METHODE AUTOMATIQUE POUR CORRIGER LA VARIATION LINGUISTIQUE LORS DE L'INTERROGATION DE DOCUMENTS XML DE STRUCTURES HETEROGENES

Ourdia Bouidghaghen(*),Mohand Boughanem(**)
yugo_doudou@yahoo.fr, bougha@irit.fr

(*)Université M'hamed Bougara-Boumerdes, 26 rue le la gare 06200, Algérie
(**)Université Paul Sabatier, IRIT-SIG-RI, 118 Route de Narbonne, 31062 Toulouse Cedex 9, France.

Mots clés :
XML, structure hétérogène, variation linguistique, recherche d'information, ontologie.
Key words:
XML, heterogeneous structures, linguistic variation, Information Retrieval ontology
Palabras clave :
XML, las estructuras heterogéneas, la variación lingüística, la Recuperación de información, la ontología

Résumé

Dans cet article, nous abordons la problématique de l'interrogation de corpus de documents XML de structures hétérogènes. En effet, vu la liberté qu'offre XML lors de la conception des DTD (Document Type Definition), des documents se rapportant au même domaine comportent des balises pouvant être différentes d'une structure à l'autre. Cette hétérogénéité peut se décliner aux niveaux de différence morphologique entre des balises sémantiquement identiques et/ou de différence dans leur agencement dans les sources disparates des documents. De plus, les langages actuels d'interrogation des documents XML, bien qu'ils soient très puissants pour la recherche du contenu des documents en se basant sur la structure, ils ne reflètent que cette dernière et ne permettent pas ainsi des requêtes sémantiques. Nous nous sommes intéressés en particulier, au problème de la variation linguistique entre les noms des balises. Nous proposons donc une méthode permettant de remédier à ce problème en procédant au regroupement automatique des balises sémantiquement proches dans la même classe. Ce regroupement est effectué en se basant sur les relations sémantiques fournies par une ressource linguistique. Ainsi, lors de la phase d'interrogation, chaque balise de la requête est alors étendue par ajout de synonymes de la classe associée.

1 Introduction

Aujourd'hui, XML est largement accepté comme format standard pour le partage et l'échange de données dans divers domaines comme les BDs, le Web, les intranets, …XML doit principalement son succès à sa flexibilité : toute personne peut écrire une DTD (Document Type Definition) pour définir la structure de ses documents sous le format XML. Une structure qui représente les informations dans la forme que la personne désire. Cependant, cette liberté de conception de DTD conduit de fait à l'élaboration de structures décrivant souvent les mêmes éléments mais avec des noms de balises différents et/ou agencés différemment. Ceci cause de réels problèmes au niveau du stockage, de l'intégration et de l'interrogation de données dans ces larges collections de documents hétérogènes.

La problématique engendrée par ce type de document dans le domaine de la recherche d'information, est liée à la nature de leur contenu. En effet, comme ces documents comportent de l'information (du texte) et des contraintes structurelles (des balises), ils ne peuvent pas être efficacement exploités par les techniques classiques de RI, qui considèrent le document comme un granule d'information indivisible.

Or, dans un document XML toute partie du document peut être considérée comme réponse potentielle à la requête de l'utilisateur. La partie concernée peut être spécifiée directement dans la requête de l'utilisateur ou calculée automatiquement par le système de RIS. Les requêtes dans les systèmes de RIS peuvent en effet avoir deux formes : une forme « contenu seulement », la requête est dans ce cas composée que de mots clés et une forme combinant la structure et le contenu.

Dans le cas où les documents ont des structures hétérogènes l'écriture d'une requête de type contenu et structure devient très difficile, car d'une part, l'utilisateur ne connaît pas forcément toutes les structures des documents et d'autre part il n'est pas possible d'exprimer la notion de synonymie structurelle dans aucun langage existant aujourd'hui.

