Vers un Système d'Annotation Automatique de Documents Historiques basé sur les Techniques des Modèles Markoviens

Ines Ben Messaoud, Haikal El Abed

Vers un Système d'Annotation Automatique de Documents Historiques basé sur les Techniques des Modèles Markoviens

Ines Ben Messaoud* — **Haikal El Abed****

** Ecole Nationale d'Ingénieurs de Tunis (ENIT), Laboratoire des Systèmes et Traitement du Signal (LSTS)*

ibmnoussa@gmail.com

*** Technische Universität Braunschweig, Institut for Communications Technology (IfN) Schleinitzstr. 22, 38106 Braunschweig, Germany*

elabed@tu-bs.de

RÉSUMÉ. *Dans notre travail nous avons réalisé un système qui représente les documents du même domaine par un schéma XML. Un algorithme de Mapping est appliqué entre le schéma de référence et un nouveau schéma spécifique pour identifier les correspondances entre les deux schémas. La modélisation des schémas XML a été réalisée en utilisant les modèles de Markov. La pertinence de Mapping est calculée selon les probabilités retournées par les modèles de Markov. Les tests ont étés effectuées sur des schémas XML représentant* 5 *domaines et variant de 60 à 100 schémas par domaine. Pour le premier modèle développé le taux de Mapping varie d'une manière croissante par rapport aux nombres communs de noeuds entre les deux schémas source et cible. Pour le deuxième modèle le taux de Mapping varie en fonction du nombre de noeuds en communs entre les deux schémas d'une manière aléatoire entre* 0.05 *et* 0.4.

ABSTRACT. *Our work first extracts XML schema which describes a specific domain. Mapping algorithm has as entries the two schemes reference schema and a specific schema. XML schemes are generated using Markov models, this model is used to calculate the Mapping efficiency. In the first model the Mapping increased according to the common number of nodes between the entries XML schemes. Mapping is pertinent when the common nodes number is over* 0.5% *of Markov model states. In the second model the Mapping changes randomly according to the in common number of nodes between* 0.05 *and* 0.4.

MOTS-CLÉS : *schéma XML, Mapping, modèles de Markov, annotation automatique*

KEYWORDS: *XML schema, Mapping, Markov Model, automatic annotation*

1. Introduction

Le domaine de recherche d'informations ainsi que l'analyse des documents a gagné une grande importance ces dernières années. La représentation des documents a évolué de la représentation classique qui est la représentation brute du document vers la représentation structurée. La représentation structurée des documents est généralement décrite en utilisant le standard XML (Extensible Markup Language) soient par la présentation des schémas XML ou les DTD (Document Type Definition). La diversité des représentations du même document dans différentes bases nécessite d'intégrer dans le processus d'annotation la tâche de Mapping qui est un algorithme appliqué entre deux schémas XML représentant les documents du même domaine.

Le problème réside au fait d'adopter un système d'annotation automatique de documents historiques qui facilitera la recherche d'information en éliminant la redondance dans la recherche et au même temps qui permettra de repérer facilement l'information recherchée.

Dans ce travail on va présenter un système de validation de Mapping entre deux schémas XML : le schéma de référence de la base d'apprentissage d'un domaine choisi et d'un nouveau schéma de la base de test. La modélisation et la validation du Mapping ont été réalisées en utilisant un système stochastique. Le but de cet article est de montrer l'application de ce système a la représentation des documents historiques.

Plusieurs travaux peuvent être mentionnés dans le domaine de l'annotation des documents textuels ou numériques. Denoyer (Denoyer 2004) a détaillé les méthodes vectorielles de représentation telles que : par vecteur binaire (on associe à chaque élément 1 s'il existe dans le document, 0 sinon), par vecteur fréquentiel (chaque élément est représenté par son nombre d'apparition dans le document), par vecteur TF-IDF qui se repose sur la loi de Zipf (chaque élément est représenté par le produit entre un facteur qui concerne le poids du terme dans le document et un autre qui concerne le poids du terme dans le corpus). La limite de ces méthodes réside dans le fait qu'elles n'informent pas sur l'ordre d'apparition des mots dans le documents.

Wisniewski et al. (Wisniewski *et al.* 2006) ont représenté un système d'extraction de structures des documents semi-structurés basé sur les techniques statistiques. Dans ce travail, ils se sont limités à la présentation des documents Web en modélisant les schémas XML avec le réseau Bayésien. les expériences ont été testées sur trois corpus différents : Corpus INEX (Fuhr *et al.* 2002), Base de données IMDb [1] et le les documents de Shakespeare. Lecerf et al.(Lecerf *et al.* 2007) ont développé un outil d'apprentissage actif pour l'annotation sémantique (ALDAI) basé sur le principe de maximum d'entropie. L'apprentissage a été testé sur un nombre limité de documents : la collection publique UCI (Newman *et al.* 1998). Antonacopoulos et al. (Antonacopoulos *et al.* 2005) représentent la structure des réalités terrains des documents numérisés en images en utilisant les schémas XML, les régions ont été présentées par des polygones pour qu'elles soient flexibles et peuvent s'adapter à différentes types de

1. http ://www.imdb.com

régions (image, texte, tableau,...). Ce travail a été basé sur une base réaliste proposée par les auteurs.

Plusieurs travaux ont été marqués dans le domaine d'analyse et reconnaissance des documents et qui se sont positionnés dans l'annotation des documents sous forme structurée soient physique ou logique.

