Modélisation multidimensionnelle d'entrepôts de documents XML répartis

Ines Ben Messaoud*, Jamel Feki*
Gilles Zufluh**

*Laboratoire Mir@cl, Faculté des Sciences Economiques et de Gestion, Université de Sfax
Route de l'Aéroport km 4, 3018 Sfax, BP. 1088 - Tunisie
{ines.benmessaoud ; jamel.feki}@fsegs.rnu.tn
**IRIT, Institut de Recherche en Informatique de Toulouse, Université Toulouse 1,
2 Rue du Doyen Gabriel Marty, 31042 Toulouse Cedex 9 – France
zurfluh@univ-tlse1.fr

Résumé. De nos jours, et avec l'ouverture des organisations sur Internet, les documents constituent une source intéressante pour les analyses décisionnelles; ils aident les décideurs à mieux comprendre l'évolution des processus métier de leur organisation. Généralement, ces documents existent sous format XML, sont géographiquement répartis et décrits par des structures différentes. Cet article propose une méthode de construction d'entrepôts de documents distribués comportant deux étapes : *i) Unification des structures des documents XML*, et *ii) Modélisation multidimensionnelle des arbres unifiés*. Plus précisément, il focalise sur l'étape de modélisation multidimensionnelle.

1 Introduction

Face à la progression rapide des technologies de l'information, les organisations se situent dans un environnement caractérisé par un volume croissant de données qui transitent par leur système d'information. Pour faire face aux différentes pressions, d'ores et déjà internes et externes, elles doivent mettre à la disposition de leurs dirigeants des supports d'assistance à la bonne prise de décisions en respectant des contraintes de délais raisonnables. Les entrepôts de données et les technologies OLAP (« *On-Line Analytical Processing* ») soutiennent de tels besoins. En effet, ils permettent d'analyser de grandes volumétries de données structurées que les organisations stockent dans des bases de données Pérez et al. (2008). Ces bases sont construites à partir de données transactionnelles extraites des systèmes d'information des entreprises. Cependant, seuls 20% des données d'un système d'information sont des données transactionnelles et peuvent être traitées par un système OLAP et les 80% restants sont constitués de documents (rapports, articles, etc.) Tseng et Chou (2006).

Par ailleurs, les documents constituent une capitalisation importante des connaissances du système de production et sont utiles pour le système de pilotage. Généralement, le contenu de ces documents est peu structuré d'où la difficulté de les intégrer dans les systèmes d'information décisionnels Ronan (2007). En conséquence, il résulte lors du processus décisionnel, que les analystes-décideurs n'arrivent pas à explorer facilement, rapidement et effi-

cacement ces documents Tseng et Chou (2006) ; cette situation risque de conduire à des décisions imparfaites. Afin de remédier à ce problème, certains chercheurs recommandent d'entreposer les documents McCabe et al. (2000) Sullivan (2001). En fait, un **E**ntrepôt de **Doc**uments (EDoc) stocke les documents issus des sources de données externes et internes à l'organisation et peut être vu comme étant un cas particulier d'entrepôt de contenu.

Par ailleurs, il arrive souvent que les documents nécessaires pour le traitement d'un besoin analytique soient géographiquement répartis ; de plus, ils peuvent être de formats hétérogènes. Particulièrement, nous nous intéressons aux documents XML (« *Extensible Markup Language* ») du fait que ce format est le plus utilisé pour la représentation et l'échange de données. Etant hétérogènes, ces documents XML sont décrits par des structures différentes. En conséquence, leurs interrogations en vue de traitements OLAP nécessite, premièrement une étape indispensable d'unification de leurs structures pour construire une vision globale de l'EDoc qui les réunit, et deuxièmement, une étape de modélisation multidimensionnelle permettant de mettre en exergue les sujets d'analyses ainsi que leurs axes associés. Cet article s'inscrit dans le cadre de construction d'un EDoc distribués intégrant, entre autre, des documents XML structurellement hétérogènes et géographiquement répartis. Nous proposons une méthode de conception d'EDoc permettant d'aboutir au schéma global des documents distribués. Cette méthode est composée de deux étapes : *Unification des structures des documents XML* et *Modélisation multidimensionnelle des arbres unifiés*. Plus précisément, nous détaillons l'étape de modélisation multidimensionnelle qui produit un schéma multidimensionnel des documents XML répartis.