Comme il n'existe aucun standard universel pour la représentation des données arbitraires sous XML, l'hétérogénéité des structures des documents est inévitable. Or, cette hétérogénéité de structures peut être seulement liée à des informations sémantiquement similaires mais codées dans des structures XML très variées :

- Variation linguistique, c'est-à-dire utilisation de différents noms de balises pour désigner un même concept dans les diverses sources d'information.
- Variation de la structuration (ou hiérarchisation) des balises, c'est-à-dire différence de leur agencement et leur nombre dans les diverses sources d'information.

La problématique considérée ici, est donc comment surpasser ces différences ? Par quel procédé est-il possible de réconcilier d'une façon automatique ces documents, pour permettre une interrogation simplifiée et une recherche efficace aboutissant à des résultats couvrant tous les documents.

C'est dans la perspective de s'affranchir de ces structures hétérogènes que se situent nos travaux **[1]**. Notre objectif est de construire un moyen permettant de manipuler les structures « similaires » de manière transparente. Nous nous sommes intéressés en particulier à résoudre la problématique de la variation linguistique. Pour cela, nous proposons une méthode permettant le matching de balises morphologiquement différentes mais qui désignent le même concept. L'idée est d'exploiter la sémantique portée par les balises XML et les relations pouvant exister entre ces balises pour faire correspondre les conditions de structure exprimées dans les requêtes avec tous les éléments présents dans la collection. La solution que nous proposons se base sur la construction et l'utilisation lors de l'interrogation d'un dictionnaire regroupant les balises sémantiquement équivalentes en utilisant une ressource linguistique, ici en l'occurrence WordNet **[2].** De façon générale, comme l'illustre la figure 1, l'approche que nous avons mise au point comprend trois étapes principales, qui sont comme suit :

(1) : La première étape correspond à l'extraction des concepts candidats pour chaque balise d'une DTD par une projection sur l'ontologie.
(2) : Dans cette étape, il s'agit d'un traitement de désambiguïsation qui permettra, sur la base d'un calcul de la proximité sémantique entre concepts, de choisir les concepts adéquats aux sens des balises telles qu'elles sont utilisées dans les documents de la collection.
(3) : La dernière étape, consiste en la construction d'un dictionnaire des synonymes. Cela se fait en définissant une entrée pour chaque concept retenu dans l'étape précédente et où sera sauvegardée de plus, la liste des balises qui lui correspond dans la collection.

Le reste de ce papier présente en détail notre approche. La section 2 présente d'abord le modèle de représentation des documents sur lequel se base notre approche pour déterminer rapidement les relations ancêtres-descendants. La section 3 présente la phase de projection sur l'ontologie. La section 4 présente la phase de désambiguïsation. La section 5 présente la dernière phase qui consiste en la construction du dictionnaire des balises synonymes. La section 6 présente un exemple montrant le bien fondé de l'approche.

