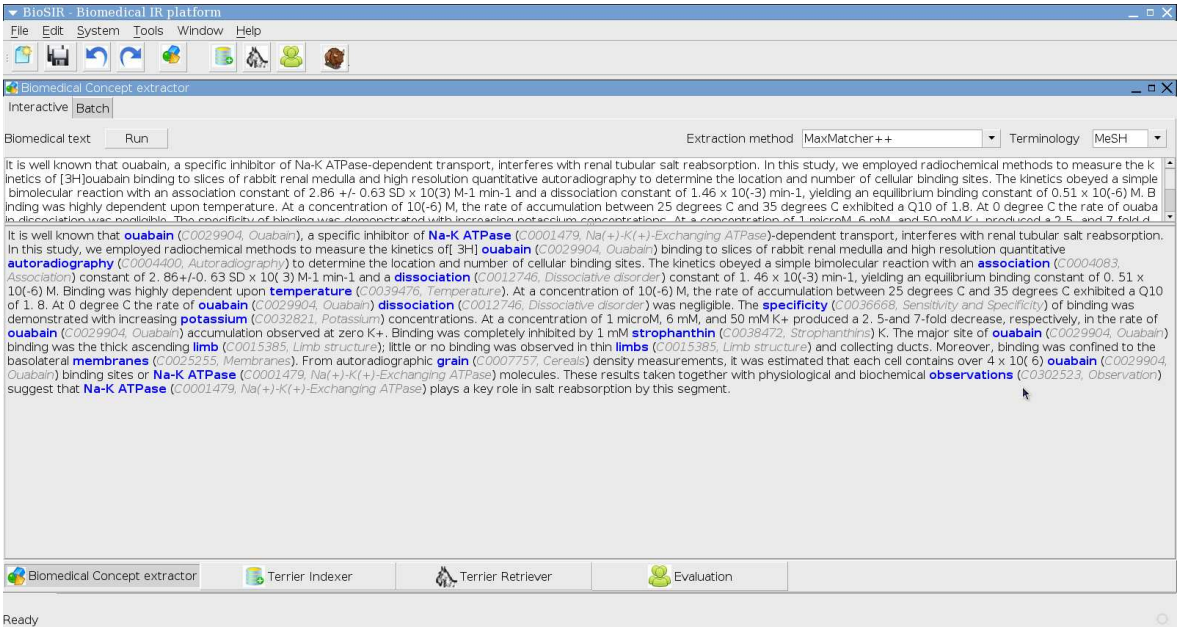


FIGURE VII.3 – BioSIR : Extraction de concepts à partir des documents

Dans un mode interactif, nos algorithmes d'extraction peuvent être exécutés en ligne de commande. Nous illustrons notre algorithme d'extraction de concepts par l'exemple suivant qui a pour but d'extraire les concepts à partir d'un texte donné passé comme argument du programme.

```
java -jar cxtractor.jar The low back pain is a common disease. \
Lung cancer is due to tobacco.
```

4. http://www.irit.fr/~Duy.Dinh/tools/cxtractor/

Les résultats affichés à l'écran sont comme suit : chaque terme identifé est mis en gras et les concepts associés sont mis entre parenthèses avec un identifiant unique et sa forme préféré en italique. La liste des concepts identifiés est également présentée par la suite. Chaque ligne correspond à un concept caractérisé par son rang, sont identifiant unique, sa forme préférée et son poids.

```
The <b>low back pain</b> (<i>C0024031, Low Back Pain</i>) is a common <b>disease</b> (<i>C0012634,
Disease</i>). <b>Lung cancer</b> (<i>C0024121, Lung Neoplasms</i>) is due to <b>tobacco</b>
 (<i>C0086707, Tobacco</i>).

0|C0024031|Low Back Pain|1.9886
1|C0024121|Lung Neoplasms|1.3257
2|C0012634|Disease|0.6629
3|C0086707|Tobacco|0.6629
```

La liste complète des options pour lancer cxtractor sur une ligne de commande est donnée comme suit :

```
Usage: java -jar cxtractor.jar [-r|--recursive] [-c|--clean] [-f|--file]
 [-d|--folder] input [-e|--doctype documentType] [-o|--output output]
 [-t|--terminology terminology] [-X|--cxMethod method]
 [-w|--wModel weightingModel] [-v|--version]

Example of usage:
java -jar cxtractor.jar -r -c -d tests -o output -X TerrierSpearmanExtractor

Option Long Option Value (y/n) Description
-r     --recursive no  Recursively processing
-c     --clean     no  Clean all previous data
-h     --help      no  Print this usage information
-f     --file      yes Extracting concepts from a file
-t     --terminology yesTerminology used
-w     --wModel    yes Weighting model  (PL2 by default)
-X     --cxMethod  yes Extraction method (MaxMatcherExtractor by default)
-d     --folder    yes Extracting concepts from a directory
-e     --doctype   yes Document type (file, trec, html)
-o     --output    yes Output directory
-v     --version   no  Version number
```

Les documents d'entrée peuvent avoir un des formats suivants : .txt, .html, ou les formats de TREC. (à configurer dans le fichier de configuration /config/settings.properties.sample). Lors de l'exécution, les paramètres de configuration sont chargés automatiquement.

Nous donnons ci-dessous un exemple de documents sous le format de TREC. Chaque document TREC contient des balises particulières, par exemple :

- **DOC** représente le document,
- **DOCNO** représente l'identifiant unique du document,
- **TITLE** correspond au titre du document,
- **ABSTRACT** correspond au résumé du document.

<DOC>
<DOCNO>11096424</DOCNO>
<TITLE>- Prenatal radiation-induced limb defects mediated by Trp53-dependent apoptosis in mice.</TITLE>
<ABSTRACT>- We reported previously that in utero radiation-induced apoptosis in the predigital regions of embryonic limb buds was responsible for digital defects in mice. To investigate the possible involvement of the Trp53 gene, the present study was conducted using embryonic C57BL/6J mice with different Trp53 status. Susceptibility to radiation-induced apoptosis in the predigital regions and digital defects depended on both Trp53 status and the radiation dose ; i.e., Trp53 wild-type (Trp53(+/+)) mice appeared to be the most sensitive, Trp53 heterozygous (Trp53(+/-)) mice were intermediate, and Trp53 knockout (Trp53(-/-)) mice were the most resistant. These results indicate that induction of apoptosis and digital defects by prenatal irradiation in the later period of organogenesis are mediated by the Trp53 gene. These findings suggest that the wild-type Trp53 gene may be an intrinsic genetic susceptibility factor that is responsible for certain congenital defects induced by prenatal irradiation. </ABSTRACT>
</DOC>

Afin de faciliter l'analyse et le traitement des concepts extraits, nous utilisons le format suivant pour sauvegarder les concepts candidats identifiés :

<DOCNO> DOC1
rank|CUI|concept name (preferred/non-preferred terms)|score
rank|CUI|concept name (preferred/non-preferred terms)|score
....
rank|CUI|concept name (preferred/non-preferred terms)|score

<DOCNO> DOC2
rank|CUI|concept name (preferred/non-preferred terms)|score
rank|CUI|concept name (preferred/non-preferred terms)|score
....

3 Expansion conceptuelle de documents

Les concepts extraits dans les étapes précédentes à partir de chaque document sont utilisés pour étendre le contenu textuel du document. La figure VII.4 présente une fenêtre permettant d'exécuter l'expansion conceptuelle de documents. Les paramètres d'entrée sont le chemin vers la collection originale, le chemin vers le fichier contenant les concepts extraits, appelés *kernel*, et le chemin vers le répertoire de sortie. Il est possible de spécifier le nombre de concepts à étendre pour chaque document, e.g., les nombres minimum et maximum de termes à étendre ou le pas d'itération pour des objectifs d'entraînement.

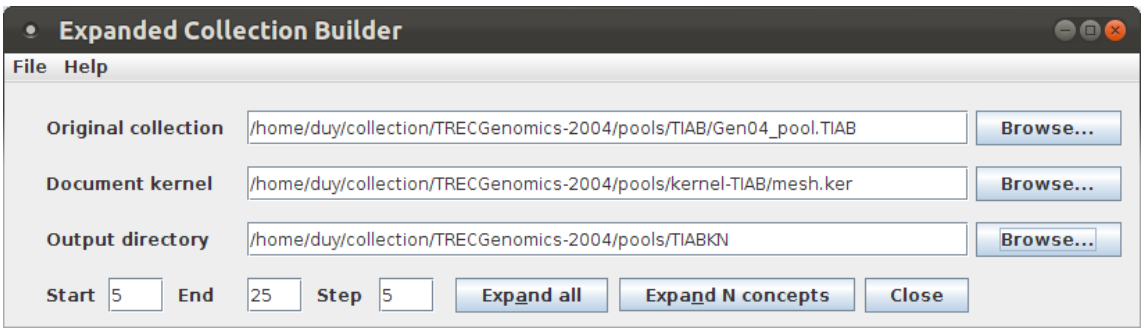


FIGURE VII.4 – BioSIR : Expansion documentaire conceptuelle

Cette fonction peut être également appelée via une ligne de commande comme suit :

```
java -jar docXpander.jar --recursive --number NConcepts --file fileName \
--kernel kernelFileName --output outputPath
```

Le tableau VII.2 illustre deux documents issus de TREC Genomics qui sont étendus par les concepts extraits en utilisant un algorithme d'extraction de concepts particulier. Les champs $TITLE$ et $ABSTRACT$ représentent le contenu textuel du document ; le champs $KERNEL$ représente les termes (préférés) désignant les concepts candidats extraits.

Afin de mieux exploiter les capacités de la plateforme de calcul OSIRIM, nous avons développé plusieurs modules en Shell qui jouent le rôle de "connecteurs" entre la plateforme OSIRIM et notre plateforme de RI biomédicale BioSIR. Ainsi, la prise en charge (e.g., soumission, gestion, parallélisation, etc.) des tâches en RI par la plateforme OSIRIM devient plus aisée. Par exemple, le script Shell suivant vise à exécuter l'expansion de documents pour deux collections TREC Genomics 2004 et 2005 en utilisant les concepts extraits à partir d'une terminologie ou de plusieurs terminologies.

<DOC>
<DOCNO>**11096424**</DOCNO>
<TITLE>*Prenatal radiation-induced limb defects mediated by Trp53-dependent apoptosis in mice.*</TITLE>
<ABSTRACT>
We reported previously that in utero radiation-induced apoptosis in the predigital regions of embryonic limb buds was responsible for digital defects in mice. To investigate the possible involvement of the Trp53 gene, the present study was conducted using embryonic C57BL/6J mice with different Trp53 status. ... These findings suggest that the wild-type Trp53 gene may be an intrinsic genetic susceptibility factor that is responsible for certain congenital defects induced by prenatal irradiation.
</ABSTRACT>
<KERNEL>
Apoptosis ; Limb Buds ; Genetic Predisposition to Disease ; Congenital Abnormalities ; Radiation ; Mice ; Extremities ; Genes ; Periodicity ; Pregnancy ; Organogenesis ;
</KERNEL>
</DOC>

<DOC>
<DOCNO>**11096458**</DOCNO>
<TITLE>*Modulation of double-stranded RNA-mediated gene induction by interferon in human umbilical vein endothelial cells.*</TITLE>
<ABSTRACT>
Endothelial cells respond to double-stranded RNA (dsRNA) with expression of a number of important immunomodulatory and inflammatory response genes, including adhesion molecules, cytokines, and antiviral genes. These studies demonstrate that priming with class I IFN can enhance the response to dsRNA through the heightened expression of genes that contribute to both the cellular response to viral infection and the host immunologic response. University, Atlanta, GA 30322, USA.
</ABSTRACT>
<KERNEL>
Vascular Cell Adhesion Molecule-1 ; Tumor Necrosis Factor-alpha ; Intercellular Adhesion Molecule-1 ; E-Selectin ; RNA, Double-Stranded ; Interleukin-6 ; Endothelial Cells ; Umbilical Veins ; Protein Kinases ;
</KERNEL>
</DOC>

TABLEAU VII.2 – Documents étendus par des termes préférés désignant les concepts extraits

```
#!/bin/bash
# Duy Dinh - University of Toulouse
# Setup document expansion for both TREC GEnomics 2004 and 2005
# last update: 09 March 2012
# Root folder of TREC Genomics collections
root="/osirim/sig1/CORPUS-TRAV/MEDICAL/dinh"

# TREC Genomics years
years="2004 2005"

# Terminologies used
terminologies="GO MESH SNOMED"

# Voting techniques used
votingTechniques="CombANZ CombMIN CombMED CombMAX CombMNZ
CombRank CombRCP CombSUM"

# number of expanded concepts
nConcepts="5 10 15 20 25 30 35 40 45 50"

# convert string variables to array of elements
arrYears=($years)
arrTerminologies=($terminologies)
arrVotingTechniques=($votingTechniques)
arrNConcepts=($N)

verifyDirectory(){
  local dirName=$1
  if !(test -d "$dirName")
  then
    echo "*** Making directory $dirName"
    mkdir "$dirName"
  fi
}

# setup mono/multi-terminology based Document Expansion
setup_DE(){
  o=$1 # short name of directory containing output results
  kn=$2 # short kernel name


  if [ "$shortKernelPath" == "" ]
  then
    shortKernelPath="kernel"
  fi

  id=0 # loop id
  # for each trec collections
```

```
  i=0
  while [ $i -lt ${#arrYears[@]} ]
  do
    year=${arrYears[$i]}

    echo "*** Processing TREC GENOMICS $year *** "
    echo "*** Configuring document expansion"

    # count number of kernel files in kernel directory
    size=0
    for f in `find $root/Genomics-$year/$kn -type f -iname '*.ker'`
    do
      (( size = size + 1 ))
    done
        # original Collection
    c="${root}/Genomics-${year}/TIAB/classic/collection"

    # for each number of expanded concepts
    j=0
    while [ $j -lt ${#arrNConcepts[@]} ]
    do
      N=${arrNConcepts[$j]}
      echo "[$id] Number of expanded concepts: $N"

      # eXpanded Collection
      xc="${root}/Genomics-${year}/${o}"
      verifyDirectory "$xc"
      xc="${xc}/${N}"
      verifyDirectory "$xc"

      # build and submit a job for document expansion using parameters
      sh docXpander.sh -r -n $N -d $c -k "$root/Genomics-$year/$kn/" \
                  -o $xc -l $id -s $size

      id=$(( id + 1 ))

      (( j=j+1 ))
    done # end number of extracted concepts used for DE

      (( i=i+1 ))
  done # end TREC collections
}

# mono-terminology based Document Expansion
setup_DE "TIABKN" "kernel"
#setup_DE "TIABKN-BM25" "kernel-mesh-bm25"
# multi-terminology based Document Expansion
#setup_DE "TIABKN-Comb" "kernel-Comb"
```

4 Évaluation de requêtes

Après avoir généré les structures d'index de la collection, nous pouvons entamer la recherche d'information pour récupérer les documents pertinents vis-à-vis d'une requête particulière afin d'évaluer les performances du système de RI. Dans la phase d'évaluation, la requête originale peut être étendue ou reformulée en utilisant les termes extraits à partir des premiers documents retournés lors de la première phase de recherche ou par des termes préférés ou des acronymes désignant les concepts extraits à partir de la requête.

Nous pouvons lancer une requête en *mode interactif* ou plusieurs requêtes en *mode d'évaluation*.

Le mode interactif permet de visualiser de manière rapide les premiers résultats retournés par le SRI (*cf.* la figure VII.5). Il est également utile de visualiser la différence entre les premiers résultats obtenus par chaque modèle de RI. L'interface dans la figure VII.6 permet de spécifier le modèle de pondération (weighting model). Cela permet de visualiser rapidement les premiers résultats en fonction du modèle de pondération sous l'onglet "Interactive". Il est possible d'évaluer un nouveau modèle de pondération en utilisant cette interface. En mode d'évaluation, nous devons spécifier un ensemble de requêtes pour évaluer et éventuellement le répertoire de sortie qui contient les résultats de la RI (*cf.* la figure VII.7).

Nous avons également entamé le développement d'une première interface Web de notre système prototype BioSIR qui est accessible en Intranet (*cf.* les figures VII.8 et VII.9).