Cet article est structuré comme suit : la section 2 étudie les travaux les plus populaires en matière d'unification des DTDs (« Document Type Definition ») et de modélisation multidimensionnelle des documents. La section 3 donne un aperçu général de notre méthode proposée pour la construction d'entrepôt de documents EDoc répartis. La section 4 décrit brièvement sa première étape appelée unification des structures des documents XML. Quant à la section 5, elle détaille l'étape de modélisation multidimensionnelle. Finalement, la section 6 synthétise ce travail et envisage ses travaux futurs.

2 État de l'art

Le format XML permet la représentation et l'échange des données sur le Web et au sein des organisations (W3C-XML, 2008). Il représente un format de texte simple et flexible. Selon ce format, le contenu d'un document est encapsulé dans des balises qui forment la structure hiérarchique du document. Ce type de document permet le stockage des données dans un formalisme auto descriptif par l'intermédiaire d'une description de structure : DTD ou XSchema. Généralement, ces structures sont différentes même pour des documents appartenant à un même domaine. Elles méritent d'être unifiées afin de construire une description commune pour l'ensemble des documents, puis modélisées de façon multidimensionnelle pour pouvoir expliciter leurs intérêts décisionnels vis-à-vis des besoins des décideurs.

Dans cette section, nous présentons d'abord les travaux les plus populaires traitant l'unification des DTDs puis, nous décrivons brièvement les travaux les plus pertinents relatifs à la modélisation multidimensionnelle de documents.

La technique d'unification des DTDs permet d'aboutir à une DTD unifiée, à partir de plusieurs DTDs, tout en réduisant la perte de sens des DTDs initiales Yoo et al. (2005). Plu-

sieurs travaux se sont intéressés à cette problématique : Júnior et Mello (2008) et Yoo et al. (2005).

En effet, les auteurs de Júnior et Mello (2008) proposent un *processus d'intégration* des instances hétérogènes des documents XML appartenant à un même domaine. Ce processus comporte deux phases : *Définition de similarité des documents* et *Unification*. La première phase compare chaque paire d'instance de documents pour stipuler un score de similarité. Ce score est basé sur des métriques et un dictionnaire pour vérifier les synonymes des termes. Cette phase classifie les documents en plusieurs sous-ensembles sémantiquement similaires, tandis que la phase d'unification génère une représentation unifiée pour chaque sous-ensemble de documents similaires. Elle repose sur une ontologie de domaine et un dictionnaire pour la dénomination des éléments du document XML résultat. Cependant, nous constatons que la comparaison des documents peut être couteuse en temps de traitement si des documents se partagent une même structure. Elle peut être améliorée par une comparaison à un niveau plus abstrait, c'est-à-dire celui des structures des documents. Notons que ce processus d'intégration n'était pas prévu dans un contexte d'entrepôt de documents répartis où il fallait résoudre des problèmes spécifiques comme la construction d'un schéma de répartition des données dans les sites de stockage.

De même, dans Yoo et al. (2005), les auteurs proposent un algorithme pour l'unification des DTDs de documents XML ayant des structures similaires et appartenant à un même domaine (e.g., DTDs d'articles de revues scientifiques). Leur algorithme reçoit en entrée un ensemble de DTDs et génère en sortie une DTD unifiée. En fait, une DTD unifiée joue le rôle d'un schéma conceptuel global pour un domaine représenté par des documents XML. L'algorithme proposé utilise les automates finies et la structure arborescente des documents. Il comporte quatre étapes appelées : *Prétraitement des DTDs*, *Représentation des DTDs*, *Génération d'une DTD unifiée*, et *Post-traitement*. Initialement, le prétraitement résout les ambigüités des noms des éléments des DTDs et ceci en se basant sur une table nommée « Element Name Resolution Table » qui fournit les synonymes dans le domaine. Ensuite, la deuxième étape transforme les DTDs sous forme d'arbre et d'automates. Puis, ces derniers sont fusionnés pour créer une DTD unifiée. Finalement, la validité syntaxique d'une DTD unifiée est vérifiée par un parseur de DTD. Cependant, la table « Element Name Resolution Table » risque d'être incomplète et peut affecter la qualité du résultat. Encore une fois, le travail proposé n'était pas dans le contexte des entrepôts documents répartis.

