Une approche de recherche sémantique dans les documents semi-structurés

Rami Harrathi, Sylvie Calabretto

LIRIS UMR 5205 - INSA de Lyon,
7 avenue Jean Capelle,
69621 Villeurbanne cedex, France,
{rami.harrathi, sylvie.calabretto}@insa-lyon.fr

Résumé. Dans cet article, nous proposons une approche de recherche d'information sémantique des documents semi-structurés. L'idée centrale de notre travail est que l'utilisation de ressources sémantiques externes telles que les thesaurus et les ontologies peut améliorer l'efficacité du processus de recherche. Ainsi, nous proposons d'utiliser le modèle vectoriel sémantique où une partie d'un document ainsi que la requête sont représentées par deux vecteurs de concepts. Nous proposons également d'utiliser les mesures de similarité sémantique pour l'évaluation d'une mesure de pertinence. La mesure proposée se base sur la comparaison entre des graphes sémantiques.

Mots clés: recherche d'information, documents (semi-)structurés, XML, ressource sémantique, ontologie.

1 Introduction

La Recherche d'Information (RI) dans les documents semi-structurés (documents XML) consiste à identifier les éléments XML (les nœuds de l'arbre XML) les plus pertinents par rapport à une requête donnée. La majorité des approches proposées dans la littérature sont des adaptations des modèles traditionnels (vectoriel, probabiliste, de langue, etc.).

Ces adaptations visent à tenir compte de la structure et à attribuer des scores de pertinence aux nœuds des documents XML en tenant compte de certaines spécificités des documents XML. Ainsi, les différents types de modèles (vectoriel, probabiliste, de langue, etc.) sont étendus de diverses façons pour tenir compte de la structure. Ainsi, on ajoute des paramètres supplémentaires pour ajuster les formules classiques comme : le nombre d'enfants d'un élément [1], son type, la fréquence de ce type d'élément dans la collection [2, 3], l'importance d'un terme dans les autres éléments du même type [2].

La majorité des approches proposées dans la recherche des documents semi-structurés (documents XML) reposent sur des systèmes d'indexation à base de mots clés ou encore sur les termes. Les seules informations utilisées concernant ces termes sont leurs fréquences d'apparition dans les documents, ou dans les éléments du document (en fonction du niveau de granularité). Ainsi, ces approches ne prennent pas en considération le sens du mot (terme).

L'indexation par des mots clés est généralement imprécise [4]. Cette imprécision est due au fait que les termes d'indexation présentent une forte ambigüité. En effet, le sens d'un mot clé peut varier selon le contexte dans lequel il apparait (phénomène de polysémie). Aussi, ces approches ne prennent pas en compte la synonymie. Par conséquent, dans ces systèmes, il est impossible de trouver des parties des documents représentées par un mot M_1 synonyme d'un mot⟦ M⟧_2, où M_2 représente une requête. Par conséquent, il se peut qu'un système de RI basé sur les mots ne renvoie pas un élément pertinent, c'est-à-dire un élément qui satisfait la requête.

Un moyen pour améliorer les performances des systèmes de RI sur les documents semi-structurés [5] est la prise en compte de la sémantique des termes d'indexation. Ce type d'indexation passe du niveau des mots au niveau des concepts (les sens des mots) pour mieux décrire le contenu du document et de la requête. Ces approches utilisent des ressources sémantiques (thésaurus, ontologies, etc.) dans les phases d'indexation et de recherche.

Dans cet article nous proposons une approche de recherche d'information sémantique des documents semi-structurés. Nous présentons cet article de la manière suivante : dans la section 2 nous décrivons les travaux connexes ; dans la section 3, nous présentons notre approche pour la recherche d'information sémantique des documents semi-structurés. Dans la section 4, nous présentons un

plan d'expérimentation que nous comptons mettre en pratique prochainement, et nous concluons dans la section 5.

2 Travaux connexes

La recherche d'information sémantique dans les documents semi-structurés s'intéresse principalement à la représentation des documents et des requêtes par des taxonomies de concepts. Les systèmes d'indexation et de recherche par les concepts proposés dans la littérature nécessitent de disposer de ressources sémantique afin d'extraire des concepts à partir des textes, et un modèle de mesure de similarité entre concepts [6].

2.1 Recherche d'information sémantique dans les documents semi-structurés

Les documents semi-structurés sont caractérisés par la présence d'une structure organisant leurs contenus textuels. Ainsi, les systèmes d'indexation et de recherche de documents semi-structurés par les concepts se divisent en trois approches correspondant à trois manières de tenir compte de la structure et du contenu textuel lors de l'indexation.

2.1.1 Approches orientées structure

Les approches orientées structure proposent d'indexer uniquement la structure. Le processus d'indexation consiste à représenter les noms des éléments (les noms des balises) par des concepts en utilisant une ontologie de noms. Par exemple les balises comme "university" et "school" ou "car" et "automobile" sont indexées par le même concept. L'intérêt de cette approche est de supporter les requêtes vagues, par exemple si on veut chercher un élément dont le nom est "university", le système de recherche peut retourner les éléments "university" et "school".

Parmi ces approches, on peut citer le système CXLEngine [7]. Ce système utilise une ontologie de noms (ontology label) pour faire correspondre les noms des balises aux concepts dans la hiérarchie de l'ontologie.

Dans [8], les auteurs proposent d'utiliser une ontologie pour gérer des documents de structures hétérogènes. Dans cette approche, la grammaire DTD (Document Type Definition) associée à un document XML est indexée par une DTD de concept (Ontology DTD).

2.1.2 Approches orientées structure et contenu

Les approches orientées structure et contenu consistent à indexer la structure et le contenu textuel par des concepts en utilisant une ontologie. Dans [9], un document XML est considéré comme un ensemble de paires (concept élément, valeur) où "valeur" désigne l'index du contenu textuel qui est représenté par un ensemble de concepts pondérés. Le score de pertinence attribué à un élément est

calculé par une fonction de similarité entre l'ensemble de concepts de la requête et la valeur de l'élément.

Le système de recherche XXL [10], permet l'interrogation de documents XML. Le moteur de recherche XXL présente une architecture s'appuyant sur 3 structures d'index [10] :

– Index des noms des éléments et des attributs : les noms sont indexés par des concepts. Cet index permet l'accès aux nœuds parents, descendants et ancêtres d'un nœud donné. Il permet de calculer la distance entre ces deux nœuds.