Concernant l'expansion conceptuelle de requêtes, l'interface graphique dans la figure VII.10 permet de spécifier les différents paramètres afin d'affiner la liste des concepts biomédicaux extraits à partir du texte libre ou d'un ensemble de requêtes (chacune sur une ligne) dans le champ de saisie en haut. Les concepts identifiés, présentés à droite de la barre verticale '|', sont ajoutés à la fin du texte pour l'expansion documentaire conceptuelle ou à la fin de la requête pour l'expansion conceptuelle de la requête (zone de texte en bas). Les différentes méthodes d'extraction de concepts peuvent être expérimentées en sélectionnant dans la liste déroulante de la boîte combinée, à savoir la méthode basée sur la RI (*cf.* la section 3), la méthode basée sur la corrélation d'ordre de mots (*cf.* la section 3.2.2) et la méthode d'extraction MaxMatcher (Zhou *et al.*, 2006b). La case à cocher "Column header" est sélectionnée si le traitement est effectué ligne par ligne et que chacune possède un identifiant unique (e.g., dans le cas du traitement d'un ensemble de requêtes). La case à cocher "Include all terms" indique que si tous les termes (y compris les termes non-préférés, les synonymes, les abréviations, etc.) sont ajoutés dans la liste des termes ex-

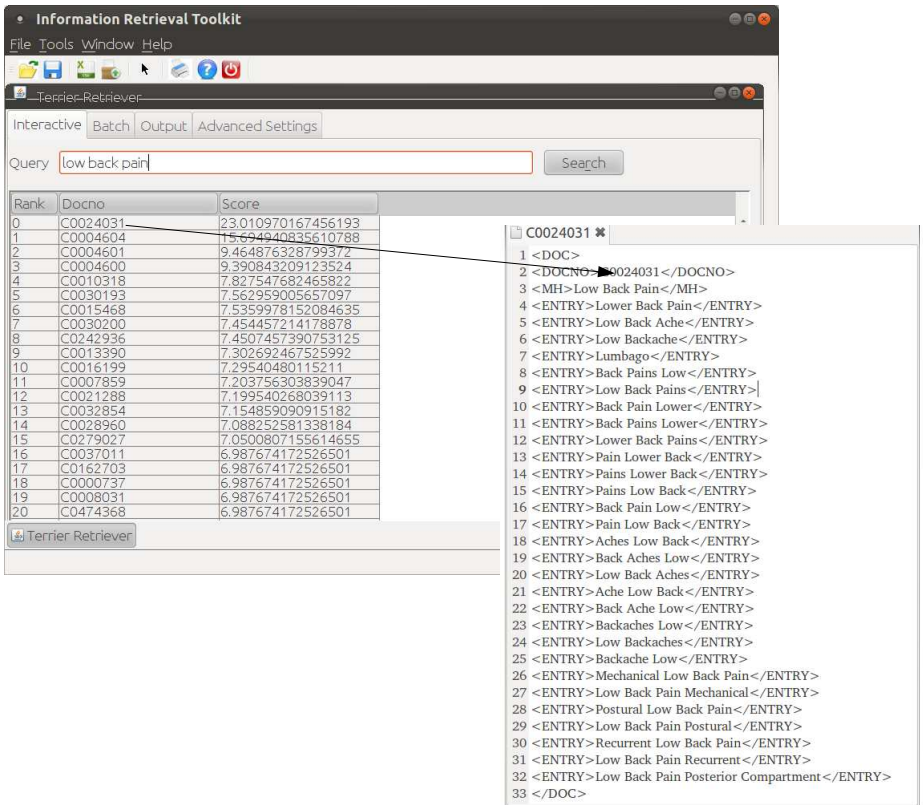


FIGURE VII.5 – BioSIR : Recherche d'information (mode interactif)

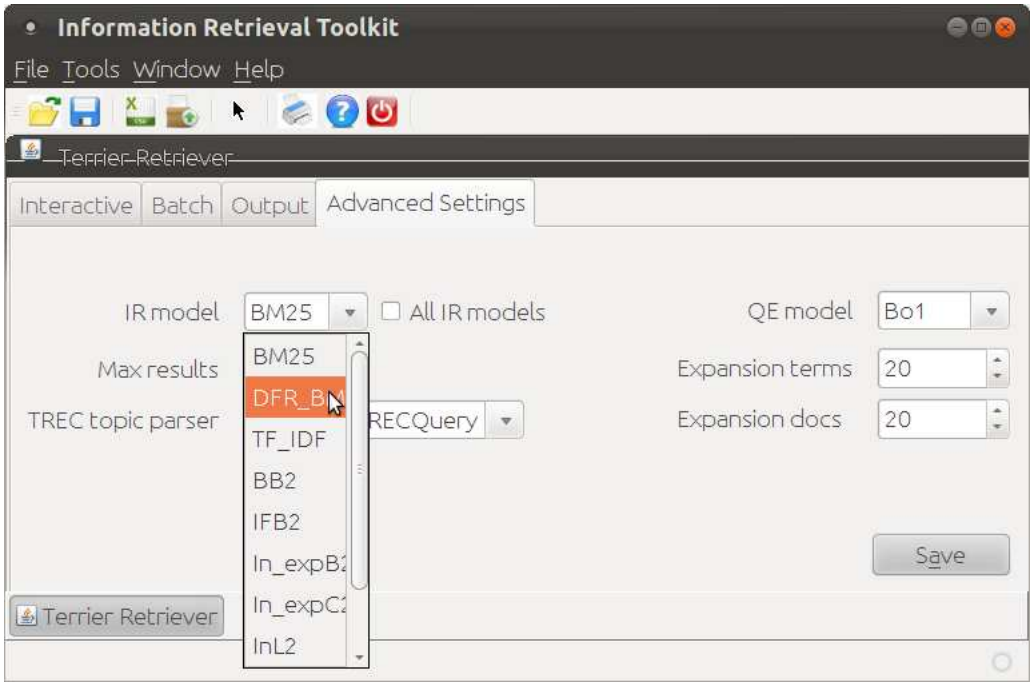


FIGURE VII.6 – BioSIR : Configuration des modèles de RI

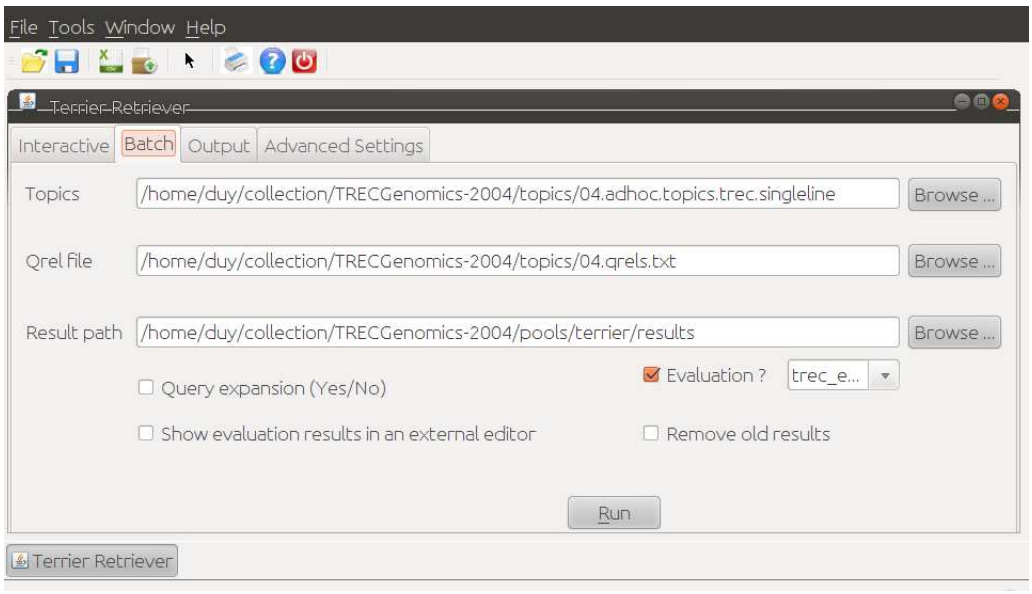


FIGURE VII.7 – BioSIR : Recherche d'information conceptuelle biomédicale (mode d'évaluation)

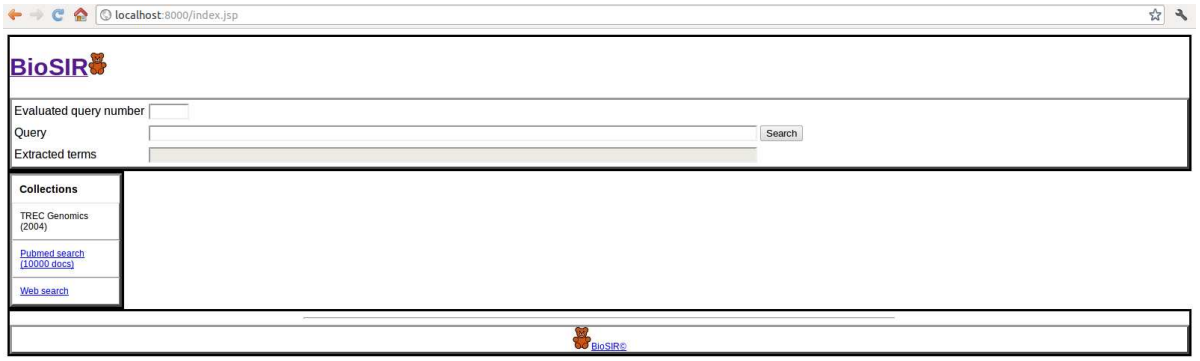


FIGURE VII.8 – Interface Web de BioSIR (accessible en Intranet)

traits ou non. La case à cocher "Allow non-positive score" permet de filtrer les concepts ayant un score non positif. La case à cocher "Exact matching" permet de retenir uniquement les termes ayant une meilleure corrélation de l'ordre de mots ($\rho = 1$, sélectionnée uniquement pour la méthode d'extraction de concepts basée sur la mesure de corrélation de Spearman).

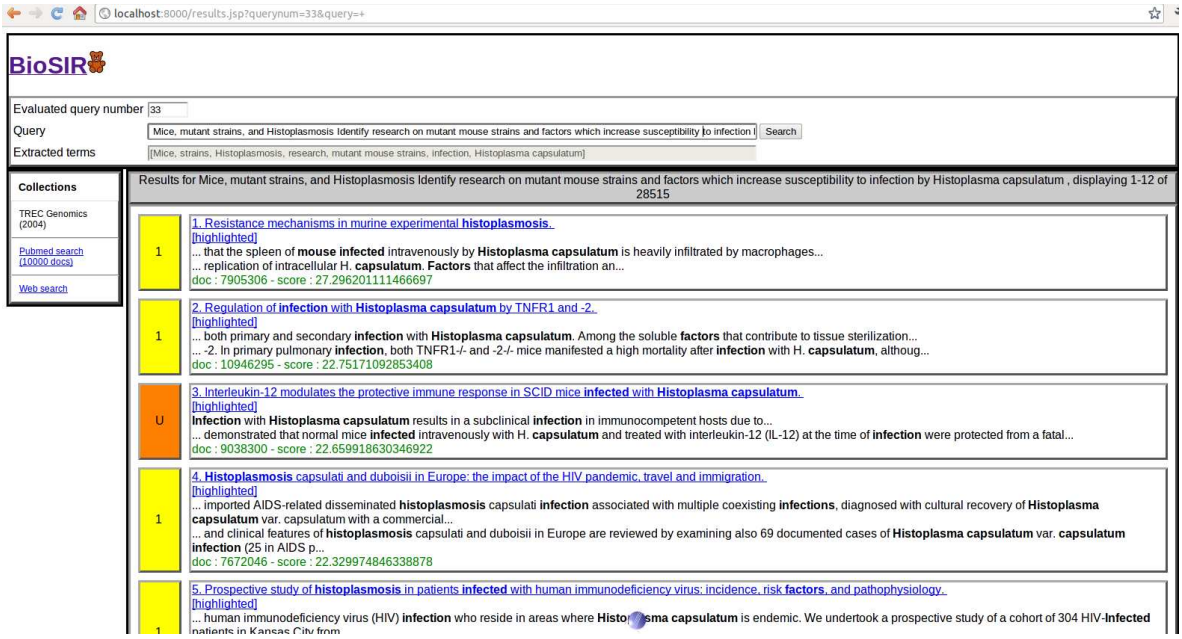


FIGURE VII.9 – Résultats de recherche avec BioSIR

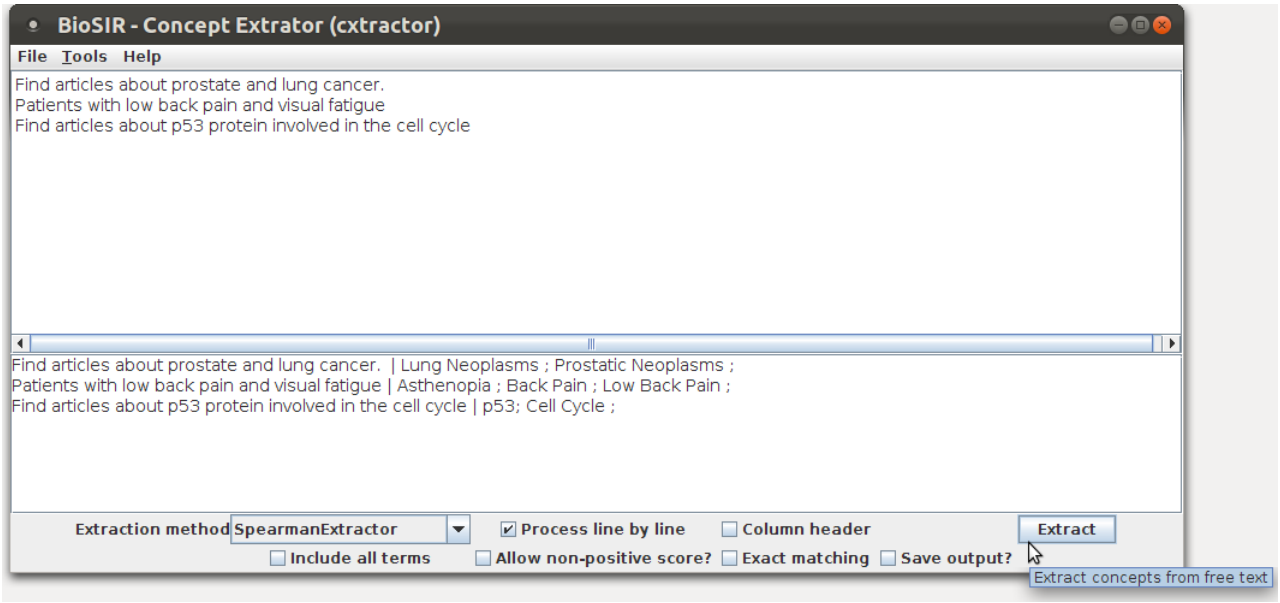


FIGURE VII.10 – BioSIR : Expansion de la requête par des concepts biomédicaux, noms de gènes ou de protéines

5 Outils d'évaluation

En mode graphique

L'interface graphique dans la figure VII.11 permet de calculer les mesures d'évaluation qui sont définies par la campagne d'évaluation TREC et puis de les afficher dans une fenêtre. L'avantage de cette interface est de pouvoir visualiser directement les valeurs liées aux performances de la RI. De plus, nous pouvons également évaluer les résultats qui sont enregistrés dans différents fichiers dans un ou plusieurs répertoires. En spécifiant le nom du répertoire racine contenant les résultats, l'outil permet de calculer les mesures d'évaluation pour chaque résultat et puis combiner les mesures sous forme de colonne ou tableau afin de faciliter le calcul des mesures statistiques comme les t-tests par exemple.

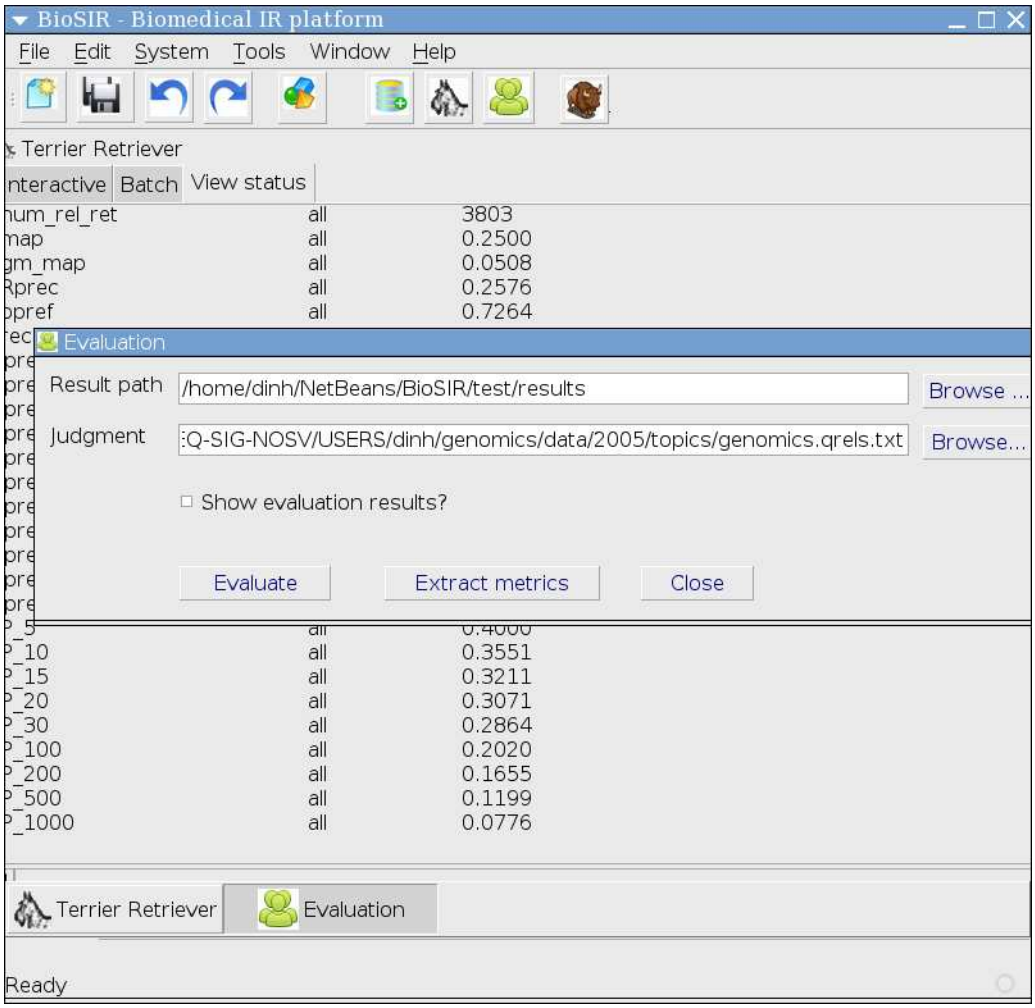


FIGURE VII.11 – BioSIR : Évaluation des performances de la RI

En mode ligne de commande

Le script Shell suivant, intitulé evaluation.sh [5], a pour objectif d'évaluer les performances des résultats qui sont sauvegardés dans plusieurs fichiers d'un répertoire donné. Chaque fichier résultat obtenu est donné au programme **trec_eval** pour calculer les mesures d'évaluation pour un ensemble de requêtes ou éventuellement pour chaque requête. L'intérêt de ce script porte sur la simplicité et la rapidité en spécifiant les arguments sur une ligne de commande. Étant donnée une liste de fichiers résultats, le calcul des mesures de performances s'effectue par le script evaluation.sh.

```
#!/bin/bash
# Duy Dinh - IRIT - University of Toulouse
# 24 March 2012
# Batch Evaluation using trec_eval program
#*********************************************************

args=$* # get all arguments

usage(){
  echo "Usage:"
  echo "sh $0 -d result_path -r qrels_path -e fileExt -t trec_eval [-p]"
  echo " -d result_path  directory containing search results"
  echo " -r qrels_path   query relevance judgements"
  echo " -e fileExt  results file extension, default value: 'res'"
  echo " -t trec_eval    trec_eval absolute file name"
  echo " -p      evaluate results per query"
  echo " -h      print usage information"
}

treceval="trec_eval"
resultsPath="$1" # directory containing results directories
qrelsPath="$2" # qrel file
fileExt="$3" #results' file extension
treceval="$4" # job directory containing trec_eval program
perQueryAnalysis="0"

# batch trec evaluation
evaluate_results(){
# search and evaluate all results under a given directory
  for result in `find ${resultsPath}* -iname "*.${fileExt}"`
  do
    echo "Evaluating file '$result'"
    "$treceval" -c -M1000 $qrelsPath $result > "$result.trec_eval"
```

5. sourceforge.net/p/irtoolkit/discussion/evaluation/thread/6ec6e485/?limit=250&page=0#a5f3

```
34      if [ "$perQueryAnalysis" == "1" ]
35      then
36        "$treceval" -c -M1000 -q $qrelsPath $result > "$result.trec_eval.q"
37          fi
38      done
39  }
40
41  # verify if a given directory exists, unless create a new one
42  verifyDirectory(){
43    local dirName=$1
44    if !(test -d "$dirName")
45    then
46      mkdir "$dirName"
47    fi
48  }
49
50  # main procedure
51  main(){
52    # parse options
53    while getopts ":d:e:r:t:hp" args; do
54    case $args in
55      d)
56        resultsPath=$OPTARG;;
57      h)
58        usage
59        exit;;
60      e)
61        fileExt=$OPTARG;;
62      r)
63        qrelsPath=$OPTARG;;
64      t)
65        treceval=$OPTARG;;
66      p)
67        perQueryAnalysis=1;;
68    esac
69    done
70
71    if !(test -d "$resultsPath")
72    then
73      echo "ERROR: Input directory '$resultsPath' does not exist. "
74          echo "Please make sure that input directory is valid!"
75      usage
76      exit
77    fi
78
79    if !(test -e "$qrelsPath")
80    then
81      echo "ERROR: Relevance judgements are not found at location '$qrelsPath'"
```

```
    usage
    exit
  fi

  if !(test -e "$treceval")
  then
    echo "ERROR: trec_eval is not found at location '$treceval'"
    usage
    exit
  fi

  if [ "$fileExt" == "" ]
  then
    fileExt="res"
  fi

  evaluate_results
}

# call the main program with some arguments
if (test $# -gt 0)
then
    main $args
else
    usage
fi
```

