REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mouloud MAMMERI de Tizi-Ouzou

Faculté de Génie Electrique et Informatique

Département d'Informatique

Mémoire de fin d'études

En vue de l'obtention du diplôme de Master en Informatique

Spécialité : Conduite de projets informatiques

THEME :

Extraction de relations entre concepts dans le domaine biomédical

Proposé et Dirigé par :

Mme F. AMIROUCHE

Réalisé par :

Melle DAHMOUN Amina

Promotion 2012-2013

A mes très chers parents, à ma sœur Sihem et à mon frère Amine,

Je dédie ce modeste travail

Remerciements

Et vient le moment d'écrire les remerciements... Quand je pense à toutes les personnes sans qui ce travail n'aurait probablement pas pu être achevé, je me dis que ce n'est pas pour rien que cette page précède la présentation de tout travail de thèse.

Mes remerciements, les plus vifs, ma profonde gratitude et mes respects s'adressent à ma Mme Amirouche Fatiha pour avoir accepté de m'encadrer, pour les conseils et orientations tant précieux qu'elle m'a prodigué durant ce Mémoire de Master.

Je tiens à exprimer toute ma reconnaissance à Melle Azzoug Wassila et Melle Cherdioui Sabrina pour leur assistance et le temps qu'elle m'ont consacré qui m'ont permis de progresser sans cesse.

J'adresse également mes sincères remerciements et mes respects aux membres du jury pour m'avoir fait l'honneur de juger ce travail.

Je ne remercierai jamais assez mes chers parents pour m'avoir toujours encouragée et m'avoir inculquée le goût du savoir et de l'ambition.

Je remercie également ma sœur Sihem et mon frère Amine pour leur soutient, leur patience et leur encouragement.

Que mes chères amis, auprès de qui j'ai toujours trouvé l'encouragement et le soutien moral, trouvent ici ma profonde gratitude.

Sommaire

Liste des tableaux

Table des figures

Introduction générale

La médecine est un domaine de connaissance très actif en matière de recherche scientifique. En effet, l'avancée des techniques médicales et l'amélioration des soins sont considérées depuis bien longtemps comme une priorité par de nombreuses institutions publiques ou privées. Il en résulte une conséquence logique qui est la variation et le développement des sources d'information et des outils d'aide à l'accès à l'information médicale et biomédicale. A titre d'exemple, MEDLINE (*Medical Literature Analysis and Retrieval System Online*) est la base de données bibliographique de premier ordre, développée par la NLM (*US National Library of Medicine*). Elle contient plus de 19 millions de références d'articles en science de la vie, notamment de la biomédecine. Un trait distinctif de MEDLINE est que les documents sont indexés manuellement ou automatiquement avec les concepts du thésaurus MeSH (*Medical Subject Headings*).

Les spécialistes du domaine médicale se trouvent alors face à des volumes d'informations riches et diversifiés et commencent à imaginer des systèmes qui exploitent ces informations pour extraire de nouvelle connaissances, déterminer des relations de cause à effet, des relations de co-occurrence, etc.… Par ailleurs, le grand public, demande à ce que ces informations soient mises à leur disposition et partagées. Face à ces différents attentes, le traitement automatique d'information biomédicale dans ses différentes formes (la recherche d'informations biomédicales, des analyses statistiques, etc.…) est devenu une nécessité, surtout que les systèmes informatiques médicaux ont connu une grande évolution de point de vue de leur architecture, de la qualité et de la diversité des services autour du stockage de l'information, l'accès à l'information pour une médecine basée sur des niveaux de preuves et d'aide à la décision pour l'amélioration de la qualité des soins. L'information biomédicale utilisée comme support pour les tâches de recherche, d'extraction d'information et de connaissances, concerne principalement la littérature médicale et les dossiers des patients.

C'est dans ce cadre que s'inscrit notre travail, nous nous positionnons dans le cadre spécifique de l'exploitation de l'information contenue dans la littérature biomédicale, en vue de concevoir des systèmes informatiques médicaux et des systèmes d'aide à la décision en médecine permettant de promouvoir l'expérience collective des praticiens. On s'intéresse plus précisément au problème de détection de relations pertinentes entre informations biomédicales. En effet, notre objectif est de proposer une approche d'extraction de relations

de co-occurrence entre concepts biomédicaux, qui peuvent être révélatrices de relations plus spécifiques à faire vérifier par des spécialistes du domaine.

Notre mémoire s'articule autour de quatre chapitres :

- Le premier chapitre introduit l'indexation et la recherche d'information dans le domaine biomédical.
- Le deuxième chapitre présente un état de l'art sur l'extraction de relations entre concepts biomédicaux.
- Le troisième chapitre présente notre contribution en décrivant dans un premier temps notre approche d'extraction de relations de co-occurrence entre concepts biomédicaux, et dans un second temps, la méthode que nous proposons pour validé notre approche.
- Le quatrième chapitre présente l'évaluation expérimentale de notre approche.

Et enfin, nous conclurons en proposant des perspectives de recherche permettant d'améliorer notre proposition.

I. Introduction à l'indexation et à la recherche d'information biomédicale

1. Introduction

Historiquement, la croissance du volume de données textuelles comme les livres et les articles dans les bibliothèques durant des siècles a imposé de définir des mécanismes efficaces pour les localiser. Les premières techniques, comme l'abstraction (abstracting), l'indexation et l'utilisation des catégories de classification ont marqué la naissance de la "Recherche d'Information" comme discipline de recherche.

Le but de la recherche d'information est de retrouver les documents qui satisfont un besoin utilisateur. Si l'utilisateur juge qu'un document répond à son besoin, le document est dit pertinent. Dans un Système de Recherche d'Information (SRI), L'utilisateur exprime son besoin d'information sous forme d'une requête. Le SRI tente de trouver tous les documents pertinents et de rejeter les documents qui ne sont pas pertinents. Dans la pratique, l'ensemble des documents renvoyés par un SRI pour une requête est composé d'un sous-ensemble de documents pertinents et un sous-ensemble de documents non pertinents. Ces sous-ensembles déterminent la performance d'un SRI.

La recherche d'information touche à tous les domaines, en particulier au domaine biomédical. En général, le texte biomédical est exprimé dans les documents en utilisant le langage naturel qui est source d'ambigüité, ce qui présente un obstacle pour la tâche de recherche et d'indexation d'information dans ce domaine. Ainsi, l'accessibilité et l'indexation sont particulièrement confrontées aux problèmes de synonymie, homonymie et à la présence d'acronymes. C'est pour cela que l'utilisation des données médicales et l'accès à une information concise sont devenus des enjeux majeurs pour les professionnels de la santé mais aussi pour le grand public. Pour remédier à ces problèmes et faciliter l'accès aux informations, plusieurs terminologies biomédicales ont été développées et des techniques spécifiques d'indexation et de recherche d'information ont été mises en œuvre.

Dans ce chapitre, nous présentons les concepts de base de la recherche d'information. Ensuite, nous nous positionnons dans le contexte de la recherche d'information médicale, et nous présentons d'abord la typologie des documents médicaux ainsi que les ressources biomédicales existantes, puis nous présentons les étapes d'indexation des documents médicaux, et nous détaillons les techniques de recherche d'information biomédicale, ainsi que les campagnes d'évaluation des systèmes de recherche d'information biomédicale.

2. Recherche d'information : Généralités

La recherche d'information est traditionnellement définie comme étant l'ensemble des techniques permettant de sélectionner à partir d'une collection de documents, ceux qui sont susceptibles de répondre aux besoins de l'utilisateur exprimés via une requête. Gérer des textes, implique stocker, rechercher et explorer des documents pertinents.

Plusieurs concepts clés s'articulent autour de cette définition :

- Collection de documents : La collection de documents constitue l'ensemble des informations exploitables et accessibles. Elle est constituée d'un ensemble de documents.

- Document : Le document constitue l'information élémentaire d'une collection de documents. Il est constitué par un texte, un morceau de texte, une image, une bande de vidéo etc., qui peut être retourné en réponse à une requête (ou besoin en information) d'un utilisateur.

- Besoin en information : le besoin en information d'un utilisateur peut avoir différents types. Trois types de besoin utilisateur ont été définis par [Ingwersen, 92] :

✓ Besoin vérificatif : l'utilisateur cherche une donnée particulière et sait souvent comment y accéder, comme par exemple, la recherche d'un document avec une adresse connu.

✓ Besoin thématique connu : l'utilisateur cherche à trouver une nouvelle information concernant un sujet ou un domaine connu.

✓ Besoin thématique inconnu : dans ce cas l'utilisateur cherche de nouveaux concepts ou de nouvelles relations hors des sujets ou domaines qui lui sont familiers.

- Requête **:** La requête constitue l'expression du besoin en information de l'utilisateur. Elle représente l'interface entre le SRI et l'utilisateur. Une requête est généralement exprimée par un ensemble de mots clés (exemple : système SMART et OKAPI), mais elle peut être également exprimée en langage naturel (exemple : système SMART), booléen (exemple : système DIALOG) ou graphique (exemple : système issu du projet NEURODOC).

- Pertinence **:** un document pertinent est un document qui doit contenir l'information que l'utilisateur recherche.

Dans ce sui suit, nous présentons le processus de recherche d'information et nous passerons en revue les modèles piliers de la RI.

2.1. Le processus de recherche d'information

De manière générale, la recherche dans un SRI consiste à mettre en correspondance la source d'information et le besoin de l'utilisateur afin de retourner à ce dernier un ensemble de résultats pertinents vis-à-vis de sa requête. Pour retrouver les documents pertinents pour une requête donnée, le processus de recherche consiste à comparer la représentation de chaque document de la collection à la représentation de la requête. Lorsque ces représentations sont proches, le document est sélectionné en réponse à la requête. Le processus comporte ainsi trois principales étapes : la représentation des documents et requête, et la mise en correspondance (ou appariement) dans laquelle s'effectue le calcul de pertinence.

L'indexation (ou processus de représentation) a pour rôle d'extraire les descripteurs à partir d'un document ou d'une requête. Le descripteur est une liste de termes significatifs qui doit couvrir au mieux le contenu sémantique d'un document ou d'une requête. Un poids est généralement associé à chacun de ces termes pour différencier leur degré de représentativité dans la source d'information correspondante. L'ensemble des termes reconnus par le SRI est rangé dans une structure appelée dictionnaire constituant le langage d'indexation.

L'appariement document-requête permet de sélectionner l'ensemble des documents potentiellement pertinents pour une requête. Il existe deux méthodes d'appariement : l'appariement exact et l'appariement approché. Dans la première méthode, on récupère les documents qui correspondent exactement à la requête spécifiée. Les documents retournés ne sont pas triés. Dans la seconde méthode, on calcule un score de pertinence RSV (Q, D) (Retrieval Status Value) entre la requête indexée Q et les descripteurs du document D. Les documents qui correspondent au mieux à la requête sont alors retournés à l'utilisateur. Les documents retournés sont triés selon leur score de pertinence vis-à-vis de la requête.

Un processus supplémentaire de reformulation de la requête est quelques fois mis en œuvre. Ce processus a pour objectif de générer une nouvelle requête plus adéquate que celle initialement formulée par l'utilisateur. La reformulation de requête consiste alors à modifier la requête initiale par ajout de termes significatifs et /ou réestimation de leur poids. Si les termes rajoutés proviennent des documents de la collection, on parle de réinjection de pertinence (relevance feedback) si le processus est supervisé, et de pseudo réinjection de pertinence si le processus est automatique. Si les termes rajoutés ne proviennent pas des documents de la

collection, on parle d'expansion de requête. Dans ce cas, la reformulation de requête peut aussi être basée sur des ressources linguistiques externes telles que les ontologies ou les thésaurus, ou sur des techniques d'association de termes telles que les règles d'association ou la cooccurrence.

Le fonctionnement général d'un SRI est donné à travers le processus de recherche communément appelé processus en U [Belkin, 92], présenté en figure1.1.

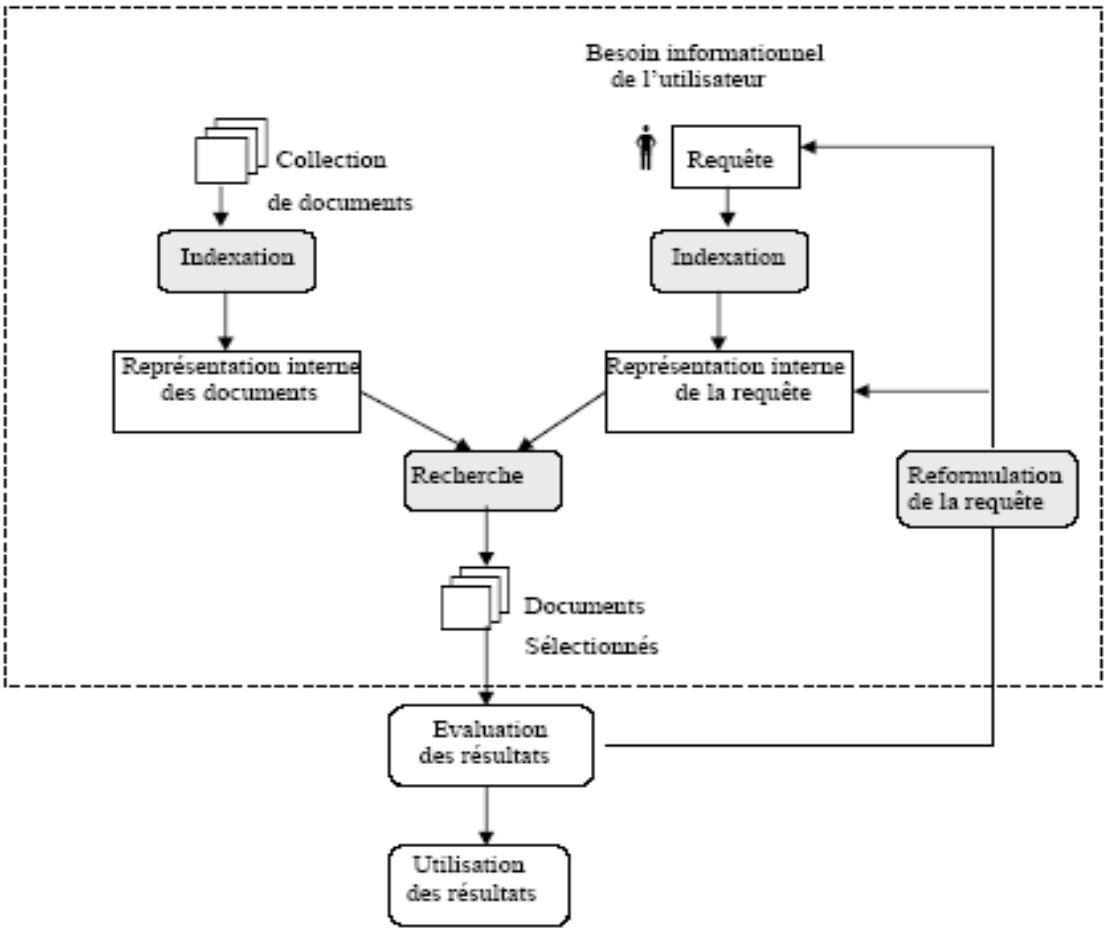


Figure I.1. Le processus en U de la RI

2.2. Les différents modèles de Recherche d'Information

Si c'est l'indexation qui permet de déterminer les termes représentatifs des documents et requêtes, c'est le modèle qui assure leur interprétation dans un formalisme de représentation propre et qui permet de mesurer la pertinence d'un document vis-à-vis d'une requête. Les modèles de recherche et représentation d'information sont basés sur un processus

de mise en correspondance entre requêtes utilisateurs et documents de la collection. Le mécanisme de recherche détermine alors, sur la base d'un degré de pertinence supposé des documents, ceux qui répondent au besoin de l'utilisateur. Les modèles de recherche peuvent être classés en trois principales catégories :

- les modèles ensemblistes dont le modèle booléen,
- les modèles algébriques dont le modèle vectoriel,
- les modèles probabilistes.

La figure I.2 présente une classification des différents modèles de RI.

Le modèle booléen a été le premier modèle utilisé en RI. Il se base sur l'utilisation des opérateurs logiques de l'algèbre de Boole « AND », « OR » et « NOT » pour la représentation des requêtes. Le document est représenté par un ensemble de termes. Le processus de recherche mis en œuvre par le système consiste à effectuer des opérations sur des ensembles de documents définis par la présence ou l'absence de termes d'index afin de réaliser un appariement exact avec la requête. Ce modèle présente le principal avantage de sa facilité de mise en œuvre. Cependant, il n'est pas possible de classer les documents par ordre de pertinence. Pour palier ce problème, des versions étendues de ce modèle (modèle booléen étendu, modèle flou) ont été proposées. Ces versions intègrent principalement la pondération des termes d'index.

Le modèle vectoriel est sans doute le modèle le plus utilisé en RI. Ce modèle préconise la représentation des documents et requêtes par des vecteurs de poids, dans l'espace engendré par les N termes d'indexation. Dans ce modèle, le degré de pertinence d'un document relativement à une requête est perçu comme le degré de corrélation entre les vecteurs associés. Ceci nécessite alors la définition d'une fonction de calcul de similarité entre vecteurs mais également d'une fonction de pondération des termes d'index. L'avantage de ce modèle réside particulièrement dans l'ordonnancement et le classement des documents sélectionnés selon leurs pertinences. Cependant, l'inconvénient majeur de ce modèle réside dans le fait que l'association entres les termes d'indexation n'est pas prise en compte.

Le modèle probabiliste s'appuie sur la théorie des probabilités. Le processus de recherche se traduit par le calcul de la probabilité de pertinence d'un document relativement à une requête. Un des inconvénients de ce modèle est l'impossibilité d'estimer ses paramètres si des collections de tests ne sont pas disponibles.

