Structuration sémantique des documents XML : Expérimentations et évaluation

Salma Ben Meftah* — Kaïs Khrouf* — Jamel Feki* — Chantal Soulé-Dupuy**

** Laboratoire MIRACL, Université de Sfax, B.P. 1088, 3018 Sfax, Tunisie.*
*** Laboratoire IRIT, Université Toulouse I Capitole, 2 rue du Doyen Gabriel Marty, 31042 Toulouse Cedex 9, France.*

{Jamel.Feki, Salma.BenMefteh}@fsegs.rnu.tn, Khrouf.Kais@isecs.rnu.tn, Chantal.Soule-Dupuy@ut-capitole.fr

RÉSUMÉ. *La norme XML permet la représentation d'un document selon un découpage logique qui ne reflète généralement pas la sémantique de son contenu. Il serait donc intéressant de compléter la structure logique des documents XML par une structure sémantique. L'objet de cet article est alors d'évaluer l'approche d'extraction de structures sémantiques pour les documents XML que nous effectuons sur un échantillon de documents issus de la collection ImageClef 2010 tout en utilisant le thésaurus MeSH (Medical Subject Headings).*

ABSTRACT. *The XML standard represents a logical structuring of documents that generally do not reflect the semantics of the content. It would be interesting to complete the logical structure of an XML documents with a semantic structure. The purpose of this paper is to evaluate the approach of extraction of semantic structures for XML documents; this evaluation is performed on a sample of documents from the ImageClef2010 collection and using the MeSH (Medical Subject Headings) thesaurus.*

MOTS-CLÉS : *Documents XML, Structures sémantiques, Expérimentations.*

KEYWORDS: *XML Documents, Semantic structures, Experiments.*

1. Introduction

Le développement d'Internet a largement augmenté le nombre de documents et les volumes de données disponibles et échangées à travers le Web. Dans ce contexte, XML est devenu le format standard pour les documents. Ainsi, un nombre de plus en plus important de documents devient disponible dans ce format. Les documents XML orienté-texte, constituent des versions électroniques des documents papiers (i.e. les articles scientifiques, les rapports internes) et qu'ils sont riches en données textuelles. Les balises utilisées pour ce genre de documents présentent généralement un découpage structurel, c'est-à-dire logique comme par exemple les balises *Contenu*, *Section* et *Paragraphe*. Cependant, ces balises ignorent complètement la sémantique de leur contenu et, par conséquent, ne facilitent pas les traitements dédiés à la compréhension du document et des concepts qu'il véhicule. Or, ce genre de traitement pourrait être réalisé s'il existait une structure sémantique des documents. Face à cette situation, la proposition d'une approche permettant de déterminer la structure sémantique d'un document XML orienté-texte est une voix vers un traitement plus significatif du contenu. Egalement, le développement d'un outil logiciel automatisant l'approche est un besoin urgent.

Dans notre contexte, nous définissons la structure sémantique d'un document XML comme une structure dérivée de la structure logique du document et permettant de décrire les concepts présents dans son contenu textuel. Pour capturer la sémantique d'un document, nous recourons aux techniques d'indexation sémantique. En effet, si un document se rapporte au logiciel Windows, Unix, Vista et si les exigences de l'utilisateur porte sur « système d'exploitation », un tel problème peut être résolu par l'indexation sémantique qui consiste à affecter un ensemble de concepts à un document. Pour valider notre approche de structuration sémantique, nous utilisons un échantillon de 1000 documents extraits du corpus ImageClef 2010 et la ressource sémantique MeSH (Medical Subject Headings).

Cet article est structuré comme suit. Nous présentons dans la section 2 un état de l'art sur les travaux abordant l'indexation sémantique en se basant sur des ressources spécialisées ou générales. Ensuite, dans la section 3, nous décrivons notre approche de structuration sémantique des documents XML orienté-texte. Dans la section 4, nous présentons le corpus ImageClef 2010 et la ressource sémantique MeSH (contexte médical de nos expérimentations). Enfin, nous détaillons les différents résultats obtenus.

2. Etat de l'art

L'indexation sémantique vise à s'appuyer sur des ressources sémantiques pour représenter les documents. Nous classifions les travaux de la littérature selon le type

de ressources utilisées pouvant être *spécialisées* (Dinh & Tamine, 2010) et (Bevan & al., 2012) ou *générales* (Baziz & al., 2007) et (Boubekeur & Azzoug, 2013).