Le script R suivant permet de visualiser la distribution des valeurs de plusieurs groupes de données obtenue par la méthode d'estimation par noyau de Parzen-Rozenblatt (Bowman et Azzalini, 1997). Ceci permet d'avoir un aperçu sur les performances de chaque requête en comparant les valeurs de performances (e.g., MAP) obtenues par plusieurs méthodes ou approches de RI (appelées *runs*).

```
#-----------------------------------------------------------------------
# Display kernel distribution of MAP values obtained by different runs
# for each query
# @author: Duy Dinh
# @date: 22 March 2012
#-----------------------------------------------------------------------
library(sm)

year=2005
# IR model
model="BM25"
#model="LGD"
```

```
outputImage=TRUE

# result directory
path = ".../collection/TRECGenomics"
suffix="pools/TIABMH/results-TIABMH/R"

path=paste(path, year, sep="-")
path=paste(path, suffix, sep="/")

# get file names from directory
fileFilter=paste("* - ", model, "$", sep="")
# full names
files = list.files(path, full.names=TRUE, pattern=fileFilter)
# short names
names = list.files(path, pattern=fileFilter)

# output image
outputImg = paste(path, model, sep="/")
outputImg = paste(outputImg, "-Genomics-", year, sep="")
outputImg = paste(outputImg, "eps", sep=".")

# start graphics drawing
if (outputImage == TRUE){
  postscript(file=outputImg, horizontal=FALSE, width=10, height=5)
  #png(filename=outputImg)
}

xAxisLabel = "MAP"
yAxisLabel = "Density"

xAxisRange = c(0.0, 1.0)
yAxisRange = c(0.0, 5.0)
n=length(files)

# compares the density of multiple groups
kernel.density.compare <- function (groups, title){
  size=length(groups)
  if (size > 0){
      plot(density(groups[[1]]),
          xlim=xAxisRange, ylim=yAxisRange, ylab=yAxisLabel,
          lty=lineTypes[1], col=lineColors[1],
          main = title)
      for (i in 1 : size){
        lines(density(groups[[i]]),
                xlim=xAxisRange, ylim=yAxisRange, lty=lineTypes[i], col=lineColors[i])
      }
  }
}
```

```
61
62 if (n > 0){
63
64   # distribute plots
65   par(mfrow=c(1,n))
66
67   for (i in  1 : n){
68
69     # read MAP values from a csv file (header in the first line)
70     table = read.table(files[i], header=TRUE)
71
72     # run names
73     runNames = c(model, "RI", "MaxMatcher", "PubMed ATM",
74         "Spearman appr.", "Spearman exact")
75
76     # line color
77     lineColors = c("red","brown", "orange", "green","blue", "black")
78
79     # line types
80     lt=length(runNames)+2
81     lineTypes=c(2:lt)
82     groups=list(table$Baseline,table$RI, table$MaxMatcher, table$PubMed,
83                 table$Spa, table$Spe)
84     kernel.density.compare(groups, names[i])
85
86 # Add a legend for which line types, colors represent appropriately the
87 # corresponding runs
88     legend("topright", runNames, lty=lineTypes, col=lineColors, bty="n")
89   }
90 }
91
92 if (outputImage == TRUE){
93   dev.off()
94 }
```

Conclusion générale

"Science is wonderfully equipped to answer the question 'How ?' but it gets terribly confused when you ask the question 'Why ?' "
–Erwin Chargaff

Synthèse

Les travaux présentés dans cette thèse se situent dans le contexte général de la recherche d'information et plus particulièrement dans le cadre de la RI biomédicale. Nous y avons présenté plusieurs approches de RI conceptuelle/sémantique en se basant sur des ressources termino-ontologiques. Dans ce large contexte, nous avons posé plusieurs questions de recherche :

1. Quels sont les granules d'information liés aux concepts dans une ou des ressources termino-ontologiques qu'on peut exploiter pour améliorer les performances de la RI biomédicale ?
2. Est-ce que les algorithmes d'extraction de concepts permettent d'améliorer les performances d'un système de recherche d'information ? Si oui, dans quelles conditions ?
3. Quel est l'intérêt de l'intégration d'une ou de plusieurs terminologies dans un processus de RI biomédicale et comment faire pour les exploiter de manière efficace ?

Pour répondre à ces questions, nous avons proposé des contributions liées à l'indexation et à la recherche d'information basées sur l'utilisation des ressources termino-ontologiques. Puis, nous avons mené plusieurs expérimentations pour montrer l'efficacité de nos solutions proposées. Nos contributions présentées dans cette thèse ont porté sur trois volets principaux : (1) la résolution de l'ambiguïté (désambiguïsation) des termes MeSH orientée domaine et son impact sur un processus de RI, (2) l'extraction de concepts basée sur la pertinence et la corrélation des contextes documentaires et terminologiques et (3) une approche de RI biomédicale multi-terminologique basée sur la fusion des concepts biomédicaux issus des ressources termino-ontologiques.

1. Concernant la résolution de l'ambiguïté dans les textes biomédicaux, nous avons proposé deux méthodes de désambiguïsation de termes ambigus en termes de domaines biomédicaux définis dans la terminologie MeSH. Les termes désambiguïsés sont associés à des domaines appropriés afin de traduire leur similarité sémantique avec le texte ainsi que leur spécificité dans la description sémantique du document et de la requête. En se basant sur nos méthodes de désambiguïsation, nous avons proposé et évalué notre approche d'indexation conceptuelle/sémantique basée

sur le sens des termes désignant les concepts biomédicaux. Le sens du terme ambigu est reflété par son propre domaine dans l'architecture poly-hiérarchique de la terminologie MeSH. De plus, notre modèle d'indexation et d'appariement conceptuel/sémantique proposé prend en compte l'adéquation du sens des termes ambigus ainsi que leur spécificité dans le document et dans la requête.

Bien que les concepts (les plus spécifiques) soient importants, il ne faut pas les isoler de leur contexte, c-à-d le document ou la requête. Cela signifie qu'il ne faut pas se baser uniquement sur les concepts pour représenter les contenus textuels car, les autres mots jouant le rôle du contexte sémantique sont aussi importants dans la description du document et de la requête. Nous avons donc combiné la représentation textuelle des mots simples du document (resp. de la requête) et la représentation conceptuelle des termes préférés désignant les concepts biomédicaux identifiés du document (resp. de la requête).

2. Concernant l'évaluation des méthodes d'extraction de concepts biomédicaux pour la RI biomédicale, nous avons proposé une nouvelle méthode d'extraction de concepts qui est essentiellement basée sur la combinaison de deux mesures de similarité : thématique et structurelle. La similarité thématique entre deux instances textuelles peut être calculée par la mesure Cosinus. Concernant la similarité structurelle, nous avons choisi la mesure de corrélation de Spearman permettant de traduire la corrélation d'ordre de mots communs entre un morceau de texte (e.g., document ou requête) et un concept dans la ressource termino-ontologique. Dans le cadre la recherche d'information biomédicale, notamment TREC Genomics, notre algorithme d'extraction de concepts issus du thésaurus MeSH est capable d'identifier également les acronymes et les abréviations qui sont définis dans l'ontologie de gènes OMIM. Les concepts MeSH permettent de représenter les sujets sémantiques du texte tandis que l'ontologie OMIM permet d'identifier la plupart des acronymes de gènes et de protéines.

   Nous avons évalué l'efficacité de notre méthode d'extraction de concepts sur les documents ainsi que sur les requêtes. Les résultats montrent que les concepts extraits sont utiles pour améliorer les performances de la RI via l'expansion documentaire et/ou l'expansion conceptuelle de requêtes. Cependant, le nombre de concepts extraits est un paramètre expérimental. Nous n'avons pas testé ce paramètre sur plusieurs collections pour vérifier la stabilité ainsi que la portabilité de notre méthode d'extraction de concepts pour la RI biomédicale. En effet, le nombre de concepts extraits à partir du document ou de la requête peut être varié en fonction de la nature de la collection (e.g., littérature biomédicale, dossiers patients, ...) ou de la tâche (e.g., essai clique, recherche d'information, ...).

3. Concernant l'indexation et la recherche d'information sémantique basée sur l'utilisation d'une ou de plusieurs ressources termino-ontologiques, notre approche de RI utilise deux méthodes d'extraction de concepts différentes : l'une consiste à extraire les concepts qui sont prédéfinis dans une terminologie uniquement tandis que l'autre vise à extraire et fusionner les concepts issus de plusieurs terminologies. Ces concepts sont d'abord extraits en utilisant chaque terminologie séparément et sont fusionnés par la suite en utilisant les algorithmes de vote de l'état-de-l'art qui ont été utilisés en RI. Nous exploitons les concepts ainsi identifiés comme un moyen pour représenter les sujets sémantiques du document via l'ajout des termes préférés désignant les concepts biomédicaux dans le contenu textuel des documents. Dans un contexte global (i.e., terminologies biomédicales), nous supposons que les concepts extraits permettent de "normaliser" les mots-clés utilisés dans le document via les termes préférés désignant les concepts. De cette manière, les problèmes liés à la synonymie, à l'utilisation des abréviations ou des acronymes dans le document ou dans la requête sont automatiquement résolus.

   Les différents techniques de fusion de données ou modèles de vote ont été utilisés pour fusionner les concepts candidats issus de multiples ressources termino-ontologiques du domaine biomédical. Sur les collections TREC Genomics, les résultats obtenus par notre approche de RI mono-terminologique (qui est basée sur la combinaison de l'expansion de documents et la reformulation de requêtes) dépassent ceux qui sont obtenus par l'approche de RI classique (qui est basée uniquement sur la reformulation de requêtes). Lorsque les techniques de fusion sont appliquées pour sélectionner les meilleurs concepts, les résultats de notre approche de RI multi-terminologique ont donné de meilleurs résultats que la baseline ainsi que l'approche mono-terminologique.

   À travers les expérimentations menées dans le cadre de la RI biomédicale sur les collections TREC Genomics, nous avons pu identifier les facteurs les plus pertinents ayant un impact significatif sur l'efficacité de la recherche biomédicale IR, à savoir (1) le choix du modèle de pondération des termes, (2) le modèle d'expansion de requêtes en utilisant un nombre approprié de termes extraits à partir d'un ensemble de meilleurs documents retournés par le modèle de pondération de termes et (3) l'expansion de documents avec une quinzaine de termes désignant des concepts du domaine issus d'une terminologie unique ou de plusieurs terminologies. La prise en compte de ces facteurs dans les modèles d'appariement sémantique permettent d'améliorer les performances de la RI biomédicale.

Perspectives

Nos contributions dans le cadre de la RI biomédicale peuvent bénéficier de plusieurs perspectives pour nos futurs travaux sur le court terme ainsi que sur le long terme :

1. Perspectives à court terme

Sur le court terme, nos perspectives portent essentiellement sur les volets suivants :

1. Indexation et recherche d'information de dossiers médicaux : cette tâche consiste à traiter les informations spécifiques aux patients comprenant des données structurées, semi-structurées ou narratives portant sur des faits observables, données factuelles, historique des patients, décisions médicales contenues dans les comptes-rendus (CR) de consultation, CR d'anatomie pathologique, CR opératoire, CR d'imagerie etc. Nous envisageons de mener des expérimentations portant sur les points clés suivants :

   (a) Représentation des dossiers patients comme des sous-collections de documents : l'idée est de représenter un dossier patient médical comme un résumé sémantique du profil du patient défini par les documents ou compte-rendus qui le constituent. Avec cette représentation, nous pouvons définir un nouveau schéma de pondération sémantique qui tient compte de la distribution des mots-clés/termes dans le dossier patient par rapport à la distribution des mots-clés/termes désignant les concepts dans la collection de dossiers de patients. De plus, ce schéma de pondération peut intégrer et combiner plusieurs facteurs liés aux concepts comme la spécificité des concepts dans les documents et leur centralité dans les dossiers patients.

   (b) Proposition des modèles d'indexation et de recherche d'information multi-terminologique : du fait que les concepts peuvent être définis dans une ou plusieurs terminologies, ce serait intéressant de tenir compte de la distribution des concepts dans chaque terminologie ainsi que dans toutes les terminologies afin de sélectionner les concepts les plus liés aux contenus textuels. Dans notre approche de RI multi-terminologique, nous avons sélectionné manuellement les terminologies (en l'occurrence MeSH, SNOMED, GO) pour extraire les concepts et indexer les documents. Nous envisageons de sélectionner et d'intégrer automatiquement les terminologies les plus

adéquates qui sont liées au vocabulaire de la collection. Pour cela, nous pouvons définir des mesures de similarité entre la collection de documents et chacune des terminologies considérées et puis les terminologies les plus similaires à la collection seront retenues.

2. Amélioration de nos algorithmes de vote pour la sélection de concepts à partir de plusieurs terminologies en exploitant les relations sémantiques dans le réseau de concepts de l'UMLS. En effet, du fait que le réseau sémantique de l'UMLS fournit une catégorisation cohérente de tous les concepts représentés dans le méta-thésaurus UMLS ainsi qu'un ensemble de relations entre concepts, nous pouvons en exploiter pour proposer un schéma de pondération de concepts qui tient compte des poids, des rangs et des relations sémantiques dans un processus de vote. L'algorithme de vote amélioré doit prendre en compte non seulement les différentes relations sémantiques des concepts dans une terminologie mais aussi dans d'autres terminologies grâce à des informations dans les tables d'associations (mappings) de l'UMLS.

2. Perspectives à long terme

Sur le long terme, nous envisageons de mettre en place une plateforme de recherche d'information biomédicale qui donne accès à différents types d'information biomédicale, y compris la littérature scientifique et les dossiers médicaux de patients anonymisés en garantissant la confidentialité des patients. Cette plateforme est basée sur les caractéristiques suivantes :

1. Les modèles d'appariement flexible document-requête qui considère :
   (a) les différents types d'utilisateurs qui émettent la requête, par exemple, les médecins, les professionnels de santé, etc.
   (b) le type de documents, par exemple, les résumés d'articles, les textes intégraux, les comptes-rendus médicaux, les dossiers patients médicaux, etc.
   (c) la nature de la requête en utilisant les patrons de besoins cliniques dans un modèle PICO [6].
2. un mécanisme d'accès à l'information biomédicale pertinente et sécurisée : les données sensibles comme le nom, le prénom, les coordonnées liées aux patients doivent être encryptées ou supprimées afin de protéger les patients. Cela permettrait de faire des analyses et statistiques sur une population de patients ou un groupe de patients afin de faciliter la recherche menée par les chercheurs dans des essais cliniques.

6. P : Patient - I : Intervention - C : Control - O : Outcome

Annexe A

Ensemble de 50 requêtes de TREC Genomics 2004 [7]

1 Ferroportin-1 in humans. Find articles about Ferroportin-1, an iron transporter, in humans.
2 Generating transgenic mice. Find protocols for generating transgenic mice.
3 Time course for gene expression in mouse kidney. What is the time course of gene expression in the murine developing kidney.
4 Gene expression profiles for kidney in mice. What mouse genes are specific to the kidney.
5 Protocols for isolating cell nuclei. Articles are relevant if they describe methods for subcellular fractionation of nuclei.
6 FancD2. Find articles about function of FancD2.
7 DNA repair and oxidative stress. Find correlation between DNA repair pathways and oxidative stress.
8 Correlation between DNA repair pathways and skin cancer. Genes and proteins (pathways) common to DNA repair, oxidative diseases, skin-carcinogenesis, and UV-carcinogenesis.
9 mutY. Find articles about the function of mutY in humans.
10 NEIL1. Find articles about the role of NEIL1 in repair of DNA.
11 Carcinogenesis and hairless mice. Find articles regarding carcinogenesis induced in hairless mice.
12 Genes regulated by Smad4. Find articles describing genes that are regulated by the signal transducing molecule Smad4.
13 Role of TGFB in angiogenesis in skin. Documents regarding the role of TGFB in angiogenesis in skin with respect to homeostasis and development.
14 Expression or Regulation of TGFB in HNSCC cancers. Documents regarding TGFB expression or regulation in HNSCC cancers.
15 ATPase and apoptosis. Find information on role of ATPases in apoptosi.
16 AAA proteins. How do AAA proteins mediate interaction with lipids or DNA and what is their functional impact.
17 DO1 antibody. Determine binding affinity of anti-p53 monoclonal antibody DO1.
18 Gis4. Properties of Gis4 with respect to cell cycle and/or metabolism.

7. http://sourceforge.net/p/irtoolkit/discussion/evaluation/thread/e7239f30/

19 Comparison of Promoters of GAL1 and SUC1. What similarities and differences exist between the upstream promoter regions of GAL1 and SUC1 ? Are there co-repressors or co-activators ? If so, are they regulated by SNF1.
20 Substrate modification by ubiquitin. Which biological processes are regulated by having constituent proteins modified by covalent attachment to ubiquitin or ubiquitin-like proteins.
21 Role of p63 and p73 in relation to DNA damage. Do p63 and p73 cause cell cycle arrest or apoptosis related to DNA damage.
22 Relative response of p53 family members to agents causing single-stranded versus double-stranded DNA breaks. Does p53 respond differently to different DNA-damaging agents ? Do they respond differently to single-strand versus double-strand breaks.
23 Saccharomyces cerevisiae proteins involved in ubiquitin system. Which Saccharomyces cerevisiae proteins are involved in the ubiquitin proteolytic pathway.
24 Mouse peptidoglycan recognition proteins (PGRP). Find all reports describing mouse peptidoglycan recognition proteins (PGRP).
25 Cause of scleroderma. Identify studies that include genome-wide scans and microarray analysis in the investigation of scleroderma.
26 Function of BUB2/BFA1 in the process of cytokinesis. Retrieval of information regarding the role of BUB2 and BFA1 in cytokinesis in yeast.
27 Role of autophagy in apoptosis. Experiments establishing positive or negative interconnection between autophagy and apoptosis.
28 Proteases that function in both apoptosis and autophagy cell death. Studies that investigate similarities in morphological changes among apoptosis and autophagy processes.
29 Phenotypes of gyrA mutations. Documents containing the sequences and phenotypes of E. coli gyrA mutations.
30 Regulatory targets of the Nkx gene family members. Documents identifying genes regulated by Nkx gene family members.
31 TOR signaling in neurofibromatosis. Reports that provide possible links between neurofibromatosis and TOR signaling.
32 Xenograft animal models of tumorogenesis. Find reports that describe xenograft models of human cancers.
33 Mice, mutant strains, and Histoplasmosis. Identify research on mutant mouse strains and factors which increase susceptibility to infection by Histoplasma capsulatum.
34 Gene products of Cryptococcus important to fungal survival. Articles reporting experiments allowing annotation of gene products of Cryptococcus.
35 WD40 repeat-containing proteins. What is the function of proteins containing WD40 repeats.
36 RAB3A. Background information on RAB3A.
37 PAM. What research is being done on peptide amidating enzyme, PAM.
38 Risk factors for stroke. Information concerning genetic loci that are associated with increased risk of stroke, such as apolipoprotein E4 or factor V mutations.
39 Hypertension. Identify genes as potential genetic risk factors candidates for causing hypertension.

