Projection de requêtes pour une recherche d'information intelligente sur le Web

Lina F.Soualmia[1,2] et Stéfan J.Darmoni[1,2]

[1]Laboratoire PSI-CNRS FRE 2645-INSA de Rouen-76131 Mont Saint-Aignan
{lina.soualmia,stefan.darmoni}@chu-rouen.fr
[2]CISMeF & L@STICS-CHU de Rouen- 76031 Rouen
http://www.cchu-rouen.fr/cismef/

Résumé : La recherche d'information sur le Web demeure problématique malgré l'existence de nombreux moteurs de recherche et de sites catalogues. Le Web doit faire face aux problèmes d'exhaustivité et de précision en recherche d'information. Le projet CISMeF (Catalogue et Index de Sites Médicaux Francophones) a été développé afin de faciliter l'accès à l'information de santé disponible sur l'Internet. La problématique d'aujourd'hui se veut aussi être une *recherche d'information intelligente* dans l'infrastructure du Web Sémantique, une extension du web actuel qui permettrait de rendre interprétable le contenu des ressources par les hommes mais aussi par les machines, grâce à des ontologies et des méta-données. La recherche d'information dans CISMeF est fondée sur une terminologie semblable à une ontologie et un ensemble de méta-données qui nous permettent de placer le projet à cheval entre le Web actuel qui est informel, et le Web Sémantique de demain. Nous proposons dans cet article d'utiliser trois types de ressources (base de connaissances morphologiques, base de règles d'association et ensemble de règles d'inférence) afin de donner les moyens à la recherche d'information de devenir intelligente. Nous détaillons les traitements nécessaires pour la construction automatique des ces ressources qui se basent sur le traitement du langage naturel, le data mining et le raisonnement sur les descriptions de concepts.

Mots-clés : Recherche d'information, Traitement du langage naturel, Extraction de connaissances, Ontologies, Web Sémantique.

1 Introduction

La quantité d'information disponible sur le Web est importante et elle ne cesse de croître. Les catalogues et les moteurs de recherche en ligne (Yahoo, Google, Lycos…etc) permettent d'effectuer des requêtes par mot clé et de les affiner à l'aide d'opérateurs booléens. Avec des requêtes plus fines, le moteur peut renvoyer plus de documents et ainsi augmenter le rappel. Cependant, la tâche la plus lourde revient à l'utilisateur qui doit fouiller dans cette masse d'information pour sélectionner les documents qui lui seront les plus utiles. Les résultats ne sont pas tous pertinents et l'information retrouvée n'est pas complète. La recherche plein texte n'est pas toujours

efficace : les fautes de frappe, les variantes lexicales et les synonymes sont considérés comme étant des termes différents.

Aujourd'hui, la problématique qui se pose est celle d'une *recherche d'information intelligente* sur le Web. Les moteurs de recherche actuels ne peuvent pas traiter *intelligemment* les pages HTML, langage le plus répandu sur le Web. Le Web Sémantique (Berners-Lee et al. 2001) est un espace d'échange qui reste à construire. Un de ses intérêts est d'une part d'apporter suffisamment de renseignements sur les ressources, en ajoutant des annotations sous la forme de *méta-données* et d'autre part, de décrire leur contenu de manière à la fois formelle et signifiante à l'aide d'une *ontologie* pour être interprétables aussi bien par les humains que par les machines. Cet espace doit être formalisé, le Web actuel étant informel. En effet, il est composé principalement de pages HTML écrites à la main ou générées automatiquement pour un traitement humain. Les ontologies et les méta-données sont donc deux éléments principaux pour la construction de l'infrastructure du Web Sémantique (Laublet et al. 2002). Une ontologie est une modélisation partagée d'un domaine pour améliorer la communication et éliminer les ambiguïtés entre personnes, entre personnes et applications ou entre applications. Elle est composée d'une hiérarchie de concepts, de relations entre concepts et d'un ensemble de règles ou de contraintes. Les méta-données font référence à une information descriptive des ressources du Web. Leur première utilité est la recherche d'information.

Plusieurs projets plus ou moins récents se basent sur l'utilisation d'une ontologie pour décrire formellement des ensembles de documents ou de ressources. SHOE (Luke & Heflin, 2000) est l'un des précurseurs du Web Sémantique. Des ontologies et un langage basé sur HTML sont utilisés pour annoter sémantiquement des pages Web. OntoSeek (Guarino et al. 1999) est un moteur de recherche de pages Web qui utilise l'ontologie terminologique WordNet. Les documents et les requêtes sont représentés à l'aide de graphes conceptuels. Dans CoMMA (Gandon et al. 2002) des annotations sous forme de fichiers RDF (Lassila & Swick, 1999) sont associées à des documents d'entreprise en fonction des concepts de l'ontologie O'CoMMA.

