Contexte et sémantique pour une indexation de documents semi-structurés

Haïfa ZARGAYOUNA

LIMSI/CNRS-Université Paris 11
Bâtiment 508. BP 133, F-91403 Orsay Cedex.
haifa.zarg-ayouna@limsi.fr
http://www.limsi.fr/Individu/zarga

Catégorie Jeune Chercheur

RÉSUMÉ. *Les documents semi-structurés comme les documents XML présentent l'avantage de posséder une structure explicite qui facilite leur présentation et leur exploitation dans différents contextes. Cependant, très souvent, la majeure partie de l'information reste contenue dans les champs textuels. Il est donc devenu primordial de concevoir des méthodes permettant d'exploiter à la fois la structure et le contenu textuel de ces documents. Les techniques classiques de Recherche d'Information (RI) n'utilisent pas ou peu la structure des documents alors que les langages de requête issus de la communauté Bases de Données (BD) n'exploitent pas le contenu textuel et ne permettent pas une présentation des résultats par ordre de pertinence. De plus en plus de chercheurs essaient de combiner les approches de RI et de BD pour pallier leurs limites respectives. Dans ce travail, nous présentons une structure d'index qui permet des requêtes structurées et une présentation des résultats par ordre de pertinence. Pour cela, nous avons étendu le modèle vectoriel de Salton pour une vue bi-dimensionnelle du document en adaptant le calcul du TF-IDF. Par ailleurs, nous proposons d'utiliser une ontologie reliée aux termes du corpus pour modéliser la notion de voisinage sémantique à l'aide d'un calcul de similarité entre termes. Cette indexation permet donc une recherche contextuelle (par la structure) et sémantique (par l'ontologie).*

MOTS-CLÉS : *XML, Indexation, Contexte, Modèle Vectoriel, Sémantique, Similarité, Ontologie.*

KEYWORDS: *XML, Indexing, Context, Vector Model, Semantic, Similarity, Ontology.*

1. Introduction

XML est de plus en plus reconnu comme un format standard de documents et l'on peut penser que dans un futur proche, un nombre important de documents et de données seront disponibles en format XML. L'avantage de ces documents est qu'ils possèdent une structure qui facilite leur présentation, ainsi que leur interprétation et leur exploitation dans des contextes présentant différents besoins. Cependant, très souvent, la majeure partie de l'information reste contenue dans les champs textuels, l'utilisation exclusive de la structure n'est donc pas suffisante.

Il est devenu primordial de concevoir des méthodes permettant d'exploiter à la fois la structure et le contenu textuel de ces documents. Ces méthodes doivent permettre de considérer les termes présents dans un document en fonction de leur contexte d'apparition, c'est à dire l'endroit où ils apparaissent dans la structure XML.

Par exemple un ensemble de comptes-rendus médicaux peut être structuré en utilisant des balises <info-patient>,<antecedents>, <symptomes>, <traitement>. L'occurrence d'un nom de médicament dans la partie du texte balisée par <antecedents> signifiera que le patient a été précédemment traité avec ce médicament, alors que l'occurrence de ce nom de médicament dans la partie du texte balisée par <traitement> signifiera que l'on préconise ce médicament pour traiter le patient.

Par ailleurs, la nécessité est apparue de tenir compte de différentes relations entre les termes : synonymie, hyperonymie, hyponymie..., et plus généralement d'essayer de repérer dans un texte l'occurrence d'un **concept** plutôt que simplement d'un **terme**. Les termes sont reliés à des concepts organisés dans une ontologie. Plusieurs travaux s'intéressent à la construction d'ontologies, et proposent soit des techniques automatiques ou semi-automatiques d'extraction de connaissances ou de classification, soit des méthodologies d'acquisition de connaissances, en particulier des méthodes de construction d'ontologies à partir de textes utilisant des outils de Traitement Automatique de la Langue [SZU 02].

En effet, l'utilisation d'ontologies générales telles que Wordnet se révèle souvent inadaptée ou insuffisante pour une catégorie de documents et il est alors nécessaire d'élaborer une ontologie spécifique à un domaine et souvent à une application. Il est intéressant de remarquer qu'une ontologie construite à partir de textes permet de guider la recherche d'informations et la fouille de ces textes, mais que celles-ci peuvent également être exploitées ensuite pour enrichir l'ontologie dans un processus cyclique.

Dans cet article, nous proposons une méthode d'indexation de documents XML qui étend le modèle vectoriel de Salton [SAL 71] en exploitant d'une part la structure des documents, et d'autre part une ontologie du domaine. Dans la section 2 nous présentons les approches issues des communautés de recherche d'information et de bases de données pour exploiter les documents XML. Nous décrivons ensuite notre modèle d'indexation en précisant dans la section 3.1 comment il utilise la structure XML des documents, et dans la section 3.2 comment il exploite une ontologie associée. Ces deux sections sont illustrées par un exemple issu d'un travail sur des comptes-rendus

d'hospitalisation extraits du corpus Menelas. Nous présentons brièvement un exemple de recherche et d'évaluation de notre système. Nous conclurons en précisant les limites et les perspectives de notre approche.