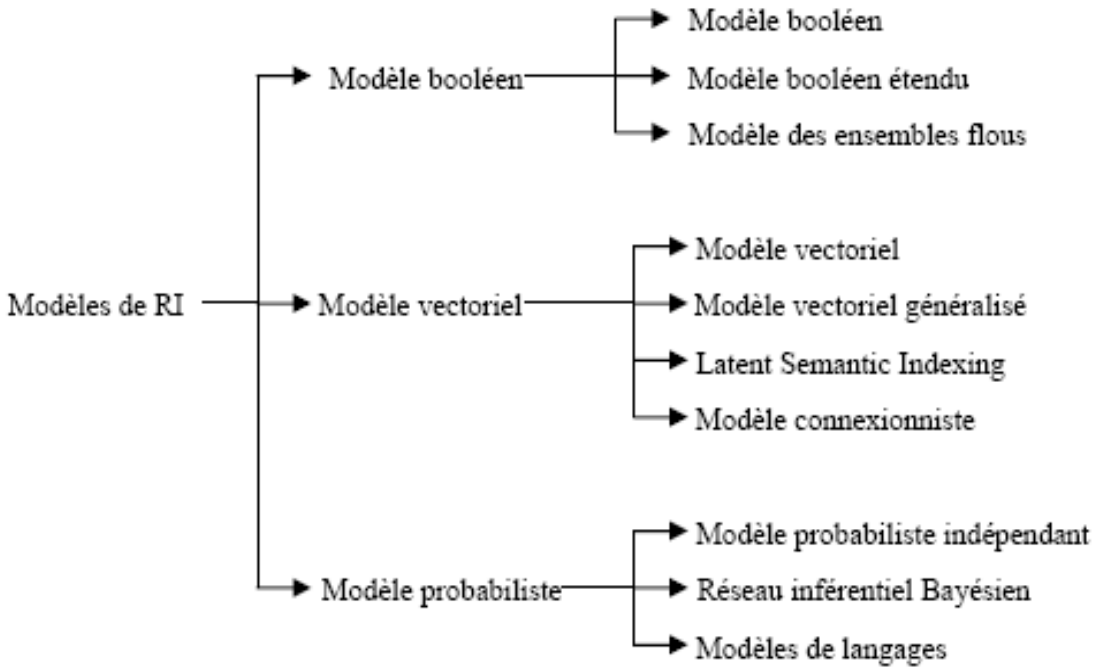


Figure I.2 : Les différents modèles de recherche d'information

3. Indexation et recherche d'information dans les documents biomédicaux

La diversité de la littérature biomédicale avec l'augmentation de documents qui la constituent d'une part, et l'informatisation des systèmes d'information biomédicaux d'autre part ont créé la nécessité d'un traitement automatique de cet ensemble d'information biomédicales à des fins diverses comme la recherche d'information, la classification, les analyses statistiques, etc. Une étape primordiale requise dans chacune de ces tâches est l'indexation. Le but étant de créer une représentation à base de termes d'index (mots clés ou concepts) permettant de repérer et de retrouver facilement l'information dans un ensemble de document.

Dans le domaine biomédical, la terminologie utilisée est normalisée. La standardisation du langage biomédical a donné naissance à des ressources terminologiques à l'exemple des terminologies GO (pour la représentation des gènes et protéines), ICD (pour la représentation des pathologies) et SNOMED (pour la représentation des actes médicaux), du

thésaurus MeSH (représentation des concepts biomédicaux) et du méta-thésaurus UMLS. L'indexation des documents biomédicaux s'appuie sur ces terminologies comme principale source d'identification du vocabulaire d'indexation dans les documents biomédicaux.

Dans ce qui suit, nous présentons la typologie des documents médicaux ainsi que les principales ressources terminologiques biomédicales. Puis nous détaillerons la tâche d'indexation des documents biomédicaux. Nous parlerons aussi des techniques de recherche d'information médicale. Enfin, nous conclurons ce chapitre par une section qui présente les campagnes d'évaluations de recherche d'information médicale.

3.1. Typologie des informations biomédicales

Selon (Hersh, 2008) les informations biomédicales se déclinent en deux catégories principales : l'information spécifique au patient et les connaissances du domaine biomédical.

- L'information spécifique au patient vise à informer les médecins, les infirmiers, et les administrateurs de l'état de santé du patient. Ces informations peuvent être présentées sous forme de résultats de laboratoire ou compte rendus médicaux. Elles constituent le dossier médical patient.
- Les connaissances du domaine biomédical peuvent être subdivisées en deux catégories la littérature primaire et la littérature secondaire.
    - La littérature primaire est une recherche originale qui apparaît dans des revues, des livres, rapports et autres sources.
    - La littérature secondaire fondée sur les connaissances se compose de commentaires et / ou de synthèses de la littérature primaire, par exemple, livres, articles de revues dans le journal et autre publication.

La littérature biomédicale est regroupée essentiellement dans 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. Elle est produite et gérée par la *National Library of Medicine* (NLM) département du NIH (*National Institutes of Health* dans le *Maryland*). MEDLINE contient plus de 20 millions d'articles (citations) couvrant tous les domaines biomédicaux (biochimie, biologie, médecine clinique, éthique, pharmacologie, toxicologie). Chaque document dans MEDLINE est annoté manuellement ou semi-automatiquement par des termes MeSh (champ MeSh).

MEDLINE est accessible en ligne via le portail PubMed. La table I.1 montre un exemple de citation enregistrée dans la base MEDLINE.

```
PMID- 11473808
OWN - NLM
STAT- MEDLINE
DA  - 20010727
DCOM- 20011101
LR  - 20061115
PUBM- Print
IS  - 0968-4328
VI  - 33
IP  - 1
DP  - 2002
TI  - Gene transfer into retinal ganglion cells by in vivo electroporation:
a new approach.
PG  - 1-6
AB  - We developed a new in vivo electroporation method to deliver genes
into retinal ganglion cells (RGCs). Efficiency and degree of tissue damage
were evaluated using green fluorescent protein (GFP) gene and TUNEL.
AD  - Department of Anatomy, Yokohama City University School of Medicine,
3-9 Fukuura, Kanazawa-ku, 236-0004, Yokohama, Japan. dezawa@med. FAU -
Takano, M
AU  - Takano M
PT  - Journal Article
MH  - *Gene Transfer Techniques
MH  - In Situ Nick-End Labeling
MH  - Luminescent Proteins/*genetics/metabolism
MH  - Microscopy, Confocal
MH  - Rats
MH  - Retina/ultrastructure
```

Table I.1. Exemple d'une citation de MEDILNE

3.2. Ressources terminologiques biomédicales

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). La structure et le contenu d'une terminologie sont crées en fonction de l'utilisation qui doit en être faite. La terminologie peut aussi rendre compte des relations qui peuvent exister entre les termes. Les relations spécialisation-généralisation permettent de hiérarchiser les termes du plus général au plus spécifique. A l'intérieur d'une terminologie, les termes désignés des concepts. Un concept peut être désigné par plusieurs termes différents (synonymes). Un concept est identifié par un code numérique ou alphanumérique unique qui ces reflète sa position dans la hiérarchie des concepts (terminologie). Il existe plusieurs déclinaisons de terminologies :

- Vocabulaire contrôlé : un vocabulaire contrôlé est la forme la plus élémentaire d'une terminologie. La signification des termes n'est pas forcément définie et il n'y a pas nécessairement d'organisation logique des termes entre eux.
- Thésaurus : un thésaurus est un vocabulaire contrôlé et organisé selon trois types de relations : la relation hiérarchique (spécialisation-généralisation, tout-partie), la relation d'équivalence (synonymes) et la relation d'association pour les sujets connexes.
- Classification : 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.
- Nomenclature : c'est une terminologie qui vise à recenser tous les termes d'un domaine et qui fournit un éventail plus varié et plus précis de concepts.
- Ontologie : une ontologie est une description formelle d'un domaine à travers ses concepts et les relations qui existent entre eux.

Dans ce qui suit nous présentons les deux principales ressources biomédicales, à savoir le thésaurus MeSH (*Medical Subject Headings*) et le metathesaurus UMLS.

3.2.1. Le thesaurus Mesh

Le MeSH est un thésaurus numérisé. Il a été développé par la *National Library of Medicine* (NLM), principalement pour indexer la base bibliographique MEDLINE. Le MeSH est une liste structurée de termes médicaux organisés en 16 arborescences indépendantes correspondant chacune à une catégorie MeSH. Au fur et à mesure que l'on descend dans la hiérarchie, les termes sont de plus en plus spécifiques. Ces termes sont appelés « descripteurs » car ils expriment de manière précise et spécifique le contenu d'un document. Les descripteurs sont au nombre de 23 000 (en 2005). Par exemple la catégorie « A » correspond à l'anatomie (*Anatomy*), la catégorie « B » aux organismes (*Organisms*), la catégorie « C » aux noms de maladies (*Diseases*), etc. Chacune de ces catégories contient plusieurs sous catégories qui constituent les différents niveaux de la hiérarchie. Par exemple « C01 » pour la catégorie « Infections bactériennes et mycoses » (*Bacterial Infections and Mycoses*), « C02 » pour « Maladies virales » (*Virus Diseases*) ou encore « C03 » pour « Maladies parasitaires » (*Parasitic Diseases*).

MeSH comprend essentiellement des termes, des concepts, des descripteurs, des relations et des qualificatifs. Dans ce qui suit, nous détaillons ces éléments et leurs règles.

- Terme: un terme est un mot ou ensemble de mots exprimant une notion.

- Concepts : un concepts comprend un ou plusieurs termes synonymes et porte le nom d'un de ces termes, dit terme préféré.
- Relation : dans MeSH, il existe trois 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 et peut véhiculer plusieurs sens :
    1. Relation « est une partie de » (méronymie),
    2. Relation « est un type de » (hyponymie),
    3. Relation « est sémantiquement proche de » (aboutness),
- Descripteur : un descripteur est constitué d'un ou plusieurs concepts de significations proches et porte le nom d'un de ces concepts, dit préféré. C'est le terme choisit lors de l'indexation comme descripteur parmi un ensemble de termes équivalents. Les autres termes présentent une relation sémantique avec le concept préféré, c'est-à-dire soit une relation hiérarchique (générique ou spécifique) soit une relation associative (associé). Enfin, tous les termes d'un descripteur sont des équivalents documentaires ou des termes d'entrée qui sont renvoyés au descripteur dans le cadre de l'indexation et de la recherche documentaire. Certains descripteurs ont plusieurs localisations, au sein de la même catégorie ou de catégories différentes, ainsi que plusieurs codes alphanumériques représentant chacun une localisation.
- Qualificatif : ce sont les termes qui permettent de préciser un aspect particulier d'un descripteur. Ils sont toujours associés à un descripteur, mai tous les qualificatifs ne sont pas associables à tout les descripteurs, en faite, seule une sélection de qualificatifs peut être associée à un descripteur donné.

La table I.2 montre un extrait de l'arborescence C de MeSH.

C<Diseases>

C02<Virus Diseases>

C02.081<Arbovirus Infections>

C02.109<Bronchiolitis, Viral>

C02.182<Central Nervous System Viral Diseases>

C02.182.500<Encephalitis>

C02.182.550<Meningitis>

C02.182.600<Myelitis>

C02.256<DNA Virus Infection>

C02.290<Encepahalitis, Viral>

C02.928<Tumor Virus Infections>

C10<Nervous System Diseases>

C18<metabolism et nutrition, maladies>

Table I.2. Extrait de l'arborescence C de MeSH

Le MeSh est caractérisé par la poly hiérarchie, certains descripteurs peuvent apparaître dans plusieurs branches de l'arborescence (voir Figure I.3), c'est-à-dire qu'un même terme peut appartenir à plusieurs catégories du MeSH et par conséquent, il peut donc avoir plusieurs identifiants.

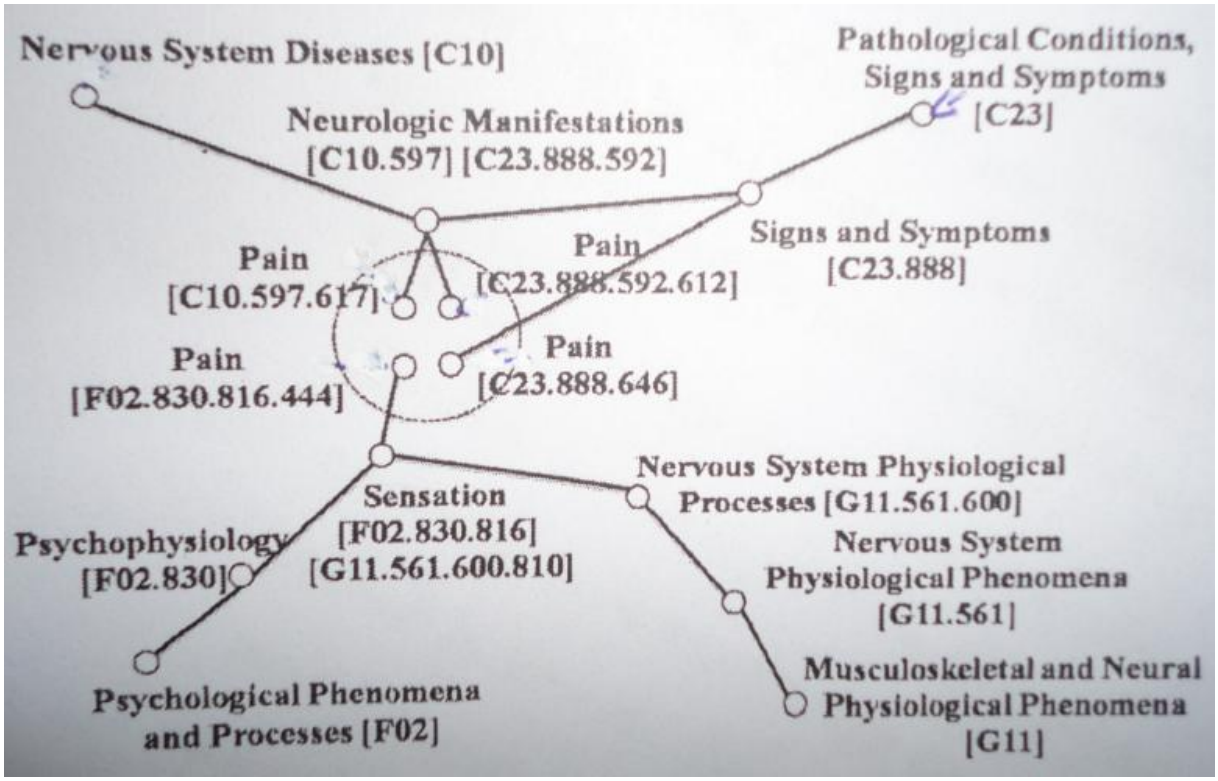


Figure I.3. Le concept Pain de MeSH

3.2.2. Le metathesaurus UMLS

La NLM a proposé la conception et le développement d'un système de langage médical unifié, UMLS « *Unified Medical Langage System* » afin d'améliorer l'accès à l'information médicale provenant de sources différentes en permettant aux différentes banques de données de communiquer avec un langage de référence commun.

L'UMLS tente de regrouper tous les thésaurus, nomenclatures et classifications existantes utilisées pour la gestion des données en santé, les bases de données bibliographiques et les dossiers patients. L'UMLS est un système qui conjugue trois bases de connaissances : le métathesaurus (qui regroupe tout les termes), le réseau sémantique (qui regroupe toutes les relations) et le SPECIALIST Lexicon (qui contient les informations morphologiques, syntaxiques et orthographiques).

- Le métathesaurus : constitue la base unifiée des concepts médicaux. Il comprend des synonymes, des variations lexicales et des concepts associés afin de dresser la liste de tout le vocabulaire des expressions médicales disponibles. UMLS regroupe les différents termes synonymes (issus de différentes terminologies) sous un même concept. Pour chaque concept,

il lui attribue un terme préférentiel, éventuellement des termes synonymes, des variantes lexicales et un identifiant unique (le CUI).

- Le réseau sémantique : alors que le metathesaurus fournit une liste de tout le vocabulaire des expressions médicales disponibles, le réseau sémantique apporte une structure à ces termes. Cette structure permet notamment de fournir une catégorisation cohérente à tous les concepts ainsi que les relations entre eux. Le réseau fournit des informations sur l'ensemble de types sémantiques, ou catégories, ou catégories, qui peuvent être affectés à des concepts et il définit l'ensemble des relations qui peuvent exister entre eux. Il contient 134 types sémantiques et 54 relations.
- Le SPECIALIST Lexicon : il contient les informations syntaxiques et morphologiques nécessaires au traitement automatique de la langue anglaise. Chaque entrée possède une forme de base (le lemme), une catégorie syntaxique, un identifiant unique et éventuellement des variantes orthographiques.

3.3. Indexation des documents biomédicaux

L'indexation en RI biomédicale a pour objectif de faciliter l'accès à la littérature biomédicale en affectant à chaque document une liste de termes désignant des concepts issus d'une ou de plusieurs terminologies biomédicales (Névéol *et al.*, 2006; Darmoni *et al.*, 2009). Ainsi l'indexation à base de concepts (ou conceptuelle) est venue comme une solution pour l'indexation classique, elle permet de mieux organiser, de structurer et surtout de faciliter l'accès à l'information biomédicale.

L'indexation des documents biomédicaux se base sur deux principales étapes :

- l'identification des concepts médicaux,
- et l'extraction des relations entre ces concepts.

3.3.1. Identification des concepts

L'identification des concepts est une étape primordiale dans le processus d'indexation. Son objectif est de trouver les termes pertinents du document qui sont liés aux concepts du domaine. Les approches de reconnaissance de termes dans le domaine biomédical sont subdivisées en trois catégories : les approches basées sur les règles, les approches basées sur le dictionnaire et les approches basées sur l'apprentissage supervisé et statistique.

L'approche d'identification de concepts basée sur les règles nécessite la création et la maintenance de patron d'extraction. Ces patrons sont développés manuellement par des experts du domaine pour extraire tous les concepts du document. Les travaux qui se base sur cette approche intègrent des propriétés orthographiques (exemple : majuscule, minuscule, accents…), lexicales (exemple : masculin, féminin, pluriel, singulier….) et morphosyntaxique (exemple : lemme, racine…) dans les règles. Par ailleurs, ces derniers sont enrichies par l'utilisation des constituants de termes (exemple : préfixe, suffixes, abréviations, acronymes…) pour mieux identifier les concepts.

Les méthodes de reconnaissance de termes basées sur le dictionnaire utilisent des ressources terminologiques pour localiser les termes dans le document. La précision de la reconnaissance de termes dépend complètement de la qualité et de l'exhaustivité du dictionnaire. Les problèmes qui peuvent se poser avec cette méthode sont liés à la variation des termes utilisés dans le document comme la synonymie, polysémie, abréviations, acronymes.

Plusieurs travaux d'identification de concepts se basent sur l'apprentissage supervisé. Parmi les méthodes d'apprentissage, les vecteurs de support (SVM) sont les plus puissants et efficaces pour la reconnaissance des entités nommés. La reconnaissance d'entités nommées est une sous tâche de l'activité d'extraction d'information dans des corpus documentaires. Elle consiste à rechercher un mot ou un groupe de mot catégorisables dans des classes telles que les noms de personnes, noms d'organisations ou d'entreprise, etc.