– Index du contenu d'élément : permet de retrouver les éléments dans lesquels un terme apparaît. La pertinence des termes est calculée par le TF-IEF (Term Frequency - Inverse Element Frequency).

– Index ontologie : permet de retrouver des mots reliés sémantiquement à un mot donné. Il calcule pour cela une similarité qui peut être restreinte à un certain type de liens. A partir de cette valeur une mesure de similarité peut être calculée entre deux concepts.

Le langage XXL permet d'interroger les documents XML avec une syntaxe proche de la syntaxe SQL. En effet, il est basé sur les langages de requêtes tels que XML-QL et XQuery auxquels il ajoute un opérateur de similarité sémantique noté « ~ ». Cet opérateur permet d'exprimer des conditions de similarité sémantique sur les éléments ainsi que sur leur contenu textuel. L'évaluation de la requête se base sur un calcul de similarité dans une ontologie.

2.1.3 Approches orientées contenu

Dans les approches orientées contenu, on indexe uniquement le contenu textuel. Dans [11], les auteurs proposent d'utiliser les graphes conceptuels pour indexer les nœuds feuilles (porteuses du contenu textuel). Les index des autres nœuds sont obtenus par l'agrégation des index (graphes conceptuels) des nœuds fils en utilisant l'opérateur de jointure maximale entre les graphes conceptuels. Le mécanisme d'interrogation proposé se base sur l'opérateur de projection. L'approche proposée présente des limites. La jointure entre deux graphes conceptuels nécessite la présence d'un concept plus spécifique commun entre les deux graphes à joindre. Par la suite il est impossible de construire les index. Dans ce modèle, le traitement des résultats se fait de manière booléenne. Par conséquent, il est impossible d'attribuer un score de pertinence à un élément. L'indexation par les graphes conceptuels consiste à extraire les concepts et relations entre concepts à partir du texte, ce qui est très difficile. Cette difficulté est essentiellement due à l'absence d'une ressource sémantique riche en termes de relations.

[12] propose une indexation sémantique des documents XML. Ainsi le contenu textuel (nœuds feuilles dans l'arbre XML) est indexé par un ensemble de concepts en utilisant la ressource sémantique WordNet[11] . Cette approche présente une

[11] Wordnet :http://wordnet.princeton.edu/

extension de mesure de similarité entre deux concepts en se basant sur la mesure de Wu-Palmer [15]. La mesure de similarité entre concepts définie précédemment est utilisée pour désambiguïser le sens des termes en favorisant le sens rattaché au concept qui maximise la densité du réseau sémantique. L'originalité de l'approche consiste principalement dans la mesure de similarité utilisée pour enrichir la méthode de pondération des termes. Le score de pertinence d'un élément est calculé en utilisant le modèle vectoriel.

XOntoRank [13] est un système de recherche de documents médicaux. Ce système utilise le thesaurus sémantique SNOMED[12] pour l'extraction et la pondération des concepts à partir du contenu textuel. L'évaluation du score de pertinence d'un élément XML se base sur le principe de propagation de pertinence. Ainsi, le score est calculé à partir des scores des éléments fils en utilisant une fonction d'agrégation.

2.2 Mesures de similarité sémantique

L'objectif des mesures de similarité sémantique est d'évaluer la proximité sémantique entre les concepts (auxquels les termes des requêtes et documents sont rattachés). En recherche d'information, les mesures de similarité jouent un rôle important, en particulier dans le processus de désambiguïsation des concepts, la pondération des concepts et l'évaluation de la pertinence. De nombreuses approches ont été proposées pour évaluer la similarité sémantique entre deux concepts. Ces approches se divisent [14] en trois catégories : les approches basées sur les arcs, les approches basées sur le contenu informationnel et les approches hybrides.

2.2.1 Approches basées sur les arcs

Ce type de mesure s'appuie sur la structure de la ressource sémantique en proposant un comptage plus ou moins élaboré du nombre d'arcs séparant deux concepts. Ces mesures se servent de la structure hiérarchique de l'ontologie pour déterminer la similarité sémantique entre les concepts. Parmi les travaux classifiés sous cette bannière on peut citer :

La mesure de Wu-Palmer [15]. Dans une ontologie, la similarité est définie par rapport à la distance qui sépare deux concepts dans la hiérarchie et également par leur position par rapport à la racine. La similarité entre c_1 et c_2 est :

$$\mathrm{sim}_{\mathrm{WPalmer}}(c_1, c_2) = \frac{2*\mathrm{prof(c)}}{\mathrm{dist}(c_1,c)+\mathrm{dist}(c_2,c)+ 2*\mathrm{prof(c)}}. \quad [\mathbf{1}]$$

[12] Snomed: http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctfiles.html

Où *c* est le concept le plus spécifique qui subsume les deux concepts c_1 et c_2, *prof(c)* est le nombre d'arcs qui sépare c de la racine et $dist(c_i,c)$ le nombre d'arcs qui séparent c_i de *c*.

La mesure de Zargayouna [12]. Cette mesure de similarité est inspirée de celle de [15]. Le lien père-fils est ainsi privilégié par rapport aux autres liens de voisinage en adaptant la mesure de Wu-Palmer. L'adaptation de la mesure est faite au travers de la fonction de calcul du degré de spécialisation d'un concept (spec) qui mesure sa distance par rapport à l'anti-racine.

$$\mathrm{sim}_{\mathrm{Zargayouna}}(c_1, c_2) = \frac{2*\mathrm{prof(c)}}{\mathrm{dist}(c_1,\mathrm{c}) + \mathrm{dist}(c_2,\mathrm{c}) + 2 * \mathrm{prof}(\mathrm{c}) + \mathrm{spec}(c_1,c_2)} \quad [2]$$

$$\mathrm{spec}(c_1, c_2) = \mathrm{prof_b(c)} * \mathrm{dist}(c_1, \mathrm{c}) * \mathrm{dist}(c_2, \mathrm{c})$$

Où $prof_b(c)$ correspond au nombre maximum d'arcs qui séparent le plus petit ancêtre commun du concept « virtuel » représentant l'anti-racine.