2. Travaux existants

L'avènement de XML pose de nouveaux défis à l'indexation et la recherche de documents (ou données). Deux communautés, RI (Recherche d'Information) et BD (Bases de données), se sont intéressées à l'élaboration de moteurs de recherche pour les documents XML en essayant d'y intégrer leur domaines d'expertise respectifs. La communauté RI centre son approche sur l'utilisation d'outils de TAL (Traitement Automatique de la Langue), ce qui lui permet un traitement plus efficace sur le contenu textuel. La communauté BD exploite la définition formelle de ses langages de requêtes, permettant ainsi une prise en compte des caractéristiques structurelles.

Cependant, les techniques issues des deux communautés présentent des limites.

– Les techniques de recherche d'information traditionnelles sont basées sur une représentation linéaire des documents, elles possèdent les limites suivantes : (i) elles ignorent la structure du document, et (ii) procèdent à des requêtes plates (par mots clés).

– Les langages de requêtes XML issus de la communauté bases de données (XQL, XOQL, Xquery[W3C 01], etc.) sont d'usage limité dans le cadre d'une RI : (i) ils requièrent une connaissance de la structure du document, (ii) procèdent à un appariement exact et (iii) ne présentent pas les résultats par ordre de pertinence.

Deux sortes de documents XML sont traités en fonction de la communauté. La communauté RI utilise des documents orientés «texte», celle de BD des documents orientés «données».

Les langages issus de la communauté BD sont plus adaptés à une utilisation des documents XML pour un échange de données dans une forme structurée, comme les EDI [1] classiques. Les documents sont orientés «données» et par ce fait présentent un contenu textuel assez faible. Le traitement sur le texte ne dépasse généralement pas le simple appariement de mots clés.

Ces communautés vont de plus en plus vers une hybridation de leurs outils pour la prise en compte de la structure ainsi que des traitements plus fins sur le texte. Ainsi, lors de la première campagne d'évaluation INEX [SIT 03] 14 groupes de recherche présentaient des outils spécifiques à XML (contre 5 groupes en BD et 20 en RI).

Les documents ne sont plus considérés en tant qu'entités atomiques, mais comme des agrégats d'objets en corrélation qui peuvent être recherchés séparément [CHI 01]. Des travaux tels que [LAL 00] se basent sur la théorie de l'évidence de Dempster-Shafer. Un opérateur d'agrégation permet de calculer la pertinence d'un document.

1. Electronic Data Interchange : Echange électronique de données

Des travaux similaires [BOR 00] calculent la pertinence d'un document par une agrégation floue du degré de pertinence de ses composantes. D'autres travaux reposent sur les modèles bayésiens [PIW 02]. Un réseau bayésien est construit par document et par requête, la probabilité conditionnelle de la pertinence d'un sous-texte par rapport au texte global est calculée pour chaque nouvelle requête et est approximée par un modèle de TF-IDF. XIRQL [FUH 02] est un système qui se greffe sur XQL et incorpore la notion de prédicats vagues, il prend aussi en considération le poids des termes. XXL [THE 02], est un moteur de recherche issu des travaux en base de données. Il utilise des index ainsi que des ontologies. Il ajoute de nouveaux opérateurs pour un appariement approximatif des mots. L'écriture de la requête reste proche de SQL.

L'émergence du Web Sémantique a accru l'intérêt pour les ontologies. La construction d'une ontologie générale est abandonnée et les travaux semblent s'orienter de plus en plus vers un web sémantique par domaine, comme l'a confirmé la tenue du premier Workshop du Web Sémantique Médical [WSM 03], où l'accent a été mis sur la nécessité des ontologie du domaine.

Notre travail comporte deux volets : un volet qui traite la structure du document et s'inspire des techniques utilisées en BD, essentiellement sur les arbres. Un autre volet traite l'aspect sémantique en intégrant une ontologie du domaine.

3. Description du modèle d'indexation

Nous proposons d'indexer des corpus textuels spécialisés composés d'un ensemble de documents partageant des éléments de structure mais pouvant avoir des structures différentes (validées ou pas par des DTDs). Nous prenons comme exemple de référence dans cet article des comptes-rendus d'hospitalisation extraits du corpus qui a été construit durant le projet Menelas [ZWE 95][2]. Ce corpus a fait l'objet [HAB 01] d'un encodage suivant le TEI XML Corpus Encoding Standard. Nous avons ajouté pour nos exemples des balises de contenu qui sont plus pertinentes pour notre application.

3.1. *La structure : une notion de contexte*

Les documents XML offrent la possibilité de délimiter leur contenu en un ensemble d'unités sémantiques. Toutefois, ce découpage reste arbitraire et une même unité sémantique peut être représentée selon des structures différentes.

Néanmoins, la structure offre un apport sémantique non négligeable, dans [SCH 02] il est même affirmé que : *«Ignorer la structure du document revient à ignorer sa sémantique»*.

2. Nous remercions Pierre Zweingenbaum et Natalia Grabar pour nous avoir fourni ce corpus ainsi que diverses références s'y reportant.