L'identification des concepts est la tâche la plus difficile dans l'indexation des documents biomédicaux et chaque approche présente sa propre vision du problème. Pour profiter des avantages de chacune, plusieurs outils d'indexation proposent de combiner plus qu'une méthode d'identification de concepts dans le même système d'indexation.

3.3.2. Extraction de relations entre concepts

Après l'identification des concepts vient l'étape d'extraction de relation entre ces concepts. L'extraction de relation tente en général d'identifier diverses relations dans un ensemble de document. Pour pouvoir extraire ces relations les chercheurs ont proposé différents types d'approches.

Un premier type d'approches, dit statistiques, se base sur la co-occurrence de termes spécifiques et/ou des techniques d'apprentissage automatique pour l'extraction de relations. L'apprentissage automatique consiste dans ce cas à définir un ensemble d'attributs pertinents

pour la détection de relations (types des entités médicales sources et cibles, mots entre les deux entités médicale, etc.), puis entrainer un classifieur sur un corpus d'entraînement ou des relations ont été annotées, et enfin, utiliser le modèle ou les règles de classifications apprises grâce à ce corpus pour extraire les relations à partir d'autres corpus.

Une deuxième catégorie d'approches, dite linguistique, se base sur des patrons pour extraire les relations. Un patron peut être vu comme un modèle de phrase qui identifie une forme particulière d'expression de la relation à extraire. Le processus d'extraction consiste alors à rechercher des correspondances entre les patrons et les phrases du corpus ciblé pour l'extraction.

Une troisième catégorie d'approches, dites hybride, combine les deux premières pour profiter de leurs avantages respectifs et éviter certains de leurs inconvénients.

Le deuxième chapitre sera entièrement consacré à la présentation de cette étape, ainsi qu'à la présentation des différents travaux effectués dans ce cadre.

3.4. Techniques de recherche d'information dans les documents médicaux

Dans le contexte de la recherche d'information biomédicale, l'efficacité des systèmes de recherche est influencée par le degré de chevauchement des termes entre les requêtes des utilisateurs et les documents pertinents. Quand un utilisateur cherche l'information dans une collection de documents, il peut formuler la requête en utilisant d'autres expressions pour mentionner la même information dans le document. Cela cause un problème d'incompatibilité des termes qui donne des résultats de recherche pauvres. Dans le domaine biomédical, les documents contiennent de nombreuses expressions différentes ou des variantes de termes pour un même concept, comme la synonymie ('*cancer*', '*tumor*' sont des synonymes du concept '*neoplasm*'), les abréviations (AMP est synonyme de *Adenosime Monophosphate*), ou encore les variations lexicales tel que la différence dans le cas d'inflexion singulier-pluriel .Pour remédier à ce problème, plusieurs travaux se sont intéressés à l'expansion de la requête et l'expansion des documents.

L'objectif de l'expansion de requête est d'augmenter la performance de recherche en augmentant la probabilité de chevauchement des termes entre la requête et les documents qui sont susceptibles d'être pertinents pour répondre au besoin de l'utilisateur.

La phase d'expansion de document vise à accroître le degré de chevauchement des mots entre la requête des utilisateurs et les documents observés. L'expansion document peut contribuer à renforcer la sémantique du document en élargissant le contenu du document avec les termes les plus informatifs.

La différence entre l'expansion du document et l'expansion de la requête concerne le moment de l'expansion ; l'expansion du document se réalise lors de l'indexation tandis que l'expansion de la requête est exécutée lors de la recherche.

3.5. Evaluation des systèmes de recherche d'information biomédicale

Des campagnes d'évaluation destinées à stimuler et favoriser l'émergence de nouveaux SRI ont été menées, très régulièrement depuis plusieurs années. On citera parmi les plus importantes le programme américain TREC.

Le *Text REtrieval* Conference (TREC) est un programme conçu comme une série d'ateliers dans le domaine de la Recherche d'information. Ce programme est soutenu conjointement par le National Institute of Standards and Technology (NIST) et par l'*Advanced Research* and *Development Activity* (ARDA) Center du Département de la Défense des États-Unis.

La campagne d'évaluation TREC a eu lieu pour la première fois en 1992, elle fournit des cadres d'évaluation des approches de recherche d'information ainsi qu'un forum pour comparer les résultats obtenus par les différentes équipes 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, etc. Il existe plusieurs pistes de recherche, nous nous intéressons particulièrement à la piste suivante : TREC Genomics pour la RI de la littérature biomédicale.

3.5.1. TREC Genomics

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. Les documents de MEDLINE ont été extraits pour développer des collections tests dédiées à l'évaluation des approches de RI dans TREC Genomics. 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*). Les requêtes ont été construites en se basant sur les besoins d'informations des scientifiques dans le domaine biomédical.

3.5.2. Protocole d'évaluation TREC

Pour chaque requête, les 1000 premiers documents restitués par le système sont examinés et des précisions sont calculées à différents points (à 1, 2, 5, 10, 15, 30, 100 et1000 premiers documents restitués). La précision à x (exemple précision à 5) définit le taux de documents pertinents parmi les x premiers documents retrouvés.

Une précision moyenne *MAP* est ensuite calculée pour chaque requête. Il s'agit de la moyenne des précisions de chaque document pertinent pour cette requête. La précision d'un document est la précision à *x*, tel que *x* est le rang de ce document dans l'ensemble des documents pertinents retrouvés.

Finalement, les précisions moyennes pour l'ensemble des requêtes sont calculées permettant d'obtenir une mesure de la performance globale du système.

4. Conclusion

Nous avons introduit en première partie de ce chapitre la recherche d'information. La deuxième partie a été entièrement consacrée à l'indexation et recherche d'information biomédicale. Nous avons présenté les documents du domaine biomédical. Le traitement des documents biomédicaux revient principalement à traiter le problème d'ambiguïté venant de l'utilisation du langage naturel. Puis nous avons présenté les ressources biomédicales qui tente de standardisé le lexique utilisé dans le domaine. Chacune d'elles à un objectif, par exemple, UMLS vise à collecter et unifier les différentes ressources en une seule, le thésaurus MeSH sert à l'indexation de la littérature biomédicale. Ces différentes terminologies ont constitué le vocabulaire de l'indexation. Nous avons expliqué le processus d'indexation les techniques de recherche d'information dans les documents médicaux. Et enfin nous avons présenté la campagne d'évaluation de RI biomédicale (campagne TREC).

Actuellement, dans le cadre de la recherche d'information dans les documents biomédicaux, la découverte de relations pertinentes entre concepts biomédicaux est

l’occupation majeurs de nombreux chercheurs, qui tentent avec différents moyens de trouver la meilleure approche afin d’extraire des relations susceptibles d’aider les acteurs du domaine biomédical, de les orienter et couvrir leur besoin en information. Nous présenterons dans le chapitre suivant les différentes techniques d’extraction de relations entre concepts dans le domaine biomédical.

II. Etat de l'art sur l'extraction de relation entre concepts biomédicaux

1. Introduction

Une vaste quantité de connaissances biomédicales est contenue dans des documents en texte libre. Les recherches actuelles pour extraire la sémantique de ces sources se concentrent sur l'identification des concepts médicaux (gènes, protéines, maladies, etc.) à partir de texte libre et sur l'extraction de relations entre les concepts identifiés.

L'extraction de relations tente en général d'identifier diverses relations dans un ensemble de documents. Dans des corpus de spécialités, l'extraction de relations implique principalement des entités appropriées à la spécialité. Cette particularité engendre des relations plus spécifiques entre les concepts. Dans le domaine médical, il peut s'agir de trouver diverses interactions entre gênes, de trouver la relation existante entre une maladie et un traitement, ou encore une maladie et un examen cliniques. Il existe pour cela trois types d'approches :

- les approches statistiques,
- les approches linguistiques,
- et les approches hybride.

Dans ce présent chapitre, nous présentons les différentes approches et techniques d'extraction de relation entre concepts biomédicaux.

2. Les approches d'extraction de relations entre concepts biomédicaux

2.1. Les approches statistiques

Cette première catégorie d'approche se décline en deux sous catégories :

- les approches qui se basent sur des mesures statistiques,
- et les approches qui se basent sur apprentissage automatique.

2.1.1. Approches qui se basent sur des mesures statistiques

Parmi les approches qui se basent sur des mesures statistiques, il existe des approches qui se basent sur la cooccurrence des concepts pour l'extraction des relations entre ces concepts, et d'autre qui se base sur le degré de corrélation de ces concepts.

L'idée principale des approches qui se basent sur la cooccurrence est que les termes qui co-occurrent ensemble ont de fortes chances d'être lies. La notion de cooccurrence fait

référence au phénomène général par lequel des mots sont susceptibles d'être utilisés dans un même contexte. Autrement dit, on considère qu'il y a cooccurrence lorsque la présence d'un mot dans un texte donne une indication sur la présence d'un autre mot. Deux mots considérés co-occurrents sont associés fortement l'un à l'autre, le contexte commun dans lequel ils apparaissent est large : il peut s'agir d'un paragraphe, d'un texte ou d'une collection de textes selon l'application, selon l'usage prévu pour cette information. Dans le cadre de l'approche statistique du traitement de la langue naturelle, plusieurs façons d'attribuer un score d'association à une paire de mots ont été élaborées. Parmi ces mesures, certaines sont directement tirées de la discipline des statistiques (Test du x^2) alors que d'autres sont nées dans le domaine de la recherche d'information (le ratio de vraisemblance,l'information mutuelle).

Pour les approches qui se basent sur la corrélation, l'idée principale est d'étudié l'intensité de la liaison qui peut exister entre concepts biomédicaux. Cette étude est faite à l'aide des logiciels statistiques tel que le logiciel R.

Dans ce qui suit, nous présentons quelque approche d'extraction de relations entre concepts dans le domaine biomédical qui se basent sur des méthodes utilisant la cooccurrence ou la corrélation de concepts.

2.1.1. Les travaux de (Stapley et al., 2000)

Les auteurs ont conçu un système d'extraction de relation entre deux gènes, qui s'appuie sur leur fréquence d'apparition. Si la fréquence entre deux gènes est signifiante, les gènes sont nécessairement en relation. Deux noms de gènes peuvent apparaître dans un même texte pour plusieurs raison, comme par exemple, deux gènes sont reliés par une relation physiologique (une interaction physique entre ces deux gènes). Le système qu'ils ont proposé peut aussi déterminer la nature des relations existences entre les gènes.

2.1.2. Les travaux de (Stephens et al., 2001)

Le système proposé se fonde sur des statistiques de cooccurrence pour repérer les relations intervenant entre les gènes. La méthode proposée se compose de trois étapes :

1. La première étape consiste à représenter chaque document en un vecteur de pondération. Chaque document d_i est converti en un vecteur de dimension M, ou M désigne le nombre de terme dans le thésaurus. La fonction de pondération utilisée est la suivante :

$$W_i[k] = T_i[k]* \log (N/n[k])$$

Ou: $W_i[k]$: le poids du k^{ieme} terme dans le document

$T_i[k]$: le nombre d'occurrence du terme T_k dans le document i ;

$\log (N/n[k])$: la fréquence inverse du terme T_k dans la base ;

N : le nombre de document dans la base ;

n[k] : le nombre de document contenant T_k dans la base.

2. Après cette étape, les auteurs ont procédé au calcul d'une mesure d'association. Cette mesure prend en paramètre le poids de deux termes de gènes K et L.

$$\text{association } [K] [L]= \sum w_i[K] * w_i[L] \quad K = 1...m, L = 1...m$$

3. En suite, la troisième étape consiste à trouver le type de relation intervenant entre chaque paire de gène. Une paire de gène est en relation si la mesure d'association (co-occurrence) entre ces deux gènes est supérieure à un seuil donné. Pour cela les auteurs proposent le score suivant :

$$\text{Score } [k][L][m]= \sum p_i$$

Pour trouver le type de relation entre les termes de gène, les auteurs ont utilisé un thésaurus qui contient toute les relations possibles entre ces termes. Ils ont appliqué les phrases contenant les cooccurrences des gènes à ce thésaurus, puis à chaque fois qu'une relation est trouvée le score est incrémenté. La relation qui obtient le maximum des scores est retenue comme la relation intervenant entre ces deux termes de gènes.

2.1.3. Les travaux de (Eya Znaidi et al., 2011)

Les auteurs ont proposé une méthode de recherche de corrélations d'informations pertinentes dans un corpus de documents biomédicaux. Ils ont abordé le problème de recherche de corrélations pertinentes comme un problème d'analyse statistique multidimensionnelle sur un corpus d'informations annotées par les mesures classiques d'importance telles que la fréquence des termes (tf), leur fréquence documentaire inverse (idf), ainsi que les relations sémantiques traduites dans les terminologies utilisées, comme

source d'évidence. Les concepts qu'ils ont utilisés sont extraits de MeSH, ces concepts représentaient le noyau sémantique des documents.

Dans leur approche, les auteurs ont exploité la notion de rareté de concepts dans les documents. Pour chaque concepts rare, ils ont identifié les concepts associés les mieux corrélés. Leur algorithme est composé de trois étapes principales (figure II.1) :

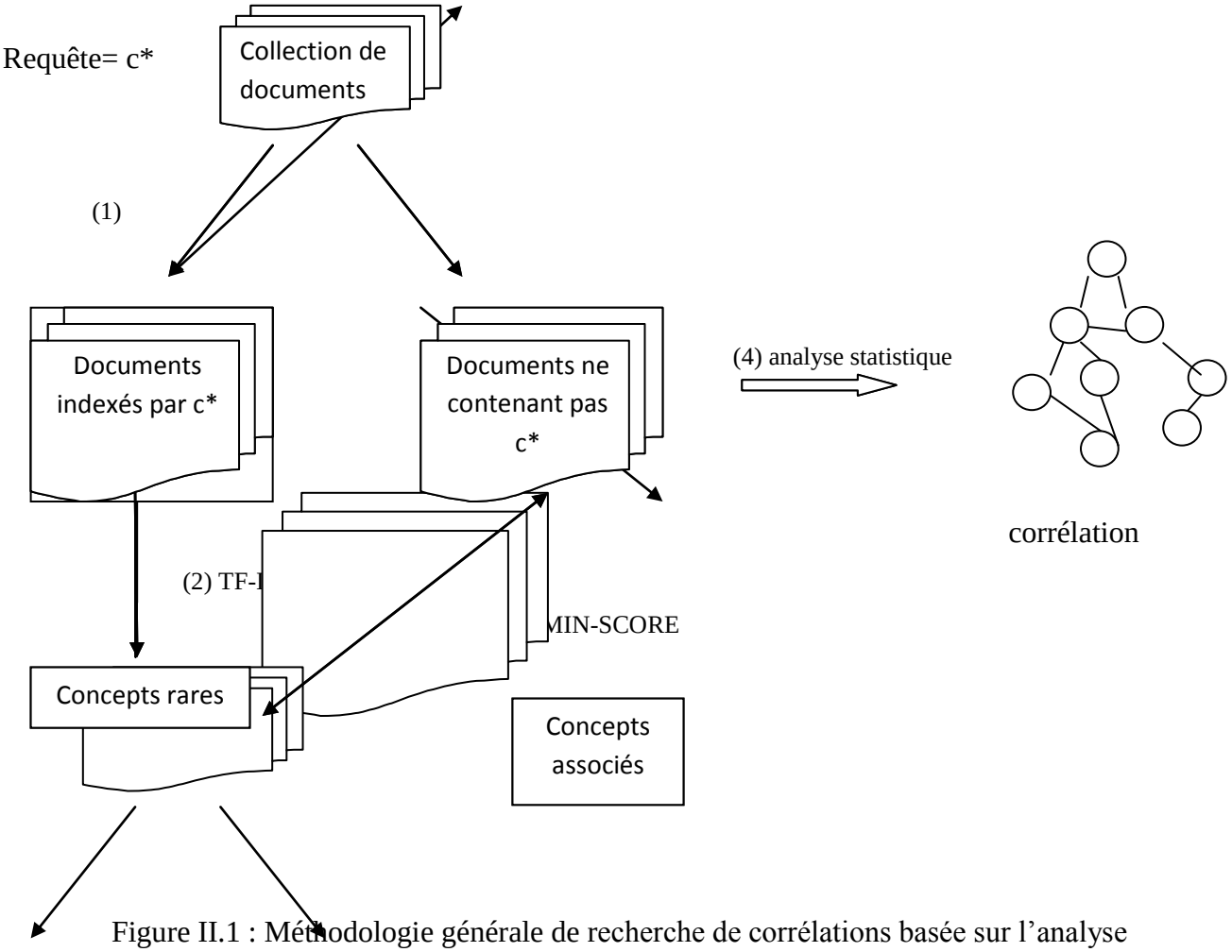


Figure II.1 : Méthodologie générale de recherche de corrélations basée sur l'analyse statistique

1. Recherche de concepts rares : elle consiste à trouver l'information rare correspondant au concept de début. Cela permet d'identifier les concepts les mieux corrélés au concept de la requête mais qui ne sont pas fréquent dans la littérature. Les auteurs ont proposé la formule statistique tf * idf pour trouver ces concepts rares, c'est une mesure statistique qui calcule la fréquence de deux concepts dans la littérature ainsi que son importance dans tous les documents de la littérature.

2. Identification des concepts associés aux concepts rares : la notion de concepts associés permet d'étudier les corrélations entre la littérature des concepts rares et les autres concepts, l'approche décrite ici est bâti sur la mesure statistique MIM-score qui se base sur la probabilité d'apparition de deux concepts dans l'ensemble de documents.

3. Analyse statistique multidimensionnel : générer les nuages de points qui représentent les corrélations pertinentes en utilisant le logiciel R statistique. Ce logiciel prend en entrée un tableau de données (concepts rares, concepts associés), il permet de générer deux tableaux, le tableau disjonctif complet qui révélera les concepts de liaisons, et le tableau de Brut. Ce logiciel génère en suite, les nuages de points représentant les corrélations pertinentes.

2.1.4. les travaux de [H.Abdoune et al., 2011]

L'objectif des auteures était d'analyser les cooccurrences de termes majeurs présents dans les notices de la base MEDLINE pour tenter de mesurer l'intérêt que peuvent présenter leurs associations. Leur méthode se base sur quatre étapes :

1. La première étape consiste à identifier dans la table MRCOC du méta thésaurus de l'UMLS, tous les couples de termes MeSH qui sont co-occurrents dans les notices de MEDLINE. Dans cette table, on trouve pour chaque couple (CUI_1, CUI_2) des attributs de cooccurrence qui qualifient le premier concept CUI_1, et une fréquence de cooccurrence avec CUI_2.