40 Antigens expressed by lung epithelial cells. To identify the antigens expressed by lung epithelial cells and the antibodies available.
41 Mutations in the Cystic Fibrosis conductance regulator gene. What phenotypes have been described resulting from mutations in the Cystic Fibrosis conductance regulator gene.
42 Genes altered by chromosome translocations. What genes show altered behavior due to chromosomal rearrangements.
43 Sleeping Beauty. Studies of Sleeping Beauty transposons.
44 Proteins involved in the nerve growth factor pathway. Create a list of all the nerve growth factor pathway proteins.
45 Mental Health Wellness-1. What genetic loci, such as Mental Health Wellness 1 (MWH1) are implicated in mental health.
46 RSK2. What human biological processes is RSK2 known to be involved in.
47 Human gene BCL-2 antagonists and inhibitors. Research the human gene BCL-2 to determine if there are antagonists and inhibitors inside of a cell.
48 Human homologues of C. elegans UNC genes. What is the focus of studies involving the members of the human UNC gene family.
49 Glyphosate tolerance gene sequence. Find reports and glyphosate tolerance gene sequences in the literature.
50 Low temperature protein expression in E. coli. Find research on improving protein expressions at low temperature in Escherichia coli bacteria.

Concepts identifiés par le service ATM de PubMed à partir de 6 premières requêtes dans TREC Genomics 2004 [8]

Les termes mis en gras sont identifiés et annotés par le service ATM de PubMed. La barre verticale ('|') les sépare de la requête originale.

1 Ferroportin-1 in humans Find articles about Ferroportin-1, an iron transporter, in humans | **Ferroportin-1**[All Fields] AND (**"iron"**[MeSH Terms] OR "**iron**"[All Fields]) AND ("**membrane transport proteins**"[MeSH Terms] OR ("**membrane**"[All Fields] AND "**transport**"[All Fields] AND "**proteins**"[All Fields]) OR "**membrane transport proteins**"[All Fields] OR "**transporter**"[All Fields]) AND ("**humans**"[MeSH Terms] OR "**humans**"[All Fields])

2 Generating transgenic mice Find protocols for generating transgenic mice | ("**Nat Protoc**"[Journal] OR "**CSH Protoc**"[Journal] OR "**protocols**"[All Fields]) AND **generating**[All Fields] AND ("**mice, transgenic**"[MeSH Terms] OR ("**mice**"[All Fields] AND "**transgenic**"[All Fields]) OR "**transgenic mice**"[All Fields] OR ("**transgenic**"[All Fields] AND "**mice**"[All Fields]))

8. La liste complète est ici `http://sourceforge.net/p/irtoolkit/discussion/evaluation/thread/e7239f30/`

3 Time course for gene expression in mouse kidney What is the time course of gene expression in the murine developing kidney | ("**time**"[MeSH Terms] OR "**time**"[All Fields]) AND **course**[All Fields] AND ("**gene expression**"[MeSH Terms] OR ("**gene**"[All Fields] AND "**expression**"[All Fields]) OR "**gene expression**"[All Fields]) AND ("**mice**"[MeSH Terms] OR "**mice**"[All Fields] OR "**mouse**"[All Fields]) AND ("**kidney**"[MeSH Terms] OR "**kidney**"[All Fields]) AND ("**time**"[MeSH Terms] OR "**time**"[All Fields]) AND **course**[All Fields] AND ("**gene expression**"[MeSH Terms] OR ("**gene**"[All Fields] AND "**expression**"[All Fields]) OR "**gene expression**"[All Fields]) AND ("**mice**"[MeSH Terms] OR "**mice**"[All Fields] OR "**murine**"[All Fields]) AND **developing**[All Fields] AND ("**kidney**"[MeSH Terms] OR "**kidney**"[All Fields])

4 Gene expression profiles for kidney in mice What mouse genes are specific to the kidney | ("**gene expression profiling**"[MeSH Terms] OR ("**gene**"[All Fields] AND "**expression**"[All Fields] AND "**profiling**"[All Fields]) OR "**gene expression profiling**"[All Fields] OR ("**gene**"[All Fields] AND "**expression**"[All Fields] AND "**profiles**"[All Fields]) OR "**gene expression profiles**"[All Fields]) AND ("**kidney**"[MeSH Terms] OR "**kidney**"[All Fields]) AND ("**mice**"[MeSH Terms] OR "**mice**"[All Fields]) AND ("**mice**"[MeSH Terms] OR "**mice**"[All Fields] OR "**mouse**"[All Fields]) AND ("**genes**"[MeSH Terms] OR "**genes**"[All Fields]) AND **specific**[All Fields] AND ("**Kidney**"[Journal] OR ("**the**"[All Fields] AND "**kidney**"[All Fields]) OR "**the kidney**"[All Fields])

5 Protocols for isolating cell nuclei Articles are relevant if they describe methods for subcellular fractionation of nuclei | ("**Nat Protoc**"[Journal] OR "**CSH Protoc**"[Journal] OR "**protocols**"[All Fields]) AND **isolating**[All Fields] AND ("**cell nucleus**"[MeSH Terms] OR ("**cell**"[All Fields] AND "**nucleus**"[All Fields]) OR "**cell nucleus**"[All Fields] OR ("**cell**"[All Fields] AND "**nuclei**"[All Fields]) OR "**cell nuclei**"[All Fields]) AND **Articles**[All Fields] AND **relevant**[All Fields] AND **describe**[All Fields] AND ("**methods**"[Subheading] OR "**methods**"[All Fields] OR "**methods**"[MeSH Terms]) AND **subcellular**[All Fields] AND ("**dose fractionation**"[MeSH Terms] OR ("**dose**"[All Fields] AND "**fractionation**"[All Fields]) OR "**dose fractionation**"[All Fields] OR "**fractionation**"[All Fields] OR "**chemical fractionation**"[MeSH Terms] OR ("**chemical**"[All Fields] AND "**fractionation**"[All Fields]) OR "**chemical fractionation**"[All Fields]) AND **nuclei**[All Fields]

6 FancD2 Find articles about function of FancD2 | **FancD2**[All Fields] AND **Find**[All Fields] AND **articles**[All Fields] AND ("**physiology**"[Subheading] OR "**physiology**"[All Fields] OR "**function**"[All Fields] OR "**physiology**"[MeSH Terms] OR "**function**"[All Fields]) AND **FancD2**[All Fields]

Liste des 10 premières requêtes dans TREC Genomics 2004 qui sont étendues par les termes préférés de l'UMLS identifiés par l'outil MetaMap [9]

1 Ferroportin-1 in humans. Find articles about Ferroportin-1, an iron transporter, in humans. | **SLC40A1 gene ; Homo sapiens ; Finding ; Article ; SLC40A1 gene ; Iron ; Membrane Transport Proteins ; Dietary Iron ; Membrane Transport Proteins ; Ferrum metallicum, Homeopathic preparation ; Membrane Transport Proteins ; Homo sapiens ;**

2 Generating transgenic mice. Find protocols for generating transgenic mice. | **Generation (action) ; Mice, Transgenic ; Finding ; Protocols documentation ; Generation (action) ; Mice, Transgenic ;**

3 Time course for gene expression in mouse kidney. What is the time course of gene expression in the murine developing kidney. | **Time course ; Gene Expression ; Mouse Kidney ; Time course ; Gene Expression ; Mus ; Kidney ; Mus ; Both kidneys ; Mus ; Entire kidney ; Murine ; Kidney ; Murine ; Both kidneys ; Murine ; Entire kidney ;**

4 Gene expression profiles for kidney in mice. What mouse genes are specific to the kidney. | **Gene Expression Profiles ; Kidney ; Both kidneys ; Entire kidney ; House mice ; Laboratory mice ; Mus ; MICE gene ; Specific qualifier value ; Entity Determiner - specific ; Kidney ; Both kidneys ;**

5 Protocols for isolating cell nuclei. Articles are relevant if they describe methods for subcellular fractionation of nuclei. | **Protocols documentation ; isolate - substance ; Cell Nucleus ; Article ; Relevance ; described ; Methods ; Methodology ; Techniques ; subcellular fractionation ; Cell Nucleus ;**

6 FancD2. Find articles about function of FancD2. | **FANCD2 gene ; FANCONI ANEMIA, COMPLEMENTATION GROUP D2 ; Finding ; Article ; physiological aspects ; Function ; Function Axis ; Mathematical Operator ; FANCD2 gene ; FANCONI ANEMIA**

7 DNA repair and oxidative stress. Find correlation between DNA repair pathways and oxidative stress. | **DNA Repair ; Oxidative Stress ; Oxidative Stress Analysis ; Finding ; Correlation ; DNA Repair Pathway ; Oxidative Stress ; Oxidative Stress Analysis ;**

8 Correlation between DNA repair pathways and skin cancer. Genes and proteins (pathways) common to DNA repair, oxidative diseases, skin-carcinogenesis, and UV-carcinogenesis. | **Correlation ; DNA Repair Pathway ; Malignant neoplasm of skin ; Genes ; Proteins ; Biochemical Pathway ; Pathway (interactions) ; Common (qualifier value) ; shared attribute ; DNA Repair ; Oxidative ; Disease ; Skin Carcinogenesis ; Ultraviolet Rays ; Carcinogenesis ; Microvolt ; Carcinogenesis ;**

9 mutY. Find articles about the function of mutY in humans. | **Finding ; Article ; physiological aspects ; Function ; Function Axis ; Mathematical Operator ; Homo sapiens ;**

10 NEIL1. Find articles about the role of NEIL1 in repair of DNA. | **NEIL1 gene ; Finding ; Article ; Social Role ; Generic Role ; NEIL1 gene ; Wound Healing ; Surgical repair ; Repair - Remedial Action ; DNA ;**

9. La liste complète est ici `http://sourceforge.net/p/irtoolkit/discussion/evaluation/thread/e7239f30/`

Liste des 7 premières requêtes dans TREC Genomics 2004 qui sont étendues par les termes préférés de MeSH identifiés par l'outil MTI avec la configuration par défaut (avec un filtrage basique) [10]

1 Ferroportin-1 in humans. Find articles about Ferroportin-1, an iron transporter, in humans. | **Iron ; Formates ; Bicarbonates ; Methanol ; Carbon Dioxide ; Formate Dehydrogenases ; Acidosis ; Kinetics ; Macaca ; Aldehyde Oxidoreductases ; Macaca mulatta ; Ribulose-Bisphosphate Carboxylase ; Hydrogen-Ion Concentration ; Oxygen ; Vitreous Body ; Eye Diseases ; Alcaligenes ; Pseudomonas ; Chromatography, Gas ; Haplorhini ; Ion Transport ; NAD ; Carboxy-Lyases ; Half-Life ; Thermodynamics ; Oxalates ; Pyrazoles ; Renal Dialysis ; Electron Transport Complex IV ; Time Factors ; Disease Models, Animal ; Drug Compounding ; Species Specificity ; Medication Errors ; Bile ; Immunodiffusion ;**

2 Generating transgenic mice. Find protocols for generating transgenic mice. | **Mice, Transgenic ; NADP ; Phosphorus Isotopes ; NAD ; Nicotinamide Mononucleotide ; Adenosine Monophosphate ; Molecular Conformation ; Binding Sites ; Nucleotides ; Structure-Activity Relationship ; Flavin-Adenine Dinucleotide ; Coenzymes ; Fourier Analysis ; Isocitrate Dehydrogenase ; Oxidation-Reduction ; Magnetic Resonance Spectroscopy ; Guanine Nucleotides ; Deuterium ; Flavin Mononucleotide ; Models, Molecular ; Hydrogen-Ion Concentration ; Protein Binding ; Kinetics ; Substrate Specificity ; Niacinamide ; Adenosine ; Catalysis ; Nucleic Acid Conformation ; L-Lactate Dehydrogenase ; Hydrogen ; Amino Acid Substitution ; Mathematics ; Hydrolysis ; Ribose ; Protein Structure, Secondary ; Phosphorus ; Mutagenesis, Site-Directed ; Transferases ; Muscles ; N-Glycosyl Hydrolases ; Models, Structural ; Myocardium ; Glycoside Hydrolases ; Guanosine ; Multienzyme Complexes ; Models, Biological ; Time Factors ; Histidine ; Aldehyde Dehydrogenase ; Recombinant Proteins ; Pyridines ; Aza Compounds ; Chickens ; Spleen ; Computers ;**

10. La liste complète est ici `http://sourceforge.net/p/irtoolkit/discussion/evaluation/thread/e7239f30/`

3 Time course for gene expression in mouse kidney. What is the time course of gene expression in the murine developing kidney. | **Carbonic Anhydrase Inhibitors ; Carbonic Anhydrases ; Binding Sites ; Protein Binding ; Zinc ; Cadmium ; Metals ; Cobalt ; Hydrogen-Ion Concentration ; Acetazolamide ; Carbon Isotopes ; Mercury ; Protein Conformation ; Metallothionein ; Copper ; Kinetics ; Magnetic Resonance Spectroscopy ; Histidine ; Apoenzymes ; Protons ; Nitrogen Isotopes ; Chromium ; Vanadium ; Phenanthrolines ; Spectrophotometry, Ultraviolet ; Azides ; Cyanides ; Structure-Activity Relationship ; Ligands ; Deuterium ; Potassium ; Spectrophotometry ; Electron Spin Resonance Spectroscopy ; Amino Acid Sequence ; Molecular Weight ; Crystallization ; Isotopes ; Sulfonamides ; Spectrometry, Fluorescence ; Superoxide Dismutase ; Iodides ; Dialysis ; Amides ; Chlorides ; Gene Expression ; Chemical Phenomena ; Tryptophan ; Chromatography, Ion Exchange ; Tosyl Compounds ; Isoenzymes ; Imidazoles ; Temperature ; Acetates ; Erythrocytes ; Freezing ; Insulin ; Chromatography, DEAE-Cellulose ; Organogenesis ; Mathematics ;**

4 Gene expression profiles for kidney in mice. What mouse genes are specific to the kidney. | **Cerebroside-Sulfatase ; Leukodystrophy, Metachromatic ; Sulfatases ; Arylsulfatases ; Lysosomes ; Mucolipidoses ; Glucuronidase ; Mucopolysaccharidoses ; Cells, Cultured ; Kidney ; Cathepsin D ; Fibroblasts ; Pinocytosis ; Galactosylceramidase ; Endopeptidases ; beta-Galactosidase ; Cathepsins ; Glycosaminoglycans ; Sphingomyelin Phosphodiesterase ; Neuraminidase ; Chondro-4-Sulfatase ; Electrophoresis, Cellulose Acetate ; Retroviridae ; Gene Expression ; Endocytosis ; Chorionic Villi ; Transfection ; Cysteine Endopeptidases ; Ammonium Chloride ; Organoids ; Hexosaminidases ; Histocytochemistry ; Phenotype ; Subcellular Fractions ; Genetic Vectors ; Substrate Specificity ; Placenta ; Metabolism, Inborn Errors ; Cerebrosides ; Enzyme Activation ; Kinetics ; Amniotic Fluid ; Consanguinity ; Staining and Labeling ; Receptors, Drug ; Chloroquine ; Hydrogen-Ion Concentration ; Leukocytes ; Isoelectric Focusing ; Mutation ; Age Factors ; Catechols ; Chromatography, Thin Layer ; Ferrocyanides ; Chromatography, DEAE-Cellulose ; Pregnancy Complications ;**

5 Protocols for isolating cell nuclei. Articles are relevant if they describe methods for subcellular fractionation of nuclei. | **Hemerythrin ; Metalloproteins ; Protein Conformation ; Binding Sites ; Protein Binding ; X-Ray Diffraction ; Models, Molecular ; Annelida ; Macromolecular Substances ; Amino Acid Sequence ; Invertebrates ; Software ; Structure-Activity Relationship ; Circular Dichroism ; Muscle Proteins ; Amino Acids ; Crystallization ; Computer Simulation ; Chymotrypsin ; Cyanogen Bromide ; Cysteine ; Scattering, Radiation ; Trypsin ; Cnidaria ; Electrons ; Oxidation-Reduction ; Sulfhydryl Reagents ; Computers ; Iron ; Kinetics ; Solvents ; Muscles ; Sulfhydryl Compounds ; Peptide Fragments ; Mercury ; X-Rays ; Spectrophotometry, Ultraviolet ; Fourier Analysis ; Time Factors ; Nematoda ; Seawater ;**

6 FancD2. Find articles about function of FancD2. | **Oxyhemoglobins ; Hemoglobins ; Protein Binding ; Binding Sites ; Oxygen ; Diphosphoglyceric Acids ; Ligands ; Iron ; Spectrophotometry ; Cobalt ; Electron Spin Resonance Spectroscopy ; Heme ; Hydrogen-Ion Concentration ; Kinetics ; Myoglobin ; Spectrophotometry, Ultraviolet ; Protein Conformation ; Carbon Dioxide ; Spin Labels ; Allosteric Regulation ; Porphyrins ; Spectrophotometry, Infrared ; Structure-Activity Relationship ; Erythrocytes ; Circular Dichroism ; Molecular Conformation ; Macromolecular Substances ; Temperature ; Allosteric Site ; Oxidation-Reduction ; Apoproteins ; Half-Life ; Fluorine ; Thermodynamics ; Time Factors ; Amino Acid Sequence ; Cyanides ; Protein Multimerization ; Cyclic N-Oxides ; Histidine ; Potentiometry ; Radioisotopes ; Models, Chemical ; Fluorides ; Mathematics ; Sulfhydryl Compounds ; Drug Stability ; Phytic Acid ; Cysteine ; Phosphates ; Chromatography, Gel ; Chromatography ; Silicon Dioxide ; Azides ; Tritium ; Models, Biological ; Models, Structural ; Cetacea ; Species Specificity ; Hybridization, Genetic ;**