La mesure de Resnik [16]. La similarité est définie par rapport à la longueur des chemins qui relient deux concepts dans la hiérarchie. La similarité entre c_1 et c_2 est :

$$\mathrm{sim}_{\mathrm{ResnikEdge}}(c_1, c_2) = 2\mathrm{D} - \mathrm{len}(c_1, c_2) \quad [3]$$

Où *D* est le maximum des longueurs des chemins possibles qui relient c_1 et c_2 et $len(c_1, c_2)$ le plus petit chemin entre c_1 et c_2.

2.2.2 Approches basées sur le contenu informationnel

La notion de contenu informationnel (CI) a été pour la première fois introduite par Resnik [17]. Le contenu informationnel d'un concept traduit la pertinence d'un concept dans le corpus en tenant compte de sa spécificité ou généralité. La fréquence de concepts dans le corpus est calculée pour retrouver le contenu informationnel. Cette fréquence regroupe la fréquence d'apparition du concept lui-même ainsi que des concepts qu'il subsume. La formule est la suivante :

$$\mathrm{CI(c)} = -\log\big(\mathrm{P(c)}\big) \quad [4]$$

Parmi les mesures basées sur le contenu informationnel on peut citer :

La mesure de Resnik [17]. La similarité entre c_1 et c_2 est définie de la façon suivante :

$$\mathrm{sim}_{\mathrm{Resnik}}(c_1, c_2) = \mathrm{CI}(c) \quad [5]$$

Où c est le concept le plus spécifique qui subsume les deux concepts c_1 et c_2.

La mesure de Lin [18]. La similarité entre deux concepts est mesurée par le ratio du contenu d'information nécessaire pour mesurer la "communalité" des deux concepts sur le montant du contenu d'information nécessaire pour décrire chacun des deux concepts. La communalité entre deux concepts dépend du contenu d'information (CI) de leur concept commun le plus spécifique (LCS)

$$\mathrm{sim}_{\mathrm{Lin}}(c_1, c_2) = \frac{2 \times \mathrm{CI}(\mathrm{LCS}(c_1, c_2))}{\mathrm{CI}(c_1) + \mathrm{CI}(c_2)} \quad [6]$$

2.2.3 Approches hybrides

Ces approches sont fondées sur un modèle mixte qui combine des approches basées sur les arcs (distances) en plus du contenu informationnel qui est considéré comme facteur de décision.

La mesure de Jiang [19]. Cette mesure combine le contenu informationnel du concept le plus spécifique (dénoté par c) à ceux des concepts et le nombre d'arcs. Ainsi la similarité est :

$$\mathrm{sim}_{\mathrm{Jiang}}(c_1, c_2) = \frac{1}{\mathrm{CI}(c_1) + \mathrm{CI}(c_2) - 2 \times \mathrm{CI}(c)} \quad [7]$$

3 Une approche de recherche d'information sémantique dans les documents semi-structurés

3.1 Modélisation d'un document semi-structuré

Dans notre approche, nous adoptons le modèle DOM [20] où la structure d'un document est modélisée par un arbre de nœuds. Les nœuds de cet arbre sont typés (éléments, attributs, texte) et sont reliés par des relations de structure (parent-fils, ancêtre–descendant). Les nœuds feuilles représentent le contenu textuel du document, ils sont de type texte. Les autres nœuds sont des nœuds internes, ils sont de type élément.

Dans la suite on dénotera par :

- N_T : un nœud d'arbre de type texte ;
- N_E : un nœud d'arbre de type élément.

La figure ci-dessous donne un exemple de document XML sous une forme arborescente :

Document XML

<doc >
<titre> European Museum Guide
</titre>
<sect1 >
<par1> The Louvre Museum is located in Paris city and Paris is the Capital City of France Country…
< par1>
<par2> African Museum is located in Lyon city... </par2>
</sect1>
<sect2 >
<par3> The Alte Nationalgalerie Museum is located in Berlin City ….
</par3>
</sect2>
</doc>

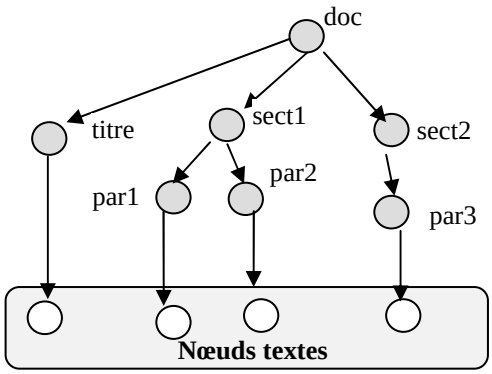

Figure 1. *Exemple de document XML sous forme arborescente.*

3.2 Vers une indexation conceptuelle du contenu textuel

Nous proposons d'utiliser le modèle vectoriel sémantique [21, 22] pour indexer le contenu textuel d'un document semi-structuré par un ensemble des concepts plutôt que par des termes. Les nœuds textes ainsi que la requête sont représentés par des vecteurs dans l'espace d'indexation. Les dimensions de l'espace d'indexation sont l'ensemble des concepts d'une ontologie Ω. Ainsi, dans un espace conceptuel d'indexation $C_\Omega=\{c_1,\ldots c_n\}$ où les c_i sont les concepts d'indexation, un nœud texte N_T^j est représenté par un vecteur de poids des concepts.

$$\overrightarrow{\mathrm{N_T^J}} = (w_{1j}, \ldots, w_{kj}, \ldots w_{nj}) \quad [8]$$

Où w_{kj} est le poids du concept c_k dans le nœud texte N_T^j. De la même façon une requête q est représentée dans l'espace d'indexation C_Ω par un vecteur des poids des concepts qui composent la requête.

$$\vec{q} = (w_1, \ldots, w_k, \ldots w_n) \quad [9]$$

3.2.1 Extraction des concepts

Le processus d'extraction des concepts consiste à détecter les concepts dans un contexte documentaire. Un contexte documentaire est défini comme une unité textuelle à l'intérieur d'un document, il peut représenter une phrase, un paragraphe ou un élément logique de la structure logique (les nœuds texte dans les documents XML).