La structuration hiérarchique du document peut être associée à la notion de contexte documentaire. Un contexte documentaire est défini comme une unité textuelle à l'intérieur d'un document (paragraphe, section, chapitre). L'objectif d'un tel contexte est de mieux prendre en compte la structure des documents [HAB 97]. Ces unités textuelles peuvent avoir des relations entre elles, des travaux [HER 03] visent à trouver ces relations et à les représenter par des annotations.

Nous modélisons la structure XML comme un arbre étiqueté où chaque élément et attribut correspond à un noeud. Nous ne faisons aucune distinction entre les éléments et les attributs.

Definition 1 *Un arbre étiqueté AE est un quadruplet <N_0,N, A, label>*

où N_0 est le noeud racine, N est un ensemble fini de noeuds ; $A \subset N \times N$ un ensemble fini d'arcs ; et label est une fonction qui associe à un noeud une étiquette.

Nous recherchons des correspondances «sémantiques» entre la structure de la requête et celle du document. Pour cela nous ne recherchons pas les balises uniquement par leurs étiquettes mais par leurs positions dans l'arbre. Ainsi, quand nous cherchons **traitements.traitement** ou **antecedents.traitements.traitement** (voir figure1) nous ne sélectionnons pas les mêmes noeuds.

Definition 2 *Un chemin est un ensemble de noeuds $n_1, .., n_n$ tel que $\forall n_i, n_{i+1}$ pour i :1..n, $\exists (n_i, n_{i+1}) \in A$*

Nous supposons que deux noeuds ayant le même label et le même chemin qui mène à la racine ont la même sémantique et réfèrent aux mêmes concepts (qu'on appelle concept de contexte).

Ainsi nous générons un modèle de balise qui regroupe ensemble les structures communes (exemple : patients.patient.symptomes.symptome dans la figure1)

Definition 3 *Nous définissons un arbre réduit par un arbre étiqueté*

AE' = <N_0, N', A', label>

*chaque noeud de N' possède un chemin unique appelé **modèle de balise**.*

Dans la figure1, l'arbre (a) est la représentation du document et l'arbre (b) est sa représentation réduite.

L'arbre réduit revient à essayer de déduire une grammaire possible pour la structure des documents. Ceci nous permet d'approximer une DTD commune.

Ces modèles de balise [3] serviront d'entrée à la structure d'index. Nous considérons ces modèles comme unités d'index (qui peuvent aussi être définies par l'utilisateur).

3. Dans ce qui suit, nous parlerons indifféremment de modèles de balises et balises

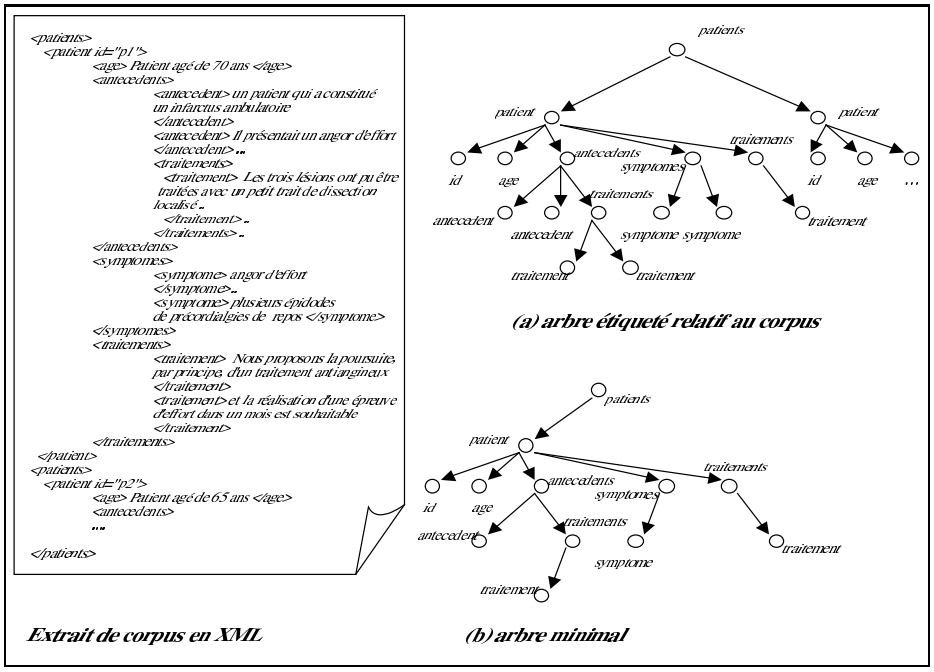

Figure 1. *Exemple d'un corpus XML avec sa représentation en arbre et représentation réduite*

Ces unités d'index peuvent servir d'unités de recherche, ainsi le document ne constitue plus une entité atomique. Il est remplacé par des unités d'information. Nous indexons ces unités selon le modèle vectoriel de Salton [SAL 71].

Le document n'est plus représenté par un vecteur mais par une **matrice** de termes et de modèles de balise.

Nous procédons à l'extraction des termes en appliquant les outils de TAL traditionnellement utilisés en RI, ce qui correspond à : (i) ignorer les mots qui appartiennent à des anti-dictionnaires (stop lists), c'est à dire les listes prédéfinies qui contiennent les mots qui peuvent être ignorés (comme les mots communs) ; (ii)appliquer les techniques de troncature (stemming) ; (iii)ne prendre en considération que les mots pleins. Les unités linguistiques seront restreintes aux catégories suivantes : noms, verbes et adjectifs ; (iv) éliminer les mots très fréquents et qui ne sont de ce fait plus représentatifs.