Smith (Smith 2009) s'est intéressé à la détection de la structure physique des pages de magazines qui ont une organisation complexe (multi-colonnes, document mixé,...). Smith a utilisé la technique des arrêtes des tableaux pour la détection des blocks appelée "Tab-Stop". Cette technique peut être utilisée dans la segmentation haut-bas ("Top-down"), l'algorithme de segmentation détecte au début les régions qui forment le document (image, texte, régions mixées) puis chaque région détectée subira à son tour d'autres segmentations (lignes, mots, caractères).

Antonacopoulos et al. (Antonacopoulos *et al.* 2005) se sont intéressés à la structuration logique des documents. Dans leur travaille ils ont défini un schéma XML qui représente la structure d'une base de documents organisée par les auteurs et qui contient des pages de magazines, des affiches de publicités et des articles techniques. La diversité des représentations structurées des documents nous a fait penser à utiliser le standard XML dans la représentation des documents et spécifiquement aux documents historiques. Un Mapping entre deux schémas différents doit être appliqué. La validité du Mapping est calculée par le système représenté par la suite qui retourne le taux de validité du Mapping. Si le taux retourné est faible le Mapping doit être reformulé pour avoir un taux plus important.

Le modèle de validation de Mapping a été décrit en utilisant les modèles de Markov hiérarchique qui est un formalisme puissant de modélisation des dépendances entre les éléments hiérarchique du schéma XML.

L'article est organisé comme suit : dans la Section 2 on différencie les différentes structurations de document. En section 3 on présente la technique du Mapping choisie. Dans la section 4 on décrit le modèle de Markov réalisé pour la modélisation des schémas XML. Le système réalisé ainsi que les tests sont énumérés dans la Section 5. Finalement en Section 6 nous discutons les résultats trouvés ainsi que l'extension du système réalisé dans d'autres domaines.

2. Système d'Annotation de Documents

2.1. *Description d'un Document*

Contrairement à la représentation classique d'un document qui ne tient compte que du contenu textuel et l'information est présentée d'une manière plate sans aucun traitement, la présentation structurelle s'intéresse à la présentation du contenu structuré, et à chaque information est associée une étiquette, soient logique ou physique.

La structure physique décrit l'organisation du document en des (Lignes, Paragraphes, Images, ...) ainsi que le format de l'information (Centre,Gauche,Droite, ...) c'est la description de l'apparence du document à l'utilisateur. Par contre la structure logique présente la signification de l'information (Titre, Nom_auteur,Date, ...).

2.2. *Présentation de l'Arbre XML*

Un arbre XML présente les balises du document XML sous formes de noeuds et la hiérarchie sous formes d'arcs entre ces noeuds. Chaque noeud de l'arbre XML correspond à une entité structurelle du document (Section, Titre, Date, Région). Chaque arc du graphe représente une relation entre deux noeuds qui décrit la structure logique du document. Ces relations peuvent être soient une relation hiérarchique (un paragraphe est contenu dans une section) ou une relation entre les éléments de même niveau dans l'arbre (une section en précède une autre).

La représentation structurée des documents différencie entre deux présentations d'informations : (1) **Une information de label** : qui est issue de l'ensemble des labels des noeuds du document. (2) **Une information d'organisation** : présentée par les arcs du graphe et permet de relier les différentes unités structurelles entre elles.

Chaque schéma XML peut être modélisé par un arbre XML, L'arbre XML est décrit par le couple $T = (N, A)$ tels que : à chaque balise du document XML on associe un noeud $no \in N$ et A est l'ensemble des relations entre les noeuds no de l'arbre XML. L'arbre XML est interprété de haut à partir de la racine vers le bas direction des feuilles et de gauche à droite.

3. Technique de Mapping

Le processus de Mapping est un algorithme qui prend comme paramètres d'entrés : le schéma cible et le schéma destination. Le Mapping est un processus qui suit le processus de Matching. Ce dernier détermine les valeurs de similarités sémantiques entre les éléments et les attributs des schémas. Les instructions de l'algorithme de Mapping sont des formules qui consistent à déterminer les relations sémantiques entre les éléments des deux schémas d'entrées. Les Mappings sont des expressions décrivant le moyen dont les instances du schéma cible sont dérivées à partir des instances du schéma source.

Les relations définies dans l'algorithme de Mapping peuvent être simples ou directs dans le cas où on correspond à chaque noeud du schéma destination un noeud du schéma source ainsi les relations sont de type $(1:1)$ ou complexes (indirects) si les relations sont des relations de fusion de type $(z:1)$ $[\text{Date} \leftrightarrow \text{Jour}, \text{Mois}, \text{Année}]$ ou de division de type $(1:z)$, ou des relations de type $(z:y)$, $[\text{Nom}, \text{Prénom}, \text{CIN}, \text{Adresse} \leftrightarrow \text{Identifiant}, \text{Coordonnées}]$, cela signifie qu'à z éléments du schéma source correspondent y éléments du schéma cible.

Les modèles de Mapping ont été divisés en deux catégories : Approches basées sur les valeurs de correspondances (Rifaieh 2004, Miller *et al.* 2001) et approches basées sur les opérateurs de transformations (Boukottaya *et al.* 2004, Zerdazi *et al.* 2005).