Dans le cadre de l'utilisation de ressources sémantiques spécialisées, (Dinh & Tamine, 2010) utilisent une méthode d'indexation sémantique adaptée aux dossiers électroniques de patients en utilisant le thésaurus MeSH. Elle est composée de deux principales étapes : *Annotation sémantique* (identification et désambigüisation des concepts) et *Génération de l'index sémantique* (enrichissement des concepts extraits par des termes qui ne correspondent pas à des entrées de MeSH). Cette méthode n'a pas été testée sur un corpus avec différentes configurations des dossiers patients électroniques. (Bevan & al., 2012) présentent une méthode de construction de graphe pondéré pour les concepts en utilisant une ontologie du domaine médical nommée *SNOMED CT*. Cette méthode prend en compte une indication sur le nombre de concepts reliés dans toute l'ontologie et non pas seulement dans le document. De cette façon, l'importance du concept est devenue plus vaste car calculée par rapport au domaine médical (au sens large) mais non par rapport au corpus lui-même (à ses spécificités).

D'autres travaux se sont intéressés à l'indexation sémantique en utilisant une ressource sémantique générale telle que Wordnet. (Baziz & al., 2007) proposent un modèle de représentation sémantique des documents et des requêtes par un réseau sémantique (ensemble de concepts reliés par des liens), en utilisant la ressource terminologique Wordnet. Cependant, ces travaux ont montré que l'indexation n'améliore les résultats que si elle est combinée avec une indexation classique basée sur les mots-clés. (Boubekeur & Azzoug, 2013) proposent également une approche d'indexation à base de concepts en utilisant WordNet et WordNetDomains. Elle extrait les concepts représentatifs des documents et des requêtes et leur assigne des poids sémantiques qui reflètent leur importance respective. Cependant, l'évaluation n'a pas été validée sur une collection de documents reconnue afin de prouver et d'évaluer leur approche.

En complément des travaux abordant l'indexation sémantique des contenus, nous proposons un découpage sémantique des documents XML en se basant sur leurs structures logiques et leurs contenus. Notre méthode de structuration sémantique des documents se base sur l'indexation sémantique et diffère des autres travaux par : *i)* Le choix d'une ontologie appropriée pour chaque document, il s'agit de déterminer la ressource sémantique qui décrit au mieux la sémantique du document et *ii)* La pondération des concepts extraits de manière à donner plus d'importance aux concepts les plus spécifiques car nous partons du constat suivant : plus le niveau auquel se situe le concept est bas dans la hiérarchie, plus l'information qu'il apporte est fine et ciblée

3. Approche de structuration sémantique

Dans cette section, nous présentons notre approche de structuration sémantique (Ben Mefteh et al., 2012). Cette approche commence par déterminer l'ontologie qui convient le mieux pour décrire la sémantique du document parmi un ensemble d'ontologies de domaines. Cette détermination passe par une phase de pondération des ontologies. Il s'agit plus précisément de pondérer les concepts des ontologies utilisées afin de donner plus d'importance aux concepts les plus spécifiques. En effet, si nous avons le choix entre un concept-père et un concept-fils pour un élément d'un document, nous optons pour le concept-fils car il présente une information plus fine et plus ciblée. Cette phase de pondération automatique des ontologies est détaillée dans (Ben Mefteh et al., 2012).

Notre approche de structuration sémantique englobe trois étapes : Choix de l'ontologie, Affectation des concepts aux éléments feuilles, et Propagation des concepts

3.1. Choix de l'ontologie

L'objet de cette phase est de déterminer, parmi un ensemble d'ontologies de domaines, celle qui convient le mieux pour décrire la sémantique du document. Généralement, un document traite un seul domaine principal et peut aborder partiellement, dans certains cas, d'autres domaines. Nous estimons que ces derniers ne sont pas importants ; c'est la raison pour laquelle, nous avons fait le choix d'affecter une seule ontologie à un document.

Ainsi, pour chaque ontologie de domaine disponible, nous calculons dans un premier temps, le poids de chacun de ses concepts C_i par rapport à chaque élément feuille E_j d'un document d. Ce poids est défini par la Formule 1.

$$PC(C_i, E_j) \quad \frac{freq(C_i, E_j)}{freq(C_i, d)} * PC(C_i, O_k) \qquad \forall\ j \qquad E_j \in d \qquad [1]$$

Où :

- $PC(C_i, E_j)$ est le poids du concept C_i par rapport à l'élément feuille E_j,
- $freq(C_i, E_j)$ est la fréquence d'apparition du concept C_i dans l'élément feuille E_j,
- $freq(C_i, d)$ est la fréquence d'apparition du concept C_i dans le document d, et
- $PC(C_i, O_k)$ est le poids du concept C_i dans son ontologie O_k.