Les techniques de TAL sont beaucoup plus riches que les outils présentés. L'apport des études morphologiques, syntaxiques et sémantiques (avec plus ou moins de robustesse) est évident pour la RI, notamment pour retrouver des unités d'index plus

pertinentes [JAC 00]. Nous préférons, dans un premier temps, mettre en oeuvre des outils robustes qui ne nécessitent pas une validation.

Les termes indexés sont lemmatisés avec le `TreeTagger` qui est un étiqueteur morphologique. Il génère pour chaque mot sa catégorie grammaticale et le lemme associé.

Le calcul du poids des termes est influencé par le contexte (l'unité d'indexation) dans lequel ils apparaissent. Ce calcul de poids s'inspire de la méthode du TF-IDF qu'on applique aux balises. Ainsi nous définissons le TF-ITDF (Term Frequency-Inverse Tag and Document Frequency[4]), de la manière suivante :

Soit T l'ensemble de tous les termes qui figurent dans le corpus,

B l'ensemble de tous les modèles de balises,

D l'ensemble de tous les documents du corpus.

$$TF - ITDF(t, b, d) = TF(t, b, d).ITF(t, d).IDF(t, b)$$
$$IDF(t, b) = log(|B|_d / TagF(t, b))$$
$$ITF(t, d) = log(|D|_b / DF(t, b))$$

$|D|_b$ est le nombre total de documents où le modèle de balise b est présent dans leur structure.

$|B|_d$ est le nombre total de balises dans le document d.

$DF(t, b)$: (Document Frequency) est le nombre de documents qui contiennent la balise b et dans laquelle le mot t apparaît au moins une fois.

$TagF(t, b)$: (Tag Frequency) est le nombre de balises dans le document d et dans lesquelles le mot t apparaît au moins une fois.

Cette formule nous permet de calculer la force discriminatoire d'un terme t pour une balise b relative à un document d.

Balises	Doc1	Doc2	Doc3	Doc4
*@antecedent	{asthme, pneumopathie, angine}	{rhumatisme}	{angoisse}	{pneumopathie}
**@symptome	{fièvre, angine, asthme}	{asthme}	–	{fièvre}
...				

Tableau 1. *Liste de mots par balises et document*

* @antecedent = patients.patient.antecedents.antecedent
** @symptome = patients.patient.symptomes.symptome

TF-IDF(asthme, doc1) = 2 * log(4/2) = 0.6

4. Par analogie à Term Frequency-Inverse Document Frequency

TF-IDF(asthme, doc2) = 1 * log(4/2) = 0.3

Nous supposons avoir cinq modèles de balise dont les deux représentés dans le tableau1.

TF-ITDF(asthme, symptome, doc1) = 1 * log(3/2) * log(5/2) = 0.06

TF-ITDF(asthme, antecedent, doc1) = 1 * log(4/1) * log(5/2) = 0.23

TF-ITDF(asthme, symptome, doc2) = 1 * log(3/2) * log(5/1) = 0.11

Le TF-ITDF(asthme, symptome, doc1) est inférieur au TF-ITDF(asthme, antecedent, doc1) parce que le terme «asthme» n'apparaît pas dans la balise antecedent des autres documents. Il est de ce fait plus *représentatif* de la balise antecedent que de la balise symptome dans le document doc1.

TF-ITDF(asthme, symptome, doc2) est supérieur au TF-ITDF(asthme, symptome, doc1) parce que le terme «asthme» apparaît aussi dans la balise antecedent. Il est de ce fait plus *spécifique* dans le document doc2.

Le terme est non seulement pondéré par sa fréquence dans la balise mais aussi par la répartition de la balise dans la base documentaire. La répartition du terme dans le document est aussi importante dans le calcul. Si un terme apparaît dans des balises différentes il est moins représentatif pour une balise donnée qu'un terme qui n'apparaît que dans cette balise.

3.2. *Le contenu : une notion sémantique*

Le TF-ITDF constitue une information numérique sur le poids du terme par rapport à un modèle de balises et un document. Aucune considération de la sémantique du terme dans le calcul de son poids n'est prise en compte. Par exemple, le poids est calculé indépendamment de l'apparition des termes synonymes dans la même balise.

Plusieurs chercheurs ont adapté le modèle vectoriel en indexant directement les concepts à la place des termes. On parle alors de vecteur de concepts [TOD 01]. Ces approches traitent essentiellement la synonymie en remplaçant les termes par leurs concepts. Nous traitons des liens plus riches entre les termes. Nous nous aidons de mesures numériques pour capter ces liens symboliques dans l'ontologie. On peut bien sûr, remplacer le terme par le concept qui lui est relatif. Ceci peut résoudre le problème de la synonymie mais ne change rien quant au besoin des autres relations de spécialisation et de généralisation.