D'autres travaux, comme McCabe et al. (2000), Tseng et Chou (2006), Ronan (2007) et Ravat et al. (2007) traitent l'analyse multidimensionnelle des documents.

Dans McCabe et al. (2000) et Tseng et Chou (2006), les auteurs utilisent le schéma en étoile pour analyser les documents. Dans McCabe et al. (2000), les auteurs distinguent les cinq dimensions suivantes : *Localisation*, *Date*, *Document*, *Catégorie* et *Terme*. Le schéma en étoile produit permet de mesurer le nombre d'occurrences des termes dans un document. Dans Tseng et Chou (2006), les auteurs différencient les trois types de dimensions suivantes : *Ordinaire* (e.g., mots-clefs), *Métadonnées* (les métadonnées extraites des documents) et *Catégorie* (e.g., catégorie de document, ontologie Wordnet). Le schéma en étoile résultat permet de compter le nombre d'occurrences des documents en fonction des dimensions du schéma (e.g., auteurs, dates). Néanmoins, à partir des schémas en étoile proposés, seules les analyses quantitatives peuvent être effectuées du fait que les indicateurs sont numériques.

Dans les travaux de Ronan (2007), Ravat et al. (2007) les auteurs proposent un modèle conceptuel multidimensionnel pour l'analyse de données documentaires appelé modèle en galaxie. Ce modèle préserve la structure des documents et les liens entre eux. Il repose sur un

seul concept : le concept *Dimension*. Un schéma en galaxie est un regroupement de dimensions liées entre elles par un ou plusieurs nœuds centraux où chaque nœud modélise les dimensions *compatibles* pour une même analyse, c'est-à-dire celles qui peuvent être étudiées ensembles. Pour ce qui est de la modélisation des données textuelles, les auteurs distinguent les attributs documentaires et les dimensions documentaires. Les attributs documentaires représentent les données issues des documents textuels (e.g., paragraphe) et permettent de décrire la structure du document au sein de la dimension documentaire. Par exemple, un mémoire de thèse est composé de paragraphes contenus dans des sections elles mêmes contenues dans des chapitres. Toutefois, le travail proposé n'aborde pas la problématique dans un environnement distribué.

Dans cette section, nous nous sommes limités à présenter succinctement les travaux qui nous semblent les plus pertinents en unification des DTDs et en modélisation multidimensionnelle des documents. Le bilan de l'état de l'art nous a permis de constater que les travaux traitant l'unification ne sont pas proposés dans le contexte des entrepôts de documents répartis. Également, dans les travaux relatifs à la modélisation multidimensionnelle des documents, nous avons noté qu'il existe deux types de modèle multidimensionnel : le schéma en étoile et le modèle en galaxie. Dans un schéma en étoile, un fait modélise un sujet d'analyse préalablement défini. Par conséquent, la spécification d'analyse est peu flexible du fait que le décideur emploi les faits comme des sujets. De plus, ces analyses sont simples parce qu'elles associent des indicateurs numériques et reposent principalement sur le comptage des documents, comme par exemple, compter le nombre des articles scientifiques écrits par l'auteur *A* et publiés dans la conférence *C*. En conséquence de cette simplicité, le contexte des articles ne peut pas être analysé Ronan (2007). Le second modèle en galaxie repose sur l'unique concept de dimension et a l'avantage d'être plus simple que le schéma en étoile.

Tenant compte de ces constatations, nous avons noté l'absence de travaux se rapportant aux entrepôts de documents répartis et aux techniques de leurs interrogations. L'objectif de ce travail est de proposer une approche de construction d'entrepôts de documents XML répartis. L'approche exploite le formalisme des arbres et utilise le modèle en galaxie pour la représentation multidimensionnelle des documents. Elle comporte deux étapes : i) unification des structures des documents XML que nous décrivons brièvement, et ii) modélisation multidimensionnelle des arbres unifiés qui sera plus développée.