Ensuite, nous calculons le poids de chaque concept C_i par rapport à tout le document d. Ce poids est égal à la somme des poids de C_i dans les différents éléments de d, selon la Formule 2.

$$PC(C_i, d) = \sum_{j=1}^{m} PC(C_i, E_j) \qquad \forall\ j \qquad E_j \in d \qquad [2]$$

Où :

- $PC(C_i, d)$ est le poids du concept C_i par rapport au document d,
- $PC(C_i, E_j)$ est le poids du concept C_i par rapport à l'élément E_j,
- m est le nombre d'éléments dans le document d.

Pour choisir l'ontologie la plus appropriée par rapport au document d, nous additionnons les poids des différents concepts appartenant à l'ontologie en question Ainsi, la Formule 3 donne le poids de l'ontologie O_k par rapport au document d. L'ontologie ayant le poids le plus élevé sera retenue pour le document.

$$PO(O_k, d) = \sum_{i=1}^{|O_k|} PC(C_i, d) \qquad [3]$$

Où :

- $PC(C_i, d)$ est le poids du concept C_i par rapport au document d (Formule 2),
- $|O_k|$ est le nombre de concepts de l'ontologie O_k.

3.2. Affectation des concepts aux éléments feuilles

L'objectif de cette phase est d'affecter un seul concept représentatif à chaque élément feuille du document, en se basant sur les poids des concepts calculés par la Formule 1. Pour un élément feuille E_k, différents cas se présentent :

– Cas 1 : Aucun concept déterminé pour E_k (Exemple : les éléments *Auteur*, *Editeur*, *Année*…). Le concept *Null* sera affecté.

– Cas 2 : Un seul concept déterminé pour E_k ; il sera retenu comme concept représentatif.

– Cas 3 : Plusieurs concepts déterminés appartenant à une même hiérarchie. Dans ce cas :

- Si les poids calculés pour ces concepts sont très proches, nous affectons à E_k le concept le plus spécifique dans la hiérarchie.
- Si les poids de ces concepts sont divergents, nous affectons à E_k le concept ayant le poids le plus élevé, indépendamment de sa position dans la hiérarchie.

– Cas 4 : Plusieurs concepts déterminés pour E_k appartenant à plusieurs hiérarchies dans l'ontologie. Dans ce cas, nous affectons à E_k le concept ayant le poids le plus élevé.

A la fin de cette phase, chaque élément feuille est associé à un et un seul concept de l'ontologie retenue. Il s'agit maintenant d'attribuer des concepts aux éléments non feuilles de la structure logique du document, par des règles d'inférence.

3.2. Inférence des concepts aux éléments non feuilles

Il s'agit maintenant de finaliser la structure sémantique en commençant par attribuer des concepts aux éléments non feuilles de la structure et ceci en procédant par inférence des concepts des feuilles vers leurs ascendants en appliquant les règles suivantes :

- Règle 1 : Un élément père ayant un seul fils recevra le même concept que son fils.
- Règle 2 : Si un élément père possède plusieurs fils dont les concepts appartiennent à une même hiérarchie de l'ontologie, alors nous associons à ce père le concept le plus générique commun aux concepts associés à ses fils.
- Règle 3 : Si un élément père possède plusieurs éléments fils dont les concepts appartiennent à plusieurs hiérarchies de l'ontologie, alors le concept attribué à ce père est l'ancêtre commun des concepts associés à ses fils.

A l'issue de cette étape d'inférence, tous les éléments de la structure logique du document sont associés soit à des concepts de l'ontologie choisie, soit à la valeur *Null* (pour des balises « non ontologiques » comme *Auteur*, *Date*…).

4. Expérimentations : Contexte médical

L'objet de cet article est d'évaluer notre approche de structuration sémantique détaillée dans (Ben Mefteh et al., 2012) sur les articles médicaux sous forme XML issus de la base ImageClef 2010 et la ressource sémantique MeSH.

4.1 Ressources terminologiques médicales

La langue des textes biomédicaux, comme tout langage naturel, est complexe et pose des problèmes de synonymie et de polysémie. En conséquence, plusieurs ressources terminologiques sont disponibles pour faire face à l'ambiguïté et à la synonymie lexicale dans la terminologie biomédicale.

Les ressources sémantiques les plus adaptées au domaine médical sont :

- SNOMED-CT : une terminologie clinique qui contient plus de 366170 concepts,
- Galen : contient plus de 25 000 concepts dont le but est de proposer des terminologies réutilisables et partageables pour le domaine médical,

- Ménélas : contient un comptage de 1 800 300 concepts et les relations avec les diverses langues,
- UML (Lindberg et al, 1993) : présente l'union de plus de quarante terminologies,
- MeSH contient plus de 25000 descripteurs et plus de 455000 mots.