L'ontologie est de plus en plus utilisée pour l'expansion de requêtes [BAZ 03]. Selon le retour de l'utilisateur, la requête est enrichie par les concepts de l'ontologie qui lui sont reliés. Son utilité s'est vue confirmée par le web sémantique.

Nous proposons d'utiliser une ontologie lexicalisée (les concepts sont reliés aux termes du corpus) lors de la phase d'indexation pour enrichir le terme par l'espace

de sens qui l'entoure. Les concepts sont aussi reliés aux balises, nous les appelons **concept de contexte**.

Cependant, contrairement aux méthodes existantes, nous ne nous restreignons pas à l'utilisation des concepts. En effet, les termes sont enrichis s'ils sont reliés aux concepts. Il est important de noter que lors de la recherche, nous pouvons aussi retrouver les termes qui ne sont pas reliés à l'ontologie. Nous calculons cette similarité lors de la phase d'indexation pour alléger le temps de traitement des requêtes.

Nous calculons une similarité entre les concepts rattachés aux termes. Ainsi, nous définissons sim une fonction qui associe à chaque couple de termes un degré de similarité sémantique en fonction de la relation entre leurs concepts respectifs.

Rada et al. [RAD 89] ont été les premiers à suggérer que la similarité dans un réseau sémantique peut être calculée en se basant sur les liens taxonomiques «is-a». Un moyen des plus évidents pour évaluer la similarité sémantique dans une taxonomie est de calculer la distance entre les noeuds comme le chemin le plus court.

Dans [RES 99], l'idée est de calculer les chemins comme ceux qui lient chaque concept à son plus proche ancêtre en haut de l'ontologie.

Nous sommes conscientes que le calcul de la mesure de similarité par restriction sur le lien «is-a» n'est pas toujours bien adapté car, dans la réalité, les taxonomies ne sont pas toujours au même niveau de granularité, des parties peuvent aussi être plus denses que d'autres. Ces problèmes peuvent être résolus en associant des poids aux liens. L'affectation de ces poids peut être basé sur : les types de liens présents, la profondeur du lien dans la taxonomie et la densité du concept par ses voisins immédiats.

Nous nous basons sur la mesure de similarité présentée par [WU 94]. Cette mesure a l'avantage d'être simple et d'avoir de bonnes performances.

$$sim(c_1, c_2) = \frac{2 * depth(c)}{depth_c(c_1) + depth_c(c_2)}$$

Où C est le concept le plus proche qui subsume[5] C_1 et C_2 (en nombre d'arcs), $depth(C)$ est le nombre d'arcs qui sépare C de la racine et $depth_c(C_i)$ avec i le nombre d'arcs qui séparent C_i de la racine en passant par C.

Cette mesure de similarité est intéressante mais présente une limite car elle vise essentiellement à détecter la similarité entre deux concepts par rapport à leur distance de leur plus petit subsumant commun. Plus ce subsumant est général, moins ils sont similaires (et inversement). Dans le cas de recherche d'information, il est à notre sens plus intuitif de ramener les concepts qui sont subsumés par les concepts de la requête que par son voisinage. La similarité doit de ce fait prendre en considération les liens père/fils et la densité des concepts.

La similarité calculée avec la mesure de similarité présentée ci-dessus entre paraclinic_sign et arterial_hypertension_sign qui est un de ses descendants (voir l'onto-

5. Un concept C1 est subsumé par C2 si C1 est un sous-concept de C2, on parle alors de père et de fils.

logie présentée dans la figure2), est inférieure à la similarité entre paraclinic_sign et clinical_sign qui est un frère.

$sim(paraclinic\_sign, arterial\_hypertension\_sign) = \frac{2*6}{6+9} = 0.8$

$sim(paraclinic\_sign, clinical\_sign) = \frac{2*5}{6+6} = 0.83$

Nous voulons favoriser les liens entre père et fils aux autres liens. Pour cela nous pénalisons les concepts qui ont un subsumant en commun par rapport à ceux de la même lignée. Nous rajoutons une mesure de spécificité qui prend en considération le degré de spécificité du concept, c'est à dire le nombre d'arcs qui le séparent de bottom[6].

$spec(c_1, c_2) = depth_b(C) * distance(C, c_1) * distance(C, c_2)$

avec $depth_b(C)$ est le nombre d'arc qui sépare C de bottom et $distance(C, c_i)$ la distance en nombre d'arc entre C et c_i

Ainsi la mesure de similarité devient :

$sim(c_1, c_2) = \frac{2*depth(c)}{depth_c(c_1)+depth_c(c_2)+spec(c_1,c_2)}$

la similarité entre paraclinic_sign et arterial_hypertension_sign reste la même tandis que la similarité entre paraclinic_sign et clinical_sign devient égale à 0.5.

Nous rajoutons cette mesure de similarité entre concepts aux termes qui leurs sont relatifs. Le poids sémantique est une fonction qui calcule le poids des termes en tenant compte de la similarité conceptuelle entre les termes du même contexte.

Ce poids noté $SemW(t, b, d)$ est calculé de la manière suivante :

$SemW(t, b, d) = TF - ITDF(t, b, d) + \sum Sim(t, t_i) * TF - ITDF(t_i, b, d)$

avec $Sim(t, t_i) > seuil$; $t_i \in T$ et seuil une valeur à déterminer empiriquement, nous la fixons à la similarité entre le concept de t et le concept contexte.