Le contexte d'application de notre problématique de recherche d'information est le projet CISMeF[1] (Darmoni et al. 2001) qui se fonde sur un ensemble de méta-données et une terminologie du domaine médical. Afin d'améliorer la recherche d'information au sein du catalogue nous étudions différentes techniques pour la projection de requêtes sur la terminologie. Les requêtes considérées sont celles qui sont saisies via une interface par les utilisateurs. Pour cela nous avons enrichi la terminologie d'une base de connaissances morphologiques en utilisant le traitement du langage naturel, d'une base de règles d'associations grâce aux techniques de data mining et enfin nous allons formaliser la terminologie, à l'aide d'un langage de représentation des connaissances, et modéliser une base de règles d'inférences pour permettre un raisonnement sur le contenu des documents. Cet article est organisé de la manière suivante : tout d'abord nous décrivons en section (2) les méta-données et la structure de la terminologie utilisée dans CISMeF. En section (3) nous expliquons le processus de recherche d'information dans le catalogue tel qu'en place aujourd'hui

[1] Catalogue et Index de Sites Médicaux Francophones ; http://www.chu-rouen.fr/cismef/

ainsi que les problèmes rencontrés. Avant de conclure, nous détaillons en section (4) les méthodes utilisées pour l'enrichissement de la terminologie et l'extraction de connaissances pour l'expansion de requêtes.

2 Vers un web sémantique médical

Le projet CISMeF a été développé en 1995 pour assister les professionnels de santé, les étudiants et le grand public dans leur quête d'information en santé sur le Web. CISMeF et Doc'CISMeF, le moteur de recherche associé, prennent en compte la diversité des utilisateurs et leur permettent de trouver des documents de qualité qui répondent à un besoin précis. De nombreuses ressources (n=11,600) sont sélectionnées en fonction de critères stricts par une équipe de documentalistes et sont répertoriées selon une méthodologie de mise à jour du catalogue. Une ressource peut être un site Web, une page Web, un document, un rapport : tout support qui contient des informations relatives à la santé. La description de ces ressources se fait à l'aide de *notices* en fonction d'un ensemble de méta-données et d'une terminologie structurée du domaine médical. Cette structure nous permet de placer le projet à cheval entre le Web informel d'aujourd'hui et le Web Sémantique de demain.

2.1 Les Méta-données

Les méta-données sont par définition des données concernant les données et dans le contexte du Web elles font référence à une information descriptive de ses ressources. LeWeb a été initialement constitué pour un traitement humain et pour cette raison qu'il est difficile de tout y automatiser. Le concept de méta-données existait avant l'avènement d'Internet mais son intérêt a grandi avec le nombre de publications électroniques et de bibliothèques virtuelles. La solution proposée par le World Wide Web Consortium (W3C) est d'utiliser les méta-données pour décrire les données disponibles sur le Web. Dans le contexte du Web Sémantique elles constituent un module fondamental et permettent notamment de faciliter la recherche d'information. Elles garantissent l'interopérabilité en assurant le partage et l'échange d'information rendant son contenu lisible et compréhensible par les machines.

Nous utilisons dans CISMeF plusieurs ensembles de méta-données parmi lesquels celui du Dublin Core (Baker, 2000) composé de 15 éléments. Les ressources indexées dans CISMeF sont décrites à l'aide de 11 éléments du Dublin Core : *auteur, date, description, format, identifiant, langage, éditeur, type de ressource, droits, sujet* et *titre*. Le Dublin Core ne permet pas de rendre compte de la qualité ou de la localisation d'une ressource. Pour pallier ce problème, 8 éléments spécifiques à CISMeF ont été définis : *institution, ville, province ou département, pays, public ciblé, type d'accès, coût* et *parrainage* de la ressource. Le type d'utilisateur est pris en compte. Pour les ressources destinées aux professionnels de santé (les lignes directrices et consensus de bonne pratique clinique) deux champs supplémentaires sont définis : *indication du niveau de preuve* et la *méthode* utilisée pour le déterminer. Pour les ressources pédagogiques ce sont onze éléments de la catégorie

« Educational » du standard IEEE 1484 qui sont rajoutés. Le format de ces méta-données est passé du langage HTML en 1995, à XML en 2000 pour permettre l'interopérabilité avec d'autres plateformes (e-learning du projet UMVF[2]). Depuis décembre 2002, le format utilisé est RDF un langage basique du Web Sémantique, et ce dans le cadre du projet européen MedCIRCLE (Mayer et al. 2003) dont le but est de qualifier la qualité des ressources d'information en santé et de guider les utilisateurs vers une information de confiance. Le vocabulaire des méta-données HIDDEL[3] est contenu dans une ontologie (représentée à l'aide du langage RDFS) et les ressources décrites en RDF en fonction des concepts de cette ontologie.

2.2 La Terminologie CISMeF

Les ressources sont indexées en fonction de la terminologie CISMeF. Celle-ci a été construite à partir des concepts du thésaurus MeSH[4] (développé depuis 1960) et de sa traduction en français fournie par l'INSERM[5]. Le MeSH dans sa version 2003 est composé d'environ 22,000 *mots clés* (comme *abdomen*, *hépatite*) et 84 *qualificatifs* (comme *diagnostic, complications, thérapeutique…*) regroupés sous la forme d'arborescences. Les mots clés correspondent à des concepts médicaux et sont organisés sous la forme de hiérarchie à 9 niveaux allant du terme le plus général en haut de la hiérarchie aux termes les plus spécifiques en bas de la hiérarchie. Par exemple le mot clé *aberration chromosomique* est plus général que le mot clé *trisomie*. Les qualificatifs, organisés également en hiérarchie, permettent de préciser le sens des mots clés en limitant leur étendue à certains aspects. Par exemple l'association du mot clé *lombalgie* et du qualificatif *diagnostic* (notée *lombalgie/diagnostic*) permet de restreindre la *lombalgie* au seul aspect *diagnostic*. Bien qu'il existe des ontologies médicales générales, comme GALEN (Rodrigues et al. 1998), ou spécifiques à un domaine comme MENELAS (Bouaud et al. 1995), c'est le MeSH qui a été choisi car il correspond aux attentes des documentalistes et il est connu des professionnels de santé.