3.1. *Algorithmes de Mapping*

Ils existent plusieurs algorithmes de Mapping, comme c'était mentionné auparavant nous avons utilisé l'algorithme proposé par (Reynaud *et al.* 2006) tout en entraînant des modifications.

3.1.1. *Ontologie*

L'ontologie est un concept utilisé dans le Web sémantique, chaque ontologie O est représentée par un couple O=(C, R) où C=$\{c_1, c_2, \ldots, c_a\}$, avec $a = |C|$, est l'ensemble de classes et R=$\{r_1, r_2, \ldots r_b\}$, avec $b = |R|$, est l'ensemble des relations entre les classes.

3.1.2. *Description de l'Algorithme*

L'algorithme de Mapping prend comme paramètres d'entrées deux arbres XML T_1 l'arbre destination et T_2 l'arbre source. L'objectif est ici de trouver les correspondances des noeuds de l'arbre T_1 et T_2. Deux procédures sont appliquées dans cet algorithme, la deuxième n'est appliquée que si la première a échoué. Soient T_1=(N_1, A_1) et T_2=(N_2, A_2) et O une ontologie qui représente les relations des Mappings effectués

1) La première procédure : Pour un noeud $c_l \in N_2$, avec $l \in [1, \ldots, |N_2|]$, s'il existe $c_f \in N_1$ un élément de même nom que c_l, avec $l \in [1, \ldots, |N_2|]$, un Mapping peut être établi entre c_l et c_f.

2) La deuxième procédure : s'applique quand la première a échoué. Elle consiste à utiliser la liste des Mappings préexistants entre O et les schémas XML des sources déjà établis. Si pour un schéma XML d'arbre T_2, un Mapping existe pour le concept c_l avec un élément $elem$ présent également dans T_1. Le domaine D_2 en commun formé par les deux noeuds c_l de T_1 et $elem$ de T_2 est défini tel que $D2 = Domaine((c_l, T_1) \cap (elem, T_2))$. Le Mapping entre c_l et $elem$ n'est pas immédiatement établi, une comparaison avec le domaine en commun entre c_l et l'ancêtre direct de $elem$ est établi. Si le domaine D_2 s'élargi d'une étape à une autre le processus est répété jusqu'à que le domaine reste constant ou il diminue, sinon le processus est arrêté et le Mapping est établi entre c_l et l'élément $elem$ ou un de ces ancêtres.

4. Modèlisation des Schémas XML en Utilisant les Modèles de Markov

4.1. *Modèle de Markov Hiérarchique*

Les H-MMs (Hierarchical Markov Models) ont été développés pour le traitement multi échelles de données séquentielles. Ils constituent une généralisation récursive

des modèles de Markov dans le sens où une séquence d'états génère une séquence de H-MM et ainsi récursivement. Ils se sont utilisés dans plusieurs domaines tels que : reconnaissance de l'écriture, de la parole et du signal. De manière formelle, un H-MM est défini par :

1) Un ensemble d'états $S = \{s_1, s_2, \ldots, s_{ne}\}$, où s_i représente le $i$ème état et D la profondeur du H-MM. Les états de profondeur d sont des états qui émettent les observations. Pour chaque état père p_m^d, on note $|p_m^d|$ le nombre de ses enfants.

2) Un ensemble de probabilités de transitions : $A^{p_m^d} = \left\{a_{i,j}^{p_m^d}\right\}$ tel que $a_{i,j}^{p_m^d} = P(s_j|s_i)$ pour $i, j \in [0 \ldots |p_m^d|]$, avec $a_{ij}^{p_m^d}$ est la probabilité de transition entre les états s_i et s_j.

3) Un ensemble de probabilités initiales : $\pi^{p_m^d} = \left\{\pi^{p_m^d}(s_i)\right\}$ avec $\pi^{p_m^d}(s_i) = P(s_i = q_1)$, avec q_1 signifie le premier état des noeuds issus du même père, représente la probabilité que s_i soit le premier état qui apparaît dans la séquence du noeud p_m^d.

4) Nous avons ajouté au H-MM le vecteur de probabilités initiales $\Pi = \{\Pi_{s_i}\}$, telle que Π_{s_i} détermine la probabilité que l'état s_i soit la racine du document.

5) Un ensemble de probabilités d'émission $B^{p_m^d} = \left\{b^{p_m^d}(v_k)\right\}$, pour V l'ensemble du vocabulaire et $k \in [1 \ldots |V|]$, telle que $b_i^{p_m^d} = P(v_k|s_i)$ est la probabilité d'observer le symbole v_k du vocabulaire V dans l'état s_i.

Un modèle de Markov hiérarchique discret (HD-MM) est défini par $\lambda = (A^{p_m^d}, \pi^{p_m^d}, \Pi)$, par contre un modèle de Markov hiérarchique caché (HH-MM) est défini par $\lambda = (A^{p_m^d}, \pi^{p_m^d}, \Pi, B^{p_m^d})$.

4.2. *Modèles d'Arbres de Schémas XML*

Dans notre travail, nous nous sommes limités à l'étude de deux modèles de structures d'arbre XML tels que dans le modèle *parent - fils* chaque noeud no n'est relié qu'à son père (chaque noeud ne peut avoir qu'un seul père) et inversement chaque parent n'est relié qu'à ses fils (un noeud parent peu être relié à plusieurs fils). Le deuxième modèle *parent - frère*, où chaque noeud no est relié à son père et à ses frères (noeud appartenant au même niveau dans la hiérarchie de l'arbre et possédant le même noeud père).