Afin d'extraire les concepts, on analyse les textes à l'aide d'un analyseur morphosyntaxique [31]. Cet analyseur fournit des mots segmentés, étiquetés syntaxiquement et lemmatisés. L'énumération des termes candidats vise à repérer les séquences des mots susceptibles d'être des labels de concepts dans l'ontologie. Dans cette étape des patrons peuvent être utilisés pour extraire seulement les syntagmes nominaux. Après cette étape seuls les termes qui ont une correspondance dans l'ontologie sont retenus. Dans ce cas, on considère un terme comme étant le label d'un concept. La dernière étape de l'extraction des concepts consiste à identifier ces derniers à partir des termes. Un terme pouvant dénoter plusieurs concepts, cette étape nécessite une désambiguïsation des termes. La figure ci-dessous décrit les étapes du processus d'extraction des concepts.

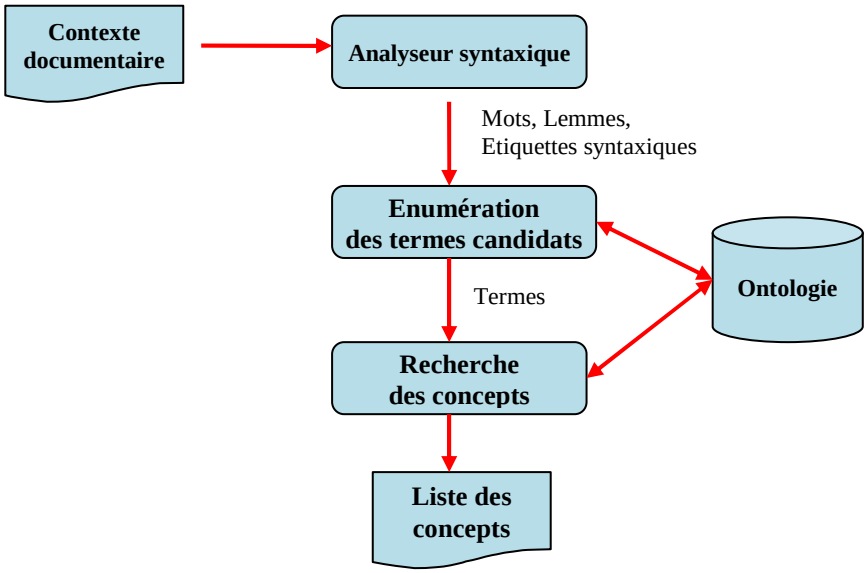

Figure 2. *Processus d'extraction des concepts [31]*.

3.2.2 Désambiguïsation des termes

On définit un contexte d'apparition CA, comme le contexte documentaire dans lequel les termes apparaissent ensemble. Le contexte d'apparition peut être une phrase ou un nœud texte.

$$CA = \{t_1, \dots, t_k, \dots t_n\} \quad [\mathbf{10}]$$

On dénote par $C_\Omega(t_k)$ l'ensemble des concepts de l'ontologie Ω ayant comme label le terme t_k.

$$C_\Omega(t_k) = \{c_k^1, \dots, c_k^i, \dots c_k^m\} \quad [\mathbf{11}]$$

Où c_k^i est le ième concept dénoté par le terme t_k et m est la cardinalité de l'ensemble $m = | C_\Omega(t_k) |$.

La désambiguïsation vise à sélectionner un seul concept parmi l'ensemble des concepts dénotés par un terme. Autrement dit, il faut sélectionner une seule combinaison des concepts (notée CB_{CA}) parmi les combinaisons possibles des concepts du contexte d'apparition *CA* .

$$CB_{CA} = \left\{ c_1^{i_1}, \dots, c_k^{i_k}, \dots c_n^{i_n} \right\},\ avec\ 1 \leq i_k \leq | C_\Omega(t_k) | \quad [\mathbf{12}]$$

Le nombre de combinaisons possibles est : $\prod_1^{k=n} | C_\Omega(t_k) |$.

En prenant exemple sur les travaux de Baziz [23], nous avons utilisé le contexte d'apparition pour la désambiguïsation ainsi que l'hypothèse de Harris [24] selon laquelle les mots qui apparaissent dans des contextes similaires tendent à avoir des sens proches. De cette façon, on sélectionne la combinaison des concepts dans laquelle les concepts sont très proches. La proximité sémantique entre les concepts peut être évaluée par l'utilisation des mesures de similarités sémantiques. La mesure de similarité est généralement une fonction à deux paramètres : les deux concepts considérés. Dans notre travail, nous proposons la définition suivante pour la similarité sémantique : soient une ontologie Ω, C_Ω l'ensemble des concepts de cette ontologie et c_1, c_2 deux concepts de C_Ω. Une fonction sim_Ω est une fonction de similarité définie sur C_Ω de la façon suivante :

$$sim_\Omega : \begin{cases} C_\Omega \times C_\Omega \longrightarrow [0,1] \\ (c_1, c_2) \longrightarrow sim_\Omega(c_1, c_2) \end{cases}$$

$\forall c \in C_\Omega,\ sim_\Omega(c,c) = 1$

sim_Ω est symétrique : $\forall c_1, c_2 \in C_\Omega,\ sim_\Omega(c_1, c_2) = sim_\Omega(c_2, c_1)$

$sim_\Omega(c_1, c_2)=0$ signifie que c_1 n'est pas similaire à c_2

$sim_\Omega(c_1, c_2)=1$ signifie que c_1 est fortement similaire à c_2

Pour sélectionner une combinaison des concepts, on doit calculer la similarité globale entre les concepts de cette combinaison. Soit $CB_{CA}=\{c_1,\ldots, c_k,\ldots c_n\}$ une combinaison de concepts d'un contexte d'apparition CA, la similarité globale est définie comme la moyenne des similarités *MS* entre les concepts.

$$MS(CB_{CA}) = \frac{2.\sum_{i=1}^{i=n}\ \sum_{j=i+1}^{n} sim_\Omega(c_i, c_j)}{n.(n-1)} \quad [\mathbf{13}]$$

La désambiguïsation consiste à sélectionner la combinaison des concepts CB_{CA}^{max} dont la moyenne de similarités entre les concepts est maximale ($max=ArgMax(MS(CB_{CA}^{i}))$, où CB_{CA}^{i} est la ième combinaison de concepts du contexte d'apparition *CA*.