Les mots clés ont été regroupés dans CISMeF en fonction de spécialités médicales (n=66) intitulés *métatermes* (Cancérologie). Ce sont des super-concepts qui permettent une vision plus globale concernant une spécialité en offrant un niveau supplémentaire d'abstraction. Les métatermes permettent de connaître l'ensemble des termes MeSH qui sont répartis dans plusieurs arborescences mais qui concernent une même spécialité. Une hiérarchie de *types de ressources* (n=127) a été modélisée et elle permet de décrire la nature de la ressource (*cours, information patient*). Les métatermes et les types de ressources permettent d'exprimer des requêtes complexes dans CISMeF comme des *'recommandations en cardiologie'* ou encore des *'cours en virologie'* ce qui n'est pas possible avec la structure actuelle du MeSH.

---

[2] Université Médicale Virtuelle Francophone ; http://www.umvf.prd.fr

[3] Health Information Disclosure Description Evaluation Language ; http://www.medcircle.org

[4] Medical Subject Headings. Le MeSH est produit par la US-National Library of Medicine pour la base documentaire Medline.

[5] Institut National de la Santé et de la Recherche Médicale
http://dicdoc.kb.inserm.fr:2010/basimesh/mesh.html

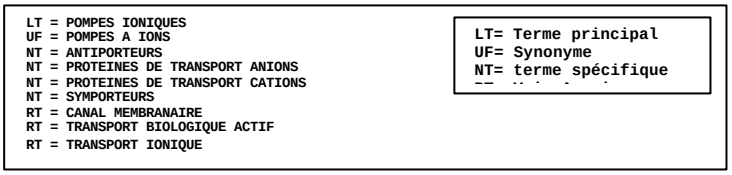

Fig. 1 – Exemple de termes et de relations dans les fichiers texte du MeSH

A partir du MeSH fournit sous la forme de fichiers texte (Fig.1) seules les relations du type *'est-un'* et *'partie-de'* sont utilisées pour définir des liens père-fils dans la hiérarchie des mots clés CISMeF (*n*=7,435 soit ~34% du MeSH). Ces liens hiérarchiques sont exploités pour la recherche d'information et la navigation dans le catalogue. Par exemple le mot clé *Oreille* initialement défini comme étant *partie-de* du mot clé *Tête*, est défini dans la terminologie CISMeF par *Oreille* est fils de *Tête*. Les fichiers MeSH sont traités automatiquement pour renseigner la terminologie CISMeF afin qu'elle soit exploitable au niveau du site.

3 Recherche d'information

3.1 Processus de recherche d'information

La structure de la terminologie est exploitée pour l'indexation des ressources, la visualisation et la navigation dans les hiérarchies des termes du domaine, la recherche de ressources par le moteur Doc'CISMeF. Différents modes de recherche d'information sont possibles. La recherche *simple* permet à l'utilisateur de saisir une requête en texte libre en français ou en anglais. La recherche *avancée* engage des recherches plus pointues à l'aide d'un formulaire contenant des listes déroulantes et permet de combiner plusieurs champs (mots clés, titre, année…etc.) avec des opérateurs booléens (ET, OU, SAUF). La recherche *logique* s'effectue à l'aide d'un langage de requêtes associé, des opérateurs booléens et des caractères spéciaux.

La recherche simple telle qu'en place aujourd'hui se base sur les relations de hiérarchie entre termes. Si le terme (un mot ou une expression) saisi par l'utilisateur est un terme existant dans la terminologie, le résultat de la requête est l'union de toutes les ressources indexées par ce terme et par ses fils directs ou indirects de toutes les hiérarchies dans lesquelles il peut se trouver. Par exemple une requête sur le terme *tumeur* va renvoyer comme réponse l'ensemble des ressources rattachées à *tumeur* mais également celles rattachées à *tumeur colon*, *tumeur rectum*…etc. De même qu'une requête sur *tête* va renvoyer les ressources rattachées à *tête* mais également à *oreille, nez*…etc. Si le terme saisi par l'utilisateur n'est pas un terme réservé, une recherche sur tous les autres champs de méta-données est effectuée, voire en plein texte sur tous les documents indexés en risquant de retomber dans les mêmes problèmes de bruit qu'un moteur de recherche plein texte. Ce type de recherche simple nécessite donc une bonne connaissance des termes de CISMeF, ce qui n'est pas évident pour un utilisateur novice.

3.2 Problèmes de la recherche d'information

La (ou les) requête(s) saisie(s) par l'utilisateur correspond(ent) rarement à la formulation exacte effectivement utilisée pour l'indexation. Nous avons extrait les requêtes des utilisateurs, à partir des logs du serveur http du moteur Doc'CISMeF, et déterminé le type de requête employé ainsi que le nombre de réponses obtenu, entre le 15/08/2002 et le 06/02/2003 (Table 1). 1,522,776 requêtes ont été extraites. 892,591 requêtes (58.62%) ont été soumises via l'interface de recherche simple et 365,688 (40.97% des requêtes simples) ne renvoient aucune réponse. Une analyse plus fine des requêtes simples (Table 2) nous a permis de déduire que 12.01% des réponses sont nulles, non pas du fait qu'elles correspondent à des requêtes erronées (ce sont bien des termes réservés), mais du fait qu'aucune ressource ne leur est rattachée.