2. La deuxième étape consiste à définir le support (CUI_1, CUI_2) comme la probabilité d'apparition simultanée d'un couple de termes MeSH (CUI_1, CUI_2) dans une notice. Le support est calculé comme suit :

$$\text{Support}(CUI_1, CUI_2) = P(CUI_1 \wedge CUI_2)$$

3. La troisième étape consiste à accorder une confiance à l'association de deux termes MeSH. La confiance (CUI_1, CUI_2) est la probabilité que le concept CUI_2 soit présent dans une notice sachant la présence du concept CUI_1. La confiance est le rapport entre le Support (CUI_1, CUI_2) et la probabilité d'avoir le concept CUI_1.

$$\begin{aligned}\text{Confiance}(CUI_1, CUI_2) &= \text{Support}(CUI_1, CUI_2) / P(CUI_1) \\ &= P(CUI_1 \wedge CUI_2) / P(CUI_1) \\ &= P(CUI_2) / P(CUI_1)\end{aligned}$$

4. La quatrième étape consiste à mesurer l'intérêt de l'association de deux termes MeSH. Pour cela les auteurs ont introduit la notion du lift d'une association. Cette mesure est calculée comme suit :

$$\text{Lift}(CUI_1, CUI_2) = \text{Confiance}(CUI1, CUI2) / P(CUI2)$$
$$= P(CUI1 \wedge CUI2) / (P(CUI1) \times P(CUI2))$$
$$= P(CUI2 / CUI1) / P(CUI2)$$

Ce sont donc les cooccurrences dont le Lift est supérieur à 1 qui sont retenues puisqu'elles indiquent la présence d'une relation entre CUI_1 et CUI_2.

2.2. Techniques à base de SVM

Pour les méthodes qui se basent sur l'apprentissage automatique, la plupart des travaux utilisent des classifieurs à base de SVM (Supports Vector Machines) pour la classification des relations. Les SVM sont une classe d'algorithme d'apprentissage. Etant données plusieurs catégories définies (dans le cas présents, les différentes relations à reconnaitre), une telle technique s'appuie sur un ensemble d'exemples d'entraînement pour pouvoir prendre une décision de classification sur le corpus ciblé. Chaque exemple doit être décrit par un ensemble d'attributs qui doivent être suffisamment discriminant pour assurer une bonne performance de classification. Dans ce qui suit, nous présentons quelques travaux qui se basent sur cette technique.

2.2.1. Les travaux de (Roberts et al., 2008)

Le but des auteurs était d'extraire des relations entre des entités (médicament, test, problème, …etc.) et des modifieurs (mot sui qualifie une entité) dans des dossiers de patients atteints d'un cancer. Les auteurs ont ciblés sept types de relation (ex : has_indication, has_location….), et ils ont associé pour chaque types de relation deux types d'arguments ; une relation ne peut exister qu'avec certaines entités spécifiques, par exemple la relation « has_location » peut exister seulement entre une condition et un locus (un locus c'est un emplacement précis d'un gène sur le chromosome qui le porte).

Les auteurs ont considéré l'extraction de relation comme une tâche de classification, et ont proposé une méthode fondée sur les SVM. Le classifieur SVM est entrainé de sorte à pouvoir attribué un type de relation à chaque paire d'entités. Une paire d'entités est un couple

d'entités qui peuvent être ou ne pas être les arguments d'une relation. Pour chaque document ils ont construit toutes les paires d'entités possibles dans les deux cas, autrement dit, dans le cas ou la première entité représente le premier argument et la deuxième entité représente le deuxième argument, et dans le cas contraire.

Pour la classification les auteurs ont utilisé trois types d'attributs : des attributs lexicaux comme les mots formant les entités en les mots situés entre les entités, des attributs morpho-syntaxiques (ex : toute les catégories verbales sont regroupées en VB) et des attributs sémantiques comme les types des entités en relation ou les types des autres entités de la phrase. Ces attributs ont été utilisés pour l'annotation des exemples d'entrainements qui seront appliqué au corpus ciblé pour l'extraction de relation.

2.2.2. Les travaux de (Frunza et Inkpen, 2010)

L'objectif des auteurs était d'extraire des relations sémantiques qui peuvent exister entre une maladie et un traitement à partir de document extrait de la base MEDLINE. Ils se sont intéressés à trois types de relations qui sont *cure*, *prevent* et *side effect*.

Pour l'extraction des relations, les auteurs ont utilisé l'outil Weka (Hall et al., 2003), qui est un ensemble de logiciels d'apprentissage automatique. Ils ont testé six modèles pour apprendre les relations qui sont les suivant : *Decision-based models* (modéles basés sur les arbres J48), Probabilistic models (CNB), Adaptive learning (AdaBoost), Linear classifier (SVM) et Classifier (ZeroR). Les résultats obtenus montrent que les modèles probabilistes et linéaires donnent les meilleurs résultats.

2.2.3. Les travaux de (Uzuner et al., 2010)

L'objectif des auteurs était d'extraire des relations sémantiques entre des problèmes, des tests et des traitements dans des comptes-rendus médicaux. Ils ont classé les relations en six catégories (ex : Present disease-treatment relation type, Present symptom-treatment relation type, …). Ils se sont basés sur les catégories sémantiques des concepts (maladie, traitement,…) et des affirmations faites sur ces concepts (affirmation de l'annotation des concepts) pour définir les types sémantiques des relations.

Avant de procédé à la classification des relations, le corpus a d'abord été annoté manuellement (annotation des concepts, relations ...). Les phrases retenues sont uniquement celles qui contiennent les paires de concepts impliqués dans les relations. En suite, pour la

classification des relations, les auteurs ont utilisés un classifieurs SR (classifieurs de relation sémantique) composé de six classifieurs SVM multi-classe correspondants aux six catégories de relation (chaque type de relation correspond à une classe). Des modéles SVM distinct on été généré pour chaque type de relation, par exemple, un SVM est entrainé à reconnaître des relations de types maladie-symptôme.

Pour chaque phrase le classifieurs SR utilise les information de l'annotation des concepts afin de pouvoir déterminé le type de relation, par exemple, une paire de concepts qui se compose d'une « present disease » et « treatment » est une relation de type « present disease-treatment ». Le classifieurs SR utilise le type de relation pour invoquer uniquement le classifieurs SVM pour cette catégorie de relation. Un SVM est entrainé par un ensemble d'exemples représentatifs de la relation ciblée. Ces exemples sont décrits avec un ensemble d'attributs. Les attributs utilisés par les auteurs pour la classification sont les suivant : attributs de surface comme l'ordre des concepts ou la distance entre les concepts, attributs lexicaux comme les mots entre les concepts, attributs syntaxiques comme les verbes ou les mots clés.

2.2.4. Les travaux de (Minard et al., 2011)

L'objectif des auteurs était de reconnaitre différents types de relations entre des concepts (les problèmes, les traitements et les tests) dans des comptes rendus médicaux. Ils se sont intéressés à huit types de relation (ex : la relation TrWP « le traitement aggrave le problème », la relation TrCP « Le traitement cause le problème », la relation TrAP « Le traitement est administré en raison du problème », …etc.).

Pour atteindre leur objectif, les auteurs ont proposé trois approches basées sur l'apprentissage supervisé. Dans leur première approche (Minard et al., 2011b) ils ont utilisé un classifieur SVM avec des attributs de différents types pour détecter et classer les relations. Dans la deuxième approche (Minard et al., 2011a), ils ont ajouté en plus du classifier SVM une fonction Kernel qui calcul la similarité entre deux arbre et donc entre deux structure syntaxique. Et dans leur troisième approche (Minard et al., 2012), ils ont d'abord procédé à la simplification de phrases avant l'extraction de relations. Dans ce qui suit, nous allons décrire ces trois approches.

Dans (Minard et al ., 2011b) les auteurs ont considéré l'identification de relation comme une tâche de classification multi-classe, chaque catégorie de relation était considérée

comme une classe. Pour cela, ils ont utilisé un classifieur SVM avec la bibliothèque LIB-SVM. Les attributs utilisés pour la classification sont : des attributs de surface (ex : l'ordre des concepts), des attributs lexicaux (ex : les mots et leurs radicaux), des attributs syntaxiques (ex : la présence d'une préposition entre les concepts), des attributs sémantiques (ex : les types des concepts), … etc.

Dans (Minard et al., 2011a), les auteurs voulaient étudier l'apport de la syntaxe (informations syntaxique) pour l'extraction de relation. Pour cela, les auteurs ont choisi d'utiliser en plus du classifieur SVM une fonction Kernel qui mesure la similarité entre deux arbres en comptant le nombre de fragments en commun. Les attributs utilisés pour la classification étaient les même que les attributs utilisés dans (Minard et al., 2011b), auxquels les auteurs ont ajouté des informations syntaxiques. Ces informations proviennent des arbres de constituant ; ces arbre syntaxique ont été produits par l'analyseur de Charniak/McClosky (McClosky, 2010). Les phrases ont été analysées après remplacement des abréviations, normalisation des dates, âges, noms propres et nombres, et annotation des concepts. L'analyseur donne en sortie un arbre représentatif de la phrase (voir figure II.2), cet arbre est appelé le sous arbre complet. A partir de cet arbre, les auteurs ont produits un sous-arbre, le sous-arbre minimal complet (voir figure II.3) reliant les deux entités en relation. Ce sous-arbre correspond au chemin le plus court pour aller d'un concept à l'autre. Et à partir du sous-arbre minimal complet, un sou-arbre plus restreint est construit (voir figure II.4) en supprimant le contexte gauche ou droit (cela dépend de la phrase). A partir de cet arbre, les auteurs ont calculé deux attributs : la taille du chemin reliant les deux entités, et le constituant du nœud racine du sous-arbre.

La phrase traitée dans les figures est la suivante : « 2 Low back strain requiring hospitalization for in pain 2002 », dans cette phrase la taille du plus petit chemin est de sept et le constituant du nœud racine du sous-arbre minimal est NP.

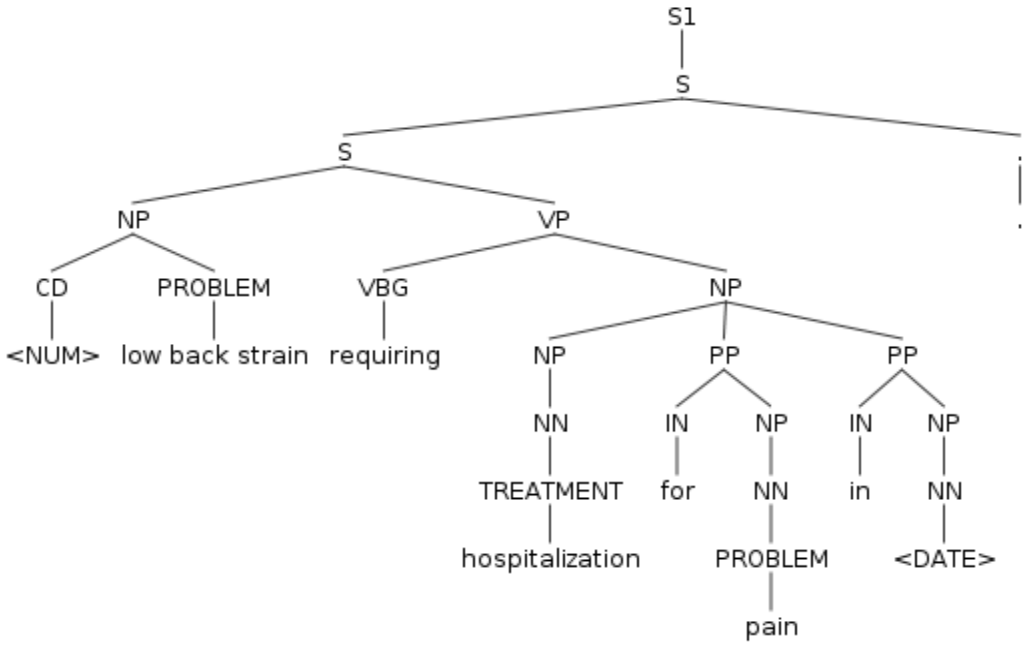


Figure II.2. Exemple d'arbre complet

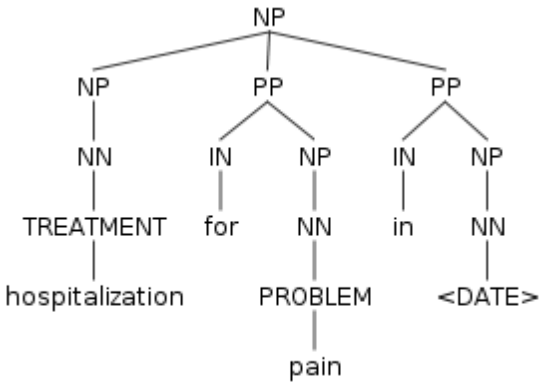


Figure II.3. Exemple du sous arbre minimal complet entre les deux entités

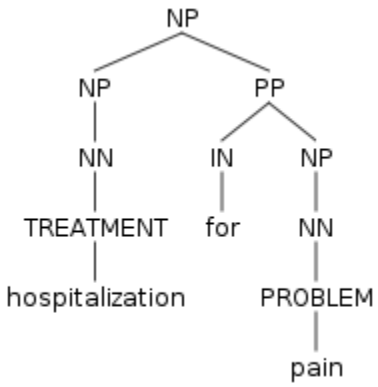


Figure II.4. Exemple du sous arbre minimal entre les deux entités

Dans (Minard et al., 2012), les auteurs ont d'abord procédé à la simplification des phrases puis l'extraire de relations. La simplification selon les auteurs, consiste à ne garder que ce qui est nécessaire à l'identification de la relation, et à supprimer les informations qui ne sont pas en rapport avec la relation ou qui peuvent perturber son identification.

Pour la simplification, les auteurs ont utilisé une méthode à base d'apprentissage, pour annoter dans les phrases les parties à garder et celles à supprimer. Quatre types d'annotation on été définis :

- L'annotation indispensable : permet de caractériser les mots qui portent l'expression de la relation.
- L'annotation utile : indique les mots qui renforcent la relation.
- L'annotation inutile : est associée aux mots n'apportant pas d'indices pour la classification de la relation.
- L'annotation « gênant » : sert à repérer les mots pouvant gêner la bonne classification de la relation.

Les auteurs ont choisi un classifier CRF (Champs Aléatoires Conditionnels) pour effectuer l'annotation des phrases. Les phrases annoté seront données en entré du classifieur SVM afin de classer que quelques phrases. Les résultats obtenus sont quasiment les même que ceux obtenus sans la simplification des phrases. Ce qui a mené les auteurs à dire que la simplification des phrases a un effet sur la classification mais ne permet pas encore de l'améliorer.

2.2. Les approches linguistiques

L'approche linguistique ou l'approche à base de patron est l'une des approches les plus utilisé pour l'extraction de relations. L'idée principale de cette approche est de définir dans un premier temps un ensemble de patrons associé à une relation. Un patron est une expression régulière décrivant un modèle de phrase ou les entités médicales sont présentes à des emplacements spécifiques. Ces patrons seront ensuite projetés sur le corpus de texte pour extraire de nouvelles relations, c'est-à-dire identifier de nouveau couple de termes correspondants à la relation spécifiée. La construction de patron est alors une étape préliminaire afin de découvrir de nouvelles relations dans un corpus. Dans ce qui suit, nous allons présenter quelques travaux qui se sont basés sur cette technique.

2.2.1. Les travaux de (Khoo et al., 2000)

L'objectif des auteurs était d'extraire des relations de causalité à partir des résumés de la base médical MEDELINE. Ce qui se traduit par un repérage dans les textes des expressions exprimant une relation de causalité entre deux entités lexicales, exemple : les passages du type « A à cause de B » ou « A est un effet de B ». L'approche développée par les auteurs pour l'extraction de relation se base sur des patrons de graphe qui représente les relations, et des arbres de dépendances syntaxiques issues de l'analyse des phrases.

Avant de procédé à l'extraction de relation, les auteurs ont d'abord analysé l'ensemble des documents pour identifier les marqueurs linguistiques employé pour indiquer des relations de causalité, comme par exemple les expressions « because of », « because » ou encore « the result was ». En suite, les auteurs ont construit un ensemble de patrons qui décrivent les différentes formes qu'une relation de causalité peut prendre. Ils ont construit un patron de sous forme de graphe conceptuel linéaire pour chaque identifiant de causalité. Par ailleurs, ils ont procédé à l'analyse syntaxique des phrases contenues dans les documents considérés. Pour cela, ils ont utilisé l'analyseur *Conexor's Funstional Dependency Grammar of English* (FDG) (http://www.conexor.fi), qui génère une représentation de la structure syntaxique de la phrase (arbre de dépendance syntaxique), qui est aussi sous forme de graphe conceptuel linéaire.

Le processus d'extraction de relation mis en place par les auteurs, consiste à faire correspondre les patrons de graphe avec les arbres de dépendance syntaxique, autrement dit, chercher les séquences communes ou les plus similaires possibles entre ces deux graphes.

2.2.2. Les travaux de (Lee et al., 2004)

Le but des auteurs était d'extraire des relations de type « traitement » entre une maladie et un traitement à partir de résumés de la base médicale MEDELINE (ces résumé concerné le domaine du cancer). La méthode proposée se base sur la construction manuelle de patron linguistique. Pour cela, ils ont d'abord identifié les phrases contenant les concepts médicament et maladie, par exemple, la phrase « *These results indicate that chronomodulated 5-FU and LV with L-OHP therapy could be an effective regimen for cases of irinotecan-resistant colon cancer* ». A partir de ces phrases, des patrons sont construit sous forme d'expression régulière prenant en compte la présence d'entités médicales à telle ou telle position de la phrase.

Un total de 224 patrons linguistiques a été manuellement construit par les auteurs. Ces patrons peuvent être regroupés dans les catégories sémantiques suivantes :

- Administration of treatment, ex: use of, using, administered.
- Treatment dosage, ex: low-dose, dose of.
- Mortality and survival, ex: mortality, extends the survival
- Therapy, ex: chemotherapy, treatment, regimen, drug.
- Clinical trial, ex: tested on, clinical trial
- Effect, ex: influence, results, effective

Une fois ces patrons construits, ils sont appliqué sur le corpus considéré pour extraire les relations ciblées.