7 DNA repair and oxidative stress. Find correlation between DNA repair pathways and oxidative stress. | **Oxidative Stress ; Adrenal Medulla ; Dopamine beta-Hydroxylase ; Catecholamines ; Tyrosine 3-Monooxygenase ; Epinephrine ; Morphine ; Insulin ; Adrenal Glands ; DNA Repair ; Reserpine ; Naloxone ; Alkaloids ; Tyramine ; Dihydroxyphenylalanine ; Dopa Decarboxylase ; Tyrosine ; Substance Withdrawal Syndrome ; Methadone ; Locus Coeruleus ; Fenclonine ; Stimulation, Chemical ; Serotonin ; Animals, Newborn ; Nicotine ; Carbon Radioisotopes ; Metaraminol ; Substance-Related Disorders ; Tritium ; Monoamine Oxidase ; Dose-Response Relationship, Drug ; Urinalysis ; Splanchnic Nerves ; Aging ; Enzyme Induction ; Organ Size ; Maternal-Fetal Exchange ; Nervous System ; Hypertension ; Kidney ; Rats, Inbred Strains ; Subcellular Fractions ; Brain ; Adenosine Triphosphate ; Inclusion Bodies ; Pyridines ; Kinetics ; Denervation ; Time Factors ; Cell Membrane ; Indoles ; Body Weight ; Drug Synergism ; Pregnancy Complications, Cardiovascular ;**

Liste des 13 premières requêtes dans TREC Genomics 2004 qui sont étendues par les termes préférés de MeSH identifiés par l'outil MTI avec un filtrage moyen [11]

1 Ferroportin-1 in humans. Find articles about Ferroportin-1, an iron transporter, in humans. | **Iron ; Cation Transport Proteins ; Ion Transport ;**

2 Generating transgenic mice. Find protocols for generating transgenic mice. | **Mice, Transgenic ;**

3 Time course for gene expression in mouse kidney. What is the time course of gene expression in the murine developing kidney. | **Peroxides ; Urea ; Gene Expression ; Organogenesis ;**

4 Gene expression profiles for kidney in mice. What mouse genes are specific to the kidney. | **Gene Expression Profiling ; Kidney ;**

5 Protocols for isolating cell nuclei. Articles are relevant if they describe methods for subcellular fractionation of nuclei. | **Cell Nucleus ; Cell Fractionation ;**

6 FancD2. Find articles about function of FancD2. | **Fanconi Anemia Complementation Group D2 Protein ; Fanconi Anemia ; DNA Repair ; Nuclear Proteins ; DNA Damage ; Cell Cycle ; Protein Binding ;**

7 DNA repair and oxidative stress. Find correlation between DNA repair pathways and oxidative stress. | **Oxidative Stress ; DNA Repair ; DNA Damage ; Reactive Oxygen Species ; Guanine ; DNA, Mitochondrial ; Urinalysis ;**

8 Correlation between DNA repair pathways and skin cancer. Genes and proteins (pathways) common to DNA repair, oxidative diseases, skin-carcinogenesis, and UV-carcinogenesis. | **DNA Repair ; Skin Neoplasms ; Cell Transformation, Neoplastic ; DNA Damage ; Proteins ; Ultraviolet Rays ; Oxidation-Reduction ; Mutation ;**

11. La liste complète est ici `http://sourceforge.net/p/irtoolkit/discussion/evaluation/thread/e7239f30/`

9 mutY. Find articles about the function of mutY in humans. | **DNA Glycosylases ; N-Glycosyl Hydrolases ; DNA Repair ; Escherichia coli ; Molecular Sequence Data ; Amino Acid Sequence ; Mutation ; Sequence Homology, Amino Acid ;**

10 NEIL1. Find articles about the role of NEIL1 in repair of DNA. | **DNA ; DNA Glycosylases ; DNA Repair ; DNA-(Apurinic or Apyrimidinic Site) Lyase ; DNA Damage ; Wound Healing ; Protein Binding ;**

11 Carcinogenesis and hairless mice. Find articles regarding carcinogenesis induced in hairless mice. | **Mice, Hairless ; Cell Transformation, Neoplastic ;**

12 Genes regulated by Smad4. Find articles describing genes that are regulated by the signal transducing molecule Smad4. | **Signal Transduction ; Smad4 Protein ; Transforming Growth Factor beta ; Trans-Activators ; DNA-Binding Proteins ; Cyclin-Dependent Kinase Inhibitor p21 ; Cyclins ; Cell Nucleus ; Genes, Tumor Suppressor ; Cell Division ; Pancreatic Neoplasms ;**

13 Role of TGFB in angiogenesis in skin. Documents regarding the role of TGFB in angiogenesis in skin with respect to homeostasis and development. | **Homeostasis ; Skin ; Physiological Processes ; Morphogenesis ; Transforming Growth Factor beta ; Cells, Cultured ;**

Liste des 20 premières requêtes dans TREC Genomics 2004 qui sont étendues par les termes préférés de MeSH identifiés par l'outil MTI avec un filtrage strict [12]

Les requêtes 6, 9, 13, 14, 15, 16, 17, 18 ne sont pas étendues car aucun concept n'est identifié par MTI.

1 Ferroportin-1 in humans. Find articles about Ferroportin-1, an iron transporter, in humans. | **Iron ;**

2 Generating transgenic mice. Find protocols for generating transgenic mice. | **Mice, Transgenic ;**

3 Time course for gene expression in mouse kidney. What is the time course of gene expression in the murine developing kidney. | **Gene Expression ;**

4 Gene expression profiles for kidney in mice. What mouse genes are specific to the kidney. | **Gene Expression Profiling ; Kidney ;**

5 Protocols for isolating cell nuclei. Articles are relevant if they describe methods for subcellular fractionation of nuclei. | **Cell Nucleus ;**

6 **FancD2. Find articles about function of FancD2.** |

7 DNA repair and oxidative stress. Find correlation between DNA repair pathways and oxidative stress. | **Oxidative Stress ; DNA Repair ;**

8 Correlation between DNA repair pathways and skin cancer. Genes and proteins common to DNA repair, oxidative diseases, skin-carcinogenesis, and UV-carcinogenesis. | **DNA Repair ; Skin Neoplasms ; Cell Transformation ;**

9 **mutY. Find articles about the function of mutY in humans.** |

12. La liste complète est ici `http://sourceforge.net/p/irtoolkit/discussion/evaluation/thread/e7239f30/`

10 NEIL1. Find articles about the role of NEIL1 in repair of DNA. | **DNA ;**
11 Carcinogenesis and hairless mice. Find articles regarding carcinogenesis induced in hairless mice. | **Mice, Hairless ;**
12 Genes regulated by Smad4. Find articles describing genes that are regulated by the signal transducing molecule Smad4. | **Signal Transduction ;**
13 **Role of TGFB in angiogenesis in skin. Documents regarding the role of TGFB in angiogenesis in skin w.r.t homeostasis and development.|**
14 **Expression or Regulation of TGFB in HNSCC cancers. Documents regarding TGFB expression or regulation in HNSCC cancers. |**
15 **ATPase & apoptosis. Find information on role of ATPases in apoptosi.|**
16 **AAA proteins. How do AAA proteins mediate interaction with lipids or DNA and what is their functional impact. |**
17 **DO1 antibody. Determine binding affinity of anti-p53 monoclonal antibody DO1. |**
18 **Gis4. Properties of Gis4 w.r.t cell cycle and/or metabolism. |**
19 Comparison of Promoters of GAL1 and SUC1. Similarities and differences exist between the upstream promoter regions of GAL1 and SUC1 ? Co-repressors or co-activators ? Are they regulated by SNF1. | **Co-Repressor Proteins ;**
20 Substrate modification by ubiquitin. Which biological processes are regulated by having constituent proteins modified by covalent attachment to ubiquitin or ubiquitin-like proteins. | **Ubiquitins ;**

Liste des 20 premières requêtes de TREC Genomics 2004 étendues par les termes MeSH donnés par MaxMatcher

1 Ferroportin-1 in humans. Find articles about Ferroportin-1, an iron transporter, in humans. | **Humans ; Transportation ; Iron ;**
2 Generating transgenic mice. Find protocols for generating transgenic mice. | **Family Characteristics ;**
3 Time course for gene expression (GE) in mouse kidney. What is the time course of [GE] in the murine developing kidney. | **Time ; Mice ; Kidney ; [GE] ;**
4 Gene expression profiles for kidney in mice. What mouse genes are specific to the kidney. | **Genes ; Mice ; Gene Expression ; Kidney ;**
5 Protocols for isolating cell nuclei. Articles are relevant if they describe methods for subcellular fractionation of nuclei. | **Methods ; Cell Separation ; Subcellular Fractions ;**
6 FancD2. Find articles about function of FancD2. |
7 DNA repair and oxidative stress. Find correlation between DNA repair pathways and oxidative stress. | **DNA Repair ; Oxidative Stress ;**
8 Correlation between DNA repair pathways and skin cancer. Genes and proteins pathways common to DNA repair, oxidative diseases, skin-carcinogenesis, and UV-carcinogenesis. | **Genes ; Disease ; Proteins ; Skin ; DNA Repair ; Skin Neoplasms ;**
9 mutY. Find articles about the function of mutY in humans. | **Humans ;**
10 NEIL1. Find articles about the role of NEIL1 in repair of DNA. | **DNA ; Role ;**

11 Carcinogenesis and hairless mice. Find articles regarding carcinogenesis induced in hairless mice. |
12 Genes regulated by Smad4. Find articles describing genes that are regulated by the signal transducing molecule Smad4. | **Genes ;**
13 Role of TGFB in angiogenesis in skin. Documents regarding the role of TGFB in angiogenesis in skin with respect to homeostasis and development. | **Homeostasis ; Documentation ; Skin ; Role ;**
14 Expression or Regulation of TGFB in HNSCC cancers. Documents regarding TGFB expression or regulation in HNSCC cancers. | **Neoplasms ; Documentation ;**
15 ATPase and apoptosis. Find information on role of ATPases in apoptosi. | **Apoptosis ; Role ;**
16 AAA proteins. How do AAA proteins mediate interaction with lipids or DNA and what is their functional impact. | **DNA ; Lipids ; Proteins ;**
17 DO1 antibody. Determine binding affinity of anti-p53 monoclonal antibody DO1. | **Antibodies ;**
18 Gis4. Properties of Gis4 w.r.t cell cycle (CC), metabolism | **Metabolism ; [CC]**
19 Comparison of Promoters of GAL1 & SUC1. Similarities & differences exist between upstream promoter regions of GAL1 & SUC1. Are there co-repressors or co-activators. Are they regulated by SNF1. | **Co-Repressor Proteins ;**
20 Substrate modification by ubiquitin. Which biological processes are regulated by having constituent proteins modified by covalent attachment to ubiquitin or ubiquitin-like proteins. | **Ubiquitination ; Proteins ; Ubiquitins ; Biological Processes ;**

Liste des 18 premières requêtes de TREC Genomics 2004 étendues par les termes UMLS donnés par MaxMatcher

1 Ferroportin-1 in humans. Find articles about Ferroportin-1, an iron transporter, in humans. | **humans ; iron transporter ; Ferroportin-1** ;
2 Generating transgenic mice. Find protocols for generating transgenic mice. | **transgenic mice ; protocols ;**
3 Time course for gene expression in mouse kidney. What is the time course of gene expression in the murine developing kidney. | **kidney ; gene expression ; time course ; mouse ;**
4 Gene expression profiles for kidney in mice. What mouse genes are specific to the kidney. | **kidney ; mouse genes ; mice ; Gene expression profiles ;**
5 Protocols for isolating cell nuclei. Articles are relevant if they describe methods for subcellular fractionation of nuclei. | **nuclei ; subcellular fractionation ; Protocols ; methods ;**
6 FancD2. Find articles about function of FancD2. | **FancD2 ; function ;**
7 DNA repair and oxidative stress. Find correlation between DNA repair pathways and oxidative stress. | **oxidative stress ; pathways ; DNA repair ; correlation ;**

8 Correlation between DNA repair pathways and skin cancer. Genes and proteins pathways common to DNA repair, oxidative diseases, skin-carcinogenesis, and UV-carcinogenesis. | **carcinogenesis ; UV ; skin ; diseases ; DNA repair ; pathways ; proteins ; Genes ; skin cancer ; Correlation ;**

9 mutY. Find articles about the function of mutY in humans. | **humans ; mutY ; function ;**

10 NEIL1. Find articles about the role of NEIL1 in repair of DNA. | **DNA ; repair ; NEIL1 ; role ;**

11 Carcinogenesis and hairless mice. Find articles regarding carcinogenesis induced in hairless mice. | **hairless mice ; carcinogenesis ;**

12 Genes regulated by Smad4. Find articles describing genes that are regulated by the signal transducing molecule Smad4. | **Smad4 ; molecule ; signal ; genes ;**

13 Role of TGFB in angiogenesis in skin. Documents regarding the role of TGFB in angiogenesis in skin with respect to homeostasis and development. | **development ; homeostasis ; respect ; skin ; angiogenesis ; TGFB ; role ; Documents ;**

14 Expression or Regulation of TGFB in HNSCC cancers. Documents regarding TGFB expression or regulation in HNSCC cancers. | **HNSCC cancers ; regulation ; expression ; TGFB ; Documents ;**

15 ATPase and apoptosis. Find information on role of ATPases in apoptosi. | **ATPases ; role ; information ; apoptosis ;**

16 AAA proteins. How do AAA proteins mediate interaction with lipids or DNA and what is their functional impact. | **impact ; DNA ; lipids ; interaction ; AAA proteins ;**

17 DO1 antibody. Determine binding affinity of anti-p53 monoclonal antibody DO1. | **DO1 ; monoclonal antibody ; anti-p53 ; affinity ;**

18 Gis4. Properties of Gis4 with respect to cell cycle and or metabolism. | **metabolism ; cell cycle ; respect ; Gis4 ; Properties ;**

Liste complètes des 50 requêtes de TREC Genomics 2004 étendues par les termes MeSH identifiés par notre méthode d'extraction de concepts basée sur la corrélation de Spearman, dénotée Spe-MeSH.

1 Ferroportin-1 in humans. Find articles about Ferroportin-1, an iron transporter, in humans. |

2 Generating transgenic mice. Find protocols for generating transgenic mice. |

3 Time course for gene expression in mouse kidney. What is the time course of gene expression in the murine developing kidney. | **Gene Expression ;**

4 Gene expression profiles for kidney in mice. What mouse genes are specific to the kidney. | **Gene Expression ; Gene Expression ;**

5 Protocols for isolating cell nuclei. Articles are relevant if they describe methods for subcellular fractionation of nuclei. | **Cell Nucleus ; Subcellular Fractions ;**

6 FancD2. Find articles about function of FancD2. |

7 DNA repair and oxidative stress. Find correlation between DNA repair pathways and oxidative stress. | **DNA Repair ; Oxidative Stress ;**

8 Correlation between DNA repair pathways and skin cancer. Genes and proteins pathways common to DNA repair, oxidative diseases, skin-carcinogenesis, and UV-carcinogenesis. | **Skin Neoplasms ; DNA Repair ;**

9 mutY. Find articles about the function of mutY in humans. |

10 NEIL1. Find articles about the role of NEIL1 in repair of DNA. |

11 Carcinogenesis and hairless mice. Find articles regarding carcinogenesis induced in hairless mice. |

12 Genes regulated by Smad4. Find articles describing genes that are regulated by the signal transducing molecule Smad4. |

13 Role of TGFB in angiogenesis in skin. Documents regarding the role of TGFB in angiogenesis in skin with respect to homeostasis and development. |

14 Expression or Regulation of TGFB in HNSCC cancers. Documents regarding TGFB expression or regulation in HNSCC cancers. |

15 ATPase and apoptosis. Find information on role of ATPases in apoptosi. |

16 AAA proteins. How do AAA proteins mediate interaction with lipids or DNA and what is their functional impact. |

17 DO1 antibody. Determine binding affinity of anti-p53 monoclonal antibody DO1. |

18 Gis4. Properties of Gis4 with respect to cell cycle and or metabolism. | **Cell Cycle ;**

19 Comparison of Promoters of GAL1 and SUC1. What similarities and differences exist between the upstream promoter regions of GAL1 and SUC1 Are there co-repressors or co-activators If so, are they regulated by SNF1. |

20 Substrate modification by ubiquitin. Which biological processes are regulated by having constituent proteins modified by covalent attachment to ubiquitin or ubiquitin-like proteins. | **Biological Processes ; Ubiquitins ;**

21 Role of p63 and p73 in relation to DNA damage. Do p63 and p73 cause cell cycle arrest or apoptosis related to DNA damage. | **Cell Cycle ; DNA Damage ;**

22 Relative response of p53 family members to agents causing single-stranded versus double-stranded DNA breaks. Does p53 respond differently to different DNA-damaging agents Do they respond differently to single-strand versus double-strand breaks. | **Family ; DNA Breaks ;**

23 Saccharomyces cerevisiae proteins involved in ubiquitin system. Which Saccharomyces cerevisiae proteins are involved in the ubiquitin proteolytic pathway. | **Saccharomyces cerevisiae ; Saccharomyces cerevisiae Proteins ;**

24 Mouse peptidoglycan recognition proteins PGRP. Find all reports describing mouse peptidoglycan recognition proteins PGRP. |

25 Cause of scleroderma. Identify studies that include genome-wide scans and microarray analysis in the investigation of scleroderma. | **Microarray Analysis ;**

26 Function of BUB2 BFA1 in the process of cytokinesis. Retrieval of information regarding the role of BUB2 and BFA1 in cytokinesis in yeast. |

27 Role of autophagy in apoptosis. Experiments establishing positive or negative interconnection between autophagy and apoptosis. |

28 Proteases that function in both apoptosis and autophagy cell death. Studies that investigate similarities in morphological changes among apoptosis and autophagy processes. | **Cell Death ;**

29 Phenotypes of gyrA mutations. Documents containing the sequences and phenotypes of E. coli gyrA mutations. |
30 Regulatory targets of the Nkx gene family members. Documents identifying genes regulated by Nkx gene family members. | **Family ;**
31 TOR signaling in neurofibromatosis. Reports that provide possible links between neurofibromatosis and TOR signaling. |
32 Xenograft animal models of tumorogenesis. Find reports that describe xenograft models of human cancers. |
33 Mice, mutant strains, and Histoplasmosis. Identify research on mutant mouse strains and factors which increase susceptibility to infection by Histoplasma capsulatum. | **Histoplasma ;**
34 Gene products of Cryptococcus important to fungal survival. Articles reporting experiments allowing annotation of gene products of Cryptococcus. |
35 WD40 repeat-containing proteins. What is the function of proteins containing WD40 repeats. |
36 RAB3A. Background information on RAB3A. |
37 PAM. What research is being done on peptide amidating enzyme, PAM. |
38 Risk factors for stroke. Information concerning genetic loci that are associated with increased risk of stroke, such as apolipoprotein E4 or factor V mutations. | **Apolipoprotein E4 ; Genetic Loci ;**
39 Hypertension. Identify genes as potential genetic risk factors candidates for causing hypertension. |
40 Antigens expressed by lung epithelial cells. To identify the antigens expressed by lung epithelial cells and the antibodies available. | **Epithelial Cells ;**
41 Mutations in the Cystic Fibrosis conductance regulator gene. What phenotypes have been described resulting from mutations in the Cystic Fibrosis conductance regulator gene. | **Cystic Fibrosis ;**
42 Genes altered by chromosome translocations. What genes show altered behavior due to chromosomal rearrangements. |
43 Sleeping Beauty. Studies of Sleeping Beauty transposons. |
44 Proteins involved in the nerve growth factor pathway. Create a list of all the nerve growth factor pathway proteins. | **Intercellular Peptides and Proteins ; Nerve Growth Factors ; Nerve Growth Factor ;**
45 Mental Health Wellness-1. What genetic loci, such as Mental Health Wellness 1 MWH1 are implicated in mental health. | **Genetic Loci ; Mental Health ;**
46 RSK2. What human biological processes is RSK2 known to be involved in. | **Biological Processes ;**
47 Human gene BCL-2 antagonists and inhibitors. Research the human gene BCL-2 to determine if there are antagonists and inhibitors inside of a cell. |
48 Human homologues of C. elegans UNC genes. What is the focus of studies involving the members of the human UNC gene family. |
49 Glyphosate tolerance gene sequence. Find reports and glyphosate tolerance gene sequences in the literature. |
50 Low temperature protein expression in E. coli. Find research on improving protein expressions at low temperature in Escherichia coli bacteria. | **Escherichia coli ;**

Liste complètes des 50 requêtes de TREC Genomics 2004 étendues par les termes MeSH identifiés par notre méthode d'extraction de concepts basée sur la corrélation de Spearman et les termes OMIM détectés par des expressions régulières, dénotée Spe-MeSH-OMIM.