Nous définissons les relations suivantes entre les noeuds de l'arbre qui représentent un document structuré :

– Parent(no) : est la fonction qui renvoie l'unique noeud parent de no.

– Precedent(no) : est la fonction qui renvoie le noeud qui précède le noeud no.

– nbenfant(no) : est la fonction qui retourne le nombre d'enfants de no.

– feuille(no) : vérifie si le noeud no est une feuille.

– Racine(no) : vérifie si le noeud no est la racine du document.

4.3. *Identification des Primitives*

Les primitives sont les noeuds du schéma de référence du domaine désigné qui sont identifiées en fonction de leurs cardinalités. On commence par déterminer tous les éléments distincts présents dans tous les schémas XML du domaine étudié ainsi que leurs cardinalités. Les primitives qui décriront le domaine étudié seront seulement celles dont la cardinalité est supérieure à un seuil local s_c déterminé pour chaque domaine.Soient V_e le vecteur des éléments présents dans les schémas XML du même domaine, $n = |V_e|$, V_c est le vecteur de cardinalités associées à chaque noeud de V_e et nb est le nombre des schémas XML qui appartiennent aux même domaine de notre base. On aura comme résultats dans cette étape deux vecteurs décrivant la hiérarchie de l'arbre XML du schéma de référence : S est le vecteur des primitives , tels que $n_e = |S|$, et P est le vecteur des noeuds pères. Le seuil s_c est déterminé en fonction de n_e, la condition que le nombre de primitives doit être supérieur ou égale à la moyenne des éléments distincts qui forment V_e doit être satisfaite, le seuil s_c est initialisé à la moyenne du nombre de schémas XML et se décrémente tant que cette condition est fausse.

4.4. *Définition du Modèle des Schémas XML avec HD-MM*

Nous allons adopter ici le modèle de HD-MM définit comme $\lambda = (A^{p_m^d}, \pi^{p_m^d}, \Pi)$ et qui décrit les probabilités de dépendance entre les noeuds d'un schéma XML, comme il été mentionné dans la section 4 on trouvera pour chaque noeud père une matrice $A^{p_m^d}$ et un vecteur $\pi^{p_m^d}$.

– Les états $S = \{s_1, s_2 \dots s_{ne}\}$ tels que $s_i \in S$,$\forall i \;\; 1 \leq i \leq ne$.

– $P = \left\{p_1^d, p_2^d, \dots, p_l^d\right\}$ est l'ensemble des noeuds parents du schéma XML tels que $\forall p_m^d \in S$ et $d \in [1 \dots D-1]$.

– Un ensemble de probabilités $A^{p_m^d} = \left\{a_{i,j}^{p_m^d}\right\}, \forall p_m^d \in P$ telle que chaque matrice $A^{p_m^d}$ est définie comme suit :

$$a_{i,j}^{p_m^d} = \begin{cases} P(s_i = \text{Precedent}(s_j|p_m^d)), \forall i \;\; 1 \leq i,j \leq \text{nbenfant}(p_m^d) \\ P(p_m^d = \text{Parent}(s_j)), \forall j \;\; 1 \leq j \leq \text{nbenfant}(p_m^d) \;\; \text{et} \;\; i = 0 \end{cases} \quad [1]$$

– Un ensemble de probabilités initiales $\pi^{p_m^d} = \left\{\pi_i^{p_m^d}\right\}, \forall p_m^d \in P.\ \pi_i^{p_m^d} = P(q_1 = s_i|p_m^d), \forall i \;\; 1 \leq i \leq \text{nbenfant}(p_m^d)$.

– le vecteur Π décrit la répartition de la probabilité que le noeud s_i soit la racine du document $s_i \in \{s_1, \dots, s_{ne}\}$. $\Pi_{s_i} = P(Racine(s_i))$

La Figure 1 décrit un exemple de schéma XML représenté par le modèle décrit ci dessus.

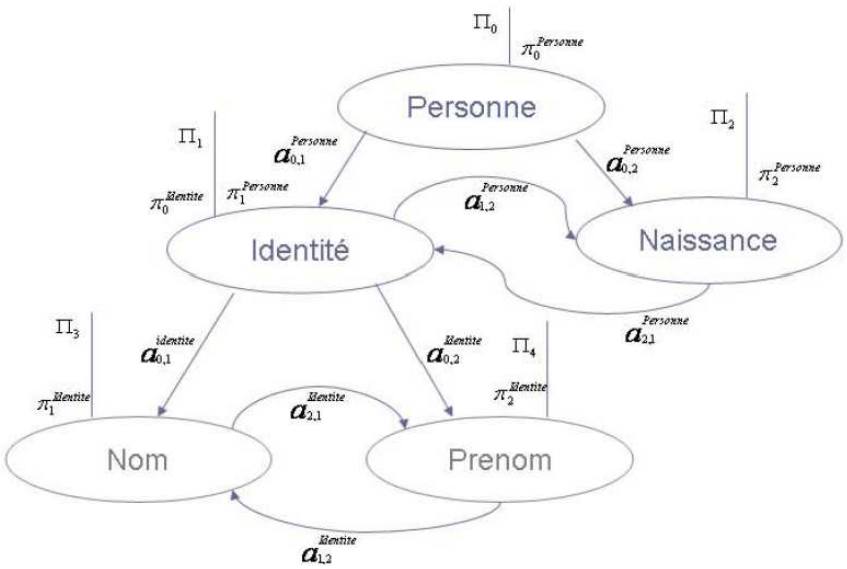


Figure 1. *Schéma "Personne .xml" représenté par le modèle de Markov discret hiérarchique HD-MM*

4.5. *Probabilités d'Observations d'une Séquence dans le H-MM*

C'est la probabilité d'observation d'un arbre XML dans une base de schémas XML d'un même domaine en se basant sur le modèle présenté dans la section 4.4.