2.2.3. Les travaux de (Mehdi Embarek et Olivier Ferret, 2008)

Les auteurs ont proposé une méthode à base de patron linguistique pour extraire des relations sémantiques entre des concepts médicaux (maladie, traitement, médicament, examen, symptôme). Ils définissent un patron linguistique comme un schéma lexico-syntaxique spécifique d'une relation sémantique intervenant entre deux entités. Ces patrons sont dit multi-niveau car ils s'appui sur des informations provenant de plusieurs niveau de traitement de textes (forme fléchi des mots, leur forme normalisé, … etc.).

Les relations considères étaient les suivantes :

- La relation « traite » : entre une maladie et un traitement
- La relation « soigne » : entre une maladie et un médicament

- La relation « détecte » : entre une maladie et un examen
- La relation « signe » : entre une maladie et symptôme

L'ensemble des concepts et des relations considérés sont représenté dans la figure 1.7 :

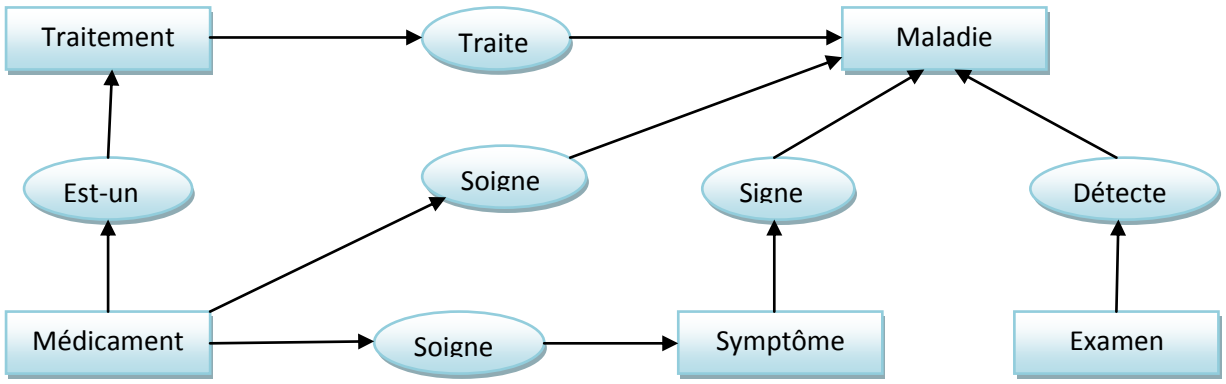


Figure II.5. Description des relations

Afin d'extraire les relations décrites ci-dessous, les auteurs ont d'abord construit des patrons linguistiques, puis, ces patrons ont été appliqué sur le corpus considéré pour identifié de nouvelle relations. Dans ce qui suit, nous allons dans un premier temps expliquer la méthode utilisée pour la construction des patrons linguistiques. Puis, dans un second temps, nous allons décrire la technique utilisée pour l'identification de nouvelle relation.

Le processus élaboré pour extraire à partir d'un corpus les patrons linguistiques caractérisant une relation est le suivant :

1. Appliquer sur le corpus considéré les règles de reconnaissance des entités médicales impliquées dans la relation cible.
2. Extraire du corpus toutes les phrases contenant les deux entités de la relation cible.
3. Sélectionner manuellement les phrases dans lesquelles la relation entre les deux entités correspond effectivement à la relation cible.
4. Réaliser l'analyse linguistique de chaque phrase sélectionnée pour en faire apparaître les différents niveaux d'information. Cette analyse est réalisée par l'analyseur LIMA.
5. Remplacer dans chaque phrase les entités par leur type (traitement, maladie, …).

6. Appliquer l'algorithme d'extraction de patron multi-niveau entre chaque couple de phrase parmi celles sélectionné précédemment.
7. Filtrer les patrons les moins significatifs.

Pour extraire les patrons linguistiques propres à chaque relation sémantique traitée, les auteurs ont utilisé l'algorithme proposé par (Pantel et al., 2004) pour apprendre des patrons multi-niveaux. Cet algorithme est composé de deux parties. La première consiste à calculer la distance d'édition minimale entre deux phrases, ce qui permet de déterminer le nombre minimum d'opérations (insertion, suppression et remplacement) à appliquer pour passer d'une phrase à l'autre. La deuxième étape extrait le patron multi-niveau le plus spécifique permettant de généraliser les deux phrases. Le processus d'extraction de patron multi-niveau est résumé dans la figureII.4. Les patrons linguistiques de chaque relation sont ensuite classés selon leur fréquence d'apparition pour ne retenir que les N premiers patrons. Dans cette approche, les auteurs ont fixé N à 50.

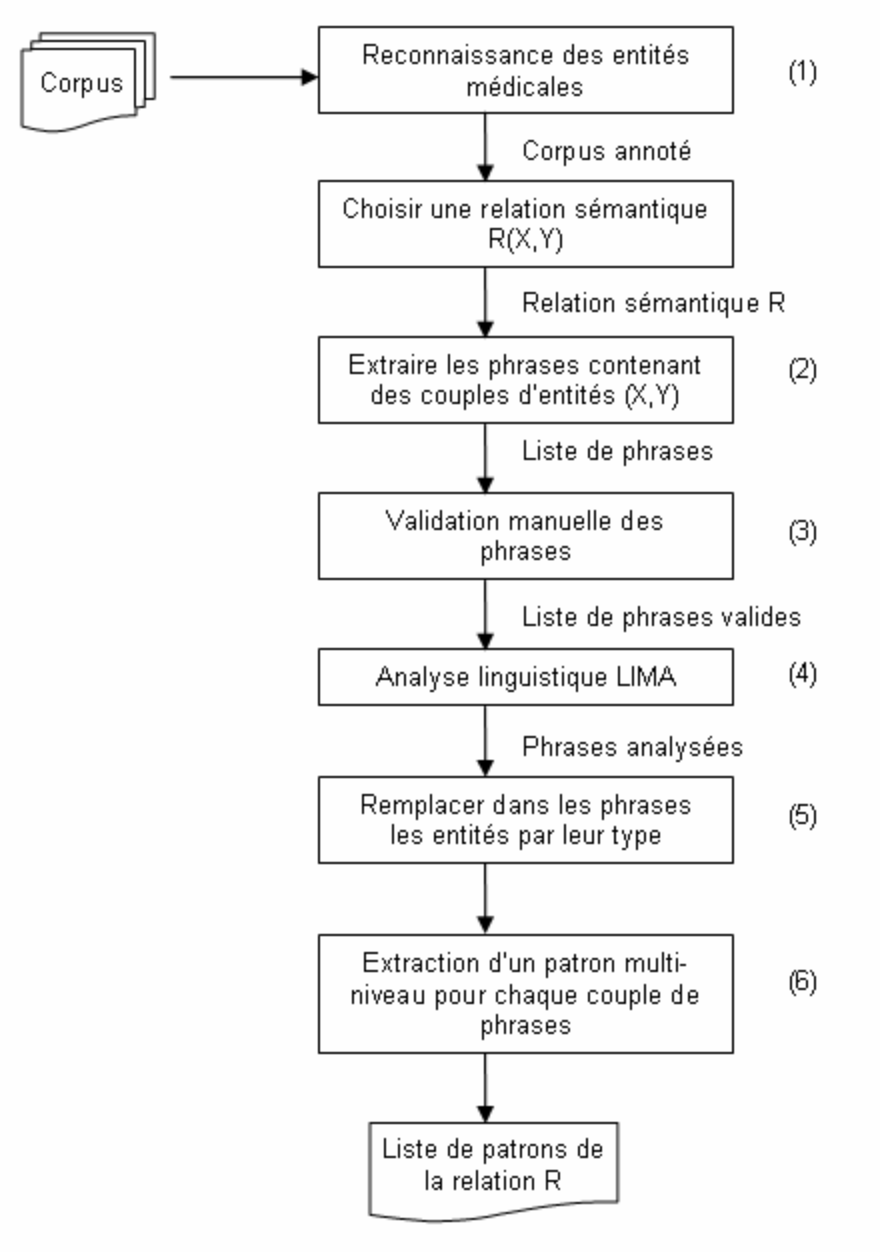


Figure II.6. Description du processus d'extraction de patron multi-niveau

Une fois que ces patrons sont construits, ils seront appliqués pour identifier les relations. Le processus que les auteurs ont suivi pour valider la présence de relations est le suivant :

1. Appliquer sur le corpus considéré les règles de reconnaissance des entités médicales impliquées dans la relation cible.
2. Extraire du corpus toutes les phrases contenant simultanément les entités de la relation cible.

3. Réaliser l'analyse linguistique de chaque phrase sélectionnée en utilisant l'analyseur LIMA.
4. Remplacer dans chaque phrase les entités par leur type.
5. Pour chaque phrase, calculer sa distance d'édition avec tous les patrons multi-niveaux de la relation. Si la distance d'édition est égal à 0 alors validé la relation.

2.3. Les approches hybrides

Une approche hybride est une approche qui combine deux méthodes différentes d'extraction de relation. Une telle méthode peu pallié à certain inconvénients des méthodes citée plus haut, à savoir les méthodes statistique et les méthodes a base de patron linguistique. Reste a trouvé une méthode efficace pour combiné deux approches différentes. Dans ce qui suit, nous allons présenter un système de question-réponse dans le domaine médical qui se base sur une approche hybride pour extraire les relations qui peuvent exister entre des entités médical.

Les travaux de (Asma Ben Abacha et Pierre Zweigenbaum, 2012)

Le but des auteurs était d'extraire des relations sémantiques entre des concepts médicaux (traitement, problème, etc.) à partir de résumé de MEDLINE. Pour cela ils ont proposé une approche hybride d'extraction de relation. Cette approche se fonde sur deux techniques différentes, la première technique utilise un ensemble de patrons linguistique, la deuxième technique se base sur une méthode d'apprentissage automatiques supervisé. Les auteurs ont ciblé trois types de relation (*cure*, *prevent* et *side effect*).

Pour l'approche linguistique, les auteurs ont utilisé un ensemble de patrons linguistiques pour extraire des relations entre les entités médicales. Ils ont ensuite modélisé ces patrons dans une ontologie qui sera liée au réseau sémantique de l'UMLS (l'UMLS précise les types sémantiques source et cible de chaque relation). Pour chaque relation ciblée du réseau sémantique de l'UMLS, un ensemble de patrons linguistique lui était associé.

Pour la construction de patrons de relations sémantiques, les auteurs ont proposé une méthode semi-automatique. Cette méthode est décrite ci-dessus :

1. Collecter automatiquement des phrases contenant au moins une source et une cible possible pour une relation R donnée, cette relation devant être définie dans le réseau

sémantique de l'UMLS comme une relation possible entre les types sémantiques de la source et de la cible.

2. Pour chaque relation ciblée R, construire un corpus de recherche de patrons : extraire du metathesaurus de l'UMLS tous les couples de concepts reliés par la relation en question. A partir de ces informations, construire des requêtes MeSH pour interroger PubMedCentral (une archive libre de plusieurs millions d'articles médicaux).
3. Une fois le corpus d'acquisition construits pour une relation R donnée, récupérer les champs utiles dans chaque article (le titre, les résumé, le corps). Les textes sont en suite segmenté en phrases, puis les phrases contenant un couple de concept (C1, C2) reliés par la relation R (dans le metathesaurus) sont gardées. A partir de ces phrases, des patrons sont manuellement construit sous forme d'expression régulière, prenant en compte la présence d'entités médicales à telle ou telle position de la phrase.

Exemple : pour la relation « traitement », ils ont construit 45 patrons (voir le tableau1.11)

Patron	**Exemple**
*** TX* provide(s) ?* (clinical) ?* benefit(s) ? in* PB***	Adjunctive clonidine provides clinical benefits in patients hospitalized with ascites.
*** PB *should be treated with* TX***	Emergent Hyperpyrexia in Children Should Be Treated With Antibiotics.

Tableau II.1 .Exemples de patrons, R = « traitement ». * : éléments contextuels, TX : Traitant, PB : Problème

Pour la deuxième méthode qui appuis sur une technique d'apprentissage automatique supervisé. Les auteurs utilisent pour la classification un classifieur SVM avec la bibliothèque LIBSVM. Leur objectif était le suivant : étant donné deux entités E1 et E2 dans une phrase, déterminer la relation qui les relie (ou l'absence de relation). Pour cela, les auteurs ont utilisé trois types d'attributs pour décrire les exemples d'entrainement : (1) des attributs lexicaux (les mots de l'entité source E1, les mots de l'entité cible E2, …), (2) des attributs morpho-syntaxiques : ce types d'attributs comporte (la catégorie morpho-syntaxique des mots), (3) des attributs sémantiques (les concepts associés à E1, les concepts associés à E2, …).

L'approche hybride combine les deux méthodes précédentes pour extraire des relations. Les auteurs ont constaté qu'ils ont obtenu de meilleurs résultats avec cette approche.

3. Conclusion

Dans ce chapitre, on a présenté les différentes techniques d'extraction de relations entre concepts dans le domaine biomédicale. Nous en distinguons : les méthodes statistiques qui se déclinent en deux techniques (les techniques qui se basent sur la co-occurrence et les techniques qui se basent sur l'apprentissage automatique), les méthodes linguistiques et les méthodes hybrides. Les approches qui se fondent sur l'apprentissage ne peuvent garantir un haut degré de précision qu'avec la disponibilité d'un grand nombre d'exemples annotés pour une relation donnée. Et l'interprétation des relations extraites nécessite un investissement humain afin de valider les informations identifiées. Les méthodes linguistiques permettent des analyses profondes du contexte d'occurrence de chaque entité médicale et de chaque relation, mais certaines relations sont indétectables avec ce genre de méthodes à cause de la grande variabilité d'expression des relations et en même temps de la structure parfois très complexes de certaines phrases. Les méthodes hybrides apportent bien un plus aux deux méthodes séparées. Elles permettent de tirer partie des avantages des deux méthodes précédentes.

La méthode que nous proposons dans le cadre de ce travail s'inscrit de la même perspective que les approches statistique, et plus exactement les approches qui se basent sur la cooccurrence de termes spécifiques. Notre méthode se déroule en deux étapes. La première étape consiste à identifier dans les documents les concepts médicaux. Dans la phrase "`the genetic material is partitioned between mother and daughter cell.`", le premier objectif est d'extraire les concepts "genetic" et "cell". La deuxième étape consiste a calculé la cooccurrence entre les concepts identifié en appliquant la mesure que nous proposons.

Le chapitre suivant présente plus en détail notre approche de recherche d'information biomédicale qui se base principalement sur une méthode d'extraction de relation de cooccurrence entre concepts biomédicaux.

III. Notre approche d'extraction de relation entre concepts biomédicaux

1. Introduction

L'objectif de notre travail est l'extraction de relations entre concepts biomédicaux. Nous avons opté pour approche statistique qui se base sur un score de cooccurrence.

Nous présentons dans ce présent chapitre, notre approche de recherche d'information biomédicale qui se base principalement sur l'extraction de relation de cooccurrence entre concepts biomédicaux.

Ce chapitre est organisé comme suit : nous présentons dans un premier temps notre approche d'extraction de relation de cooccurrence entre concepts biomédicaux. Puis dans un second temps, nous détaillons la méthode que nous proposons pour valider notre approche d'extraction de relation. Et enfin, nous conclurons ce chapitre par une description d'une illustration de notre méthode.

2. Description de notre approche d'extraction de relations entre concepts

Le but de notre approche est d'extraire des relations de cooccurrence entre concepts biomédicaux. L'idée principale d'une telle approche est que les concepts qui co-occurrent ensemble ont forcément un lien. Ce lien peut se traduire par la présence d'une relation plus spécifique entre ces deux concepts.

Avant d'extraction les relations à partir de la littérature biomédicale, nous procédons d'abord à l'identification des concepts. Pour ce faire, nous avons utilisé l'approche de Duy (Duy, 2012). Cette approche utilise le thésaurus MeSH pour extraire les meilleurs concepts représentatifs des documents. Elle est basée sur la combinaison de deux mesure de similarité, la mesure thématique (liée au thème) et la mesure structurelle (liée à la structure ou la formation des termes, notamment l'ordre des mots constituant les termes 'concepts'). Ces mesures sont calculées entre le texte et les concepts dans la terminologie. Cette méthode est implémentée dans le logiciel Cxtractor.

Pour l'extraction des relations entre concepts, nous proposons un score de co-occurrence qui calcule la probabilité que deux concepts apparaissent ensemble. Le score que nous proposons s'appui principalement sur la fréquence des concepts (la fréquence des termes préférés et la fréquence des termes non préféré) dans les documents et sur le nombre de fois ou ils apparaissent ensemble dans toute la collection.

Notre approche d'extraction de relation est basé sur l'intuition suivante : plus deux concepts apparaissent ensemble, plus ils sont susceptibles d'être liés par une relation plus spécifique. Notre but est donc de favorisé les paires (couples) de concepts qui apparaissent plus souvent ensemble. Et la fonction « exponentielle » qui est une fonction continue est strictement croissante dans l'intervalle [0, +∞], nous semble être un moyen judicieux pour ce faire.

Le score que nous proposons est le suivant :

$$\text{score [k] [l]} = \sum_{i=0}^{n} (\min (f_k, f_l)) * \exp (\text{nb(k, l)})$$

On notera :

f_k : la fréquence du concepts k dans le document i ;

f_l : la fréquence du concepts l dans le document i ;

min (f_k, f_l) : le minimum des fréquence des concepts k et l dans le document i ;

$\sum_{i=0}^{n} (\min (f_k, f_l))$: la somme des minimum des fréquences des concepts k et l dans chaque document de la collection ;

nb(k, l) : le nombre de document ou les concepts k et l apparaissent ensemble dans toute la collection;

exp (nb(k, l) : l'exponentielle de nb.

N : le nombre total des documents de la collection.

Pratiquement, deux concepts C_k et C_l sont liés par une relation de cooccurrence si leur score de cooccurrence est supérieur à un seuil S défini expérimentalement.

L'algorithme complet de notre approche est le suivant :

Algorithme de calcul du score de co-occurrence

Entrée :

Document contenant les concepts représentatifs des documents : doc

synonymes :doc3

Sortie :

Fichier contenant le score de co-occurrence de chaque couple de concepts

Variable :

F :frequence

Min : min des fréquences

Som-min : somme des min

N : nombre d'apparition de chaque couple de concepts

Score : le score de co-occurrence

C1, C2 : concepts

i : entier

```
Début
        Pour i=1 à i= nombre de document de la collection faire
            Pour concepts C1 dans doc faire
                Calcul de la fréquence F ;
                    Pour chaque couple de concepts C1 et C2 dans doc faire
                    Si  F(C1)<= F(C2)
                        Min :=F(C1)
                    Sinon
                        Min :=F(C2)
                Pour chaque couple de concepts C1 et C2 dans doc faire
                        Som-min :=Som-min+Min ;
                        N++;
                        Score:= Som-min*exp(N);
            Finpour
        Finpour
```

retourne (C1, C2, score) ;

Fin.