1 Ferroportin-1 in humans. Find articles about Ferroportin-1, an iron transporter, in humans. |

2 Generating transgenic mice. Find protocols for generating transgenic mice. |

3 Time course for gene expression in mouse kidney. What is the time course of gene expression in the murine developing kidney. | **Gene Expression ;**

4 Gene expression profiles for kidney in mice. What mouse genes are specific to the kidney. | **Expression ; Expression ;**

5 Protocols for isolating cell nuclei. Articles are relevant if they describe methods for subcellular fractionation of nuclei. | **Cell Nucleus ; Subcellular Fractions ;**

6 FancD2. Find articles about function of FancD2. | **FANCONI ANEMIA, COMPLEMENTATION GROUP D2 ; FA4 ; FANCD ; FANCD2 ; ;**

7 DNA repair and oxidative stress. Find correlation between DNA repair pathways and oxidative stress. | **Repair ; Oxidative Stress ;**

8 Correlation between DNA repair pathways and skin cancer. Genes and proteins pathways common to DNA repair, oxidative diseases, skin-carcinogenesis, and UV-carcinogenesis. | **Skin Neoplasms ; Repair ;**

9 mutY. Find articles about the function of mutY in humans. |

10 NEIL1. Find articles about the role of NEIL1 in repair of DNA. | **ENDONUCLEASE VIII-LIKE 1 ; NEIL1 ; ;**

11 Carcinogenesis and hairless mice. Find articles regarding carcinogenesis induced in hairless mice. |

12 Genes regulated by Smad4. Find articles describing genes that are regulated by the signal transducing molecule Smad4. | **MOTHERS AGAINST DECAPENTAPLEGIC, DROSOPHILA, HOMOLOG OF, 4 ; DPC4 ; SMAD4 ; ;**

13 Role of TGFB in angiogenesis in skin. Documents regarding the role of TGFB in angiogenesis in skin with respect to homeostasis and development. | **TRANSFORMING GROWTH FACTOR, BETA-1 ; TGFB ; TGFB1 ; ;**

14 Expression or Regulation of TGFB in HNSCC cancers. Documents regarding TGFB expression or regulation in HNSCC cancers. | **TRANSFORMING GROWTH FACTOR, BETA-1 ; TGFB ; TGFB1 ; ; SQUAMOUS CELL CARCINOMA, HEAD AND NECK ; HNSCC ; ;**

15 ATPase and apoptosis. Find information on role of ATPases in apoptosi. |

16 AAA proteins. How do AAA proteins mediate interaction with lipids or DNA and what is their functional impact. | **ACHALASIA-ADDISONIANISM-ALACRIMA SYNDROME ; AAA ; ;**

17 DO1 antibody. Determine binding affinity of anti-p53 monoclonal antibody DO1. |

18 Gis4. Properties of Gis4 with respect to cell cycle and or metabolism. | **Cell Cycle ;**

19 Comparison of Promoters of GAL1 and SUC1. What similarities and differences exist between the upstream promoter regions of GAL1 and SUC1 Are there co-repressors or co-activators If so, are they regulated by SNF1. | **LECTIN, GALACTOSIDE-BINDING, SOLUBLE, 1 ; LGALS1 ; GAL1 ; ;**

20 Substrate modification by ubiquitin. Which biological processes are regulated by having constituent proteins modified by covalent attachment to ubiquitin or ubiquitin-like proteins. | **Biological Processes ; Ubiquitins ;**

21 Role of p63 and p73 in relation to DNA damage. Do p63 and p73 cause cell cycle arrest or apoptosis related to DNA damage. | **TUMOR PROTEIN p63 ; TP63 ; ; TUMOR PROTEIN p73 ; TP73 ; ; Cell Cycle ; Damage ;**

22 Relative response of p53 family members to agents causing single-stranded versus double-stranded DNA breaks. Does p53 respond differently to different DNA-damaging agents Do they respond differently to single-strand versus double-strand breaks. | **Family ; Breaks ;**

23 Saccharomyces cerevisiae proteins involved in ubiquitin system. Which Saccharomyces cerevisiae proteins are involved in the ubiquitin proteolytic pathway. | **; Proteins ;**

24 Mouse peptidoglycan recognition proteins PGRP. Find all reports describing mouse peptidoglycan recognition proteins PGRP. |

25 Cause of scleroderma. Identify studies that include genome-wide scans and microarray analysis in the investigation of scleroderma. | **Microarray Analysis ;**

26 Function of BUB2 BFA1 in the process of cytokinesis. Retrieval of information regarding the role of BUB2 and BFA1 in cytokinesis in yeast. |

27 Role of autophagy in apoptosis. Experiments establishing positive or negative interconnection between autophagy and apoptosis. |

28 Proteases that function in both apoptosis and autophagy cell death. Studies that investigate similarities in morphological changes among apoptosis and autophagy processes. | **Cell Death ;**

29 Phenotypes of gyrA mutations. Documents containing the sequences and phenotypes of E. coli gyrA mutations. | **E. coli ;**

30 Regulatory targets of the Nkx gene family members. Documents identifying genes regulated by Nkx gene family members. | **Family ;**

31 TOR signaling in neurofibromatosis. Reports that provide possible links between neurofibromatosis and TOR signaling. |

32 Xenograft animal models of tumorogenesis. Find reports that describe xenograft models of human cancers. |

33 Mice, mutant strains, and Histoplasmosis. Identify research on mutant mouse strains and factors which increase susceptibility to infection by Histoplasma capsulatum. |

34 Gene products of Cryptococcus important to fungal survival. Articles reporting experiments allowing annotation of gene products of Cryptococcus. |

35 WD40 repeat-containing proteins. What is the function of proteins containing WD40 repeats. |

36 RAB3A. Background information on RAB3A. | **RAS-ASSOCIATED PROTEIN RAB3A ; ;**

37 PAM. What research is being done on peptide amidating enzyme, PAM. | **MYC-BINDING PROTEIN 2 ; PAM ; ;**

38 Risk factors for stroke. Information concerning genetic loci that are associated with increased risk of stroke, such as apolipoprotein E4 or factor V mutations. | **Apolipoprotein ; Genetic Loci ;**

39 Hypertension. Identify genes as potential genetic risk factors candidates for causing hypertension. |

40 Antigens expressed by lung epithelial cells. To identify the antigens expressed by lung epithelial cells and the antibodies available. | **Epithelial Cells ;**

41 Mutations in the Cystic Fibrosis conductance regulator gene. What phenotypes have been described resulting from mutations in the Cystic Fibrosis conductance regulator gene. |

42 Genes altered by chromosome translocations. What genes show altered behavior due to chromosomal rearrangements. |

43 Sleeping Beauty. Studies of Sleeping Beauty transposons. |

44 Proteins involved in the nerve growth factor pathway. Create a list of all the nerve growth factor pathway proteins. | **Intercellular Peptides and ; Nerve Growth Factors ; Nerve Growth Factor ;**

45 Mental Health Wellness-1. What genetic loci, such as Mental Health Wellness 1 MWH1 are implicated in mental health. | **Genetic Loci ;**

46 RSK2. What human biological processes is RSK2 known to be involved in. | **RIBOSOMAL PROTEIN S6 KINASE, 90-KD, 3 ; RSK2 ; RPS6KA3 ; MRX19 ; ; Biological Processes ;**

47 Human gene BCL-2 antagonists and inhibitors. Research the human gene BCL-2 to determine if there are antagonists and inhibitors inside of a cell. |

48 Human homologues of C. elegans UNC genes. What is the focus of studies involving the members of the human UNC gene family. | **C. elegans ;**

49 Glyphosate tolerance gene sequence. Find reports and glyphosate tolerance gene sequences in the literature. |

50 Low temperature protein expression in E. coli. Find research on improving protein expressions at low temperature in Escherichia coli bacteria. | **E. coli ;**

Annexe B

Résultats officiels dans TREC 2011 (piste TRECMed)

	Manual runs				*Automatic runs*			
	Run	**bpref**	**P@10**	**Rprec**	**Run**	**bpref**	**P@10**	**Rprec**
1	NLMManual	0.658	0.727	0.500	CengageM11R3	0.552	0.656	0.440
2	buptpris01	0.474	0.547	0.342	SCAIMED7	0.552	0.603	0.425
3	**IRITm1QE1**	0.462	0.488	0.344	UTDHLTCIR	0.545	0.603	0.422
4	SCAIMED1	0.457	0.506	0.324	udelgn	0.522	0.544	0.407
5	UCDCSIrun3	0.456	0.459	0.324	WWOCorrect	0.494	0.415	0.306
6	mayolbrst	0.426	0.279	0.220	uogTrDeNIo	0.493	0.568	0.401
7	ohsuManAll	0.379	0.582	0.328	NICTA6	0.490	0.503	0.355
8	merckkgaamer	0.275	0.459	0.247	EssieAuto	0.482	0.497	0.337

Évaluation des résultats pour les meilleurs runs *manuels vs. automatiques* dans TRECMed 2011.

Résultats officiels dans CLEF 2011 (piste ImageCLEF, tâche Case-based retrieval)

	Run	**Group**	**MAP**	**bPref**	**P10**
1	UESTC_full_indri	UESTC	0.1297	0.1212	0.1889
2	HES-SO-VS_CASE_BASED_FULLTEXT	HES-SO-VS	0.1293	0.1122	0.2000
3	UESTC_full_p2QE	UESTC	0.1199	0.1082	0.1556
4	UESTC_full_p2	UESTC	0.1179	0.1162	0.1889
5	MRIM_KJ_A_VM_Sop_T4G	MRIM	0.1114	0.1064	0.1444
6	IRIT_LGDc1.0_KLbfree_1	**IRIT**	0.1030	0.0930	0.1556
7	IRIT_CombSUMc1.0_KLbfree_1	**IRIT**	0.0947	0.0862	0.1333
8	iti-essie-manual	ITI	0.0941	0.1162	0.1667
9	IRIT_LGDc1.0_KLbfree_1_ignore_low_idf	**IRIT**	0.0937	0.0716	0.1111
10	MRIM_KJ_A_VM_Pos_T4G	MRIM	0.0911	0.0938	0.1111
11	UESTC_full_okapi	UESTC	0.0907	0.0970	0.1444
12	IRIT_CombSUMc1.0_KLbfree_2	**IRIT**	0.0874	0.0710	0.1111

13	IRIT_LGDc1.0	**IRIT**	0.0872	0.0722	0.1111
14	IRIT_CombSUMc1.0_3	**IRIT**	0.0859	0.0783	0.1444
15	UESTC_ac_okapi	UESTC	0.0835	0.0734	0.1222
16	IRIT_In_expB2c1.0_KLbfree_ignore_low_idf	**IRIT**	0.0793	0.0707	0.1444
17	IRIT_In_expB2c1.0_KLbfree_0	**IRIT**	0.0772	0.0675	0.1000
18	UESTC_ac_indri	UESTC	0.0767	0.0669	0.1111
19	iti-lucene-baseline	ITI	0.0762	0.0737	0.1444
20	UESTC_full_okapi_fb	UESTC	0.0762	0.0841	0.1333
21	IRIT_In_expB2c1.0_1	**IRIT**	0.0743	0.073	0.1111
22	UESTC_ac_p2	UESTC	0.0722	0.0628	0.1222
23	IRIT_CombSUMc1.0_2_ignore_low_idf	**IRIT**	0.0721	0.0683	0.1333
24	UESTC_ac_p2QE	UESTC	0.0677	0.0633	0.1000
25	UESTC_ac_okapi_fb	UESTC	0.05	0.0484	0.0778
26	IPL2011CaseBasedT1-C6-M0_2-RO_01-BM25F-AVG	IPL	0.0463	0.0588	0.0889
27	IPL2011CaseBasedT1-C6-M0_2-BM25F-AVG	IPL	0.0461	0.0588	0.0889
28	HES-SO-VS_CASE_BASED_CAPTIONS	HES-SO-VS	0.0437	0.054	0.1111
29	iti-lucene-baseline+expanded-concepts	ITI	0.0264	0.0252	0.0333
30	iti-lucene-baseline+expanded-concepts+cases	ITI	0.0249	0.023	0.0333
31	iti-lucene-expanded-concepts	ITI	0.0243	0.0249	0.0333
32	IPL2011CaseBasedT1-C6-M0_2-BM25F-SUM	IPL	0.0201	0.0176	0.0333
33	IPL2011CaseBasedT1-C6-M0_2-RO_01-BM25F-SUM	IPL	0.0201	0.0174	0.0333
34	iti-essie-frames	ITI	0.0174	0.0333	0.0667
35	iti-lucene-frames	ITI	0.0141	0.0239	0.0667

Résultats officiels dans CLEF 2010 (piste ImageCLEF, tâche Case-based retrieval)

	Run	**Group**	**MAP**	**bPref**	**P10**
1	baselinefbWMR_10_0.2sub	UIUCIBM	0.2902	0.3049	0.4429
2	baselinefbWsub	UIUCIBM	0.2808	0.2816	0.4429
3	runfile_hes-so-vs_case-based_fulltext.txt	HES-SO VS	0.2796	0.2699	0.4214
4	baselinefbsub	UIUCIBM	0.2754	0.2856	0.4286
5	baselinefbWMD_25_0.2sub	UIUCIBM	0.2626	0.2731	0.4000
6	IRIT_SemAnnotator-2.0_BM25_N28.res	**IRIT**	0.2265	0.2351	0.3429
7	IRIT_SemAnnotator-2.0_BM25_N28_1.res	**IRIT**	0.2193	0.2139	0.3286

8	IRIT_SemAnnotator-1.5.2_BM25_N34.res	**IRIT**	0.2182	0.2267	0.3571
9	IRIT-run-bl.res	**IRIT**	0.2103	0.1885	0.2786
10	IRIT_SemAnnotator-1.5.2_BM25_N34_1.res	**IRIT**	0.2085	0.2083	0.3143
11	IRIT_SemAnnotator-2.0_BM25_N34_1.res	**IRIT**	0.2085	0.2083	0.3143
12	ISSR_cb_cts.txt	ISSR	0.1986	0.1883	0.3071
13	ISSR_cp_ctp.txt	ISSR	0.1977	0.1873	0.3000
14	ISSR_CB_CT.txt	ISSR	0.1977	0.1873	0.3000
15	ipl_aueb_CaseBased_CTM_0.2.txt	IPL	0.1874	0.1927	0.3214
16	ipl_aueb_CaseBased_CTM_0.1.txt	IPL	0.1860	0.1897	0.3214
17	ipl_aueb_CaseBased_CT.txt	IPL	0.1841	0.1803	0.3143
18	ipl_aueb_CaseBased_CTM_0.3.txt	IPL	0.1833	0.1919	0.3143
19	ipl_aueb_CaseBased_CTM_0.4.txt	IPL	0.1809	0.1895	0.3143
20	ipl_aueb_CaseBased_CTM_0.4.txt	IPL	0.1809	0.1895	0.3143
21	ipl_aueb_CaseBased_CTM_0.5.txt	IPL	0.1716	0.1811	0.3429
22	UESTC_case_pBasic.txt	UESTC	0.1692	0.1840	0.2643
23	UESTC_case_pQE.txt	UESTC	0.1677	0.1852	0.2786
24	UESTC_case_pNw.txt	UESTC	0.1522	0.1725	0.2714
25	case_based_expanded_queries_backoff_0.1.trec	ITI	0.1501	0.1749	0.2929
26	case_based_queries_backoff_0.1.trec	ITI	0.1280	0.1525	0.2357
27	runfile_hes-so-vs_case-based_nodoublon_captions.txt	HES-SO VS	0.1273	0.1375	0.2500
28	case_based_expanded_queries_types_0.1.trec	ITI	0.1217	0.1502	0.2929
29	case_based_queries_pico_MA_0.1.trec	ITI	0.1145	0.1439	0.2000
30	case_based_queries_types_0.1.trec	ITI	0.0996	0.1346	0.2286
31	case_based_queries_terms_0.1.trec	ITI	0.0522	0.0700	0.0857
32	GE_GIFT8_case.treceval	medGIFT	0.0358	0.0612	0.0929
33	PhybaselineRelfbWMR_10_0.2sub	UIUCIBM	0.3059	0.3348	0.4571
34	PhybaselineRelfbWMD_25_0.2sub	UIUCIBM	0.2837	0.3127	0.4571
35	PhybaselineRelFbWMR_10_0.2_top20sub	UIUCIBM	0.2713	0.2897	0.4286
36	case_based_queries_pico_backoff_0.1.trec	ITI	0.1386	0.1666	0.2000
37	PhybaselinefbWMR_10_0.2sub	UIUCIBM	0.3551	0.3714	0.4714
38	PhybaselinefbWsub	UIUCIBM	0.3441	0.3480	0.4714
39	PhybaselinefbWMD_25_0.2sub	UIUCIBM	0.3441	0.3480	0.4714
40	case_based_expanded_queries_terms_0.1.trec	ITI	0.0601	0.0825	0.0857
41	C_TA_T.lst	SINAI	0.2555	0.2518	0.3714
42	C_TA_TM.lst	SINAI	0.2201	0.2307	0.3643
43	C_TAbs_TM.lst	SINAI	0.1146	0.1661	0.2643
44	C_TAbs_T.lst	SINAI	0.1076	0.1660	0.2571

Bibliographie

Abdou, S. et Savoy, J. (2008). Searching in Medline : Query expansion and manual indexing evaluation. *Information Processing Management*, 44(2): 781–789.