3.2.3 Pondération des concepts

Dans la recherche de documents semi-structurés, le poids d'un terme tend à rendre compte de son importance de manière locale au sein de l'élément et de manière globale au sein de la collection. Le poids d'un terme est évalué selon trois dimensions : la fréquence d'un terme dans le nœud texte (TF); la fréquence inverse de document pour le terme (IDF) et la fréquence inverse de l'élément pour le terme (IEF).

Une étude sur la pondération des termes [25] a montré que la combinaison de TF et IEF donne la meilleure performance. Ainsi, nous adoptons cette mesure pour calculer les pondérations des concepts. Le poids d'un concept est évalué selon deux dimensions:

- CF_i^j : la Fréquence du Concept c_j dans le nœud texte N_T^i
- $IECF_j$: la Fréquence Inverse d'Elément pour le Concept c_j

$$IECF_j = log\left(\frac{|N_T|}{|N_T^{c_j}|}\right) \quad [\mathbf{14}]$$

Où $|N_T|$ est le nombre total de nœuds textes de la collection, et $|N_T^{cj}|$ est le nombre total de nœuds textes contenant le concept c_j.

Le poids d'un concept c_j dans un nœud texte N_T^i (dénoté par W_{ij}) est donné par la formule suivante :

$$W_{ij} = CF_i^j * IECF_j \quad [\mathbf{15}]$$

3.3 Appariement nœud/requête basé sur un graphe sémantique

L'appariement nœud/requête vise à attribuer des scores de pertinence aux éléments d'un document (les nœuds de type texte et les nœuds de type élément dans l'arbre XML).

3.3.1 Score de pertinence d'un nœud de type texte

Dans notre approche, nous utilisons le modèle vectoriel sémantique où dans un espace conceptuel d'indexation, un nœud texte et une requête sont représentés par deux vecteurs de poids des concepts. Généralement, on mesure la proximité entre documents et requêtes grâce au cosinus Salton [26]. Le problème du cosinus est qu'il considère comme indépendantes des dimensions proches [27, 28]. Cependant, les concepts ont des relations sémantiques entre eux. Aussi la mesure de cosinus est incapable de détecter si la représentation sémantique d'une requête est proche de la représentation du nœud texte. Par exemple si on a une requête indexée par un seul concept $\{ c_q \}$ et de la même façon pour le nœud $\{ c_n \}$, il est impossible de retrouver ce nœud dans le cas où c_q est différent de c_n, même si les deux concepts c_q et c_n sont sémantiquement très proches.

Dans notre approche, deux représentations sémantiques sont considérées comme proches si et seulement si les concepts de la requête sont proches des concepts du nœud. Pour illustrer notre approche, on représente un nœud texte ainsi qu'une requête par un graphe (voir Figure 3) pondéré dont les nœuds sont les concepts. Chaque arête de ce graphe est affectée d'un poids représentant la similarité sémantique entre les concepts.

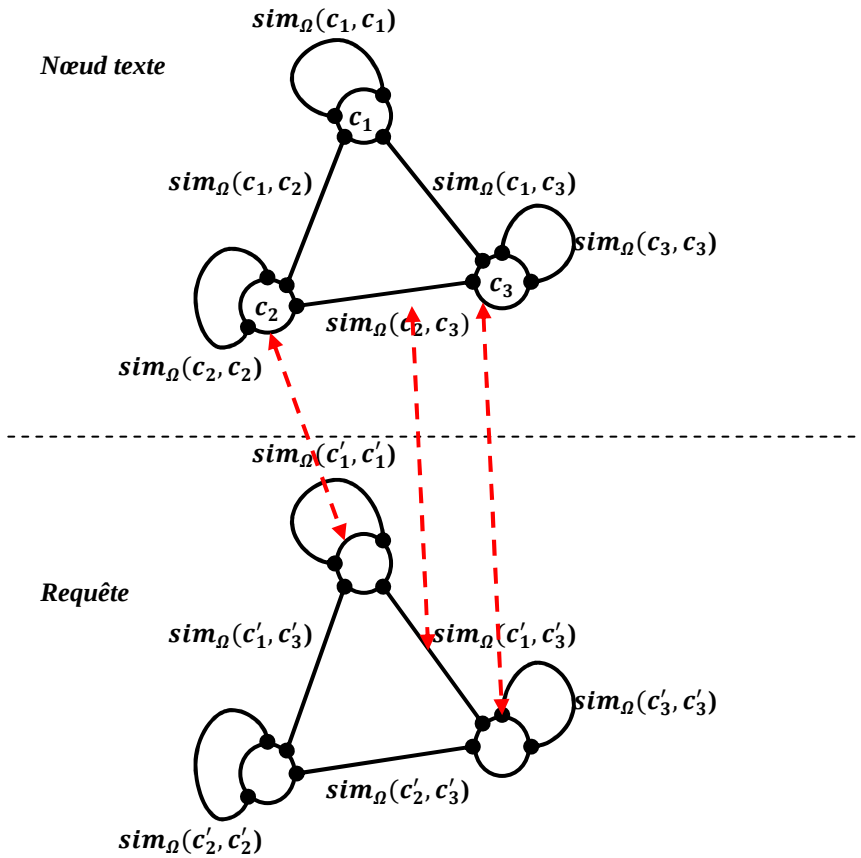

Figure 3. *Graphe sémantique d'une requête et d'un nœud texte***.**

Ainsi, le calcul de score de pertinence d'un nœud texte vis-à-vis d'une requête revient à mesurer à quel point le graphe sémantique du nœud est proche du graphe sémantique de la requête. Afin d'évaluer la formule de calcul du score, on utilise la représentation classique d'un graphe sous forme de matrice. Etant donné un vecteur sémantique d'un nœud texte $N_T^j = (w_{1j}, \dots, w_{kj}, \dots w_{nj})$, la matrice (notée $M_T{}^j$) représentant le graphe sémantique du nœud est une matrice carrée d'ordre n qui est définie de la façon suivante :

$$M_T^j[a,b] = w_{aj} * w_{bj} * sim_\Omega(c_a, c_b), pour\ 1 \leq a, b \leq n \qquad [16]$$

Où w_{aj} et w_{bj} sont les poids respectifs des concepts c_a et c_b dans le nœud texte N_T^j, $sim_\Omega(c_a, c_b)$ est la mesure de similarité sémantique définie sur l'ontologie Ω entre les deux concepts c c_a et c_b. Nous avons tenu compte des poids des concepts dans la matrice. En effet, si un concept admet un poids nul, alors il n'a pas une similarité avec les autres concepts (ce concept n'existe pas dans le graphe car son poids est nul).