Table 1. Analyse des requêtes des utilisateurs du 15/08/2002 au 6/02/2003.

	Requêtes		**Requêtes Nulles**	
Type de Requête	**Nombre**	**Pourcentage**	**Nombre**	**Pourcentage**
Simple	892 591	58.62 %	365 688	40.97 %
Autre	630 175	41.38 %	144 790	22.97 %
Total	1 522 776		510 478	

Table 2. Répartition des requêtes simples à 0 réponse

	Nombre	**Pourcentage**
Expression reconnue	43 922	12.01 %
Expression non reconnue	321 766	87.99%
Total	365 688	

Afin d'améliorer ce type de recherche d'information qui est le plus utilisé dans le catalogue, nous proposons d'appliquer et d'évaluer la contribution de trois méthodes. Nous détaillons les pré-traitements de données nécessaires.

4 Améliorer la recherche d'information

4.1 Traitement du langage naturel

Les connaissances morphologiques sont utiles pour la recherche d'information et leur apport a été démontré dans plusieurs travaux (Gaussier et al. 2000) (Savoy, 2002) pour retrouver des expressions différentes qui dénotent des notions identiques ou proches. A partir d'un mot, on peut obtenir trois formes de variations. La *flexion* produit les pluriels, féminins lorsqu'il s'agit d'un nom et les conjugaisons lorsque c'est un verbe. La *dérivation* produit la forme adjectivale d'un nom. Enfin la *composition* combine plusieurs noms. Les connaissances morphologiques (flexions, dérivations, compositions) d'un mot donné constituent sa *famille morphologique*. Il

existe pour le domaine médical un lexique de mots dérivés en langue anglaise qui est le Specialist Lexicon de l'UMLS[6] (Lindberg et al. 1993) mais aujourd'hui aucune ressource de ce type n'est disponible pour le français médical. Nous souhaitons utiliser ce type de connaissances pour améliorer la recherche d'information en réalisant une expansion de requêtes. Le but est de construire cette ressource morphologique pour la terminologie CISMeF.

Une étude préliminaire (Zweigenbaum et al. 2001) a été réalisée sur un ensemble de requêtes lancées sur Doc'CISMeF. Les résultats ont montré que l'utilisation de connaissances morphologiques amélioraient sensiblement les résultats des requêtes en diminuant le nombre de réponses nulles. La base de connaissances morphologique a été construite automatiquement dans des travaux antérieurs et l'algorithme proposé consiste à corriger la requête de l'utilisateur (dans le cas de non réponse seulement) en éliminant les « mots vides » (*comment, alors, du* …etc.) et en remplaçant chaque terme de la requête par une disjonction de tous les termes de sa famille morphologique. Par exemple le terme *Cœur* a comme flexion *Cœurs*, comme dérivation *Cardiaque*, et comme composition *Cardiovasculaire*. Si l'utilisateur saisit la requête *interaction entre médicaments et alimentation* l'algorithme permet de reconnaître le mot clé *interaction aliment médicament*. Cette première ressource n'a pas été construite en fonction des termes de CISMeF et elle contient 6,312 couples (mot | mot dérivé). Après comparaison avec le sous-ensemble des termes de CISMeF utilisés pour l'indexation des ressources, nous avons obtenu 646 termes dérivés qui couvrent 608 mots clés, 30 qualificatifs et 8 types de ressources.

Nous avons analysé au préalable la structure de la terminologie de CISMeF (Table 3) relative à la composition des termes réservés. Il nous a semblé plus logique pour cette étude de ne considérer en premier lieu que les termes utilisés pour l'indexation des ressources car même si la requête d'un utilisateur correspond à un terme réservé, le résultat sera nul s'il n'existe pas de ressource qui lui soit rattachée.

Table 3. Structure des termes utilisés pour l'indexation

Nombre de mots	Mots Clés	Qualificatifs	Types de Ressources	Termes
1	1 437	55	28	1 520
2	1 706	10	42	1 758
3	612	11	39	662
4	148	3	12	163
5	40	--	4	44
6	8	--	2	10
7	2	--	--	2
TOTAL	**3953**	**79**	**127**	**4 159**

Dans un second temps nous avons complété nos données grâce à la ressource terminologique Lexique (New et al. 2001). Elle comporte tout le lexique du français contemporain déduit à partir d'un corpus de textes, écrits entre les années 1950 et 2000, en se basant sur des calculs de fréquences d'apparition des mots contenus dans

[6] Unified Medical Language System

des pages du Web. Grâce à cette ressource terminologique, nous avons obtenu 34,710 variantes morphologiques et couvert exactement 1,300 termes de CISMeF (1,222 mots clés, 53 qualificatifs et 25 types de ressources). Toutes les variantes, y compris les verbes et subjonctifs, sont présentes et la liste est relativement complète.

L'analyse des termes composés de 2 ou plusieurs mots nous a permis de déduire que 1,935 termes étaient 'semi-couverts' (1,899 mots clés; 8 qualificatifs; 28 types de ressources). On considère qu'un terme est semi-couvert si au moins un des mots qui le composent est couvert. Par exemple *accident circulation* est un mot clé composé d'un terme dérivé du mot clé *accidents* qui a comme famille : *{accident, accidents, accidenté, accidentés, accidentées, accidentel, accidentels, accidentelle, accidentelles, accidentellement, accidenter}*. Celle-ci existant déjà dans la base grâce à l'étape de reconnaissance des termes, on considère que le terme *accident circulation* est semi-couvert. Il nous reste donc à compléter cette base de connaissances morphologiques sachant que l'appariement à plusieurs mots est plus exigeant (Zweigenbaum et al. 2001).