Dans nos expérimentations, nous avons utilisé le thésaurus MeSH qui est qualifié par (Charlet, 2002) et (Camous et al., 2006) comme un thésaurus de référence pour l'indexation des articles médicaux.

4.2 Description du thésaurus MeSH

La structure de MeSH est centrée sur les descripteurs, concepts et termes :

- un descripteur *d* se compose d'un ou plusieurs concepts dont chacun a un concept préféré. Le nom du descripteur est le nom du concept préféré. Ces descripteurs appartiennent à 16 domaines : A-Anatomie, B-Organismes, C-Maladie, etc.
- un concept *c* est représenté par un ensemble de termes synonymes. Il se représente par un nœud qui peut appartenir à plusieurs domaines et plusieurs hiérarchies.
- un terme *t* est une chaîne de caractères comprenant un mot ou un groupe de mots. Chaque concept est associé à un terme préféré qui le représente. Les autres termes synonymes sont dits termes non préférés.

Il est à noter que chacun des descripteurs constitue dans nos expérimentations une ontologie à part pour deux raisons : *i)* Un descripteur décrit un contexte particulier et dont chacun est défini par un ensemble de concepts et *ii)* Pouvoir effectuer des tests concernant la première phase de notre approche, à savoir : Choix de l'ontologie par document.

4.3 Corpus ImageCLEF 2010

L'ensemble des données d'ImageCLEF 2010 est fournie par **RSNA** (« ***R**adiological **S**ociety of **N**orth **A**merica* »). Cette base compte 77506 images. Pour chacune de ces images, un document XML est fourni, contenant les balises suivantes :

- <figureID> : l'identifiant de l'image,
- <figureURL> : l'URL de l'image,
- <caption> : la légende de l'image,
- <title> : le titre de l'article à partir duquel l'image a été extraite,
- <articleURL> : l'URL de l'article contenant l'image.

5. Expérimentations : Résultats obtenus

Pour valider nos propos, nous proposons de tester notre approche de structuration sémantique sur un échantillon de la collection de documents ImageClef 2010 et des ontologies de domaine issues du thésaurus MeSH. La base de tests utilisée est décrite dans le tableau ci-dessous :

Description	Nombre
Documents	1000
Eléments feuilles	1700
Eléments non feuilles	1000
Ontologies	128
Concepts	25186

Tableau 1. *Caractéristiques de la base de tests.*

Afin de vérifier et de valider l'apport de la pondération des ontologies, nous avons réalisé : *i)* des tests sans tenir compte des poids des concepts (Algorithme *Sans_Poids*) et *ii)* des tests en tenant compte de la pondération automatique des concepts des ontologies (Algorithme *Avec_Poids*). Dans ce qui suit, nous présentons les différents résultats obtenus.

Tableau 2 présente le nombre d'ontologies assignées aux documents selon les deux algorithmes.

Description	Sans_Poids	Avec_Poids
Nombre de documents ayant reçu une seule ontologie	316 / 1000	879 / 1000
Nombre de documents ayant reçu plus que 2 ontologies	563 / 1000	0 / 1000
Nombre de documents n'ayant reçu aucune ontologie	121 / 1000	121 / 1000

Tableau 2. *Nombre d'ontologies affectées aux documents.*

Nous remarquons, qu'avec l'algorithme *Avec_Poids*, une seule ontologie a été affectée pour 879 documents. Alors que, avec l'algorithme *Sans_Poids*, deux ou plus ontologies ont été attribuées à 563 documents ; ce qui représente une amélioration de 56,3% entre les deux algorithmes.

Dans le tableau 3, nous avons examiné l'association des ontologies aux 100 premiers documents de notre échantillon pour savoir celles qui ont été correctement associées.

Description	*Sans_Poids*	*Avec_Poids*
Ontologies correctement associées	98	83
Ontologies non correctement associées	105	1

Tableau 3. Association des ontologies aux 100 premiers documents.

Nous remarquons que l'algorithme *Sans_Poids* affecte plus d'ontologies correctement associées que l'algorithme *Avec_Poids* ; cela peut s'expliquer par le fait que certains documents abordent deux domaines à la fois (voire plus) et que l'algorithme *Avec_Poids* affecte généralement une ontologie par document. Cependant, l'algorithme *Avec_poids* permet d'écarter d'une manière très nette les ontologies non correctement associées.

Nous nous intéressons à ce stade à l'apport de la pondération des concepts des ontologies par rapport à l'affectation de concepts aux éléments feuilles.