Acr (2004). The Medstract Project – AcroMed 1.1. `http://medstract.med.tufts.edu/acro1.1/`.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.-C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P. et Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nat Biotech*, 24(5):537–544.

Agirre, E. et Rigau, G. (1996). Word sense disambiguation using conceptual density. *In International Conference on Computational Linguistics (COLING)*, pages 16–22.

Amati, G. (2003). *Probabilistic models for Information Retrieval based on Divergence from Randomness*. Thèse de doctorat, University of Glasgow.

Amati, G. et van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information System*, 20(4):357–389.

Ananiadou, S. (1994). A methodology for automatic term recognition. *In International Conference on Computational Linguistics (COLING)*, pages 1034–1038.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C. et Murzin, A. G. (2004). SCOP database in 2004 : refinements integrate structure and sequence family data. *Nucleic Acids Research*, pages 226–229.

Aronson, A. R. (1996). The effect of textual variation on concept based information retrieval. *In Proceedings of AMIA Symposium*, pages 373–377.

Aronson, A. R. (2001a). Effective mapping of biomedical text to the UMLS Metathesaurus : the MetaMap program. *In Proceedings AMIA Symposium*, pages 17–21.

Aronson, A. R. (2001b). MetaMap Evaluation. Rapport technique, US National Library Of Medicine.

Aronson, A. R., Humphrey, S. M., Ide, N. C., Kim, W., Loane, R. R., Mork, J. G., Smith, L. H., Tanabe, L. K., Wilbur, W. J., Xie, N., Demner-Fushman, D. et Liu, H. (2004a). Knowledge-Intensive and Statistical Approaches to the Retrieval and Annotation of Genomics MEDLINE Citations. *In TREC*.

Aronson, A. R., Mork, J. G., CW Gay, S. M. H. et Rogers, W. J. (2004b). The NLM Indexing Initiative's Medical Text Indexer. *In Medinfo 2004*, pages 268–272.

Aronson, A. R., Rindflesch, T. C., D, P. et D, P. (1997). Query expansion using the UMLS. *In Proceedings of AMIA Symposium*, pages 485–489.

Aronson, A. R., Thomas, R. C. et Allen, B. C. (1994). Exploiting a large thesaurus for information retrieval. pages 197–216.

Aussenac-Gilles, N., Biebow, B. et Szulman, S. (2000). Revisiting ontology design : A methodology based on corpus analysis. *In Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management*, EKAW '00, pages 172–188, London, UK, UK. Springer-Verlag.

Avillach, P., Joubert, M. et Fieschi, M. (2007). A Model for Indexing Medical Documents Combining Statistical and Symbolic Knowledge. *Proceedings AMIA Symposium*, pages 31–35.

Baeza-Yates, R. A. et Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

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

Belkin, N. J., Cool, C., Koenemann, J., Ng, K. B. et Park, S. (1995). Using Relevance Feedback and Ranking in Interactive Searching. *In TREC*.

Berners-Lee, T. et Cailliau, R. (1990). WorldWideWeb : Proposal for a HyperText Project. Rapport technique, European Laboratory for Particle Physics (CERN).

Bodenreider, O., Nelson, S. J., Hole, W. T. et Chang, H. F. (1998). Beyond synonymy : exploiting the UMLS semantics in mapping vocabularies. *Proceedings of AMIA*, pages 815–819.

Bordogna, G. et Pasi, G. (2000). Flexible querying of structured documents. *In FQAS'00*, pages 350–361.

Bordogna, G. et Pasi, G. (2001). Modeling Vagueness in Information Retrieval. *In* Agosti, M., Crestani, F. et Pasi, G., éditeurs : *Lectures on Information Retrieval*, volume 1980 de *Lecture Notes in Computer Science*, pages 207–241. Springer Berlin / Heidelberg.

Bosc, P. et Prade, H. (1996). An Introduction to the Fuzzy Set and Possibility Theory-Based Treatment of Soft Queries and Uncertain Or Imprecise Databases. *In Uncertainty Management in Information Systems*.

Boubekeur, F. (2008). *Contribution à la définition de modèles flexibles de recherche d'information basés sur les CP-Nets*. Thèse de doctorat, Université Paul Sabatier.

Boudin, F., Shi, L. et Nie, J.-Y. (2010). Improving Medical Information Retrieval with PICO Element Detection. *In Proceedings of the ECIR 2010 Conference*.

Boughanem, M. et Savoy, J., éditeurs (2008). *Recherche d'information états des lieux et perspectives*. Hermès Science Publications, http ://www.editions-hermes.fr/.

Bourigault, D., Aussenac-Gilles, N. et Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes Un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, pages 87–110.

Bowman, A. W. et Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis : The Kernel Approach with S-Plus Illustrations (Oxford Statistical Science Series)*. Oxford University Press, USA.

Bringay, S. (2006). *Les annotations pour supporter la collaboration dans le dossier patient électronique*. Thèse de doctorat, Université de Picardie Jules Verne - Amiens.

Bruce, R. et Wiebe, J. (1994). Word-Sense Disambiguation using Decomposable Models. *In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 139–146.

Buckley, C. et Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. *In SIGIR*, pages 25–32.

Buitelaar, P., Magnini, B., Strapparava, C. et Vossen, P. (2007). Domain specific word sense disambiguation, chapter 10. *In Word sense disambiguation : algorithms and applications*, pages 275–298.

Büttcher, S., Clarke, C. L. A. et Cormack, G. V. (2004). Domain-Specific Synonym Expansion and Validation for Biomedical Information Retrieval (MultiText Experiments for TREC 2004). *In* Voorhees, E. M. et Buckland, L. P., éditeurs : *TREC*, volume Special Publication 500-261. National Institute of Standards and Technology (NIST).

Calegari, S. et Pasi, G. (2011). Definition of User Profiles Based on the YAGO Ontology. *In* Melucci, M., Mizzaro, S. et Pasi, G., éditeurs : *IIR*, volume 704 de *CEUR Workshop Proceedings*. CEUR-WS.org.

Callan, J. P., Croft, W. B. et Harding, S. M. (1992). The INQUERY Retrieval System. *In Proceedings of DEXA-92, 3rd International Conference on Database and Expert Systems Applications*, pages 78–83.

Ceusters, W. et Smith, B. (2006). Strategies for referent tracking in electronic health records. *Journal of Biomedical Informatics*, 39(3):362–378.

Chapman, R. L., éditeur (1992). *Roget's International Thesaurus (5th edition)*. HarperCollins.

Chirita, P. A., Nejdl, W., Paiu, R. et Kohlschütter, C. (2005). Using odp metadata to personalize search. *In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 178–185, New York, NY, USA. ACM.

Claypool, M., Le, P., Wased, M. et Brown, D. (2001). Implicit interest indicators. pages 33–40.

Cleverdon, C. (1967). The Cranfield tests on index language devices. *In Aslib Proceedings*, pages 173–194.

Cleverdon, C. W. (1970). Progress in documentation. Evaluation of information retrieval systems. *Journal of Documentation*, 26:55–67.

Cleverdon, C. W. (1991). The significance of the Cranfield tests on index languages. *In Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '91, pages 3–12, New York, NY, USA. ACM.

Clinchant, S. et Gaussier, E. (2010). Information-based models for ad hoc IR. *In Proc. of Conference on Research and Development in Information Retrieval*, SIGIR'10, pages 234–241. ACM.

Collier, N., Nobata, C. et Tsujii, J.-i. (2000). Extracting the names of genes and gene products with a hidden Markov model. *In 18th international conference on Computational linguistics*, pages 201–207, Morristown, NJ, USA. Association for Computational Linguistics.

Cormack, G. V., Grossman, M. R., Hedin, B. et Oard, D. W. (2010). Overview of the TREC 2010 Legal Track. *In TREC 2010*.

Cornet, R. et de Keizer, N. (2008). Forty years of SNOMED : a literature review. *In Proc. of BMC Medical Informaticas and Decision Making*, pages 268–272.

Cortes, C. et Vapnik, V. (1995). Support-vector networks. *In Machine Learning*, pages 273–297.

Cowie, J. R. et Lehnert, W. G. (1996). Information Extraction. *Communications of the ACM*, 39(1):80–91.

Croft, W. et Harper, D. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, pages 285–295.

Croft, W. B. (1986). User-specified domain knowledge for document retrieval. *In Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '86, pages 201–206, New York, NY, USA. ACM.

Croft, W. B. et Harper, D. J. (1988). Document retrieval systems. chapitre Using probabilistic models of document retrieval without relevance information, pages 161–171. Taylor Graham Publishing, London, UK, UK.

Croft, W. B., Metzler, D. et Strohman, T. (2009). *Search Engines - Information Retrieval in Practice*. Pearson Education.

Cutting, D., Kupiec, J., Pedersen, J. et Sibun, P. (1992). A practical Part-of-Speech tagger. *In Proceedings of the 3rd conference on Applied Natural Language Processing*, pages 133–140.

Daoud, M. (2009). *Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche*. Thèse de doctorat, MITT.

Darmoni, S. J., Pereira, S., Sakji, S., Merabti, T., Prieur, É., Joubert, M. et Thirion, B. (2009). Multiple Terminologies in a Health Portal : Automatic Indexing and Information Retrieval. *In Artificial Intelligence in MEdicine (AIME)*, pages 255–259.

Delbecque, T. et Zweigenbaum., P. (2005). Indexation UMLS en français : une expérience. *In Journées francophones d'informatique médicale*.

Dinh, D. et Tamine, L. (2010a). Sense-Based Biomedical Indexing and Retrieval. *In NLDB*, pages 24–35.

Dinh, D. et Tamine, L. (2010b). Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients. *In CORIA*, pages 325–336.

Dinh, D. et Tamine, L. (2011a). Combining Global and Local Semantic Contexts for Improving Biomedical Information Retrieval. *In European Conference on Information Retrieval (ECIR)*, pages 375–386.

Dinh, D. et Tamine, L. (2011b). Voting Techniques for a Multi-terminology Based Biomedical Information Retrieval. *In Artificial Intelligence in MEdicine*, pages 184–193.

Dubois, D. et Prade, H. (1988). *Possibility theory : an approach to computerized processing of uncertainty*. Plenum press.

Duda, R. O., Hart, P. E. et Stork, D. G. (2001). *Pattern Classification*. John Wiley and Sons Inc.

Elkin, P., Cimino, J., Aronow, D., Payne, T. et Barnett, G. (1988). Mapping to MeSH. *In* Association, A. M. I., éditeur : *Proceedings of the 1988 Symposium on Computer Applications for Medical Care*, pages 185–190.

euG (2004). Genomic Information for Eukaryotic Organisms. `http://eugenes.org/`.

Eyre, T. A., Ducluzeau, F., Sneddon, T. P., Povey, S., Bruford, E. A. et Lush, M. J. (2006). The HUGO Gene Nomenclature Database. *Nucleic Acids Research*, 34(Database-Issue):319–321.

Fellbaum, C., éditeur (1998). *WordNet : An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, illustrated edition édition.

Foskett, D. J. (1997). Readings in information retrieval. chapitre Thesaurus, pages 111–134. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Fox, E. A. et Shaw, J. A. (1994). Combination of Multiple Searches. *In TREC 1994*, pages 243–252.

Francis, W. N. et Kucera, H. (1979). Brown Corpus Manual. Rapport technique, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Frantzi, K., Ananiadou, S. et Mima, H. (2000). Automatic recognition of multi-word terms : the C-value/NC-value method. *International Journal on Digital Libraries*, 3:115–130.

Fujita, S. (2004). Revisiting Again Document Length Hypotheses TREC 2004 Genomics Track Experiments at Patolis. *In TREC'04*.

Fukuda, K., Tsunoda, T., Tamura, A. et Takagi, T. (1998). Toward Information Extraction : Identifying protein names from biological papers. *In Proceedings of Pacific Symposium on Biocomputing*, pages 707–718.

Gaizauskas, R., Demetriou, G., Artymiuk, P. J. et Willett, P. (2003). Protein structures and information extraction from biological texts : the PASTA system. *Bioinformatics*, 19(1):135–143.

Gaizauskas, R., Demetriou, G. et Humphreys, K. (2000). Term Recognition and Classification in Biological Science Journal Articles. *In Proceedings of the Computional Terminology for Medical and Biological Applications Workshop of the 2nd International Conference on NLP*, pages 37–44.

Gale, W. A., Church, K. W. et Yarowsky, D. (1992). One sense per discourse. *In HLT '91 : Proceedings of the workshop on Speech and Natural Language*, pages 233–237.

Gaudinat, A., Boyer, C., Baujard, V. et Ruch, P. (2002). Evaluation de l'extraction de termes mesh pour les systèmes de recherche d'information dans le domaine médicale. *In Actes des 9ièmes Journées Francophones d'Informatique Médicale.*

Gliozzo, A., Magnini, B. et Strapparava, C. (2004). Usupervised domain relevance estimation for word sense disambiguation. *In Conference on empirical methods in natural language processing (EMNLP)*, pages 380–387.

Gobeill, J., Ruch, P. et Zhou, X. (2009). Query and document expansion with medical subject headings terms at medical Imageclef 2008. *In CLEF'08*, pages 736–743, Berlin, Heidelberg. Springer-Verlag.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5:199–220.

Haynes, R. B., McKibbon, K. A., Walker, C. J., Ryan, N., Fitzgerald, D. et Ramsden, M. F. (1990). Online access to MEDLINE in clinical settings. A study of use and usefulness. *Ann Intern Med*, 112(1):78–84.

Hersh, W. (2004). Who are the Informaticians ? What We Know and Should Know. *Journal of the American Medical Informatics Association : JAMIA*, 13(2):166–170.

Hersh, W. (2008). *Information Retrieval : A Health and Biomedical Perspective (Health Informatics).*

Hersh, W., Buckley, C., Leone, T. J. et Hickam, D. (1994). OHSUMED : an interactive retrieval evaluation and new large test collection for research. *In Conference on Research and Development in Information Retrieval*, pages 192–201, New York, NY, USA. Springer-Verlag New York, Inc.

Hersh, W., Cohen, A., Yang, J., Bhupatiraju, R. T., Roberts, P. et Hearst, M. (2005). TREC 2005 Genomics Track Overview. *In TREC.*

Hersh, W., Price, S. et Donohoe, L. (2000). Assessing thesaurus-based query expansion using the UMLS metathesaurus. *In Proceedings of AMIA Symposium*, pages 344–348.

Hersh, W. et Voorhees, E. (2009). Trec genomics special issue overview. *Information Retrieval*, 12(1):1–15.

Hersh, W. R. et Bhupatiraju, R. T. (2003). Trec genomics track overview. *In Proceedings of Text REtrieval Conference*, TREC'03, pages 14–23.

Hersh, W. R., Bhupatiraju, R. T. et Price, S. (2003). Phrases, Boosting, and Query Expansion Using External Knowledge Resources for Genomic Information Retrieval. *In TREC*, pages 503–509.

Hersh, W. R., Bhuptiraju, R. T., Ross, L., Johnson, P., Cohen, A. M. et Kraemer, D. F. (2004). TREC 2004 Genomics Track Overview. *In Proceedings of Text REtrieval Conference*, TREC 2004.

Hersh, W. R., Cohen, A. M., Ruslen, L. et Roberts, P. M. (2007). TREC 2007 Genomics Track Overview. *In Proceedings of Text REtrieval Conference*, TREC'07.

Hersh, W. R. et Greenes, R. A. (1990). SAPHIRE– an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. *Comput. Biomed. Res.*, 23:410–425.

Hirschman, L., Morgan, A. A. et Yeh, A. S. (2002). Rutabaga by any other name : extracting biological names. *Journal of Biomedical Informatics*, 35(4):247–259.

Hliaoutakis, A., Zervanou, K. et Petrakis, E. G. M. (2009). The AMTEx approach in the medical document indexing and retrieval application. *Data Knowledge Engineering*, pages 380–392.

Hoffart, J., Suchanek, F., Berberich, K., Kelham, E., de Melo, G., Weikum, G., Suchanek, F., Kasneci, G., Ramanath, M. et Pease, A. (2009). YAGO2 : A spatially and temporally enhanced knowledge base from wikipedia. *Communications of the ACM*, 52(4):56–64.

Hou, W. J. (2003). Enhancing performance of protein name recognizers using collocation. *In ACL-03 Workshop on Natural Language Processing in Biomedicine, Sapporo Convention*, pages 25–32.

Huang, X., Zhong, M. et Si, L. (2005). York University at TREC 2005 : Genomics Track. *In TREC*.

Humphreys, K., Demetriou, G. et Gaizauskas, R. (2000). Two Applications of Information Extraction to Biological Science Journal Articles : Enzyme Interactions and Protein Structures. *In Pacific Symposium on Biocomputing*, pages 505–516.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction-elements of cognitive theory. *Journal of documentation*, 52:3–50.

Jansen, B. J. et Spink, A. (2006). How are we searching the world wide web ? : a comparison of nine search engine transaction logs. *Inf. Process. Manage.*, 42(1):248–263.

Jiang, J. et Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *In International Conference on Research in Computational Linguistics (ROC)*, pages 19–33.

Jiang, J. et Zhai, C. (2007). An empirical study of tokenization strategies for biomedical information retrieval. *Inf. Retr.*, 10(4-5):341–363.

Jung, S., Herlocker, J. L. et Webster, J. (2007). Click data as implicit relevance feedback in web search. *Inf. Process. Manage.*, 43(3):791–807.

Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., de Herrera, A. G. S. et Tsikrika, T. (2011). The CLEF 2011 medical image retrieval and classification tasks. *In CLEF 2011 working notes*, Amsterdam, The Netherlands. Springer.

Kazama, J., Makino, T., Ohta, Y. et Tsujii, J. (2002). Tuning support vector machines for biomedical named entity recognition. *In Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*, pages 1–8. Association for Computational Linguistics.

Keizer, N. F., Abu-Hanna, A. et Zwetsloot-Schonk, J. H. M. (2000). Understanding terminological systems I : Terminology and Typology. *Methods of information in medicine*, pages 16–21.

Kelly, D. et Belkin, N. J. (2001). Reading time, scrolling and interaction : exploring implicit sources of user preferences for relevance feedback. *In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 408–409, New York, NY, USA. ACM.

Keselman, A., Browne, A. C. et Kaufman, D. R. (2008). Consumer health information seeking as hypothesis testing. *Journal of the American Medical Informatics Association : JAMIA*, 15(4):484–495.

Khan, L., Leod, D. M. et Hovy, E. (2004). Retrieval effectiveness of an ontology-based model for information selection. *The VLDB Journal*, 13:71–85.

Kim, W., Aronson, A. R. et Wilbur, W. J. (2001). Automatic MeSH term assignment and quality assessment. *Proceedings of AMIA Symposium*, pages 319–323.