Table 4. Couverture du vocabulaire

	Mots Clés	**Qualificatifs**	**Types de Ressources**	**Termes**
Nb termes couverts	1 405	54	25	**1 484**
Couverture 1 mot	97.77%	98.18%	89.28%	**97.63%**
Semi-couvertue	83.58%	78.48%	41.73%	**77.59%**
Couverture exacte	35.54%	68.35%	19.68%	**35.68%**

Les résultats que nous avons obtenus dans (Grabar et al. 2003) montrent qu'une normalisation des requêtes et de la terminologie augmente sa couverture : en effet si le mot clé est *accidenté*, la requête *Accidentè* sera nulle. Nous avons donc désaccentué et mis en minuscule tous les termes dérivés obtenus. La gestion des accents et minuscules est également effectuée sur les requêtes au niveau du prototype de recherche développé pour la réalisation des différents tests. A présent, l'algorithme permet de déduire à partir de la requête simple *douleurs dorsales* la requête logique `douleur.mc ET dorsalgie.mc`, avec mc indiquant que le terme considéré a été reconnu par l'algorithme de recherche comme étant un mot clé.

Nous avons également extrait de Lexique tous les termes qui peuvent correspondre à des mots vides. Les mots vides sélectionnés sont tous les adjectifs possessifs (*mon*), les conjonctions (*mais*), les déterminants (*du*), les interjections (*diantre*), les prépositions (*durant*), les pronoms personnels (*il*), les pronoms possessifs (*leur*) et les pronoms relationnels (*auquel*). Nous avons déterminé ainsi 873 mots vides supplémentaires aux 473 initiaux, nous donnant un total de 1,346 mots vides. Ce nombre est élevé vu que des termes comme *boum, bye, bravo* ou encore *sniff* sont considérés comme vides. La requête *douleur du bas du dos* est ainsi transformée en `douleur.mc ET dos.mc`.

En plus de connaissances morphologiques, des connaissances sémantiques sont nécessaires. Par exemple le terme médical correspondant à *fausse couche* est *avortement spontané*. Nous étudions actuellement les logs des utilisateurs et

collaborons avec des associations de patients et la Ligue Nationale contre le Cancer pour compléter la liste des synonymes CISMeF.

4.2 Data mining

Nous souhaitons découvrir de «nouvelles» connaissances à partir de la base de données CISMeF (en particulier à partir des notices et des termes) qui seront exploitées dans le processus de recherche d'information. Nous appliquons une technique de Data Mining appelée *Règles d'Association* dans le but d'extraire des associations intéressantes, non triviales, précédemment inconnues à partir de la base. Les règles d'association ont été initialement utilisées en analyse des données puis en fouille de données dans les bases de données relationnelles de grande taille (Agrawal & Srikant, 1994). Ces règles d'association utilisées dans un contexte d'expansion de requêtes permettent d'améliorer les performances de recherche d'information (Haddad et al. 2000).

Nous nous intéressons à la découverte de règles d'association booléennes. Une règle d'association booléenne RA est de la forme :

$$RA : a_1 \wedge a_2 \wedge \ldots \wedge a_i \Rightarrow a_{i+1} \wedge \ldots \wedge a_n \quad (1)$$

Elle s'interprète intuitivement de la manière suivante : si un objet possède les attributs $\{a_1, \ldots, a_i\}$ alors il a tendance à posséder également les attributs $\{a_{i+1}, \ldots, a_n\}$. Le *support* d'une règle représente son utilité. Cette mesure correspond à la proportion d'objets qui contiennent à la fois l'antécédent et le conséquent de la règle. La *confiance* représente sa précision. Cette mesure correspond à la proportion d'objets contenant le conséquent de la règle parmi ceux contenant l'antécédent.

Le processus d'extraction de connaissances est composé de plusieurs phases : la préparation des données et du contexte (sélection des objets et des attributs), l'extraction des ensembles fréquents d'attributs (*itemsets* fréquents par rapport à un seuil de support minimum), la génération des règles d'association les plus informatives à l'aide d'un algorithme de Data Mining (par rapport à un seuil de confiance minimum) et enfin l'interprétation des résultats (ou déduction de nouvelles connaissances).

Notre contexte d'extraction est le triplet $C=(O,A,R)$ avec O l'ensemble des objets, A l'ensemble des items, R une relation binaire entre O et A. Les objets sont les notices utilisées pour décrire les ressources indexées. Elles ont un identifiant unique. La relation R correspond à la relation d'indexation entre un objet et un item. Nous considérons pour l'instant deux cas différents pour les items :

- A={Mots Clés} ; A est l'ensemble des Mots clés.
- A={(Mot Clé, Qualificatif)} ; A est l'ensemble des couples (Mot Clé, Qualificatif).