Considérons l'exemple d'ontologie présentée dans la figure2.

Termes	Doc1	Doc2
apa# "anomalie pression artérielle"	0	1.39
ha# "hypertension artérielle"	0.69	0
fiévre	0.95	0
angoisse	0	0.22

Tableau 2. *Liste de termes avec leurs poids dans les documents respectifs*

Les termes présentés dans le tableau2 sont des termes qui apparaissent dans la balise **symptome**. Cette balise est rattachée au concept de contexte **sign**.

6. Le concept le plus bas de l'ontologie.

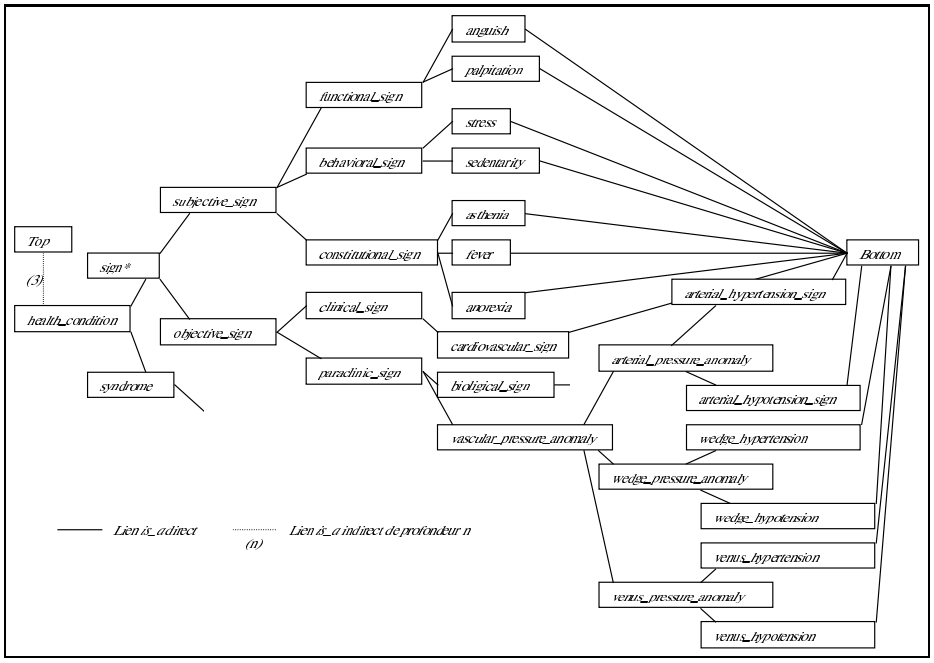

Figure 2. *Un extrait de l'ontologie Menelas*

$SemW(apa, symptome, doc1) = 0 + (sim(arterial\_pressure\_anomaly,$
$arterial\_hypertension\_sign) * 0.69) = 0.94 * 0.69 = 0.64$

Les $sim(arterial\_pressure\_anomaly, fever)$ et $sim(arterial\_pressure\_anomaly,$ $anguish)$ ne sont pas prises en considération car elles sont au dessus du seuil :

$sim(arterial\_pressure\_anomaly, sign) = \frac{2*4}{4+8} = 2/3$

$SemW(ha, symptome, doc1) = 0.69 + (sim(arterial\_hypertension\_sign,$
$arterial\_pressure\_anomaly) * 0) = 0.69$

Les poids sémantiques sont présentés dans le tableau qui suit :

Termes	Doc1	Doc2
apa# "anomalie pression artérielle"	0.64	1.39
ha# "hypertension artérielle"	0.69	1.3
fièvre	0.95	0
angoisse	0	0.22

Tableau 3. *Liste de termes avec leurs poids sémantiques*

Les poids de "anomalie pression artérielle" et "hypertension artérielle" ont été augmenté par leurs concepts proches présents dans les documents.

4. Recherche et expérimentations

Notre système d'indexation permet trois types de requêtes : par mots clés, par la structure et par structure + mots clés. Une requête structurée est sous forme XML, l'élément à retourner est marqué par ' ?'. Une matrice lui est associée comme pour les documents indexés. On peut attribuer un poids aux termes de la requête. Dans un premier temps nous attribuons des 1 et 0 pour signifier l'existence ou non des termes. Nous avons adapté une mesure de pertinence classique du modèle vectoriel. La pertinence d'une requête Q par rapport à un document D devient :

$S(D,Q) = \prod_{i=1}^{n} cosinus(b_{iQ}, b_{iD})$

Tels que b_{iQ} et b_{iD} sont les vecteurs de poids des termes dans la balise b_i de Q (requête) et D (document) respectivement. Rappelons que :

$$cosinus(V1,V2) = \frac{\sum_{i=1}^{m} v1_i.v2_i}{\sqrt{\sum_{i=1}^{m} v1_i.v1_i}.\sqrt{\sum_{i=1}^{m} v2_i.v2_i}}$$

Où $v1_i$ et $v2_i$ représentent les éléments du vecteur V1 et V2. Le produit des cosinus peut bien sûr être remplacé par une autre fonction (la somme par exemple).