De la même façon, le graphe sémantique associé au vecteur requête

$\vec{q} = (w_1, \ldots, w_k, \ldots w_n)$ est représenté par une matrice carrée d'ordre *n* qui est définie de la façon suivante :

$$M_q[a,b] = w_a * w_b * sim_\Omega(c_a, c_b), pour\ 1 \leq a, b \leq n \qquad [\mathbf{17}]$$

L'évaluation de la pertinence entre le graphe sémantique d'un nœud texte et le graphe sémantique d'une requête revient à mesurer la similarité entre les deux matrices représentant respectivement le nœud et la requête. La similarité entre les deux matrices est obtenue en utilisant la mesure de cosinus.

Définition 1. Le cosinus entre deux matrices carrées A et B de même ordre n est [29] :

$$cosinus\ (A,B) = \frac{\langle A\ , B\rangle_F}{\|A\|_F \|B\|_F} \qquad [\mathbf{18}]$$

Où $/\!/A/\!/_F$ et $/\!/B/\!/_F$ sont les normes de Frobenius de *A* et *B*. La norme de Frobenius [29] d'une matrice carrée *A* d'ordre *n* est définie de la façon suivante :

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{i,j}|^2} \qquad [\mathbf{19}]$$

$\langle A, B\rangle_F$ est le produit interne de Frobenius de A et B. Le produit interne de Frobenius [29] entre deux matrices carrées A et B de même ordre n est défini de la façon suivante :

$$\langle A, B\rangle_F = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j} * b_{i,j} \qquad [\mathbf{20}]$$

Le cosinus entre deux matrices est considéré comme une mesure de similarité [29]. Cette mesure est équivalente à la mesure du cosinus entre les vecteurs dans le modèle vectoriel.

Le score de pertinence d'un nœud texte $N_T{}^j$ vis-à-vis une requête q est obtenu en utilisant la mesure de cosinus (Formule 18) entre les deux matrices $M_T{}^j$ et M_q.

$$Score(N_T^j, q) = cosinus(M_T^j, M_T^j) = \frac{\langle M_T^j\ ,\ M_q \rangle_F}{\left\|M_T^j\right\|_F \left\|M_q\right\|_F} \qquad [\mathbf{21}]$$

3.3.2 Score de pertinence d'un nœud de type élément

Dans notre approche on indexe seulement les nœuds de type texte par des vecteurs sémantiques de concepts. Comme les documents semi-structurés possèdent une structure arborescente, les index des nœuds sont imbriqués les uns dans les autres et par conséquent, l'index d'un nœud de type élément contient les index de ses nœuds descendants de type texte [11, 30]. Ainsi, les concepts des nœuds de type texte sont propagés dans l'arbre des documents. La construction des index se base sur deux hypothèses:

– **Hypothèse 1:** les concepts apparaissant près de la racine d'un sous-arbre paraissent plus porteurs d'information pour le nœud associé que ceux situés plus bas dans le sous-arbre. Autrement dit, plus la distance entre un nœud de type texte et son ancêtre est importante, moins il contribue à sa représentation.

– **Hypothèse 2:** les concepts apparaissant plusieurs fois dans les nœuds descendants sont plus porteurs d'information pour le nœud ancêtre. Autrement dit, plus un concept apparait souvent dans tous les nœuds descendants, plus il contribue à sa représentation, même si sa fréquence dans chaque nœud est faible.

Nous modélisons l'hypothèse 1 par l'utilisation dans la fonction de propagation du paramètre $dist(N_E, N_T{}^k)$, qui représente la distance entre le nœud de type élément N_E et de ses nœuds descendants de type texte et $N_T{}^k$ dans l'arbre du document, c'est à dire le nombre d'arcs séparant les 2 nœuds.

Comme nous utilisons le modèle vectoriel sémantique pour la représentation interne des index, le vecteur d'un nœud de type élément est construit à partir des vecteurs de ses nœuds descendants de type texte en utilisant l'opérateur somme entre les vecteurs. Etant donnés un nœud de type élément N_E et un ensemble des_{NE} de ses nœuds descendants de type texte : $des_{NE} = \{ N_T{}^1, \ldots, N_T{}^k, \ldots N_T{}^m \}$, le vecteur sémantique représentant le nœud N_E en tenant compte de l'hypothèse 1 est calculé de la façon suivante :

$$\overrightarrow{\mathrm{N_E}} = \sum_{k=1}^{m} \lambda^{1-\,dist(\,N_E, N_T^k)} \times \overrightarrow{\mathrm{N_T^k}} \qquad [\mathbf{22}]$$

Où $N_T{}^k$ est le vecteur sémantique représentant le k-ième nœud descendant du nœud élément $N_T{}^k$ *et* $\lambda \in]0,1]$ est un paramètre permettant de quantifier l'importance de la distance séparant les nœuds dans la formule de propagation. Ainsi, le poids w_j du concept c_j dans le vecteur N_E est :

$$\mathrm{w_j} = \sum_{k=1}^{m} \lambda^{1-dist(N_E, N_T^k)} \times \mathrm{w_{kj}} \ , pour\ 1 \leq j \leq \mathrm{n} \qquad [\mathbf{23}]$$

Où w_{kj} est le poids du concept c_j dans le vecteur $N_T{}^k$

Nous modélisons l'hypothèse 2 par l'utilisation dans la fonction de propagation du paramètre $|c_E{}^j|$, qui représente le nombre des nœuds texte descendants de N_E contenant le concept c_j. Plus le nombre $|c_E{}^j|$ est grand, plus le concept c_j contribue dans la représentation du nœud N_E. La formule de calcul de poids en tenant compte de l'hypothèse 1 et l'hypothèse 2 est :

$$\mathrm{w_j} = \sum_{k=1}^{m} \left|\mathrm{c_E^j}\right|^{\beta} \times \lambda^{1-dist(N_E, N_T^k)} \times \mathrm{w_{kj}} \ , pour\ 1 \leq j \leq \mathrm{n} \qquad [\mathbf{24}]$$

Où $\beta \in]0,1]$ est un paramètre permettant de quantifier l'importance du nombre des nœuds texte descendants contenant le concept c_j dans la formule de propagation.