Un itemset est fréquent dans son contexte C si son support est supérieur à un seuil minimal défini au préalable. Le problème de l'extraction des itemsets fréquents est de complexité exponentielle dans la taille n de l'ensemble d'items, le nombre d'itemsets fréquents potentiels étant 2^n. Dans le premier nous avons n=7,435. Les itemsets forment un treillis (Davey & Priestley, 1994). Plusieurs algorithmes de découverte

d'itemsets fréquents ont été proposés. Le plus connu est l'algorithme Apriori (Agrawal & Srikant, 1994). Nous utilisons l'algorithme A-Close (Pasquier, 2000) dans lequel l'extraction se fait par le calcul des itemsets *fermés* fréquents avec l'opérateur de fermeture de la connexion de Galois d'une relation binaire finie (Ganter & Wille, 2000). L'espace des itemsets à étudier est ainsi réduit. L'algorithme calcule aussi les générateurs des itemsets fermés fréquents. Les générateurs d'un itemset fermé I_f sont les itemsets de taille maximale dont la fermeture est égale à I_f. Les bases pour les règles d'association sont déterminées à partir des itemsets fermés fréquents et de leurs générateurs. L'union de ces bases est un ensemble de générateurs non redondants de toutes les règles d'association non redondantes, d'antécédents minimaux et de conséquences maximales qui ne représentent aucune perte d'information. Ce sont les règles les plus utiles et les plus pertinentes.

Pour tester l'algorithme nous avons fixé le support à 5 documents et la confiance à 100%. La première étape de l'algorithme (itemsets de taille 2) nous a permis de trouver les règles suivantes :

- *hépatite C* Þ *sida* ; support = 14 (*A*={Mots Clés}).
- *sida/prévention et contrôle* Þ *condom* ; support = 6 (*A*={(MotClé,Qualificatif)}).

La seconde étape de cette étude est de déterminer toutes les autres règles d'association qui seront exploitées dans le processus de recherche d'information.

4.3 Raisonnement sur le contenu des documents

La terminologie CISMeF a la même utilité et structure qu'une ontologie terminologique (Sowa, 2000) :

- Le vocabulaire est bien connu des documentalistes et des professionnels de la santé et il correspond à celui du domaine médical.
- Chaque concept (Fig.2) a un terme préférentiel (Descripteur) pour l'exprimer en language naturel, un ensemble de propriétés, la définition en langage naturel permet quelquefois le différencier des concepts le subsumant et de ceux qu'il subsume, un ensemble de synonymes et un ensemble de règles et de contraintes (Fig.3)
- Les concepts sont organisés selon une relation de subsomption allant du concept le plus général en haut de la hiérarchie au plus spécifique en bas de la hiérarchie.

D'après l'exemple de définition de la Fig.2, le terme associé au concept ayant l'identifiant unique D006521 est *Hépatite Chronique*. Le code cat.MeSH indique à quel niveau ce concept est situé dans la hiérarchie : on peut déduire que *Hépatite Chronique* (C06.552.380.350) est subsumé par *Hépatite* (C06.552.380). La Fig.3 est un exemple de contraintes sous la forme de règles à appliquer sur les concepts. Par exemple l'association *Hépatite/induit chimiquement* est équivalente (≡) au concept *hépatite toxique*. On peut considérer cette équivalence comme une règle d'inférence qui remplacerait toute association du concept *hépatite* avec le qualificatif *induit chimiquement* par le concept *hépatite toxique*. Ces règles n'existent pas sous format

électronique mais nous comptons les modéliser et les exploiter dans le processus de recherche d'information.

```
Descripteur Francais: HEPATITE CHRONIQUE
Descripteur Americain: Hepatitis, Chronic
Code Cat MESH: C06.552.380.350
Synonymes Français: HEPATITE CHRONIQUE ACTIVE
Synonymes Américains: Chronic Hepatitis
                      Cryptogenic Chronic Hepatitis
                      Hepatitis, Chronic, Cryptogenic
Derives Americains: Hepatitis, Chronic Active
                    Active Hepatitides, Chronic
                    Active Hepatitis, Chronic
                    …
                        Hepatitis, Cryptogenic Chronic
MESH définition: A collective term for a clinical and pathological syndrome which has several causes and is
characterized by varying degrees of hepatocellular necrosis and inflammation. Specific forms of chronic
hepatitis include autoimmune hepatitis (HEPATITIS, AUTOIMMUNE), chronic hepatitis B; (HEPATITIS B, CHRONIC),
chronic hepatitis C; (HEPATITIS C, CHRONIC), chronic hepatitis D; (HEPATITIS D, CHRONIC), indeterminate
chronic viral hepatitis, cryptogenic chronic hepatitis and drug-related chronic hepatitis (HEPATITIS,
CHRONIC, DRUG-INDUCED).
Numero NLM: D006521
```

Fig.2 – Exemple de définition de concept

```
Hepatitis : C06.552.380+
Viral Hepatitis = Hepatitis,Viral Human and Hepatitis,Viral Animal
/chemically induced = Hepatitis,Toxic
/veterinary = Hepatitis,Animal or Hepatitis,Viral Animal
hepatitis parenterally transmitted= Hepatitis C
hepatitis enterally transmitted = Hepatitis E
Non-A, Non-B hepatitis = probably Hepatiis C
```

Fig.3 – Exemple de contraintes sur les concepts.

Nous envisageons d'améliorer le moteur Doc'CISMeF pour lui permettre de réaliser une recherche intelligente de ressources dans le cadre du Web Sémantique. Pour cela nous proposons une approche semblable à celle de nombreux projets, qui est l'utilisation d'une ontologie formelle définissant les concepts et les relations entre concepts et un ensemble de ressources annotées en fonction des concepts et relations de l'ontologie. Il manque à notre terminologie une dimension formelle mais sa structure est telle qu'elle sera facilement traduite dans un langage de représentation des connaissances.