Description	Sans_Poids	Avec_Poids
Nombre d'éléments feuilles attribués à un seul concept	702 / 1700	1128 / 1700
Nombre d'éléments feuilles attribués à plus d'un concept.	998 / 1700	572 / 1700

Tableau 4. *Affectation des concepts aux éléments feuilles.*

Dans la suite, nous décrivons de façon plus détaillée la dernière expérimentation réalisée. Les documents intégrés dans la base de test comprennent 2000 éléments feuilles (300 éléments métadonnées comme *figureID*, *figureURL*, *articleURL* et 1700 éléments représentant des contenus textuels, tels que *Caption* et *Title)*.

- Un seul concept a été affecté à 702 éléments feuilles par l'algorithme *Sans_Poids* et à 1128 éléments feuilles par l'algorithme *Avec_Poids*. Ce qui a apporté une amélioration de 25,05% entre les deux méthodes.
- Deux concepts ou plus ont été affectés à 998 éléments feuilles par l'algorithme *Sans_Poids*, et à 572 éléments feuilles par l'algorithme *Avec_Poi/ds*.
- Le concept *Null* à été affecté à 300 éléments représentant *figureURL*, *figureID et articleURL* (il s'agit de métadonnées).

Ces résultats s'expliquent par le fait que, dans un élément, nous pouvons trouver à la fois un concept et son concept-fils. L'algorithme *Sans_Poids* affecte ces deux concepts à l'élément en question. Par contre, l'algorithme *Avec_Poids* retient le concept fils (plus spécifique et précis) car la pondération automatique des ontologies que nous proposons donne plus d'importance aux concepts fils.

6. Conclusion

Ce travail présente une approche de structuration sémantique des documents XML à partir de leurs structures logiques et de leurs contenus. Cette structuration sémantique passe par trois étapes principales, à savoir : *i)* la détermination de l'ontologie qui sera affectée au document, celle qui décrit sa sémantique (cette étape se base sur une démarche de pondération des ontologies favorisant les concepts

spécifiques), *ii)* l'affectation, à chaque élément feuille de la structure logique du document, du concept significatif de l'ontologie retenue, et *iii)* l'inférence des concepts aux éléments non feuilles du document. Les expérimentations réalisées sur un échantillon de 1000 documents à partir de la collection ImageClef 2010 et en utilisant la ressource sémantique MeSH montrent que la pondération automatique des ontologies que nous proposons a amélioré l'affectation des ontologies aux documents et l'association des concepts aux éléments feuilles.

Plusieurs perspectives à ces travaux sont envisageables. Dans un premier temps, il est important de continuer les expérimentations sur toute la collection ImageClef 2010. Nous comptons également étendre ces travaux par la possibilité d'associer plusieurs structures sémantiques à un même document XML (multi-structuralité sémantique des documents) afin de traduire les points de vue de plusieurs lecteurs.

7. References

Baziz M., Boughanem M., Prade H. « Une approche de représentation de l'information en RI basée sur les sous-arbres », *Conférence en Recherche d'Information et Applications* (CORIA'07), Saint-Etienne, France, 28-30 mars 2007, p. 335-350.

Ben Meftah S., Khrouf k., Ben Kraiem M., Feki J., Soulé-Dupuy C., « Une ne approche pour l'extraction automatique de structures sémantiques de documents XML », *Congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision* (INFORSID'12), Montpellier, France, 29-31 Mai 2012, p. 523-538.

Bevan K., Guido Z., Peter B., Larraine S., Michael L., « Graph-based term weighting for information retrieval », *The Seventeenth Australasian Document Computing Symposium* (ADCS'12), Dunedin, New Zealand, 05-06 December 2012.

Boubekeur F., Azzoug W., « Concept-based indexing in text information retrieval», *International Journal of Computer Science & Information Technology* (IJCSIT), Vol 5, N° 1, 2013.

Camous F., Blott S. et Smeaton A. , « On combining text and mesh searches to improve the retrieval of medline documents», *Conférence en Recherche d'Information et Applications* (CORIA'06), Lyon, France, 15-17 Mars 2006, p. 271-282.

Charlet J., « Développements, résultats et perspectives pour la gestion des connaissances médicales », Mémoire d'habilitation à diriger des recherches, Université Pierre et Marie Curie, Paris, France, 2002.

Dinh D., Tamine L., « Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients », *Conférence en Recherche d'Information et Applications* (CORIA 2010), Sousse, Tunisie, 18-20 Mars 2010, p. 325-336.

Dinh D., « 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*, Université Paul Sabatier, Toulouse, France, 2012.

Lindberg D., Humphrey B., McCray A., « The unified medical language system », *in Methods Inf Med*, 1993.