– Soit la séquence $E = \{e_1, e_2, \ldots e_G\}$, tel que E présente l'ensemble de noeuds d'un nouveau schéma n'appartenant pas à la base de documents (schéma spécifique), $G = |E|$. En modèle de Markov cette séquence représente la séquence à observer en utilisant le modèle de Markov présenté dans la section 4.4.

– Soit $Ps = \left\{ps_1^d, ps_2^d, \ldots ps_K^d\right\}$ est l'ensemble des noeuds pères dans la séquence E et $K = |Ps|$.

– Soit la séquence $Q = \left\{q_1, q_i^{ps_j^d}, \ldots, q_G^{ps_j^d}\right\}$, $\forall i \in [1, \ldots, G]$ et $j \in [1, \ldots, K]$, tel que $Q = E$, Q représente la hiérarchie du schéma de référence, avec q_1 représente la racine du schéma et $q_i^{ps_j^d}$ est le $i$ème noeud fils du noeud ps_j^d,$\forall i \in \left[1, \ldots, \text{nbenfant}(ps_j^d)\right]$.

Modèle de structure "Parent-fils"

$$\begin{aligned} P(E|\lambda) &= P(q_1 = racine) \prod_{j=1}^{K} \prod_{i=1}^{\text{nbenfant}(ps_j^d)} P(q_i^{ps_j^d} | ps_j^d) \\ &= \Pi(q_1) \prod_{j=1}^{K} \prod_{i=1}^{\text{nbenfant}(ps_j^d)} a_{0,q_i^{ps_j^d}}^{ps_j^d} \end{aligned} \quad [2]$$

Modèle de structure "Parent-frère"

$$\begin{aligned} & P(E|\lambda) \\ &= P(q_1 = racine) \prod_{j=1}^{K} \left[P(q_1^{ps_j^d} | ps_j^d) \pi^{ps_j^d}(q_1^{ps_j^d}) \prod_{i=2}^{\text{nbenfant}(ps_j^d)} [P_i] \right] \\ &= P(racine) \prod_{j=1}^{K} \left[P(q_1^{ps_j^d} | ps_j^d) \pi^{ps_j^d}(q_1^{ps_j^d}) \prod_{i=2}^{\text{nbenfant}(ps_j^d)} \left[P(q_i^{ps_j^d} | ps_j^d) P_i \right] \right] \\ &= \Pi(q_1) \prod_{j=1}^{K} \left[a_{0,q_1^{ps_j^d}}^{ps_j^d} \pi^{ps_j^d}(q_1^{ps_j^d}) \prod_{i=2}^{\text{nbenfant}(ps_j^d)} \left[a_{0,q_i^{ps_j^d}}^{ps_j^d} a_{q_{i-1}^{ps_j^d}, q_i^{ps_j^d}}^{ps_j^d} \right] \right] \\ & \text{avec } P_i = P(q_i^{ps_j^d} | ps_j^d, \text{Precedent}(q_i^{ps_j^d})) \end{aligned} \quad [3]$$

5. Expérimentation et Résultats

5.1. *Base de Données*

Notre travail a été testé sur une base faite par les auteurs. Les schémas XML ont été groupés par domaine. Les domaines de notre base d'apprentissage sont $\{$ *Livre, Personne, Formation, Adresse, Société, Université, Étudiant* $\}$. Les schémas de la base sont très variés pour que les primitives extraites soient pertinentes. La Figure 2 montre le schéma de l'architecture du système réalisé.

5.2. *Expérimentation*

Le système opère suivant les trois modes principaux : le mode d'apprentissage, le mode de Mapping et la validité du Mapping en fonction des probabilités retournées. Au cours de l'apprentissage, les schémas d'apprentissage sont étiquetés par une procédure automatique de façon à réduire le temps requis pour l'apprentissage du système. Il procède ensuite à l'identification des primitives (voir la section 4.3) pour chaque domaine de la base d'apprentissage. Les matrices de Markov qui modélisent les schémas XML pour chaque domaine sont formées. Au cours du Mapping le système est amené à identifier les correspondances entre le schéma à intégrer dans une base (schéma cible) et le schéma de référence du domaine de la base choisie (schéma source).