La figure II.1 résume les principales étapes de notre approche. Nous donnons ensuite les formats des fichiers d'entrés et sortie :

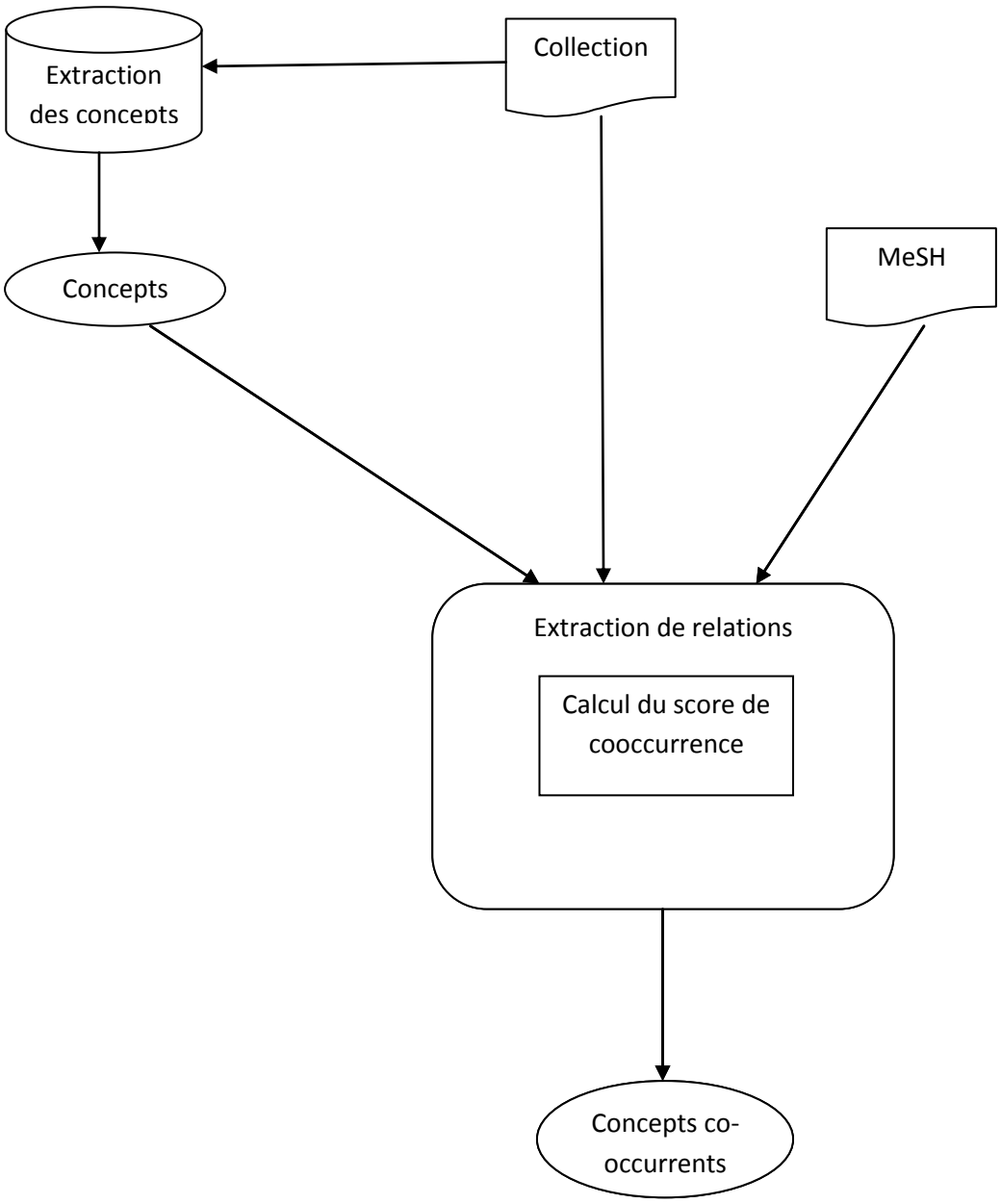


Figure III.1 : Approche d'extraction de relation de cooccurrence entre concepts biomédicaux.

En entrée de notre application d'extraction de relations entre concepts, nous avons :

- Une collection de documents extraite de la collection TREC Genomics 2004,
- Les concepts représentatifs de chaque document,
- Et le fichier MeSH.

La structure des documents de notre corpus est montrée dans la Table III.1.

```
<DOC>
<DOCNO>12346</DOCNO>
<ABSTRACT>
…….
</ABSTRACT>
</DOC>
```

Table III.1 : Structure des documents de notre corpus

Les concepts représentants des documents de la collection sont extrait avec Cxtractor (Duy et al., 2012) qui est un outil d'extraction de concepts biomédicaux. La structure du fichier contenant les concepts représentatifs de chaque document est présentée dans la table III.2.

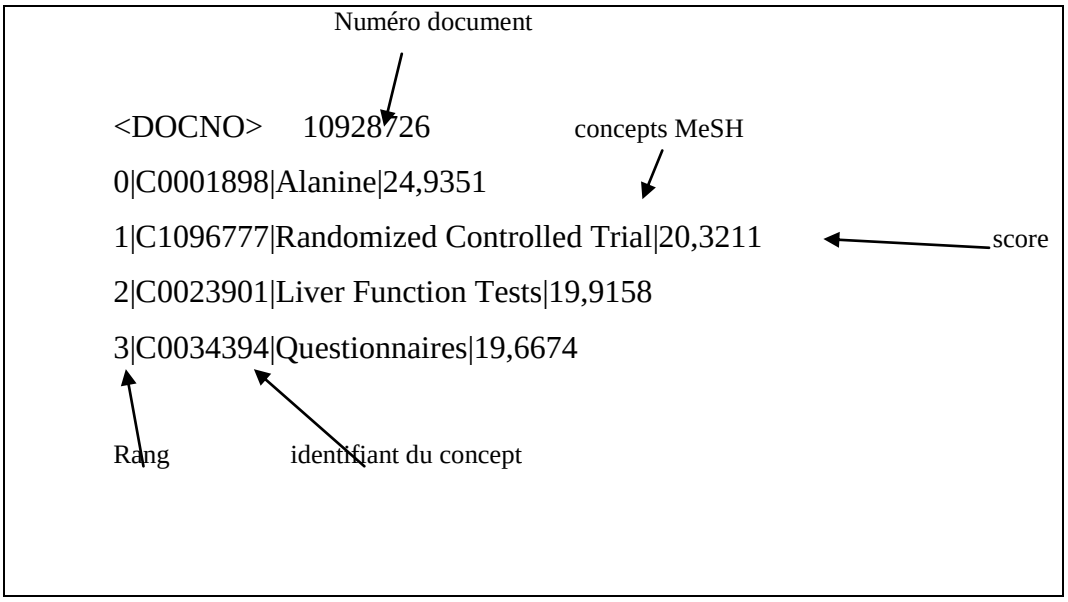


Table III.1. Extrait du fichier contenant les concepts représentatifs de chaque document

Et on a aussi le fichier MeSH qui contient tous les concepts MeSH avec leur synonymes (voir table III.4). Le concept MeSH est considéré comme un document qui contient plusieurs termes biomédicaux. Chaque concept est représenté par un terme principal dit préféré, et un ensemble de concepts appelés termes non préférés.

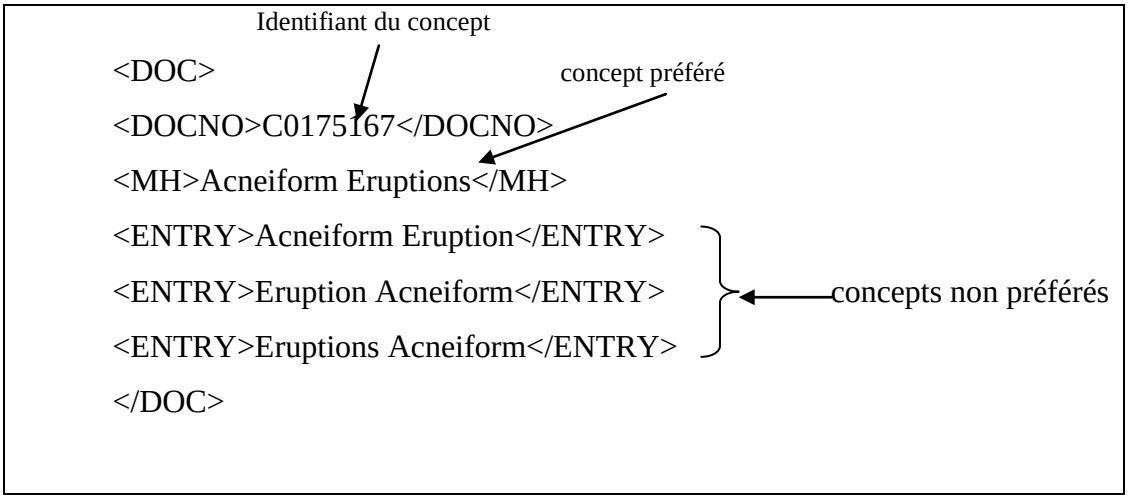


Table III.4. Extrait du fichier MeSH

En sortie de notre application, on aura un fichier contenant les résultats du calcul du score de co-occurrence. Un extrait de ce fichier est présenté dans la table III.5.

Score de cooccurrence

<DOC>10928726
C0001898|C1096777|1|1|2.718
C0001898|C0023901|1|1|2.718
C0001898|C0001948|1|1|2.718
C1096777|C0023901|1|1|2.718
C1096777|C0034394|2|2|14.778
C1096777|C0001948|1|1|2.718
…..
</DOC>

Table III.5. Extrait du fichier de sortie de notre application

3. Approche de validation

Pour valider les relations extraites en utilisant notre score de cooccurrence, nous les intégrons dans un processus de recherche. Pour ce faire nous proposons de tenir compte de ces relations dans le calcul du score d'appariement.

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. La RSV du document D pour la requête Q est classiquement calculée comme la mesure du cosinus entre les deux vecteurs D et Q. Cette mesure est globalement basée sur les termes de la requête qui apparaissent dans le document. Dans notre cas, nous proposons de prendre en compte en plus des termes de la requête qui apparaissent dans le document les termes de la requête qui sont sémantiquement liés à ceux des documents.

La mesure que nous proposons est la suivante :

$$RSV(d, q) = \alpha \left[\frac{\sum_{j=1}^{n} wdj * wqj}{\sqrt{\sum_{j=1}^{n} wdj^2} + \sqrt{\sum_{j=1}^{n} wqj^2}} \right] + \beta \left[\frac{\sum_{j=1}^{n} wdk * wql}{\sqrt{\sum_{j=1}^{n} wdj^2} + \sqrt{\sum_{j=1}^{n} wqj^2}} \right]$$

Où :

wdj, wqj : est le poid du terme tj dans le document respectivement dans la requête,

tl est un terme de la requête sémantiquement lié au terme tK du document.

Les coefficients alpha et Beta sont déterminés expérimentalement. Dans notre cas après plusieurs tests, nous avons fixé alpha à '1' et Beta à '0.06'.

4. Illustration

Nous présentons dans cette section une illustration de notre méthode d'extraction de relations à partir de documents biomédicaux. Nous présentons les résultats obtenus pour le document suivant :

```
<DOC>
<DOCNO>10929710</DOCNO>
<ABSTRACT>
Exit from mitosis must not occur prior to partitioning of
chromosomes between daughter cells. We find that the GTP binding
protein Tem1, a regulator of mitotic exit, is present on the spindle
pole body that migrates into the bud during S phase and mitosis.
Tem1's exchange factor, Lte1, localizes to the bud. Thus, Tem1 and
Lte1 are present in the same cellular compartment (the bud) only
after the nucleus enters the bud during nuclear division. We also
find that the presence of Tem1 and Lte1 in the bud is required for
mitotic exit. Our results suggest that the spatial segregation of
Tem1 and Lte1 ensures that exit from mitosis only occurs after the
genetic material is partitioned between mother and daughter cell.
</ABSTRACT>
</DOC>
```

Nous avons dans un premier temps soumis la collection à Cxtractor pour avoir les concepts représentatifs de chaque document. Le résultat obtenu pour le document « 10929710 » est le suivant :

```
<DOCNO>    10929710
0|C0037406|Social Control, Formal|25,1967
1|C1155789|Cell Nucleus Division|22,1824
2|C0026591|Mothers|20,7453
3|C0026255|Mitosis|18,4189
4|C0018748|Health Services Accessibility|17,9872
5|C0086376|GTP-Binding Proteins|16,7026
6|C0033023|Prejudice|14,8057
7|C0017337|Genes|13,9862
8|C0028574|Nuclear Family|11,0578
9|C0080129|S Phase|7,6093
10|C0008633|Chromosomes|6,5588
11|C0007634|Cells|4,1629
```

Avant de procédé au calcul du score de cooccurrence, nous avons d'abord calculé la fréquence des concepts dans les documents. Les résultats obtenus pour le document '10929710 ' sont les suivants:

Fréquence du concept dans le document

```
<DOC> 10929710
C0037406|Social Control Formal|1
C1155789|Cell Nucleus Division|1
C0026591|Mothers|1
C0026255|Mitosis|3
C0018748|Health Services Accessibility|1
C0086376|GTP-Binding Proteins|1
C0033023|Prejudice|1
C0017337|Genes|1
C0028574|Nuclear Family|2
C0080129|S Phase|1
C0008633|Chromosomes|1
C0007634|Cells|2
</DOC>
```

Le tableau suivant présente les résultats de l'exécution de notre application d'extraction de relations de cooccurrence entre concepts biomédicaux. La première colonne présente les couples de concepts et leurs min des fréquences, la deuxième colonne présente les couples de concepts et leurs scores de cooccurrence.

<DOC>10929710	<DOC>10929710
C0037406|C1155789|1	C0037406|C1155789|3|3|60.257
C0037406|C0026591|1	C0037406|C0026591|2|2|14.778
C0037406|C0026255|1	C0037406|C0026255|41|21|5407144511.3812*10
C0037406|C0018748|1	C0037406|C0018748|90|42|1565347447.3684510*101^{1}
C0037406|C0086376|1	C0037406|C0086376|17|8|50676.286
C0037406|C0033023|1	C0037406|C0033023|2|2|14.778
C0037406|C0017337|1	C0037406|C0017337|1115|308|6456174997,005631*10^{129}
C0037406|C0028574|1	C0037406|C0028574|6|5|890.479
C0037406|C0080129|1	C0037406|C0080129|28|19|4997504426.969
C0037406|C0008633|1	C0037406|C0008633|46|24|1218499617,9728 *10^{4}
C0037406|C0007634|1	C0037406|C0007634|1491|390|35344804654,38035*10^{166}
C1155789|C0026591|1	C1155789|C0026591|2|2|14.778
C1155789|C0026255|1	C1155789|C0026255|5|5|742.066
C1155789|C0018748|1	C1155789|C0018748|2|2|14.778
C1155789|C0086376|1	C1155789|C0086376|1|1|2.718
C1155789|C0033023|1	C1155789|C0033023|1|1|2.718
C1155789|C0017337|1	C1155789|C0017337|1|1|2.718
C1155789|C0028574|1	C1155789|C0028574|3|3|60.257
C1155789|C0080129|1	C1155789|C0080129|3|3|60.257
C1155789|C0008633|1	C1155789|C0008633|3|3|60.257
C1155789|C0007634|1	C1155789|C0007634|6|6|2420.573
C0026591|C0026255|1	C0026591|C0026255|2|2|14.778
C0026591|C0018748|1	C0026591|C0018748|4|4|218.393
C0026591|C0086376|1	C0026591|C0086376|2|2|14.778
C0026591|C0033023|1	C0026591|C0033023|1|1|2.718
C0026591|C0017337|1	C0026591|C0017337|11|8|32790.538
C0026591|C0028574|1	C0026591|C0028574|4|4|218.393
C0026591|C0080129|1	C0026591|C0080129|1|1|2.718
C0026591|C0008633|1	C0026591|C0008633|5|3|100.428
C0026591|C0007634|1	C0026591|C0007634|9|9|72927.755
C0026255|C0018748|1	C0026255|C0018748|17|13|7521027.664
C0026255|C0086376|1	C0026255|C0086376|3|3|60.257
C0026255|C0033023|1	C0026255|C0033023|6|5|890.479
C0026255|C0017337|1	C0026255|C0017337|56|31|1626735581,2538558*10^{7}
C0026255|C0028574|2	C0026255|C0028574|6|5|890.479
C0026255|C0080129|1	C0026255|C0080129|12|10|264317.59
C0026255|C0008633|1	C0026255|C0008633|29|21|3824565630,0013 *10
C0026255|C0007634|2	C0026255|C0007634|111|48|7788576862,42837 * 10^{13}
C0018748|C0086376|1	C0018748|C0086376|4|4|218.393
C0018748|C0033023|1	C0018748|C0033023|10|5|1484.132
C0018748|C0017337|1	C0018748|C0017337|185|77|5103249890,867865* 10^{26}
C0018748|C0028574|1	C0018748|C0028574|4|3|80.342
C0018748|C0080129|1	C0018748|C0080129|5|4|272.991
C0018748|C0008633|1	C0018748|C0008633|39|20|18921442620.982
C0018748|C0007634|1	C0018748|C0007634|336|121|1191164077,407752* 10^{46}
C0086376|C0033023|1	C0086376|C0033023|1|1|2.718
C0086376|C0017337|1	C0086376|C0017337|12|9|97237.007
C0086376|C0028574|1	C0086376|C0028574|1|1|2.718
C0086376|C0080129|1	C0086376|C0080129|1|1|2.718
C0086376|C0008633|1	C0086376|C0008633|1|1|2.718
C0086376|C0007634|1	C0086376|C0007634|27|17|652183724.347

C0033023|C0017337|1	C0033023|C0017337|13|8|38752.454
C0033023|C0028574|1	C0033023|C0028574|1|1|2.718
C0033023|C0080129|1	C0033023|C0080129|1|1|2.718
C0033023|C0008633|1	C0033023|C0008633|20|10|440529.316
C0033023|C0007634|1	C0033023|C0007634|9|6|3630.859
C0017337|C0028574|1	C0017337|C0028574|20|13|8848267.84
C0017337|C0080129|1	C0017337|C0080129|32|21|42202103503.463
C0017337|C0008633|1	C0017337|C0008633|471|137|1483764335,236554* 10^{53}
C0017337|C0007634|1	C0017337|C0007634|6288|671|1622220518,0583859* 10^{281}
C0028574|C0080129|1	C0028574|C0080129|1|1|2.718
C0028574|C0008633|1	C0028574|C0008633|4|4|218.393
C0028574|C0007634|2	C0028574|C0007634|21|15|68649364.822
C0080129|C0008633|1	C0080129|C0008633|14|8|41733.412
C0080129|C0007634|1	C0080129|C0007634|90|58|1390985042,0310936 *10^{18}
C0008633|C0007634|1	C0008633|C0007634|257|71|1757281506,0490252*10^{24}
</DOC>	</DOC>

Tableau III.2. Résultats de l'exécution de notre approche

Nous avons obtenu des scores de co-occurrence entre certain concepts très élevé tel que le score entre le concept « C0017337 » et le concept « C0007634 ». Ce qui peut être traduit par la présence d'une éventuelle relation entre ces deux concepts..