Notre système d'index a pour vocation de permettre une recherche plus précise et plus pertinente (par rapport aux méthodes classiques) pour l'utilisateur. Il possède plusieurs composantes, ce qui lui permet d'être modulaire (nous pouvons utiliser d'autres outils de TAL plus riches, d'autres mesures de similarités par exemple). Cette modularité qui constitue sa richesse pose malheureusement des difficultés d'évaluation.

Nous pouvons évaluer directement la structure d'index. Il s'agit généralement de calculer le temps d'indexation, l'espace de stockage de l'index par rapport à la taille de la base documentaire. Comme nous utilisons une ontologie, sa construction et son rattachement au corpus font partie de la phase d'indexation. Le calcul du temps de construction de l'index ne permet pas de juger de la valeur de l'index.

On peut aussi évaluer la pertinence d'un index en testant son impact sur la recherche, en utilisant les mesures de pertinence classiques de rappel et précision ou l'exhaustivité et la pertinence (comme pour INEX). La difficulté de l'évaluation de notre système est d'avoir un corpus avec des balises XML «pertinentes» (en vue d'une recherche structurée) et une ontologie associée. Quand l'ontologie est créée à partir du corpus manuellement ou par des méthodes semi-automatiques, le lien entre les termes et le concept est évident. Le problème se pose quand on dispose d'un corpus de spécialité et d'une ontologie du domaine, le lien n'est pas toujours évident.

Plusieurs variables entrent donc en compte pour l'évaluation de notre système. Son évaluation est assez délicate parce que plusieurs composantes se mêlent. Il faudrait tester la validité des termes choisis par l'index, l'utilité de l'ontologie, la pertinence de la mesure de similarité.

Dans un premier temps nous avons testé notre système avec quelques requêtes pour pouvoir juger de son efficacité par rapport à une recherche plate (sans structure) et une recherche structurée (avec ou sans mots clés).

Nous avons choisi la collection de pièces de Shakespeare (37 documents XML, 8308 Kbits) parce que son balisage est «sémantiquement» pertinent. Les documents sont découpés en actes, scènes, etc. Nous utilisons dans un premier temps WordNet comme hiérarchie de concepts. Nous pensons que l'utilisation de WordNet est adaptée ici, du fait que les termes dans le corpus sont assez généraux.

Les expérimentations sont en cours et présentent déjà des résultats encourageants. Nous présentons quelques exemples de requêtes :

Requête 1
Chercher les titres des scènes
où un des personnages est un roi :

```
<PLAY>
<ACT>
<SCENE>
<TITLE> ? </TITLE>
<SPEECH>
<SPEAKER> King </SPEAKER>
</SPEECH>
</SCENE>
</ACT>
</PLAY>
```

Requête 2
Chercher les titres des actes
où un homme parle de mort :

```
<PLAY>
<ACT>
<TITLE> ? </TITLE>
<SCENE>
<SPEECH>
<SPEAKER> Man </SPEAKER>
Death
</SPEECH>
</SCENE>
</ACT>
</PLAY>
```

Requête 3
Chercher les titres des pièces
qui ont un épilogue

```
<PLAY>
<TITLE> ? </TITLE>
<EPILOGUE>
</EPILOGUE>
</PLAY>
```

Nous orientons nos expérimentations vers une évaluation à deux phases. Ainsi dans une première phase nous pouvons tester dans le cadre de INEX 2004 l'utilité de l'adaptation du modèle vectoriel aux documents semi-structurés. Dans une deuxième phase nous évaluerons l'apport de l'utilisation de l'ontologie dans le cadre classique de RI.

5. Conclusion

Nous avons défini un modèle d'indexation de documents XML qui adapte le modèle vectoriel de Salton sur deux aspects :

– la prise en compte de la structure du document,

– l'intégration de la notion de voisinage sémantique d'un terme.

Pour la prise en compte de la structure du document, nous avons défini la notion de chemin de balise qui permet de découper le document en un ensemble d'unités d'information. Chaque chemin de balise définit un contexte d'occurrence des termes. Nous avons alors adapté le calcul du poids TF-IDF en définissant le TF-ITDF qui calcule le poids des termes par chemin de balise. Le document n'est donc plus représenté par un vecteur mais par une matrice.

Pour intégrer la notion de voisinage sémantique, nous avons utilisé une ontologie de concepts auxquels sont reliés les termes des documents. Dans un premier temps nous n'avons pris en considération que les liens de spécialisation/généralisation entre les concepts. En nous basant sur la mesure de similarité entre concepts présentée par [WU 94], nous avons proposé une nouvelle mesure telle que les descendants directs d'un concept sont considérés plus similaires au concept que ses frères. Nous avons alors défini un nouveau calcul du poids des termes SemW qui tient compte de la similarité conceptuelle entre les termes du même contexte.

La mise en oeuvre sur un ensemble de comptes-rendus d'hospitalisation extraits du corpus Menelas a permis d'illustrer les bénéfices de cette indexation par rapport au modèle vectoriel. Nous nous attachons actuellement à définir un cadre d'évaluation plus rigoureux de notre approche permettant de la comparer sur différents corpus à d'autres types d'indexation. Il sera

intéressant d'évaluer aussi séparément l'apport de la prise en compte de la structure et l'apport du voisinage sémantique.