Le score de pertinence d'un nœud élément vis-à-vis d'une requête est obtenu facilement en utilisant la formule 20.

$$Score(N_E, q) = cosinus(M_E, M_q) = \frac{\langle M_E\ ,\ M_q \rangle_F}{\|M_E\|_F \left\|M_q\right\|_F} \qquad [\mathbf{25}]$$

Où M_E et M_q sont les matrices représentant les graphes sémantiques associés au vecteur du nœud $\mathrm{N_E}$ et au vecteur du requête q .

4 Projet d'évaluation

A court terme nous proposons de valider notre approche de recherche d'information sémantique dans les documents semi-structurés. Nous utilisons dans nos expérimentations la collection de la campagne d'évaluation INEX[13]. INEX fournit une collection de documents, un ensemble de requêtes et des jugements de pertinence, c'est-à-dire les estimations humaines des éléments pertinents concernant chaque requête. La collection actuelle d'INEX est composée de 2666190 documents en anglais extraits de l'encyclopédie en ligne Wikipedia et représente un volume d'environ 50.7GB.

[13] INitiative for Evaluation of XML retrieval : http://www.inex.otago.ac.nz/

Dans notre proposition le choix d'une ressource sémantique adaptée constitue un point déterminant pour les performances de l'approche. Il est nécessaire de disposer de mesures de similarité s'appliquant sur cette ressource. Il faut que la ressource sémantique soit généraliste. En effet la collection de test fourni par INEX est de domaine général. Ainsi, nous proposons d'utiliser le thesaurus sémantique Wordnet[14] .

La première évolution à court terme consiste dans l'implémentation d'un prototype opérationnel en mesure de nous permettre d'évaluer notre approche sur la collection de documents de la campagne d'évaluation INEX. Actuellement, nous avons développé un module en Java (voir Figure 4) permettant d'extraire les concepts à partir du texte. Ce module fait appel à l'analyseur morphosyntaxique Stanford POS Tagger et au thesaurus Wordnet. Pour la désambiguïsation des termes, des mesures de similarité sémantique sur WordNet sont disponibles dans la librairie Java WordNet::Similarity .

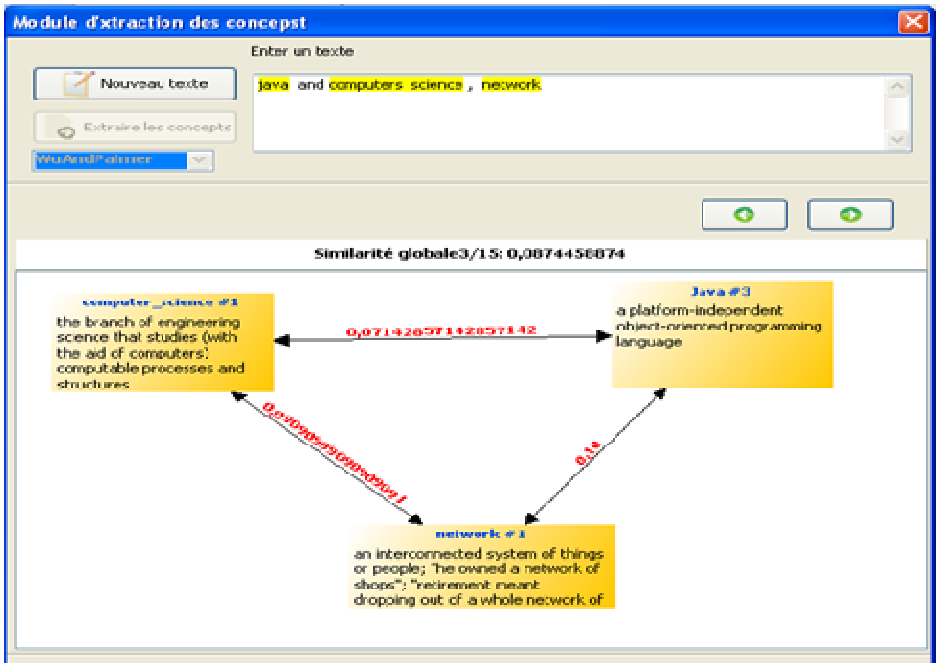

Figure 4. *Module d'extraction des concepts*.

[14] Wordnet :http://wordnet.princeton.edu/

Les résultats obtenus par ce module sont encourageants. Nous signalons que la qualité de la désambiguïsation des termes dépend fortement de la qualité de la mesure de similarité utilisée.

5 Conclusion

Nous avons présenté dans cet article un état de l'art de différentes approches de la recherche d'information sémantique dans les documents semi-structurés. Cela passe notamment par l'emploi de ressources sémantiques externes à la collection de documents, sur lesquelles il est nécessaire de disposer de mesures de similarité sémantique pour pouvoir effectuer des comparaisons entre concepts.

Nous avons proposé une représentation des nœuds de l'arbre d'un document ainsi que la requête par des vecteurs sémantiques de concepts. L'extraction des concepts se base sur un analyseur morphosyntaxique et des mesures de similarité sémantiques pour la désambiguïsation. La pondération des concepts est évaluée selon deux dimensions : la fréquence d'un concept dans un nœud et la fréquence inverse d'élément pour le concept.

La mesure de pertinence d'un nœud proposée se base sur l'évaluation de la similarité entre les graphes sémantiques du nœud et de la requête. La représentation du contenu sous forme de graphe sémantique permet d'évaluer la mesure de pertinence en utilisant la représentation matricielle des graphes. Le modèle vectoriel sémantique nous permet de rendre notre modèle d'indexation flexible du fait que le vecteur d'un nœud est construit facilement à partir des vecteurs de ses nœuds descendants.