Nos données de départ sont un sous-ensemble de mots clés, qualificatifs et types de ressources ainsi qu'un ensemble de ressources. A cela nous ajoutons les règles et contraintes sur les mots clés, le réseau sémantique de l'UMLS qui est composé de concepts médicaux (n=134) et de relations (n=54) entre les concepts ainsi qu'un ensemble de *règles métier* (n=45). Celles-ci ont été recueillies auprès d'un médecin généraliste. Elles sont de la forme *Complications (Hépatite, Cirrhose)* indiquant que le concept *Hépatite* est relié au concept *Cirrhose* par la relation *Complications.* En analysant de plus près ces relations on remarque qu'elles correspondent aux qualificatifs du MeSH et que les concepts sont les mots clés du MeSH. Ces règles permettront entre autres d'enrichir l'ontologie car la seule information qui est disponible dans la terminologie est que les concepts *Cirrhose* et *Hépatite* sont subsumés par le concept *Maladies du Foie.*

L'ontologie reposera sur un schéma RDFS qui définit les concepts (classes) qui sont les mots clés et types de ressources, les rôles (relations entre concepts) qui sont les qualificatifs et une relation de subsomption pour organiser les classes en hiérarchie. Les ressources seront annotées en fonction des concepts et des rôles de l'ontologie sous le format RDF. Les règles métier ne sont pas utilisées pour annoter

les ressources et elles ne contribuent pas à la définition de concepts. Elles seront traduites sous la forme de règles d'inférence et utilisées pour un raisonnement sur le contenu des ressources dans le processus de recherche d'information. RDFS permet de définir des hiérarchies de classes et des propriétés mais il n'intègre pas de capacités de raisonnement, comme ceux qu'offrent les systèmes basés sur des langages formels comme les Logiques de Description. L'écriture des règles d'inférence n'étant pas possible en RDFS, nous utiliserons les fonctionnalités de l'outil TRIPLE (Sintek & Decker, 2001) qui a été développé pour une recherche d'information intelligente basée sur les connaissances. Il permet de réaliser des raisonnements complexes sur des ressources RDF instances de concepts en traduisant RDF en Horn-Logic mais aussi en DAML+OIL[7]. Un accès à des éléments externes comme le classifieur RACER (Haarslev & Möller, 2001) (basé sur les Logiques de Description) offre la possibilité de bénéficier de ses mécanismes de raisonnement. Pour la recherche de ressources c'est particulièrement utile. Par exemple si une ressource *a* est instance du concept *C := Hepatite.Complications* et qu'un utilisateur recherche des ressources associées à *Cirrhose*, c'est à dire instances du concept *Cirrhose*, le système déduira que la ressource *a* est également une réponse à la requête grâce à la règle d'inférence *Hepatite.Complications Þ Cirrhose*.

```
// collection of resources
@cismef:resources {
  cismef:doc1[
    meta:title->"Document 1 is related to Hepatitis and
Aids";
    meta:author->"Toto";
    meta:keyword->HEPATITIS;
    meta:keyword->AIDS;
    meta:qualifier->COMPLICATIONS].
cismef:doc2[
    meta:title->"Document 2 is related to Accidents";
    meta:author->"Doctor E  in Risks";
    meta:keyword->ACCIDENT;
    meta:qualifier->RISKS].
cismef:doc4[
    meta:title->"Document 4 is related to Cirrhosis";
    meta:author->"Titi";
    meta:keyword->LIVERCIRRHOSIS].
}
// domain ontology
@cismefOntology {
  HEPATITIS[subClassOf -> LIVERDISEASES].
  LIVERCIRRHOSIS[subClassOf -> LIVERDISEASES].
  HEPATITIS[Complications->LIVERCIRRHOSIS].
  // ...
}
// requête toutes les ressources reliées à LIVERCIRRHOSIS//
FORALL Resource, Author <-
search(Resource,LIVERCIRRHOSIS)@search(cismef:resources,
cismefOntology) AND  Resource[meta:author -> Author].
```

```
compiling cismef.triple
running cismef.triple
***
Resource = cismef:doc1, Author = 'Toto'
Resource = cismef:doc4, Author = 'Titi'
done.
```

Fig.4 – Exemple d'ontologie, ressources, requête et fichier résultat sous TRIPLE.

L'intégration de règles d'inférences à l'outil TRIPLE nous permettra de réaliser des requêtes de niveau supérieur dans CISMeF et les relations du réseau sémantique offriront une navigation sémantique plus riche que la navigation hiérarchique actuellement en place.

5 Perspectives et Conclusion

[7] DAML+OIL joint committee. http://www.daml.org/2001/03/daml+oil-index.html.

Nous avons abordé dans cet article la problématique de la recherche d'information sur le Web. Nous avons présenté certains aspects du projet CISMeF qui est notre contexte d'application pour la recherche intelligente d'information. Nous voulons pour cela lui donner les moyens d'être intelligente en construisant une base de connaissances morphologiques, une base de règles d'associations et en formalisant la terminologie à l'aide d'un langage standard de représentation des connaissances. Les techniques de traitement automatique du langage naturel ont permis de construire une base de connaissances morphologiques. Le data mining permettra de découvrir des règles d'association entre concepts. Enfin le raisonnement sur les ontologies offrira un niveau supérieur tant au niveau de l'ontologie (vérification de la consistance et cohérence, exploitation du réseau sémantique de l'UMLS) qu'au niveau recherche d'information grâce à un ensemble de règles d'inférences. Nous pensons que la plus value est dans la combinaison de ces différentes techniques.