3 Méthode proposée

Les documents constituent une capitalisation de connaissances Ronan (2007). Ils aident les décideurs à comprendre l'évolution des activités de l'organisation au cours du temps. En fait, les auteurs de Tseng et Chou (2006) affirment que seulement 20% de l'information décisionnelle peuvent être extrait à partir des bases de données conventionnelles, alors que les 80% restant se présentent sous forme de données non numériques (*i.e.*, les documents) et ne sont pas intégrés au système d'information décisionnel. Ces documents existent sous plusieurs formes, nous nous intéressons particulièrement aux documents XML du fait que ce format est le plus utilisé pour la représentation et l'échange des données (W3C-XML, 2008).

Généralement, les documents sont stockés dans des sites distants et sont décrits par des structures hétérogènes. Par conséquent, lors d'une analyse OLAP, le décideur a besoin de récupérer des informations venant des sites de stockage des documents. De ce fait, il sera contraint de prendre en considération l'hétérogénéité structurelle des documents. Ceci con-

duit à écrire des requêtes multiples, soit autant de requêtes que de structures distinctes. Afin de surmonter les problèmes inhérents à cette complexité, le décideur a besoin d'une structure commune qui joue un rôle similaire à celui d'un schéma global dans le contexte des bases de données distribuées.

Afin d'atteindre cet objectif, nous proposons une méthode de construction d'entrepôt de documents distribués permettant de fournir une vue globale de l'entrepôt. Cette méthode comporte les deux étapes suivantes Ben Messaoud et al. (2010) :

- Unification des structures des documents XML.
- Modélisation multidimensionnelle des arbres unifiés.

La figure 1 schématise notre méthode proposée.