6 Bibliographie

1. Fuller M., Mackie E., Sacks-Davis R., and Wilkinson R: Structured answers for a large structured document collection. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pages 204–213, (1993)

2. Grabs T. , Schek H.-J, Zurich ETH: Flexible Information Retrieval from XML with PowerDB-XML.. INEX Workshop 2002: 141-148 (2002)

3. Schlieder T. and Meuss H: Querying and Ranking XML Documents. Journal of American Society for Information Science and Tecnology (JASIST), Special Topic Issue on XML and Information Retrieval, 53(6):489–503, Apr. (2002)

4. Genest D. and Chein, M: A content-search information retrieval process based on conceptual graphs", Knowledge and Information Systems ,Volume 8 , Issue 3(September 2005), pp. 292-309 , Springer-Verlag, (2005).

5. Rosso, P., Ferretti, E., Jimenez, D., Vidal, V.: Text categorization and information retrieval using wordnet senses. Proceedings of the 2nd Global Wordnet Conference (GWC 2004), Czech Republic 299-304 (2004)

6. Bellia, Z., Vincent, N., Kirchner, S., Stamon, G.: Assignation automatique de solutions à des classes de plaintes liées aux ambiances intérieures polluées. 8èmes journées d'Extraction et de Gestion des Connaissances (EGC 2008), Sophia-Antipolis (2008)

7. Taha K. and Elmasri R: CXLEngine: A Comprehensive XML Loosely Structured Search Engine, In Proceedings of the EDBT workshop, Nantes, France 2008. ACM International Conference Proceeding Series, New York, USA; Vol. 261, ISBN:978-1-59593-966-1, pp 37-42.(2008)

8. Kim M. S. and. Kong Y.-H: Ontology-DTD Matching Algorithm for Efficient XML Query, in FSKD (2), ser. Lecture Notes in Computer Science, L. Wang and Y. Jin, Eds., vol. 3614. Springer, 2005, pp.1093-1102, (2005)

9. Weikum G., Theobald M. and Schenkel R.: Exploiting structure, annotation and ontological knowledge for automatic classification of xml data, In WebDB, San Diego, CA. 2003.

10. Schenkel R., Theobald A., and Weikum G, : Semantic similarity search on semistructured data with the XXL search engine, Information Retrieval, 8(4): pp. 521–545, (2005)

11. Chiaramella Y.: Information Retrieval and Structured Documents. In Proceedings of the Third European Summer-School on Lectures on Information Retrieval (ESSIR 2000)-Revised Lectures, pp. 286-309, (2000).

12 Zargayouna, H., Salotti, S.: Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. Actes de la conférence IC'2004 (2004)

13. Farfán, F. , Hristidis V., Ranganathan A., Weiner M.,: XOntoRank: Ontology-Aware Search of Electronic Medical Records. In Proc. ICDE 2009: pp. 820-831,(2009)

14. Slimani T. Yaghlane B. B., and Mellouli K.: A new similarity measure based on edge counting. In Proceedings of world academy of science, engineering and technology, (2006).

15. Wu Z. and Palmer M.: Verb semantics and lexical selection. In 32nd. Annual Meeting of the Association for Computational Linguistics, pp. 133 –138, (1994).

16. Resnik P.: Using information content to evaluate semantic sirnilarity. In: Proc. 14th Int. Joint Conf. Artificial Intelligence, Montreal, pp. 448-453, (1995).

17. Resnik P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, Journal of Artificial Intelligence Research, pages 95–130,(1999).

18. Lin D.: An information-theoretic definition of similarity. In Proc. 15th International Conf. on Learning, Morgan Kaufmann, San Francisco, CA, pp. 296–304, (1998).

19. Jiang J. J and Conrath D. W.:Semantic similarity based on corpus statistics and lexical taxonomy. In International Conference Research on Computational Linguistics (ROCLING X) (1997).

20. Apparao,V., Byrne, S., Champion,M., Isaacs, S., Jacobs, I., Le Hors,A., Nicol,G., Robie,J., Sutor,R., Wilson,C., Wood,L. :Document Object Model (DOM). W3C recommendation, Technical Report REC-DOM-Level-1-19981001, (1998)

21. Woods W. Conceptual Indexing : a better way to organize knowledge. Technical Report SMLI TR-97-61 : SUN Micosystems, Lab. Mountain View Canada, (1997)

22. Berry, M. W., Z. Drmac, et E. R. Jessup : Matrices, vector spaces, and information retrieval. SIAM Rev. 41(2), 335–362(1999).

23. Baziz M, : Indexation Conceptuelle Guidée par Ontologie pour la Recherche d'Information .Thèse., Institut de Recherche en Informatique de Toulouse , Toulouse,2005.

24. Harris, Z., Gottfried, M., Ryckman, T., Mattick, P., Daladier, A., Harris, T.N., Harris, S.: The form of Information in Science: Analysis of an immunology sublanguage. Dordrecht : Kluwer Adademic Publishers (1989)

25. Sauvagnat k, Boughanem.M., Propositions pour la pondération des termes et l'évaluation de la pertinence des éléments en recherche d'information structurée. Dans : Actes de CORIA 2006, Lyon, 15-17 mars (2006).

26. Salton, G.; Wong, A. and Yang, C. S. A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620. (1975).

27. Ventresque A, Cerqueus T, Celton L, Hervouet G, Levin D., Lamarre P, Cazalens S: Mysins : make your semantic INformation system. EGC 2010: 629-630 (2010)

28. Ventresque A: Espaces vectoriels sémantiques : enrichissement et interprétation de requêtes dans un système d'information distribué et hétérogène. Thèse,Université de Nantes (2008)

29. Bing . S.: Sense Matrix Model and Discrete Cosine Transform., In Proceedings of AIRS 2004 (the first Asia Information Retrieval Symposium). Oct 18-20, Beijing, CHINA; LNCS AIRS Proceedings, Springer Verlag, (2004).

30. Abolhassani M. and Fuhr N :Applying the divergence from randomness approach for content-only search in XML documents. In Proceedings of ECIR 2004, Sunderland, pages 409-419, (2004).

31 Maisonnasse L: Les supports de vocabulaires pour les systèmes de recherche d'information orientés précision : application aux graphes pour la recherche d'information médicale. Thèse,Université Joseph Fourier – Grenoble I (2008)