5. Conclusion

Nous avons présenté dans ce chapitre notre approche de recherche d'information biomédicale. Nous avons décrit dans un premier temps notre approche d'extraction de relation de cooccurrence qui se base sur trois étapes : le calcul de la fréquence des concepts de chaque document, le calcul du minimum des fréquences entre chaque deux concepts et le calcul du score de co-occurrence. Puis dans un second temps, nous avons présenté la méthode que nous proposons pour valider notre approche d'extraction de relations. Et enfin, nous avons effectué une illustration de notre méthode.

Le chapitre suivant, sera entièrement consacré à l'évaluation de notre approche. Nous présentons d'abord le cadre d'évaluation choisi. En suite, nous présentons les résultats de nos expérimentations.

IV. Implémentation et évaluation

1. Introduction

Nous avons présenté dans le chapitre précédant, nos contributions pour un modèle de recherche d'information biomédicale. Dans ce chapitre, nous commençons par la présentation de l'environnement technologique. Puis, nous présentons brièvement le corpus utilisé pour l'expérimentation. Et enfin, nous présentons les résultats de l'évaluation de notre approche.

2. Environnement technologique

Pour mener à bien notre travail, nous avons choisi comme environnement de développement Eclipse sous le système d'exploitation Windows 7. Nous avons également utilisé Cxtractor pour l'extraction de concepts et Terrier pour l'évaluation de notre approche.

- **Eclipse :** Eclipse est un environnement de développement (IDE) très puissant, extensible et intégré dont le but est une plate-forme modulaire pour permettre de réaliser des développements informatiques. Il a était développer par I.B.M. Dans notre application nous utilisant Eclipse GALILEO, Avec Eclipse, on travaille toujours au sein d'un projet, un projet est un ensemble de fichiers sources java. L'IDE Eclipse stocke l'information associée à un projet dans un dossier projet (projet folder), qui en plus de nos fichiers sources et les fichiers compilés, inclus toutes les ressources nécessaires à notre projet de développement, ainsi que deux fichiers de configurations (*.classpath* et *.projet*) utilisés par l'IDE. Tous nos projets sont regroupés dans un répertoire, ***le workspace*** Eclipse, qu'on doit choisir lors du lancement d'Eclipse.

- **Cxtractor :** Cxtractor est un outil d'extraction de concepts biomédicaux. Il a été développé par (Duy, Dinh 2012) en java.

- **Terrier :** Terrier est un moteur de recherche robuste et efficace, développé par le département informatique de l'université Glasgow de Scotland. Terrier permet l'indexation classique recherche des documents pertinents pour répondre aux requêtes formulés par l'utilisateur, évaluation des résultats de la recherche. Terrier a été utilisé pour l'évaluation de notre approche.

3. Evaluation expérimentale

L'objectif de ces expérimentations est de mesurer les performances et la viabilité de notre approche d'extraction de relation de co-occurrence entre concepts biomédicaux. L'évaluation complète de notre approche consiste dans un premier temps à tester la mesure de RSV que nous proposons qui se base sur notre approche d'extraction de relation de co-occurrence, puis dans un second temps, comparé les résultats obtenus avec les résultats d'une RSV classique à base de cosinus.

3.1. Cadre d'évaluation

Dans ce qui suit, nous présentons la collection de document utilisée dans nos expérimentations, les requêtes associées à la collection ainsi que le protocole d'évaluation.

3.1.1. Description de la collection de document

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,

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 MEDLINE4, 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 IV.1. Statistiques de la collection TREC Genomics 2004

Le nombre de documents est d'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.

Exemple de document :

```
<DOC>
<DOCNO>10930667</DOCNO>
<ABSTRACT>
Mice immunised with oxidised mannan-MUC1 fusion protein (M-FP)
develop MHC restricted CD8(+) cytotoxic T cells. We now demonstrate
that in MUC1/HLA-A2 transgenic mice, IL-12 gives enhanced CTL, CTLp
and tumor protection. CTLp in MUC1 transgenic mice with M-FP were
1/55,000, and with IL-12, this increased to 1/19,000, with improved
tumor protection. Thus, IL-12 is important for effective CTL
responses to MUC1 in transgenic mice.
</ABSTRACT>
</DOC>
```

Pour notre évaluation nous avons considéré une sous collection de 5001 documents.

3.1.2. Description de l'ensemble des requêtes

Nous avons utilisé l'ensemble des 50 requêtes de la collection TREC Génomics. 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.

Exemple de requête :

```
<TOPIC>
<ID>1</ID>
<TITLE>Ferroportin-1 in humans</TITLE>
<NEED>Find articles about Ferroportin-1, an iron transporter, in
humans.</NEED>
<CONTEXT>Ferroportin1 (also known as SLC40A1; Ferroportin 1; FPN1;
HFE4; IREG1; Iron regulated gene 1; Iron-regulated transporter 1;
MTP1; SLC11A3; and Solute carrier family 11 (proton-coupled divalent
metal ion transporters), member 3) may play a role in iron
transport.</CONTEXT>
</TOPIC>
```

Des jugements de pertinence sont associés aux requêtes selon le format suivant :

Exemple de jugement de pertinence :

```
1     0     10077651   2
1     0     10084280   2
1     0     10084283   1
1     0     10088633   2
1     0     10226041   2
1     0     10228348   2
1     0     10318901   2
1     0     10343352   2
```

Tableau IV.2 ; Exemple de jugement de pertinence

3.2. Protocole d'évaluation

Notre approche d'extraction de relations est évaluée à travers la « qualité des relations produits ». Pour évaluée la qualité de ces relations, nous les intégrons dans un processus de recherche d'information, en particulier à travers notre nouveau score d'appariement. L'évaluation se fait alors sure la base des résultats de la recherche basée sur ce score proposé.

L'évaluation est effectuée selon le protocole TREC. Plus précisément, chaque requête est soumise au système de RI avec les paramètres fixés. Le système renvoie les 1000 premiers documents pour chaque requête. Les valeurs de précision P@1, P@2, P@3, P@4, P@5, P@10, P15, P20, P30, P100 et P1000, ainsi que R-Precision (précision réelle ou exacte) et MAP (précision moyenne) sont calculées.

- **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.
- **R-Precision** (précision réelle ou exacte) correspond à la précision exacte calculée sur l'ensemble des documents pertinents retournés.
- **précision moyenne** (Mean Average Precision, dénotée *MAP*) correspond à la précision moyenne calculée sur 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.

3.3. Résultats expérimentaux

Le tableau IV.2 présente les résultats obtenus pour l'ensemble des requêtes. Les résultats montrent que notre mesure de RSV qui se base sur notre approche d'extraction de relation de cooccurrence entre concepts présenté dans le chapitre précédant offre une précision moyenne plus importante que la mesure de RSV classique (la mesure du cosinus).

information	cosinus	RSV proposé
Number of queries	44	43
Retrieved	35303	42605
Relevant	1048	1026
Relevant retrieved	690	695
Average Precision	0.1058	0.1098
R Precision	0.1272	0.1299
Precision at 1 :	0.1591	0.2326
Precision at 2 :	0.1932	0.1860
Precision at 3 :	0.1742	0.1860
Precision at 4 :	0.1648	0.1860
Precision at 5 :	0.1545	0.1814
Precision at 10 :	0.1705	0.1744
Precision at 15 :	0.1545	0.1581
Precision at 20 :	0.1364	0.1430
Precision at 30:	0.1182	0.1271
Precision at 50 :	0.0995	0.1060
Precision at 100 :	0.0745	0.0742
Precision at 200 :	0.0491	0.0486
Precision at 500 :	0.0257	0.0266
Precision at 1000 :	0.0157	0.0162
Precision at 0%:	0.5106	0.4786
Precision at 10%:	0.3219	0.3290
Precision at 20%:	0.2685	0.2610
Precision at 30%:	0.2016	0.2103

Precision at 40%:	0.1758	0.1951
Precision at 50%:	0.1487	0.1551
Precision at 60%:	0.1014	0.1122
Precision at 70%:	0.0732	0.0663
Precision at 80%:	0.0276	0.0261
Precision at 90%:	0.0138	0.0099
Precision at 100%:	0.0021	0.0029

Tableau IV.3. Résultat de l'évaluation de notre approche avec la plateforme de RI Terrier.

Remarque : nous avons pris en considération uniquement les concepts biomédicaux ce qui explique les valeurs de précision obtenus.

Le tableau suivant est exprimé sous forme de deux graphes dans ce qui suit :

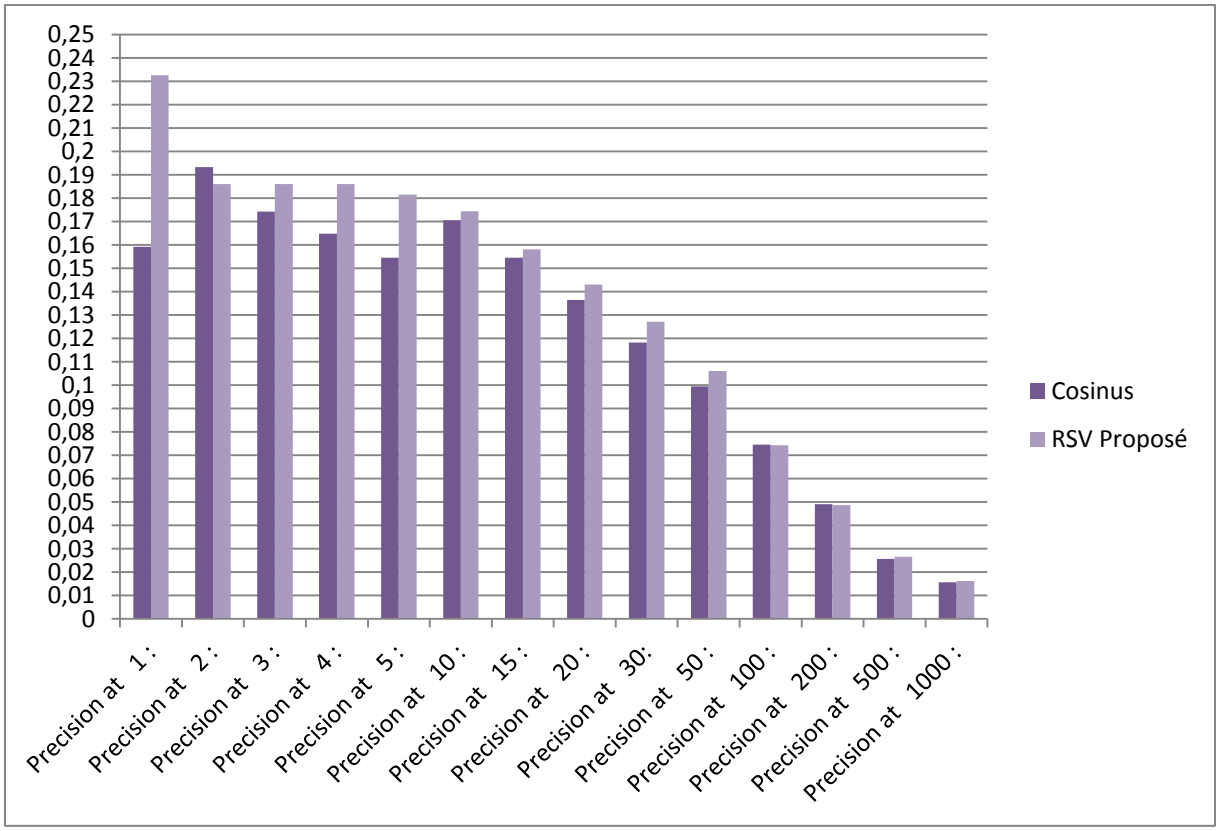


Figure IV.1. Précision @X

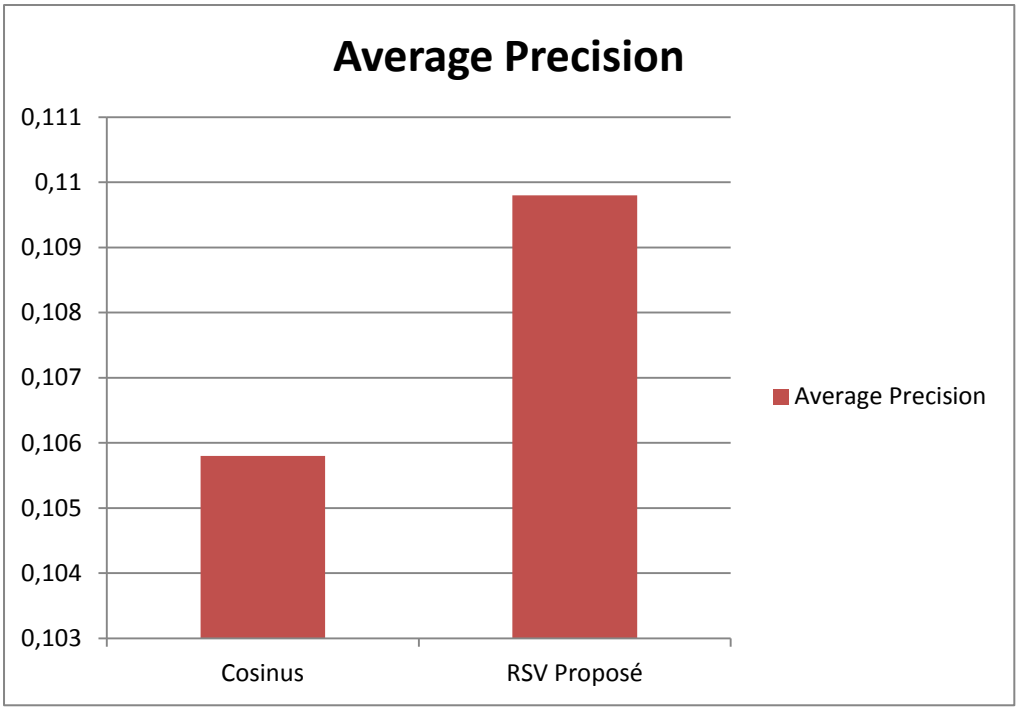


Figure IV.2. MAP

Nous constatons que notre mesure de RSV donne de meilleurs résultats que la mesure du cosinus.

4. Conclusion

Dans ce chapitre nous avons présenté notre expérimentation. L'expérimentation a été réalisée sur un corpus de 5001 documents. Selon les données statistiques il en résulte une amélioration de la précision.

V. Conclusion générale & perspectives

Le travail présenté dans ce mémoire s'inscrit dans le cadre de l'indexation et de la recherche d'information biomédicale. Nous nous somme intéressés plus particulièrement à l'extraction de relation entre concepts biomédicaux.

Notre objectif était de proposer une approche d'extraction de relations entre concepts biomédicaux. Nous avons opté pour une approche statistique qui se base sur la cooccurrence des concepts pour extraire des relations entre ces concepts. . Le score de cooccurrence que nous proposons se base principalement sur la fréquence des concepts et sur le nombre de fois ou ils apparaissent ensemble dans toute la collection. Nous avons aussi proposé une méthode pour valider notre approche d'extraction de relations. Cette méthode de validation s'inscrit dans le cadre du processus de recherche d'information. L'idée est de prendre en compte les relations extraites lors du calcul du score de pertinence doc-req. Le score de pertinence que nous proposons prend en compte en plus des termes de la requête qui apparaissent dans le document, les termes de la requête qui sont sémantiquement liés aux termes du document.

La validation des relations extraites est ainsi faite à travers l'évaluation des résultats de la recherche. Les résultats obtenus sont très encourageants relativement à une recherche classique à base de cosinus.

En perspective il serait intéressant de réaliser des tests expérimentaux complémentaires sur différentes valeurs de seuil de cooccurrence afin d'améliorer encore les résultats obtenus.

VI. Références bibliographiques

[A.BEN ABACHA et al., 2011] Asma Ben Abacha, Pierre Zweigenbaum**.** Une approche hybride pour la détection automatique des relations sémantiques entre entités médicales. LIMSI, CNRS, F-91403 Orsay, France.

[A.BEN ABACHA et al., 2012] Asma Ben Abacha, Pierre Zweigenbaum**.** Recherche de réponses précises à des questions médicales : le système de questions-réponses MEANS. Thèse de Doctorat en Informatique de l'Université PARIS-SUD LIMSI-CNRS. Juin 2012.

[A.Minard et al., 2011] A.Minard, A.Ligozat, B.Grau. Extraction de relations dans des comptes rendus hospitaliers. LIMSI-CNRS, Université Paris-Sud, ENSIIE 2011.

[A.Minard et al., 2011] A.Minard, A.Ligozat, B.Grau .Apport de la syntaxe pour l'extraction de relations en domaine médical. TALN 2011, Montpellier, 27 juin –1er juillet 2011.

[A.Minard et al., 2012] A.Minard, A.Ligozat, B.Grau .Simplification de phrases pour l'extraction de relations. LIMSI-CNRS, Université Paris-Sud, ENSIIE 2012.

[A.Minard et al., 2012] A.Minard, A.Ligozat, B.Grau. Extraction de relations en domaine de spécialité. Thèse de Doctorat en Informatique de l'Ecole Paris-Sud (edips). Laboratoire d'Informatique, de Mécanique et de Sciences de l'Ingénieur (limsi). Décembre 2012.

[Baziz, 05] Baziz M. Indexation Conceptuelle Guidée Par Ontologie Pour La Recherche d'Information. Thèse de Doctorat en Informatique de l'Université Paul Sabatier de Toulouse (Sciences). Décembre 2005.