Une des limites de notre approche, concernant l'aspect sémantique, tient au fait que nous supposons disposer d'une ontologie de concepts reliée au corpus. Rappelons que nous nous plaçons dans le cadre de l'indexation de corpus spécialisés, pour lesquels on peut supposer qu'il existe certaines ressources sur le vocabulaire du domaine. Pour utiliser le modèle présenté dans cet article, il suffit de disposer d'une structure hiérarchique entre concepts correspondant aux liens de spécialisation/généralisation. Cependant nous envisageons de travailler sur la prise en compte d'autres type de liens comme par exemple le lien de composition.

Concernant l'aspect structurel, nous considérons actuellement que la requête précise dans quel contexte (chemin de balise) doit apparaître un terme. Nous étudierons plus précisément le problème de la formulation de la requête pour assouplir cette contrainte.

6. Bibliographie

[BAZ 03] BAZIZ M., AUSSENAC-GILLES N., BOUGHANEM M., « Exploitation des Liens Sémantiques pour l'Expansion de Requêtes dans un Système de Recherche d'Information », *actes du XXIème congrès INFORSID 2003*, 2003.

[BOR 00] BORDOGNA G., PASI G., « Flexible Querying of Structured Documents », *Proceedings of the Fourth International Conference on Flexible Query Answering Systems(FQAS)*, 2000.

[CHI 01] CHIARAMELLA Y., « Information retrieval and structured documents », *Lectures on information retrieval*, Springer-Verlag New York, Inc., 2001, p. 286–309.

[FUH 02] FUHR N., GROSSJOHANN K., « XIRQL : an XML Query Language Based on Information Retrieval Concepts », *Proceedings of the 11th International Conference on Information and Knowledge Management*, 2002.

[HAB 97] HABERT B., NAZARENKO A., SALEM A., *Les linguistiques de corpus*, U Linguistique, Armand Colin/Masson, 1997.

[HAB 01] HABERT B., GRABAR N., JACQUEMART P., ZWEIGENBAUM P., « Building a text corpus for representing the variety of medical language », *Corpus linguistics, Lancaster*, 2001.

[HER 03] HERNANDEZ N., GRAU B., « What is this text about ? Combining topic and meta descriptors for text structure presentation », *ACM SIGDOC, San Francisco, USA*, 2003.

[JAC 00] JACQUEMIN C., ZWEIGENBAUM P., « Traitement automatique des langues pour l'accès au contenu des documents », *Le document en sciences du traitement de l'information*, 2000.

[LAL 00] LALMAS M., « Uniform representation of content and structure for structured document retrieval », *20th SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence*, 2000.

[PIW 02] PIWOWARSKI B., DENOYER L., GALLINARI P., « Un modèle pour la recherche d'information sur des documents structurés », *Journées internationales d'Analyse statistique des Données Textuelles*, 2002.

[RAD 89] RADA R., MILI H., BICKNELL E., BLETTNER M., « Development and application of a metric on semantic nets », *IEEE Transaction on Systems, Man, and Cybernetics*, 1989.

[RES 99] RESNIK P., « Semantic Similarity in a Taxonomy : An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language », vol. 11, 1999, p. 95-130.

[SAL 71] SALTON G., *The SMART Retrieval System - experiments in automatic document processing*, U Perntice-Hall, Inc., Englewood Cliffs, NJ, 1971.

[SCH 02] SCHLIEDER T., MEUSS H., « Querying and Ranking XML Documents », *Special Topic Issue of the Journal of the American Society of Information Science on XML and Information Retrieval*, 2002.

[SIT 03] SITEINEX, « Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX) », 2003.

[SZU 02] SZULMAN S., B.BIEBOW, AUSSENAC-GILLES N., « Structuration de terminologies à l'aide d'outils de TAL avec TERMINAE », *Revue Traitement Automatique des Langues*, vol. 43, 2002.

[THE 02] THEOBALD A., WEIKUM G., « The Index-Based XXL Search Engine for Querying XML Data with Relevance Ranking », *Extending Database Technology*, 2002, p. 477-495.

[TOD 01] TODIRASCU A., « Semantic Indexing for Information Retrieval Systems », 2001.

[W3C 01] W3C, « XQuery 1.0 : An XML Query Language », 2001, http ://www.w3.org/TR/xquery/.

[WSM 03] WSM, « Première journée Web sémantique médical », 2003, http ://www.wsm2003.org.

[WU 94] WU Z., PALMER M., « Verb Semantics and Lexical Selection », *Proceedings of the 32nd Annual meeting of the Association for Computation Linguistics*, 1994.

[YOO 01] YOON J., RAGHAVAN V., CHAKILAM V., « BitCube : A Three-Dimensional Bitmap Indexing for XML Documents », *Journal of Intelligent Information Systems*, vol. 17, 2001.

[ZWE 95] ZWEIGENBAUM P., MENELAS C., « Menelas final report », *Deliverable report AIM-MENELAS 17, DIAM-SIM/INSERM U.194*, 1995.