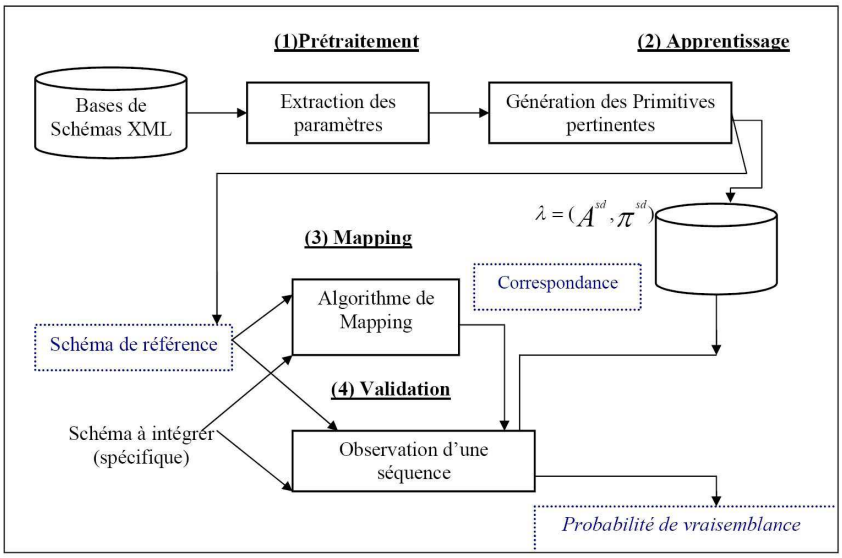


Figure 2. *Architecture du système de validation de Mapping entre les schémas XML*

Domaine	Livre	Personne	Adresse	Formation	Université	Société
Nbr schémas	100	80	60	80	60	60
V_e	11	12	14	20	12	15
S	11	6	7	10	8	9

Tableau 1. *Identification des primitives de chaque domaine à partir de la base des schémas XML*

5.3. *Résultats*

5.3.1. *Apprentissage*

Durant la première phase on identifie les états du schéma de référence de chaque domaine selon la base de schémas XML. Ces éléments (primitives) présentent par la suite la séquence S du modèle de Markov et l'ensemble N_1 des noeuds du schéma dans l'algorithme de Mapping selon l'algorithme décrit dans la section 4.3. le Tableau 1 représente le nombre des schémas utilisés et les cardinalités des vecteurs V_e et S, le Tableau 2 définit les éléments du vecteur S de chaque domaine étudié.

La deuxième étape de la phase d'apprentissage est l'identification des paramètres du modèle de Markov $\lambda = (A^{s_m^d}, \pi^{s_m^d}, \Pi)$ présentés dans la section 4.4. On applique ici le principe du maximum de vraisemblance. Pour le modèle du domaine *Livre* les paramètres du modèles de Markov sont :

Livre	Personne	Adresse	Formation	Universite	Societe
Livre	Personne	Adresse	Formation	Universite	Societe
Titre	Identite	Nomrue	Session	Nom	Departement
Titre1	Prenom	Numrue	Nom	Adresse	Personnel
Titre2	Nom	Numero	Apprenants	Diplôme	Nom
Auteur	Email	CodePostale	Apprenant	Cycle	Prenom
Nom	Telephone	Ville	Identifiant	Nom	Fonction
Prenom		Pays	Password	Niveau	Nom
Description			Nom	Specialite	Type
Date			Prenom		Adresse
Isbn			Groupes		
Prix					

Tableau 2. *primitives des domaines étudiés*

– $S = \{$Livre, Titre, Titre1, Titre2, Auteur, Nom, Prenom, Description, Date, Isbn, Prix$\}$

– La profondeur est $D = 3$.

– $P = \left\{s_1^1, s_2^2, s_3^2\right\}$, avec s_1^1=Livre, s_2^2=Titre et s_3^2=Auteur.

– Racine(Document)=Livre

Dans le reste de ce rapport on va prendre l'exemple du domaine *Livre*. Pour le modèle de ce domaine les paramètres du modèles de Markov sont décrits pour chaque noeud père s_m^d les matrices $A^{s_m^d}$,$\pi^{s_m^d}$ et Π dans le Tableau 3.

5.3.2. *Mapping*

L'algorithme du Mapping est appliqué entre deux schémas, schéma source et le schéma cible, pour l'exemple choisi ci dessous la description des deux schémas qui représentent les entrées de l'algorithme de Mapping.

– N_2={Livre, Titre, Titre1, Titre2, Auteur, Nom, Prenom, Description, Date, Isbn, Prix}

– A_2={(Livre,Titre), (Livre,Auteur),(Livre,Date),(Livre,Isbn),(Livre,Prix),(Titre,Titre1)} (Titre, Titre2), (Auteur, Nom), (Auteur, Prenom)}

– N_1={Livres, Sujet, Livre, Type, Description, Titre, Auteur, Date, Titre1, Titre2, Nom, Prenom, jj, aaaa }

– A_1={(Livres,Livre), (Livres,Sujet), (Livre,Date), (Livre,Auteur), (Livre,Titre), (Livre,Description), (Livre,Type), (Date,jj), (Date,aaaa), (Auteur,Nom), (Auteur,Prenom), (Titre,Titre1), (Titre,Titre2)}

L'ontologie utilisée à ce niveau est définie par le modèle suivant :

– $C = \{Auteur, Nom\_Auteur, Jour, An, Mois, Date\}$

Livre $A^{s_1^1} = A^{livre}$	$\pi^{s_1^1} = \pi^{livre}$	Π
$\begin{pmatrix} 0 & 0.22 & 0.22 & 0.11 & 0.17 & 0.15 & 0.11 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0.10 & 0 & 0.21 & 0.10 & 0.10 & 0.47 \\ 0 & 0.4 & 0 & 0 & 0.4 & 0.2 & 0 \\ 0 & 0.13 & 0 & 0.26 & 0 & 0.6 & 0 \\ 0 & 0.84 & 0 & 0 & 0.15 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 0.7 & 0 & 0.1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0.95 \\ 0 \\ 0 \\ 0 \\ 0.5 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$
Titre $A^{s_2^2} = A^{Titre}$	$\pi^{s_2^2} = \pi^{Titre}$	
$\begin{pmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$	
Auteur $A^{s_1^1} = A^{Auteur}$	$\pi^{s_1^1} = \pi^{Auteur}$	
$\begin{pmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$	