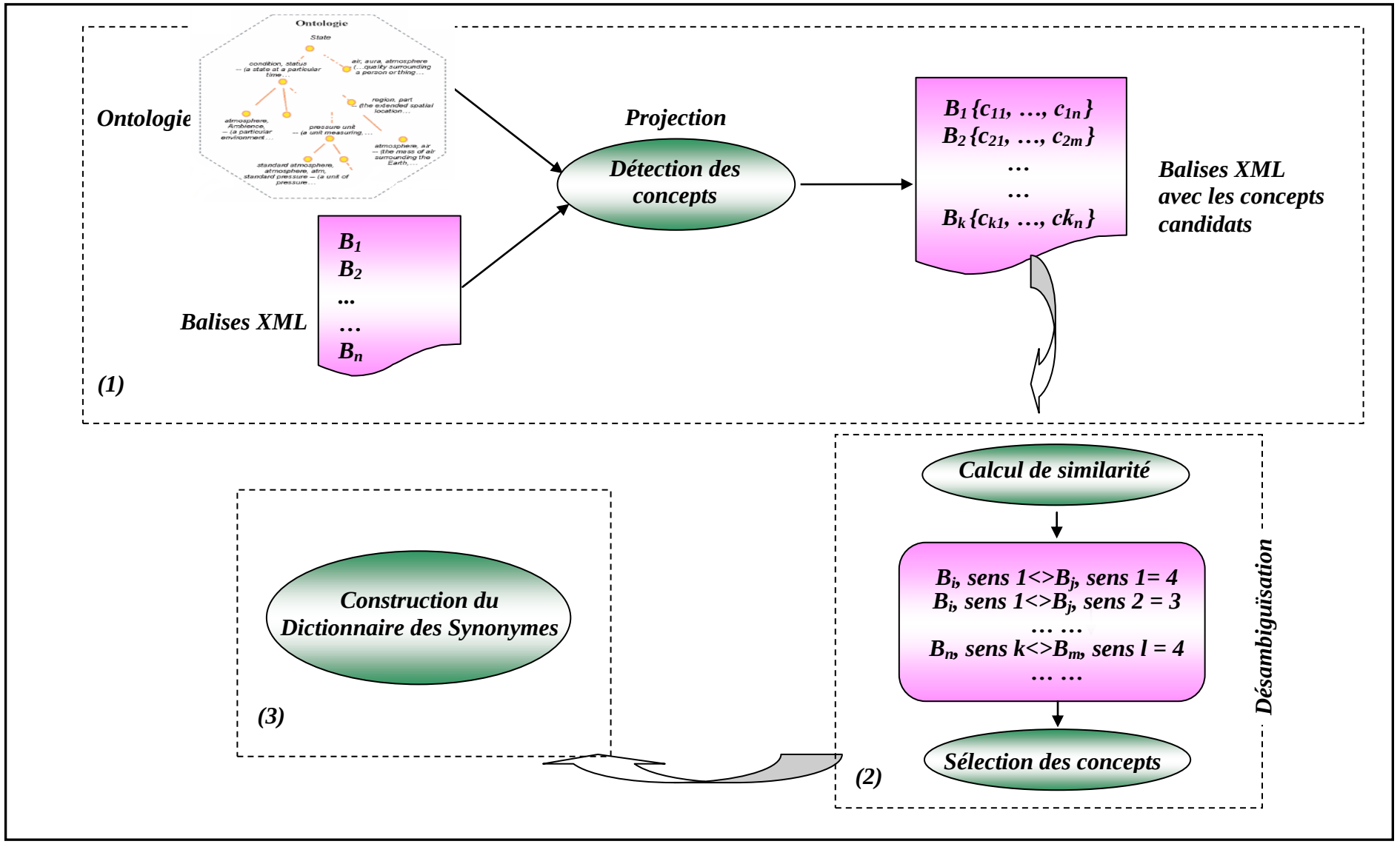

Figure 1 : Schéma général de l'approche.

2 Modèle de représentation de la structure des documents

Les documents XML possèdent des structures arborescentes décrites par des DTD. On exploitera cette structure pour représenter chaque DTD de la collection sous forme d'arbre.

Une DTD sera donc représentée par un arbre (*ds)*, défini par les ensembles *N, A* et *L* : *ds = (N, A, L).*
Avec $N = \{n_1, n_2, ...\}$ l'ensemble des nœuds éléments, $A = \{a_1, a_2, ...\}$ l'ensemble des attributs et L est un ensemble d'arcs orientés. Un arc orienté est une paire *(u, v)* formée de deux éléments des ensembles *N* ou *A* tels que :

- *u* est parent de *v*
- chaque $n_i \in N$ appartient au moins une fois à *L* en tant que premier composant d'une paire formant un arc.
- chaque $n_i \in N$, $a_i \in A$ excepté le nœud racine appartient une et une seule fois à *L* en tant que second composant d'une paire formant un arc.

Les nœuds sont ainsi reliés entre eux par des arcs qui forment les relations parent/enfant. Tous les noeuds excepté le noeud racine ont exactement un nœud parent. De cette façon, il sera facile de déterminer rapidement les relations de hiérarchie entre les balises de la même DTD.
Dans la suite, on désignera par une balise un nom d'un élément (qu'il soit composé ou simple) ou d'un attribut XML. Pour chaque balise on disposera des informations sur sa balise mère et éventuellement ses balises filles et/ou attributs.