[B.J. STAPLEY et al., 2000] B.J. Stapley, G. Benoit. BIOBIBLIOMETRICS: information retrieval and visualization from co-occurrence of gene names in Medline Abstracts. Pacific Symposium on Biocomputing 5:526-537 (2000).

[D.DINH et al., 2012] Accès à l'information biomédicale : vers une approche d'indexation et de recherche d'information conceptuelle basée sur la fusion de ressources termino-ontologiques. Thèse de Doctorat en Informatique de l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier).

[F.AMIROUCHE et al., 2008] Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets. Thèse de Doctorat en Informatique de l'Université Toulouse III - Paul Sabatier.

[F.Babinet, 2012] François Babinet. La recherche bibliographique sur Internet (Medline et PubMed). Septembre 2012.

[F.HARRATHI et al., 2009] Extraction de concepts et de relations entre concepts à partir des documents multilingues : Approche statistique et ontologique. L'Institut Nationale des Sciences Appliquées de Lyon.

[H.Abdoune et al., 2011] H.Abdoune, L.Soualmia, M.Joubert. Analyse de cooccurrences de concepts biomédicaux dans MEDLINE. LERTIM, LIM&Bio.

[Khoo et al., 2000] Christopher S.G. Khoo, Syin Chan and Yun Niu. Extracting Causal Knowledge from a Medical Database Using Graphical Patterns. Centre for Advanced Information Systems, School of Computer Engineering Nanyang Technological University Singapore 639798.

[L.Chew-Hung et al., 2004]. Automatic Identification of Treatment Relation For Medical Ontology Learning : An Exploratory Study. In I.C. Mcllwaine (ED.), Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference. Wurzburg, Germany: Ergon Verlag

[M.EMBAREK et al., 2006] Mehdi EMBAREK, Olivier Ferret. Learning patterns for building resources about semantic relations in the medical domain. CEA, LIST, Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue.

[M.EMBAREK et al., 2008] Mehdi EMBAREK, Olivier Ferret. Un système de question-réponse dans le domaine médical, Le système Esculape. Université de Paris-Est 2008.

[M.EMBAREK et al., 2010] Mehdi EMBAREK, Olivier Ferret. Adapter un système de question-réponse en domaine ouvert au domaine médical. TALN 2010, Montréal, 19–23 juillet 2010.

[MS,1999] C.D. Manning & H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[M. STEPHENS et al., 2001] M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje. DETECTING GENE RELATIONS FROM MEDLINE ABSTRACTS. Pacific Symposium on Biocomputing 6:483-496 (2001).

[M.Manser, 2012] Mounira Manser. État de l'art sur l'acquisition de relations sémantiques entre termes : contextualisation des relations de synonymie. Université Paris 13. Juin 2012.

[Natalie Clairoux, 2012] PUBMED: une interface d'interrogation en accès libre de MEDLINE. Université de Montréal. Automne 2012.

[O.Frunza et al., 2010] Oana Frunza, Diana Inkpen. Identifying and classifying semantic relations between medical concepts in clinical data, i2b2 challenge (2010).

[O.Frunza et al., 2010] Oana Frunza, Diana Inkpen. Extraction of Disease-Treatment Semantic Relation from Biomedical Sentences. School of Technology and Engineering University of Ottawa, ON, Canada, K1N6N5. ACL 2010.

[O.Uzuner et al., 2010] O.Uzuner, J.Mailoa, R.Ryan, T.Sibanda. Semantic Relations for Problem-Oriented Medical Records. University at Albany, State University of New York.

[Pierre Claveirole, 2004] CDRMG / UNAFORMEC. MeSH.

[R. Barbara et herast, 2004] Barbara Rosario, Marti A. Hearst. Classifying Semantic Relations in Bioscience Texts. SIMS UC Berkeley. 2004.

[Roberts et al., 2008] A.Roberts, R.Gaizauskas, M.Hepple, Y.Guo. Mining clinical relationships from patient narratives. *from* Natural Language Processing in Biomedicine (BioNLP) ACL Workshop 2008 Columbus, OH, USA. 19 June 2008.

[S.RÉHEL, 2005] SIMON RÉHEL. Catégorisation automatique de textes et cooccurrence de mots provenant de documents non étiquetés. FACULTÉ DES SCIENCES ET DE GÉNIE UNIVERSITÉ LAVAL QUÉBEC. Janvier 2005

Annexe 1 : Cxtractor

Cette annexe est principalement consacrée à la présentation de Cxtractor (Version 1.0.3) une plateforme de RI de haute performance et évolutive qui a comme tâche principale l'extraction de concepts biomédicaux à partir de textes.

1. Introduction

Cxtractor a été développé par (Duy Dinh, 2012) de l'université de Toulouse 3 Paul Sabatier, il est intégré dans la plateforme BioSIR (**Bio**medical **S**emantic **I**nformation **R**etrieval). C'est un logiciel open source entièrement écrit en java. Dans ce logiciel sont implémentées plusieurs méthodes d'extraction de concepts à partir des documents biomédicaux, à savoir les méthodes d'extraction basées sur les modèle de RI (ex : TF_IDF, BM25, etc). Il utilise plusieurs terminologies pour l'extraction de concepts biomédicaux (ex : MeSH, SNOMED, GO ou UMLS).

2. Installation de Cxtractor

Cxtractor est téléchargé à partir du site http://www.softpedia.com/progDownload/Cxtractor-Download-230918.html . Télécharger cxtractor-1.0.3.zip et cxtractor-ressources.zip. Puis extraire ces deux fichiers compressés. Et enfin, Copier et coller les répertoires 'lib' et 'nlpdata' dans le répertoire 'cxtractor-1.0.3/cxtractor'.

3. Structure de Cxtractor

Terrier contient un ensemble de répertoires, ils sont structurés comme suit :

- bin\ : contient l'ensemble des scripts nécessaires pour démarer Cxtractor.
- config\ : contient les fichiers de configuration de Cxtractor (le fichier settings properties sample contient la plupart des propriétés de configuration de Cxtractor)
- doc\ : contient la documentation relative à Cxtractor.

- examples\ : contient des exemples de document donnés en entrée pour l'extraction de concepts.
- lib\ : contient les différentes librairies externes utilisées par Cxtrator.
- nlpdata\ : contient les différentes ressources qu'utilise Cxtractor pour extraire les concepts (Mesh, UMLS, SNOMED, etc)
- output\ : contient les résultats lors de l'extraction.
- screenstrats\ : contient quelques captures d'écran.
- scripts\ : contient un fichier en script shell.
- src\ : contient le code source java des programmes de Cxtractor.
- tests\ : contient les collections qu'on donne pour l'extraction de concepts.

4. Lancement de Cxtractor sur une ligne de commande

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]

Exemple d'utilisation:

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.

Exemple de 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

Avec:

rank: représente le rang du concept

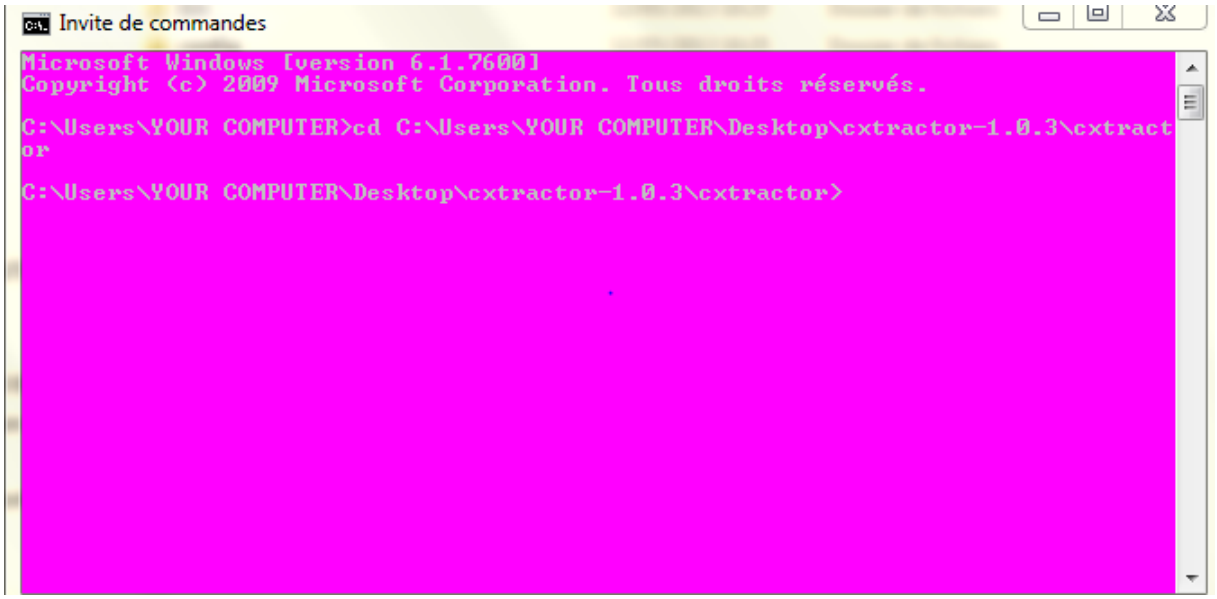
CUI : l'identifaint unique du concept

concept name (preferred/non-preferred terms): la forme préféré du concepts

score : le poid du concepts dans le document.

5. Exemple d'exécution

Avant de lancer l'exécution placez le dossier contenant les documents à partir des quels on veut extraire les concepts dans le répertoire '….\cxtractor-1.0.3\cxtractor\tests\trec ', puis faire un 'cd' vers '...../cxtractor-1.0.3/cxtractor'.



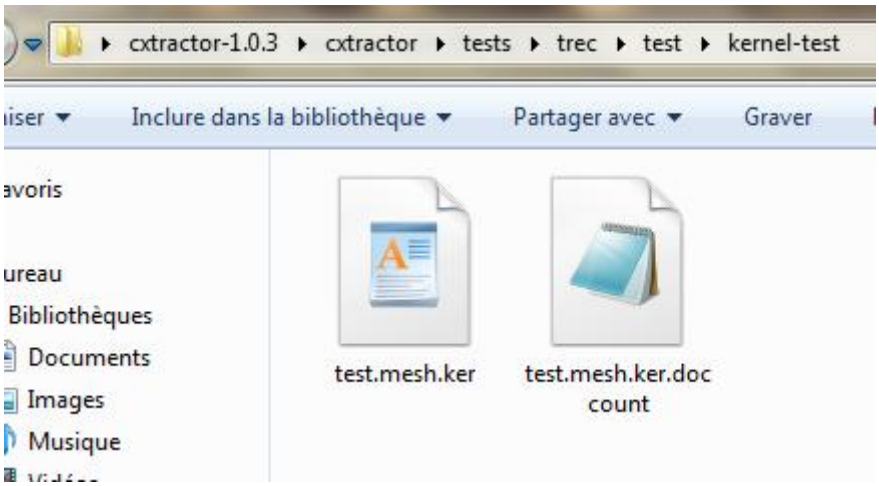
En suite lancez la commande : 'java -jar cxtractor-1.0.3.jar -r -c -d tests/trec/test' test représente le dossier contenant les documents à partir des quels on veut extraire les concepts. Cette commande va utiliser par défaut le thésaurus MeSH pour l'extraction de concepts. Si on veut utiliser une autre terminologie on peut spécifier dans la commande la terminologie à utilisé exemple : java -jar cxtractor-1.0.3.jar -r -c -d tests/trec/ -o output/ -t snomed.

```
C:\Users\YOUR COMPUTER\Desktop\cxtractor-1.0.3\cxtractor>java -jar cxtractor-1.0
.3.jar -r -c -d tests/trec/test
INFO - Loading document lengths for document structure into memory
INFO - Structure meta reading lookup file into memory
INFO - Structure meta reading reverse map for key docno directly from disk
INFO - Structure meta loading data file into memory
INFO - Time to intialise index : 1.263
***** INPUT: tests/trec/test
***** OUTPUT: C:\Users\YOUR COMPUTER\Desktop\cxtractor-1.0.3\cxtractor\tests\tre
c\test/kernel-test/test.mesh.ker

[Tue Aug 20 12:51:09 CEST 2013] Navigating directory 'tests/trec/test' ...
[Tue Aug 20 12:51:09 CEST 2013] Navigating directory 'C:\Users\YOUR COMPUTER\Des
ktop\cxtractor-1.0.3\cxtractor\tests\trec\test\kernel-test' ...
[Tue Aug 20 12:51:09 CEST 2013] Total of 0 processed documents in directory 'C:\
Users\YOUR COMPUTER\Desktop\cxtractor-1.0.3\cxtractor\tests\trec\test\kernel-tes
t'

[Tue Aug 20 12:51:09 CEST 2013] Processing file C:\Users\YOUR COMPUTER\Desktop\c
xtractor-1.0.3\cxtractor\tests\trec\test\test.txt ...
[Tue Aug 20 12:51:10 CEST 2013] Number of processed documents : 1
                                Trying to process a bunch of 100 documents. Pleas
e wait ...
[Tue Aug 20 12:51:10 CEST 2013] Total of 1 processed documents in directory 'tes
ts/trec/test'
```

Les résultats de l'extraction seront sauvegardé dans le répertoire '\cxtractor1.0.3\cxtractor\tests\trec\test' :



Le fichier 'test.mesh.ker' contient les résultats de l'extraction :

```
<DOCNO>    11096424
0|C0026809|Mice|27,5774
1|C0162638|Apoptosis|25,9135
2|C0282505|Limb Buds|19,9813
3|C0314657|Genetic Predisposition to Disease|19,0689
4|C0851346|Radiation|16,4739
5|C0000768|Congenital Abnormalities|14,5548
6|C0017337|Genes|13,9862
7|C0242290|Organogenesis|12,8244
```

```
8|C0031084|Periodicity|10,8325
9|C0015385|Extremities|8,3114
```

Et le fichier ‘test.mesh.ker.doc count’ contient le nombre de document traité, dans notre cas c’est ‘1’.

Annexe 2 : Terrier

Cette annexe est principalement consacrée à la présentation de Terrier (Version 3.5) une plateforme de RI de haute performance et évolutive qui permet le développement rapide et à grande échelle de nouvelles applications de recherche d'information.

1. Introduction

Terrier, TeRabyte RetrlEveR a été développé par le département informatique de l'université de Glasgow. C'est un logiciel open source entièrement écrit en java. Il est utilisé avec succès pour la recherche Ad hoc, la recherche sur le web et la recherche inter-langage dans les environnements centralisés et distribués.

Terrier offre plusieurs modèles de pondération de documents et d'expansion de requêtes basées sur le Framework DFR (Divergence From Randomness). Comme tous les moteurs de recherche Terrier possède les principales facettes suivantes :

- Indexation : permet l'extraction des termes des différents documents du corpus.
- Recherche : permet de générer des résultats aux requêtes formulées par les utilisateurs.
- Evaluation des résultats de la recherche.

2. Installation de Terrier

Pour pouvoir utiliser Terrier il est nécessaire d'installer une JRE. La JRE ou la JDK peuvent être téléchargé à partir du site http://www.oracle.com/technetwork/java/javase/downloads/index.html.

Terrier est téléchargé à partir du site http://www.terrier.org//downloads.html. Après avoir téléchargé Terrier, créer un nouveau répertoire et dézippé Terrier dans ce dernier.

3. Structure de Terrier

Terrier contient un ensemble de répertoires et ils sont structurés comme suit :

- bin\ : contient les scripts nécessaires pour démarrer Terrier.
- doc\ : contient la documentation relative à Terrier.

- ect\ : Fichiers de configuration de terrier, le fichier terrier.properties.sample contient la plupart des propriétés de configuration de terrier.
- lib\ : contient les classes compilées de Terrier et les différentes librairies externes utilisées par Terrier.
- share\ : contient la liste des mots vides (stop word list).
- src\ : contient le code source de Terrier.
- var\index : contient les structures de données.
- var\results : contient les résultats de la recherche.
- licenses\ : contient les informations sur la licence des différents composants inclus dans Terrier.

4. Lancement de Terrier

Dans cette section, nous décrivons l'utilisation de Terrier pour l'indexation, la recherche et l'évaluation :

- Indexation :
  - Supprimer le contenu du dossier index dans var dans terrier (s'il est plein)
  - Initialisation de terrier pour une nouvelle indexation (collection ou corpus Trec)
    - ✓ ..\terrier-3.5\bin**trec-setup** <path de la collection ou le corpus à indexer >
  - Maintenant nous pouvons indexer la collection TREC :

  >> .\bin\trec_terrier –i
- Recherche :
  - Récupérer les documents pertinents pour la requête
  - Ajouter la propriété trec.topics dans **terrier.properties:**
    - ✓ trec.topics=<Path vers le fichier txt contenant les requêtes>
  - Dans l'invité de commandes exécuter:
    - ✓ ..\terrier-3.5\bin**trec_terrier –r**
  - Le fichier résultat (.res) est dans ..\terrier-3.5\var\result\ x.res
  - Par défaut, le nombre max de documents retourné par terrier est 1000 documents.
  - Configuration dans **terrier.properties ,** le nombre max de documents à retourner:

- ✓ trec.output.format.length=200

- Evaluation :

  Selon le protocole TREC.

  - Pour chaque requête, les 1000 premiers documents restitués par le système sont examinés
  - Les précisions P@x à différents points x (x = 5, 10, 20, 50, 100, 200, 500, 1000) ainsi que la précision moyenne Average Precision sont calculées.
  - Spécifier dans **terrier.properties** le chemin vers le Rel_Ass, qui contient les documents pertinents pour chaque requête.
    - ✓ trec.qrels=<path vers fichiers Rel_Ass.qrels>
  - Dans l'invité de commandes exécuter:
    - ✓ ..\terrier-3.5\bin**trec_terrier -e**
  - Le fichier évaluation (.eval) est dans :
    - ✓ ..\terrier-3.5\var\result\ x.eval 59

Voici un extrait du fichier d'évaluation (.eval).

```
Number of queries  = 1
Retrieved          = 720
Relevant           = 40
Relevant retrieved = 36

Average Precision: 0.4508
R Precision      : 0.5000

Precision at   1 : 1.0000
Precision at   2 : 1.0000
Precision at   3 : 0.6667
Precision at   4 : 0.5000
Precision at   5 : 0.6000
Precision at   10 : 0.5000
Precision at   15 : 0.5333
Precision at   20 : 0.5500
Precision at   30 : 0.5333
Precision at   50 : 0.4600
Precision at   100 : 0.3300
Precision at   200 : 0.1750
Precision at   500 : 0.0720
Precision at   1000 : 0.0360
```