Tableau 3. *Les matrices du modèle de Markov pour les noeuds pères du domaine Livre(Livre, Titre et Auteur)*

– $R = \{r_1, r_2, r_3, r_4\}$; $r_1 = (Auteur, Nom\_Auteur)$,$r_2 = (Date, jour)$,$r_3 = (Date, an)$, $r_4 = (Date, mois)$

5.3.3. *Validation du Mapping*

Le système retourne la probabilité d'observation d'une séquence d'états E formée par les noeuds du schéma XML (voir section 4.5). On détermine les deux probabilités d'observation du schéma source et du schéma cible en tenant compte des correspondances retournées par l'algorithme de Mapping. On définit le quotient R :

$$R = \frac{\text{Probabilité du schéma source}}{\text{Probabilité du schéma cible}} = \frac{P(E|\lambda)\Big|^{source}}{P(E|\lambda)\Big|^{cible}}$$

Selon la valeur du R on déterminera la nature du Mapping, si $R \approx 1$ donc le Mapping est pertinent si $R \approx 0$ le Mapping n'est pas pertinent donc les correspondances retournées sont fausses sinon il est moyen.
Dans notre travail nous avons distingué deux modèles, on définit le rapport R pour chaque modèle :

Modèle de structure parent

$$R = \frac{\Pi(q_1) \prod_{j=1}^{K} \prod_{i=1}^{\text{nbenfant}(ps_j^d)} a^{ps_j^d}_{0,q_i^{ps_j^d}} \Big|^{source}}{\Pi(q_1) \prod_{j=1}^{K} \prod_{i=1}^{\text{nbenfant}(ps_j^d)} a^{ps_j^d}_{0,q_i^{ps_j^d}} \Big|^{cible}} \quad [4]$$

Modèle de structure parent- frère

$$R = \frac{\Pi(q_1) \prod_{j=1}^{K} \left[a^{ps_j^d}_{0,q_1^{ps_j^d}} \pi^{ps_j^d}(q_1^{ps_j^d}) \prod_{i=2}^{\text{nbenfant}(ps_j^d)} \left[a^{ps_j^d}_{0,q_i^{ps_j^d}} a^{ps_j^d}_{q_{i-1}^{ps_j^d}, q_i^{ps_j^d}} \right] \right] \Big|^{source}}{\Pi(q_1) \prod_{j=1}^{K} \left[a^{ps_j^d}_{0,q_1^{ps_j^d}} \pi^{ps_j^d}(q_1^{ps_j^d}) \prod_{i=2}^{\text{nbenfant}(ps_j^d)} \left[a^{ps_j^d}_{0,q_i^{ps_j^d}} a^{ps_j^d}_{q_{i-1}^{ps_j^d}, q_i^{ps_j^d}} \right] \right] \Big|^{cible}} \quad [5]$$

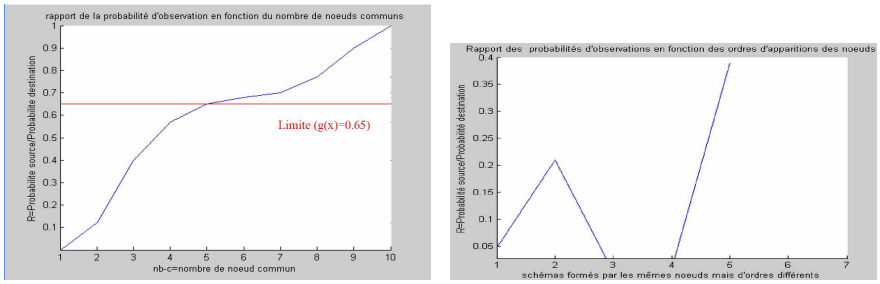
5.3.4. *discussion des Résultats*

Le but de notre travail est de valider le Mapping réalisé entre deux schémas XML source et cible, en utilisant le modèle de Markov décrit. Pendant la phase de reconnaissance nous avons distingué deux modèles : le modèle de structure *parent-fils* et le modèle de structure *parent-frère*. En phase de test on a appliqué l'algorithme de Mapping entre plusieurs schémas du même domaine. Les schémas sont choisis tels que le nombre de noeuds en commun entre le schéma source et cible varie ainsi que le l'ordre de disposition des noeud frère issus du même père change. A chaque itération on calcule le rapport R des probabilités d'observation des deux schémas. On définit par la suite la fonction f telle que :

$$f \begin{cases} f : \mathbb{N} \rightarrow [0 \dots 1] \\ f(nbc) = R \end{cases} \quad [6]$$

Modèle de structure parent-fils

Selon les tests établis nous avons remarqué que la probabilité varie en fonction du nombre de noeuds communs entre les deux schémas : le schéma à intégrer et le schéma du modèle du domaine étudié. Plus le nombre de noeuds communs entre les deux schémas augmente plus les probabilités d'observation des deux schémas sont proches plus le rapport R tend vers 1. f est une fonction croissante qui a un maximum en 1 quand $nbc = ne$ c'est-à-dire quand le schéma de référence et le schéma spécifique sont identiques ou le schéma spécifique est un sous schéma du schéma référence,selon



(a) (b)

Figure 3. *Variation de la probabilité en fonction : (a) du nombre de noeuds communs, (b) des ordres d'apparitions des noeuds*

la variation de f on peut juger la performance de l'algorithme du Mapping. La Figure 3(a) montre que f est croissante cela signifie que le Mapping est performant quand le nombre de noeuds communs nbc augumente. Le Mapping est valide si $nbc \geq 0.5 \times ne$. la Figure 3(a) montre que le rapport R tend de 0.6 à 1.