3 Projection sur l'ontologie

Toutes les balises XML identifiées dans la collection sont projetées sur l'ontologie pour obtenir les concepts auxquels elles sont associées. Les nominations des balises sont généralement sous forme de noms ou d'abréviation de noms. Pour les balises abrégées, il faudra d'abord interroger un dictionnaire des abréviations pour avoir les noms appropriés.
Cependant, comme chaque nom de balise peut avoir plusieurs sens, et ainsi correspondre à plusieurs synsets (ou concepts) de l'ontologie, des mesures de similarité entre les différents sens des noms, sont calculées en vue de sélectionner, pour chaque balise, le meilleur sens correspondant dans l'ontologie.

La mesure de similarité entre deux noeuds représente une valeur condensée résultant de la comparaison de deux sens possibles pour deux termes (donc deux concepts candidats) en utilisant la distance entre les positions des deux concepts candidats dans l'ontologie ou encore les relations sémantiques de l'ontologie. Cette valeur n'a pas de sens précis mais exprime le degré du lien entre les deux concepts candidats. Nous l'explicitons dans la section suivante.

4 Le traitement de désambiguïsation

Se déroule en deux phases :

1) Calcul de la similarité entre concepts

L'ontologie n'offre pas une quantification des liens sémantiques entre les différents concepts qu'elle définit. Pour cela, diverses mesures permettant de calculer la valeur de la proximité sémantique entre concepts sont proposées dans la littérature.
On peut distinguer :

1. Les mesures se basant sur le chemin (Path based measures) entre deux concepts à comparer, telles que définies par exemple dans **[3] [4] [5].**
2. Les mesures se basant sur la notion de contenu d'information (Information Content IC), telles que définies dans **[6]**.
3. Les mesures se basant sur une combinaison du chemin et du contenu d'information **[7]** ou sur l'algorithme de Lesk adapté à WordNet dans **[8].**

Nous décrirons la mesure de Lesk adaptée à WordNet dans **[8]**. Elle représente le nombre de mots communs entre deux concepts. Formellement elle est décrite comme suit : étant donné un ensemble de relations R = {R_1, R_2, …., R_n} et deux mots b_i et b_j auxquels sont affectés deux sens $S_\alpha^{\,i}$ et $S_\beta^{\,j}$. La similarité sémantique entre $S_\alpha^{\,i}$ et $S_\beta^{\,j}$, notée : Sim ($S_\alpha^{\,i}$, $S_\beta^{\,j}$) est définit comme suit :

$$Sim(S_\alpha^i, S_\beta^j) = \sum_{l,k \in \{1,\ldots,n\}} \left\| R_k(S_\alpha^i) \cap R_l(S_\beta^j) \right\|$$

Les relations dépendent de l'ontologie utilisée. Dans le cas de WordNet, on trouve par exemple : les relations de synonymie, d'hypéronymie, d'hyponymie, de méronymie et d'holonymie plus les relations de glossaire et les relations de domaines, …

L'utilisation de ce nombre relativement élevé de relations a pour but de couvrir au maximum les différents types de liens que deux concepts peuvent partager.

De plus, il faudra identifier le contexte de chaque balise qui servira pour le choix du sens qui lui soit le plus approprié. Une première idée, serait de considérer le cotexte formé de toutes les balises de la DTD à laquelle elle appartient, donc on pourra penser à évaluer la proximité sémantique avec chacun des concepts relatifs à ces balises. Une autre façon est de considérer le contexte local d'une balise en se restreignant à l'ensemble formé de sa balise mère et éventuellement la liste de ses balises filles et attributs. Dans ce cas, il suffira de calculer la proximité sémantique avec les concepts relatifs à ce seul ensemble.