Kraaij, W., Westerveld, T. et Hiemstra, D. (2002). The Importance of Prior Probabilities for Entry Page Search. *In Proceedings of the 25th*

annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '02, pages 27–34, New York, NY, USA. ACM.

Kraft, D. H., Pasi, G. et Bordogna, G. (2007). Vagueness and uncertainty in information retrieval : how can fuzzy sets help ? *In Proceedings of the 2006 international workshop on Research issues in digital libraries*, IWRIDL '06, pages 3 :1–3 :10. ACM.

Krauthammer, M. et Nenadic, G. (2004). Term identification in the biomedical literature. *J. of Biomedical Informatics*, 37:512–526.

Krauthammer, M., Rzhetsky, A., Morozov, P. et Friedman, C. (2000). Using BLAST for identifying gene and protein names in journal articles. *In Gene*, pages 245–252.

Krovetz, R. et Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141.

Kwok, K. L. (1996). A new method of weighting query terms for ad-hoc retrieval. *In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 187–195, New York, NY, USA. ACM.

Lam, W., Ruiz, M., Ruiz, M., Srinivasan, P. et Srinivasan, P. (1999). Automatic Text Categorization and Its Application to Text Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 11:865–879.

Le, D. T. H., Chevallet, J.-P. et Dong, T. B. T. (2007). Thesaurus-based query and document expansion in conceptual indexing with umls. *In RIVF'07*, pages 242–246.

Leacock, C. et Chodorow, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. *An Electronic Lexical Database*, pages 265–283.

Leacock, C., Towell, G. et Voorhees, E. (1993). Corpus-based statistical sense resolution. *In Proceedings of the workshop on Human Language Technology*, HLT '93, pages 260–265. Association for Computational Linguistics.

Lee, J. H. (1997). Analyses of multiple evidence combination. *SIGIR Forum*, 31:267–276.

Lenoir, P., Michel, J. R., Frangeul, C. et Chales, G. (1981). Réalisation, développement et maintenance de la base de données A.D.M. *Médecine informatique*, (6):51–56.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone. *In Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.

Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Lin, D. (1998). An information-theoretic definition of similarity. *In Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Liu, S., Liu, F., Yu, C. et Meng, W. (2004). An effective approach to document retrieval via utilizing WordNet and recognizing phrases. *In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'04, pages 266–272, New York, NY, USA. ACM.

LL04 (2004). National Center for Biotechnology Information - LocusLink Home Page. `http://www.ncbi.nih.gov/LocusLink/`.

Lou, B. (1995). *Users Reference Guide British National Corpus Version 1.0*. Oxford University Computing Services, UK.

Lu, Z., Kim, W. et Wilbur, W. J. (2009). Evaluation of query expansion using MeSH in PubMed. *Information Retrieval*, 12(1):69–80.

Ma, Z., Pant, G. et Sheng, O. R. L. (2007). Interest-based personalized search. *ACM Trans. Inf. Syst.*, 25(1).

Maglott, D., Ostell, J., Pruitt, K. D. et Tatusova, T. (2005). Entrez Gene : Gene-centered information at NCBI. *Nucleic Acids Research*, 33 (Database Issue):54–58.

Mallery, J. C. (1988). *Thinking About Foreign Policy : Finding an Appropriate Role for Artificially Intelligent Computers*. Thèse de doctorat, M.I.T. Political Science Department.

Manning, C. D., Raghavan, P. et Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Maron, M. E. et Kuhns, J. L. (1960). On Relevance, Probabilistic Indexing and Information Retrieval. *J. ACM*, 7:216–244.

McInnes, B. T., Pedersen, T. et Carlis, J. (2007). Using UMLS Concept Unique Identifiers (CUIs) for Word Sense Disambiguation in the Biomedical Domain. pages 746–750.

Mckusick, V. A. (1998). *Mendelian Inheritance in Man : A Catalog of Human Genes and Genetic Disorders*. Johns Hopkins University Press.

Mihalcea, R., Tarau, P. et Figa, E. (2004). PageRank on semantic networks with application to word sense disambiguation. *In International Conference on Computational Linguistics (COLING)*, pages 1126–1132.

Miller, D. R. H., Leek, T. et Schwartz, R. M. (1999). A hidden Markov model information retrieval system. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 214–221, New York, NY, USA. ACM.

Miller, G. A. (1995). WordNet : A Lexical Database for English. *Communications of the ACM*, 38:39–41.

Miller, G. A., Leacock, C., Tengi, R. et Bunker, R. T. (1993). A semantic concordance. *In Proceedings of the workshop on Human Language Technology*, HLT '93, pages 303–308. Association for Computational Linguistics.

Montague, M. et Aslam, J. A. (2001). Relevance score normalization for metasearch. *In Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 427–433, New York, NY, USA. ACM.

Morgan, A., Hirschman, L., Yeh, A. et Colosimo, M. (2003). Gene name extraction using FlyBase resources. *In Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine*, volume 13, pages 1–8.

Morris, J. et Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 1:21–48.

Nakov, P., Schwartz, A., Stoica, E. et Hearst, M. (2004). BioText Team Experiments for the TREC 2004 Genomics Track. *In In The thirteenth Text Retrieval Conference, TREC 2004. National Institute of Standards and Technology.*

Narayanaswamy, M., Ravikumar, K. E. et Vijay-Shanker, K. (2003). A Biological Named Entity Recognizer. *In Pacific Symposium on Biocomputing*, pages 427–438.

Navigli, R. (2009). Word sense disambiguation : A survey. *ACM Comput. Surv.*, 41:10 :1–10 :69.

Névéol, A., Rogozan, A. et Darmoni, S. (2006). Automatic indexing of online health resources for a french quality controlled gateway. *Information Processing and Management*, 42(3):695–709.

Névéol, A., Shooshan, S. E., Humphrey, S. M., Mork, J. G. et Aronson, A. R. (2009). A recent advance in the automatic indexing of the biomedical literature. *Journal of biomedical informatics*, 42(5):814–823.

Niwa, Y. et Nitta, Y. (1994). Co-occurrence vectors from corpora vs. distance vectors from dictionaries. *In Proceedings of the 15th conference on Computational linguistics - Volume 1*, COLING '94, pages 304–309. Association for Computational Linguistics.

Oard, D. et Kim, J. (1998). Implicit Feedback for Recommender Systems. *In in Proceedings of the AAAI Workshop on Recommender Systems*, pages 81–83.

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M. et Thornton, J. (1997). CATH - a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109.

Ounis, I., Amati, G., V., P., He, B., Macdonald, C. et Johnson (2005). Terrier Information Retrieval Platform. *In Proceedings of the 27th European Conference on IR Research (ECIR 2005)*, volume 3408 de *Lecture Notes in Computer Science*, pages 517–519. Springer.

Ounis, I., de Rijke, M., Macdonald, C., Mishne, G. et Soboroff, I. (2006). Overview of the TREC-2006 Blog Track. *In Text Retrieval Conference*.

Paice, C. D. (1984). Soft evaluation of Boolean search queries in information retrieval systems. *Inf. Technol. Res. Dev. Appl.*, 3:33–41.

Pereira, S., Massari, P., Buemi, A., Dahamna, B., Serrot, E., Joubert, M. et Darmoni, S. J. (2009). F-MTI : outil d'indexation multi-terminologique : application à l'indexation automatique de la SNOMED. *In Risques et technologies de l'information pour les pratiques médicales : comptes rendus des treizi mes journées francophones d'informatique médicale (JFIM)*, volume 17 de *Informatique et santé*, pages 57–67, France.

Pereira, S., Névéol, A., Kerdelhué, G., Serrot, E., Joubert, M. et Darmoni, S. (2008). Using multi-terminology indexing for the assignment of MeSH descriptors to health resources. *Proceedings AMIA Symposium*, pages 586–590.

Platt, J. C., Cristianini, N. et Shawe-taylor, J. (2000). Large margin dags for multiclass classification. *In Advances in Neural Information Processing Systems 12*, pages 547–553.

Ponte, J. M. et Croft, W. B. (1998). A language modeling approach to information retrieval. *In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

Porter, M. F. (1997). *An algorithm for suffix stripping*, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Pouliquen, B. (2002). *Indexation de textes médicaux par extraction de concepts, et ses utilisations*. Thèse de doctorat, Université de Rennes I.

R Development Core Team (2008). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rada, R., Mili, H., Bickneli, E. et Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transaction Systems Man and Cybernetics*, 19:17–30.

Ren, A., Du, X. et Wang, P. (2009). Ontology-Based Categorization of Web Search Results Using YAGO. *In CSO (1)*, pages 800–804.

Resnik, P. (1999). Semantic Similarity in a Taxonomy : An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *J. Artif. Intell. Res. (JAIR)*, 11:95–130.

Rijsbergen, C. (1979). *Information retrieval, Second edition*. Butterworths.

Robertson, S. (2002). Introduction to the Special Issue : Overview of the TREC Routing and Filtering Tasks. *Inf. Retr.*, 5(2-3):127–137.

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. et Gatford, M. (1994). Okapi at TREC-3. *In Proceedings of the Third Text REtrieval Conference*, TREC'94.

Robertson, S. et Zaragoza, H. (2009). The Probabilistic Relevance Framework : BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

Robertson, S. E. (1977). The Probability Ranking Principle in IR. *Journal of Documentation*, 33(4):294–304.

Robertson, S. E. (1991). On term selection for query expansion. *J. Doc.*, 46(4):359–364.

Robertson, S. E. et Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. SIGIR '94, pages 232–241, New York, NY, USA. Springer-Verlag New York, Inc.

Robertson, S. E., Walker, S. et Hancock-Beaulieu, M. (1998). Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive. *In Proceedings Text REtrieval Conference*, TREC-7, pages 199–210.

Rocchio, J. (1971). *Relevance Feedback in Information Retrieval*, pages 313–323.

Ruch, P. (2006). Automatic assignment of biomedical categories : toward a generic approach. *Bioinformatics*, 22(6):658–664.

Ruch, P., Baud, R. H. et Geissbühler, A. (2003). Learning-Free Text Categorization. *In AIME*, pages 199–208.

Sakji, S. (2010). *Recherche d'information et indexation automatique des médicaments à l'aide de plusieurs terminologies de santé*. Thèse de doctorat, Université de Rouen.

Salton, G. (1970). Evaluation problems in interactive information retrieval. *Information Storage and Retrieval*, 6(1):29–44.

Salton, G. (1989). *Automatic text processing : the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Salton, G. (1991). The SMART Information Retrieval System after 30 years - Panel. *In SIGIR*, pages 356–358.

Salton, G., Fox, E. A. et Wu, H. (1983). Extended boolean information retrieval. *Commun. ACM*, 26:1022–1036.

Sanderson, M. (1994). Word sense disambiguation and information retrieval. *In SIGIR*, pages 142–151.

Schardt, C., Adams, M. B., Owens, T., Keitz, S. et Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC medical informatics and decision making*, 7(1):16+.

Schiemann, T., Leser, U. et Hakenberg, J. (2008). Word Sense Disambiguation in Biomedical Applications : A Machine Learning Approach. *In Information Retrieval In Biomedicine*, pages 142–161.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Seki, K., Costello, J. C., Singan, V. R. et Mostafa, J. (2004). TREC 2004 Genomics Track Experiments at IUB. *In NIST Special Publication 500-261 : The Thirteenth Text REtrieval Conference Proceedings (TREC*.

Sheather, S. J. et Jones, M. C. (1991). A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):683–690.

Shen, D., Zhang, J., Zhou, G., Su, J. et lim Tan, C. (2003). Effective adaptation of a hidden Markov model-based named entity recognizer for biomedical domain. *In In : Proceedings of NLP in Biomedicine, ACL*, pages 49–56.

Sieg, A., Mobasher, B. et Burke, R. (2007). Web search personalization with ontological user profiles. *In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 525–534, New York, NY, USA. ACM.

Silberztein, M. (1999). Text Indexation with INTEX. *Computers and the Humanities*, 33(3):265–280.

Sinclair, J., éditeur (1995). *Collins Cobuild English Dictionary*. HarperCollins.

Singhal, A. et Pereira, F. (1999). Document expansion for speech retrieval. *In SIGIR'99 Conference on Research and Development in Information Retrieval*, pages 34–41, New York, NY, USA. ACM.

Smeaton, A. F., Over, P. et Kraaij, W. (2006). Evaluation campaigns and TRECVid. *In MIR '06 : Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA. ACM Press.

Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M. et Boydell, O. (2005). Exploiting Query Repetition and Regularity in an Adaptive Community-Based Web Search Engine. *User Modeling and User-Adapted Interaction*, 14(5):383–423.

Snowball, R. (1997). Using the clinical question to teach search strategy : fostering transferable conceptual skills in user education by active learning. *Health Libr Rev*, 14:167–173.

Sohn, S., Kim, W., Comeau, D. C. et Wilbur, W. J. (2008). Optimal training sets for bayesian prediction of mesh assignment. *In Journal American of Medicine*, pages 546–553.

Sonnenburg, S., Rätsch, G., Schölkopf, B. et Rätsch, G. (2006). Large scale multiple kernel learning. *JOURNAL OF MACHINE LEARNING RESEARCH*, 7.

Srinivasan, P. (1996). Query expansion and medline. *Information Processing and Management*, 32:431–443.

Stokes, N., Li, Y., Cavedon, L. et Zobel, J. (2009). Exploring criteria for successful query expansion in the genomic domain. *Information Retrieval*, 12(1):17–50.

Sumner, R. G., Jr., Yang, K., Akers, R. et Shaw, W. M. (1998). Interactive Retrieval using IRIS : TREC-6 Experiments. *In TREC-6*, pages 711–734.

Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. *In International Conference on Information and Knowledge Base Management*, pages 67–74.

Tamburis, O. (2006). The specific role of ICT : different perspectives between traditional healthcare service and e-healthcare service. *IJEH*, 2(3):250–262.

Thirion, B., Robu, I. et Darmoni, S. J. (2009). Optimization of the PubMed Automatic Term Mapping. *In* Adlassnig, K.-P., Blobel, B., Mantas, J. et Masic, I., éditeurs : *MIE*, volume 150 de *Studies in Health Technology and Informatics*, pages 238–242. IOS Press.

Tranchevent, L.-C., Barriot, R., Yu, S., Van Vooren, S., Van Loo, P., Coessens, B., De Moor, B., Aerts, S. et Moreau, Y. (2008). Endeavour update : a web resource for gene prioritization in multiple species. *Nucleic Acids Research*, 36(suppl 2):W377–W384.

Trieschnigg, D. (2010). *Proof of Concept : Concept-based Biomedical Information Retrieval*. Thèse de doctorat, University of Twente.

Trieschnigg, D., Kraaij, W. et Schuemie, M. (2006). Concept based document retrieval for genomics literature. *In* Voorhees, E. et Buckland, L. P., éditeurs : *Fifteenth Text REtrieval Conference, TREC 2006*, volume SP 500 de *NIST Special Publication*, Gaithersburg, MD, USA. National Institute of Standards and Technology (NIST).

Trieschnigg, D., Pezik, P., Lee, V., de Jong, F., Kraaij, W. et Rebholz-Schuhmann, D. (2009). Mesh up : effective mesh text classification for improved document retrieval. *Bioinformatics*, 25(11):1412–1418.

Tuason, O., Chen, L., Liu, H., Blake, J. A. et Friedman, C. (2004). Biological nomenclatures : a source of lexical knowledge and ambiguity. *Pac Symp Biocomput*, pages 238–249.

Uzuner, Ö., Katz, B. et Yuret, D. (1999). Word Sense Disambiguation for Information Retrieval. *In AAAI/IAAI*, page 985.

Villanueva, E. V., Burrows, E. A., Fennessy, P. A., Rajendran, M. et Anderson, J. N. (2001). Improving question formulation for use in evidence appraisal in a tertiary care setting : a randomised controlled trial. *BMC Med Inform Decis Mak*, 1.

Voorhees, E. M. (1993). Using WordNet to Disambiguate Word Senses for Text Retrieval. *In SIGIR*, pages 171–180.

Voorhees, E. M. (2002). The Philosophy of Information Retrieval Evaluation. *In Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, CLEF '01, pages 355–370, London, UK, UK. Springer-Verlag.

White, R., Ruthven, I. et Jose, J. M. (2002). The use of implicit evidence for relevance feedback in web retrieval. *In Proceedings of the 24th BCS-IRSG European Colloquium on IR Research : Advances in Information Retrieval*, pages 93–109, London, UK, UK. Springer-Verlag.

Wilbur, W. J. (2003). PubMed Related Citations Algorithm. Rapport technique.

Wu, S. et McClean, S. (2006). Performance prediction of data fusion for information retrieval. *Inf. Process. Manage.*, 42:899–915.

Xu, J. et Croft, W. B. (1996). Query Expansion Using Local and Global Document Analysis. *In Conference on Research and Development in Information Retrieval*, pages 4–11.

Yang, Y. et Chute, C. (1994a). Words or concepts : the features of indexing units and their optimal use in information retrieval. *In Proceedings of SCAMC'94*, pages 685–689.

Yang, Y. et Chute, C. G. (1994b). An example-based mapping method for text categorization and retrieval. *ACM Trans. Inf. Syst.*, 12(3):252–277.

Yu, S., Tranchevent, L.-C. C., De Moor, B. et Moreau, Y. (2010). Gene prioritization and clustering by multi-view text mining. *BMC bioinformatics*, 11(1):28+.

Zhai, C. et Lafferty, J. (2001a). A study of smoothing methods for language models applied to Ad Hoc information retrieval. *In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 334–342, New York, NY, USA. ACM.

Zhai, C. et Lafferty, J. (2001b). Model-based feedback in the language modeling approach to information retrieval. *In Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 403–410, New York, NY, USA. ACM.

Zhao, J., yen Kan, M., Procter, P. M., Zubaidah, S., Yip, W. K. et Li, G. M. (2010). Improving Search for Evidence-based Practice using Information Extraction. *In AMIA Annual Symposium*, pages 937–941.

Zhou, W., Yu, C., Smalheiser, N., Torvik, V. et Hong, J. (2007a). Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. *In Proceedings of Conference on Research and Development in Information Retrieval*, SIGIR'07, pages 655–662.

Zhou, X., Hu, X. et Zhang, X. (2007b). Topic Signature Language Models for Ad hoc Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 19:1276–1287.

Zhou, X., Hu, X., Zhang, X., Lin, X. et yeol Song, I. (2006a). Context-Sensitive Semantic Smoothing for the Language Modeling Approach to Genomic IR. *In Proc. of Conference on Research and Development in Information Retrieval*, SIGIR'06, pages 170–177. ACM.

Zhou, X., Zhang, X. et Hu, X. (2006b). MaxMatcher : Biological Concept Extraction Using Approximate Dictionary Lookup. *In PRICAI*, volume 4099, pages 1145–1149.

Zhou, X., Zhang, X. et Hu, X. (2006c). Using Concept-Based Indexing to Improve Language Modeling Approach to Genomic IR. *In ECIR*, pages 444–455.