Modèle de structure parent-frère

Le rapport des probabilités des schémas de référence et le schéma spécifique n'est pas proportionnel au nombre de noeuds communs comme pour le cas du modèle de structure parent. f varie aléatoirement, nbc peut augmenter mais f diminue. La Figure 3(b) montre que le Mapping varie aléatoirement en fonction de nbc de 0.05 à 0.4.

5.4. *Discussions*

Les résultats montrent que pour les modèles réalisés, le premier modèle "parent-fils" est performant pour la validation du Mapping, l'autre modèle "parent-frère" ne donne pas de bonnes résultats car l'ordre d'apparition des noeuds n'est pas primordiale dans la représentation des schémas XML et des documents d'une façon générale. Notre système peut être étendu pour l'annotation automatique des documents historiques. Plusieurs travaux ont été réalisés pour la représentation des documents historiques avec les schémas XML. Notre système peut être intégré dans la phase de la réalisation d'un schéma de référence à toute la base des schémas XML qui représentent le même document historique. Le modèle de Markov qui décrit l'annotation de la base de schémas XML est défini. Le Mapping est appliqué entre le schéma de référence et un nouveau schéma à intégrer dans la base. La validation du Mapping est vérifiée par la dernière phase de système, si le Mapping n'est pas valide les correspondances entre les deux schémas sont modifiées.

6. Conclusions

Nous avons décrit un système de validation des algorithmes de Mapping des schémas XML. Une description de la structure des schémas XML en utilisant une méthode stochastique a été testée. Plusieurs tests ont été réalisés en identifiant deux modèles. Les résultats du premier modèle encouragent à tester le système sur l'annotation automatique des documents historiques et sa validation.

7. Bibliographie

Antonacopoulos A., Karatzas D., Bridson D., « Ground Truth for Layout Analysis Performance Evaluation », *in* S. B. . Heidelberg (ed.), *ICDAR*, p. 302-311, Août, 2005.

Boukottaya A., Vanoirbeek C., Paganelli F., Khaled O., « Automating XML documents Transformations : A conceptual modelling based approach », *First Asian-Pacific conference on Conceptual modelling (APCCM)*, p. 81-90, Janvier, 2004.

Denoyer L., Apprentissage et inférence statistique dans les bases de documents structurés : Application aux corpus de documents textuels, PhD thesis, Université de Paris 6, Décembre, 2004.

El Abed H., Märgner V., « Comparison of Different Preprocessing and Feature Extraction Methods for Offline Recognition of Handwritten Arabic Words », *9th International Conference on Document Analysis and Recognition (ICDAR)*, vol. 2, p. 974-978, 2007.

El Abed H., Märgner V., « Base de Données et Compétitions - Outils de Développement et d'Évaluation de Systèmes de Reconnaissance de Mots Manuscrits Arabes », *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, p. 103-108, 2008.

Fuhr N., Govert N., Kazai G., Lalmas M., « INEX : Initiative for the Evaluation of XML Retrieval », *Conference on Research and Development in Information Retrieval*, 2002.

Lecerf L., Chidlovskii B., « Apprentissage actif pour l'annotation de documents », *Conférences Francophone en Recherche d'Information et Applications*, p. 93-107, Mars, 2007.

Miller R. J., Hernbndez M. A., Haas L. M., Yan L., Ho C. T. H., Fagin R., Popa L., « The Clio Project : Managing Heterogeneity », *ACM SIGMOD Record 30*, p. 78-83, Mars, 2001.

Newman D., Hettich S., Blake C., Merz C. (eds), *UCI repository of machine learning databases*, http //www.ics.uci.edu/mlearn/MLRepository.html, 1998.

Reynaud C., Safar B., « Mappings pour l'intégration de documents XML », *Atelier Modélisation des connaissances*, Janvier, 2006.

Rifaieh R. D., utilisation des ontologies contextuelles pour le partage sémantique entre les systèmes d'information dans l'entreprise, PhD thesis, L'institut National des sciences Appliquées de Lyon, Décembre, 2004.

Smith R., « Hybrid Page Layout Analysis via Tab-Stop Detection », *ICDAR*, p. 241-245, Juillet, 2009.

Wisniewski G., Denoyer L., Maes F., Gallinari P., « Modèle probabiliste pour l'extraction de structures dans les documents semistructurés Application aux documents Web », *Conférences Francophone en Recherche d'Information et Applications*, p. 169-180, Mars, 2006.

Zerdazi A., Lamolle M., « Hyperschema XML un modèle dintégration par enrichissement sémantique de schémas XML », *MajecSTIC 05*, p. 143-150, Novembre, 2005.