L'évaluation de l'apport de chacune des méthodes se fera de deux manières. Tout d'abord par une expansion automatique (enrichissement) des requêtes pour élargir le champ de la recherche en utilisant chacune des ressources (base morphologique, règles d'association, et ontologie formelle) séparément puis conjointement. Les requêtes considérées sont celles du fichier log dont le nombre de réponses est nul. Ensuite par une expansion interactive : nous demanderons à un échantillon d'utilisateurs abonnés au site d'évaluer, pour chaque requête qu'ils poseront, « l'utilité » des différentes suggestions de requêtes enrichies apportées par les différentes méthodes. Cette évaluation (expansion automatique ou interactive) à échelle réelle permettra d'établir une base de règles ou un protocole pour l'application des méthodes en fonction du type de requête posé.

Références

AGRAWAL R. & SRIKANT R. (1994) Fast Algorithms for Mining Association Rules in Large Databases. *Proceedings VLDB Conference*, p.478-499.

BAKER T. (2000) A Grammar of Dublin Core. *Digital-Library Magazine.* 6(10).

BERNERS-LEE T., HENDLER J. & LASSILA O. (2001). The Semantic Web. *Scientific American* p.35-43.

BOUAUD J., BACHIMONT B., CHARLET J. & ZWEIGENBAUM P. (1995) Methodological Principles for Structuring an « Ontology ». *Proceedings of IJCAI conference*.

DARMONI SJ., THIRION B., LEROY JP., DOUYÈRE M. & al. (2001). A Search Tool based on 'Encapsulated' MeSH Thesaurus to Retrieve Quality Health Resources on the Internet. *Medical Informatics & the Internet in Medicine*, 26 (3):165-178.

DAVEY BA. & PRIESTLEY HA. (1994) Introduction to Lattices and Order. *Cambridge University*

GANDON F., DIENG-KUNTZ R., CORBY O. & GIBOIN A. (2002) Web Sémantique et Approche Multi-Agents pour la Gestion d'une Mémoire Organisationnelle Distribuée. *Journées Ingénierie des Connaissances*, p.15-26.

GANTER B. & WILLE R. (2000) Formal Concept Analysis : Mathematical Foundations. *Springer-Verlag*.

GAUSSIER E., GREFENSTETTE G., HULL D. & ROUX C. (2000) Recherche d'Information en Français et Traitement Automatique des Langues. *TAL*, 41(2) : p.473-493.

GRABAR N., ZWEIGENBAUM P., SOUALMIA LF., & DARMONI SJ. (2003) Matching Controlled Vocabulary Words. *Medical Informatics Europe* p. 445-450.

GUARINO N., MASOLO C. & VETERE G. (1999) Ontoseek : Content-Based Access to the Web. *IEEE Intelligent Systems* 14(3).

HAARSLEV V. & MÖLLER R. (2001) Description of the RACER System and its Applications. *Proceedings Internationl Workshop on Description Logics.*

HADDAD MH., CHEVALLET JP. & BRUANDET MF. (2000) Relations between Terms Discovered by Association Rules. *Practices of Knowledge Discovery in Databases.*

LASSILA O. & SWICK R. (1999) Resource Description Framework (RDF) Model and Syntax Specification. *W3C Candidate Recommendation 1999*.

LAUBLET P., REYNAUD C. & CHARLET J. (2002). Sur Quelques Aspects du Web Sémantique. *Actes des deuxièmes assises nationales du GdRI3*, p.59-78.

LINDGBERG DAB., HUMPHREYS BL. & MCCRAY AT. (1993) The Unified Medical Language System. *Methods of Information in Medicine.*

LUKE S. & HEFLIN J. (2000) SHOE Project Specification.

MAYER MA., DARMONI SJ., FIENE M., KÖHLER C., & al. (2003) MedCIRCLE - Modeling a Collaboration for Internet Rating, Certification, Labeling and Evaluation of Health Information on the Semantic World-Wide-Web. *Medical Informatics Europe* p.667-672.

NEW B., PALLIER C., FERRAND L. & MATOS R. (2001) Une Base de Données Lexicales du Français Contemporain sur Internet: LEXIQUE, *L'Année Psychologique*, p. 447-462.

PASQUIER N. (2000) Data Mining, : Algorithmes d'Extraction et de Réduction des Règles d'Association dans les Bases de Données. *Thèse de doctorat*, Université Clermont-Ferrand II.

RODRIGUES JM., TROMBERT-PAVIOT B., BAUD R. & al. (1998) GALEN-In-Use : using Artificial Intelligence Terminology Tools to Improve the Linguistic Coherence of a National Coding System for Surgical Procedures. Cesnik et al. (eds). *MedInfo'1998*.

SAVOY J. (2002) Morphologie et Recherche d'Information. *Cahier de Recherche en Informatique,* CR-I-2002-01, Université de Neuchatel.

SINTEK M. & DECKER S. (2001) TRIPLE- An RDF Query, Inference and Transformation Language. *Proceedings of Deductive Databases and Knowledge Management Workshop.*

SOWA JF. (2000) Ontology, Metadata and Semiotics. *ICCS*

ZWEIGENBAUM P., GRABAR N. & DARMONI SJ. (2001) Apport de Connaissances Morphologiques pour la Projection de Requêtes sur une Terminologie Normalisée. *TALN*.