2) Sélection des concepts

A cette étape, nous connaissons pour chaque balise son sens représenté par l'ensemble des concepts associés (les synsets de WordNet), noté : $S_i = \{S_1^i, S_2^i, ..., S_n^i\}$, ainsi que les valeurs de sa proximité sémantique calculées avec les balises du même contexte. Il reste uniquement à choisir pour chaque balise le meilleur concept parmi tous les sens candidats extraits de l'ontologie. Le principe de la désambiguïsation consiste à supposer que, parmi les différents concepts candidats (sens) pour une balise donnée, le plus adéquat (vraisemblable) est celui qui a le plus de liens avec les autres concepts du même contexte qu'elle. En généralisant cette règle à toutes les balises, on se retrouve avec des balises qui se désambiguïsent mutuellement et de manière globale par rapport au contexte de chaque DTD.

Pour formaliser cette idée, on affecte à chaque concept candidat (ou sens d'une balise) un score (*C_score*). Le score d'un concept candidat est égal à la somme des valeurs de similarité qu'il a obtenu avec les autres concepts candidats des balises de son contexte sauf ceux qui sont dans le même ensemble de sens que le sien : pour une balise b_i, le score de son $k^{ième}$ sens est alors calculé comme suit :

$$C\_score(S_k^i) = \sum_{\substack{j \in [1..m],\ j \neq i \\ l \in [1..n]}} Sim(S_k^i, S_l^j) \ (*)$$

Où *m* représente le nombre des balises formant le contexte d'une balise et *n* le nombre de sens qui est propre à chaque balise b_i. Le meilleur concept (synset) S_i retenu est celui qui représente au mieux le sens de la balise b_i. C'est celui qui maximise *C_score* :

$$S_i = Best\_score(b_i) = \underset{k=1..n}{ArgMax} \ \left\| C\_score(S_k^i) \right\|$$

Le concept ainsi sélectionné, représentera une entrée dans le dictionnaire des synonymes qui sera construit dans l'étape suivante.

5 Construction du dictionnaire des concepts

La dernière étape de l'approche concerne la construction du dictionnaire des concepts. Pour chaque concept sélectionné à l'issue de la phase de désambiguïsation, on lui crée une entrée dans le dictionnaire des balises synonymes. On lui associe de plus, la liste des balises le référençant dans la collection. Pour chaque balise, on gardera une référence vers son concept.

Ainsi, pour chaque requête d'un utilisateur contenant des conditions de structure formulées dans les termes d'une quelconque DTD de la collection, il sera possible de chercher pour chaque balise qui y figure, le concept correspondant dans le dictionnaire et d'identifier la liste des balises synonymes pour étendre la requête aux autres documents de la collection qui suivent d'autres DTD et les inclure dans la recherche.

6 Un exemple

Nous illustrons les étapes précédentes en les appliquant sur deux simples DTD (représentées dans la figure ci-dessous sous forme d'arbres) :

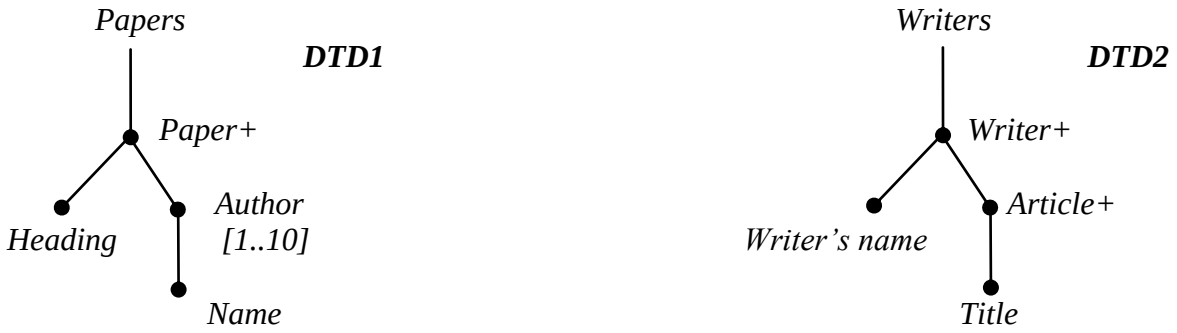

Figure 2 : Deux DTD différentes décrivant le même domaine.

En premier lieu, toutes les balises contenues dans les documents (représentés ici par les DTD), sont extraites et projetées sur l'ontologie WordNet pour avoir leurs sens. Vu la polysémie des sens, les noms de balises se voient attribuer plusieurs synsets, par exemple la balise "name" possède 6 sens, la balise "paper" 7 sens, etc. (comme le montre la figure 3 suivante).

The noun name has 6 senses (first 6 from tagged texts)
1. (698) **name** -- (a language unit by which a person or thing is known; "his name really is George Washington"; "those are two names for the same thing")
2. (44) **name** -- (by the sanction or authority of; "halt in the name of the law")
3. (26) **name** -- (a person's reputation; "he wanted to protect his good name")
4. (15) **name**, figure, public figure -- (a well-known or notable person; "they studied all the great names in the history of France"; "she is an important figure in modern music")
5. (6) **name**, gens -- (family based on male descent; "he had no sons and there was no one to carry on his name")
6. (2) **name**, epithet -- (a defamatory or abusive word or phrase)

The noun paper has 7 senses (first 6 from tagged texts)
1. (31) **paper** -- (a material made of cellulose pulp derived mainly from wood or rags or certain grasses)
2. (21) composition, **paper**, report, theme -- (an essay (especially one written as an assignment); "he got an A on his composition")
3. (12) newspaper, **paper** -- (a daily or weekly publication on folded sheets; contains news and articles and advertisements; "he read his newspaper at breakfast")
4. (5) **paper** -- (a scholarly article describing the results of observations or stating hypotheses; "he has written many scientific papers")
5. (4) **paper** -- (medium for written communication; "the notion of an office running without paper is absurd")
6. (2) newspaper, **paper**, newspaper publisher -- (a business firm that publishes newspapers; "Murdoch owns many newspapers")
7. newspaper, **paper** -- (the physical object that is the product of a newspaper publisher; "when it began to rain he covered his head with a newspaper")

Figure 3 : Les différents sens que peut avoir un mot comme "name" ou "paper".

En second lieu, les similarités sémantiques sont calculées entre tous les concepts candidats, c'est-à-dire : les sens possibles des balises identifiées précédemment, en utilisant les différentes mesures de similarité sémantique. Ici, et en guise d'illustration, nous avons utilisé la mesure de Lesk décrite dans (la section 4) avec les relations de synonymie et de glossaire.

Par exemple, (comme on le voit dans la figure 4), la première ligne de la deuxième colonne veut dire que la similarité entre le sens1 du nom "heading" et le sens4 du nom "paper" est égale à 4.

author#n#1 <> name#n#1=3	paper#n#4<>heading#n#1=4
author#n#1 <> name#n#3=1	paper#n#4<>heading#n#3=0
author#n#2 <> name#n#1=1	paper#n#5<>heading#n#1=1
author#n#2 <> name#n#3=1	paper#n#5<>heading#n#3=0
… …	… …
title#n#1 <> article#n#1=3	writer #n#1<> article #n#1=4
title#n#1<> article#n#4=0	writer #n#1<> article #n#4=0
title#n#6 <> article#n#1=0	writer #n#2<> article #n#1=1
title#n#6 <> article#n#4=0	writer #n#2<> article #n#4=0
… …	… …

Figure 4 : la similarité calculée entre les concepts.

Puis, vient l'étape de sélection du concept qui représente au mieux le sens d'une balise. Pour chaque concept candidat, son sens ayant le plus grand score cumulé calculé avec la formule (*), est retenu comme le concept approprié. Les résultats pour notre exemple sont illustrés ci-dessous :

name#n#1=16	paper#n#4= 25
article#n#1=22	heading#n#1=17
title#n#1= 20	writer #n#1=30
author#n#1=32	writer's name #n#1=19

Figure 5 : Le meilleur score cumulé des concepts retenus.

Par exemple pour la balise "article" possédant 4 sens, le sens1 qui est sélectionné, correspond effectivement au sens approprié dans le contexte de la DTD à laquelle elle appartient. De même pour toutes les autres balises.

Enfin, on crée pour chaque concept retenu une entrée dans le dictionnaire des balises où on gardera une liste de ses balises synonymes. Par exemple la balise "author" est identifiée comme synonyme de la balise "writer", elles correspondent au même concept de l'ontologie, à savoir "writer#n#1", elles seront insérées dans la liste de ses synonymes. On fera de même pour toutes les autres balises, on obtiendra à la fin de cette opération l'ensemble des classes suivantes :

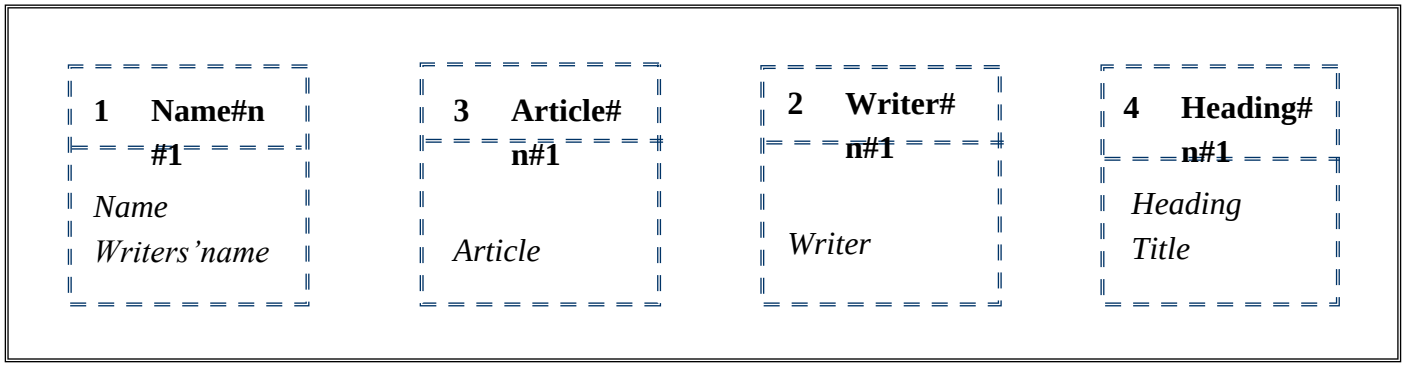

Figure 6 : Ensemble des concepts insérés dans le dictionnaire des synonymes.

Ainsi, un utilisateur pourra utiliser l'un ou l'autre vocabulaire des DTD pour formuler ses requêtes, il suffira de garder un pointeur pour chaque balise vers son concept dans le dictionnaire, et le système se chargera d'aller chercher ses éventuels synonymes dans le dictionnaire, pour lancer une recherche dans les termes des autres DTD.

7 Bibliographie

[1] Ourdia Bouidghaghen, *Prise en compte de l'hétérogénéité structurelle en recherche d'information semi-structurée*, mémoire de magister de l'université de M'HAMED BOUGARA, Boumerdes, Avril 2007.
[2] A.G. Miller, WordNet, *A lexical Database for English*, ACM 38 (11), 39-41, 1995.
[3] Rada, R., Mili, H., Bicknell, E., and Blettner, M. "*Development and application of a metric on semantic nets*". IEEE Transaction on Systems, Man, and Cybernetics, 19(1):17–30.
[4] Leacock, C., Miller, G. A., and Chodorow, M. "*Using corpus statistics and WordNet relations for sense identification*". Comput. Linguist.24, 1(Mar.98), 147-165.
[5] Jiang J.and Conrath D. "*Semantic similarity based on corpus statistics and lexical taxonomy*".In Proceedings on International Conference on Research in Computational Linguistics, Taiwan, 1997.
[6] 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 (JAIR), 11, pp. 95-130, 1999.
[7] D. Lin. "*An information-theoretic definition of similarity*". In Proceedings of 15th International Conference on Machine Learning, 1998.
[8] S. Patwardhan,S. Banerjee, and T. Pedersen : *Using measures of semantic relatedness for word sense disambiguation*. In Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics CICLING, Mexico City, 2003.