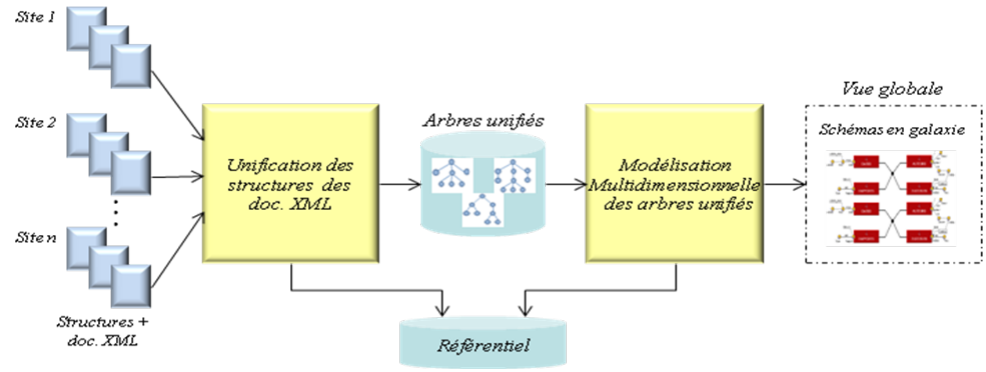

FIG. 1 – *Construction d'un entrepôt de documents distribués.*

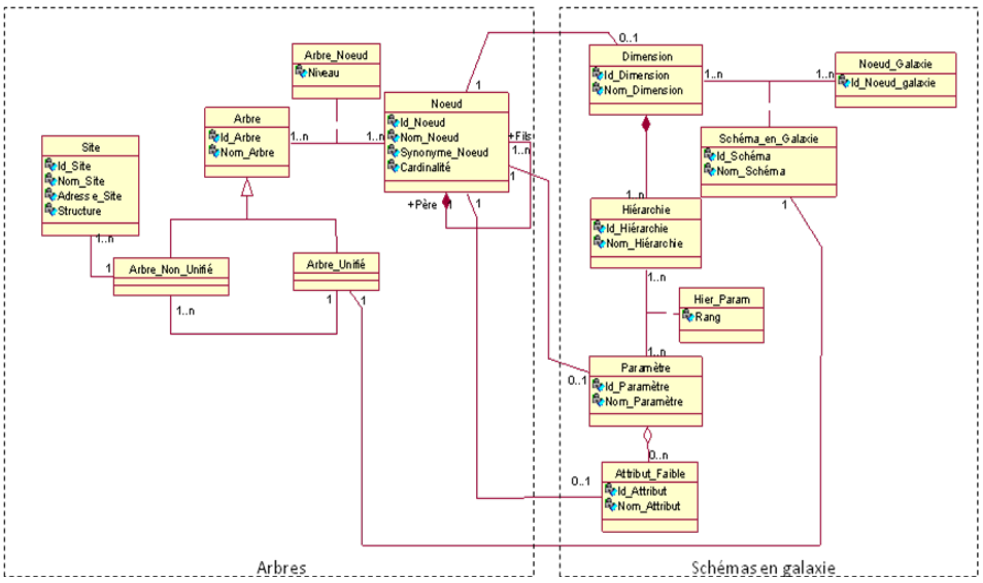

FIG. 2 – *Référentiel des arbres et des schémas en galaxie.*

L'unification des structures des documents XML définit des structures communes pour représenter les documents XML appartenant aux différents sites. Ces structures sont présentées sous forme d'arbres et sont sauvegardées dans un référentiel. Quant à l'étape de modélisation multidimensionnelle, elle conçoit des schémas multidimensionnels en galaxie à partir des arbres générés dans l'étape précédente. Ces schémas constituent des vues globales des documents distribués et sont enregistrés dans un référentiel qui mémorise les informations relatives au site d'hébergement du document XML (i.e., nom, adresse). Ce référentiel établit le lien entre les arbres unifiés issus de la première étape et, les schémas multidimensionnels produits dans la deuxième étape. La figure 2 est un diagramme de classes de ce référentiel.

Dans les sections suivantes, nous décrivons brièvement la première étape et nous focalisons sur la deuxième étape de notre méthode : la modélisation multidimensionnelle des arbres unifiés.

4 Unification des structures des documents XML

Souvent, les structures des documents XML diffèrent même pour les documents appartenant à un même domaine. Il est alors crucial de définir une structure unifiée pour travailler sur un ensemble de documents XML hétérogènes. Cette structure unifiée assiste le décideur à exprimer son besoin en spécifiant une seule requête qui touche les différentes structures de manière transparente.

Nous proposons une méthode d'unification des structures des documents XML appartenant à un même domaine. Cette méthode se base sur l'agencement arborescent des documents XML et comporte les trois étapes suivantes :

- Représentation arborescente.
- Génération des arbres unifiés.
- Validation des arbres unifiés.

Dans les sous-sections suivantes, nous donnons un bref aperçu de ces étapes. Pour plus de détails et d'exemples, le lecteur peut se référer à Ben Messaoud et al. (2011).

4.1 Représentation arborescente

Elle traduit la structure d'un document XML selon le formalisme des arbres comme dans Yoo et al. (2005), Lee et al. (2002). Le choix des arbres est motivé par leur facilité de compréhension par le décideur/concepteur. En fait, les arbres encouragent les décideurs à participer dans la prochaine étape de validation du résultat d'unification.

4.2 Génération des arbres unifiés

Il s'agit de comparer les arbres résultats de l'étape précédente et de générer un ensemble d'arbres unifiés. Cette génération s'effectue comme suit :

- Traitement sémantique des nœuds des arbres.
- Calcul de similarité inter arbres.
- Production des arbres unifiés.

Initialement, le traitement sémantique résout les ambigüités sémantiques des noms des nœuds des arbres provenant de sites différents en utilisant l'ontologie *Wordnet*. Ainsi, les nœuds ayant le même sens sont substitués par un nom unique. Par exemple, si nous avons

trois arbres A_1, A_2, A_3 et si A_1 et A_3 possèdent un nœud commun nommé *Writer* et A_2 possède un nœud *Author*. Ces trois nœuds ont le même sens et seront agrégés en un seul nœud nommé *Author* par exemple.

Ensuite, nous définissons une matrice de similarité (notée *MS*). Elle est inspirée de la matrice présentée dans Feki (2004) utilisée pour l'intégration des schémas multidimensionnels. Cette matrice permet de déterminer les arbres les plus prioritaires à fusionner, c'est-à-dire, ceux ayant des structures étroitement similaires. Chaque cellule de *MS décrit* la similarité *Sim(i, j)* de l'arbre en ligne *i* et de l'arbre en colonne *j* de la matrice *MS*. Cette similarité est calculée par la formule (1).

$$Sim(i, j) = \begin{cases} 0.75 \; si \; n_i = c_{i,j} \; et \; n_i < n_j \\ \dfrac{c_{i,j}}{q} \; \text{sinon} \end{cases} \qquad (1)$$

Où :

n_i et n_j sont respectivement le nombre de nœuds dans les arbres *i* et *j*,

$c_{i,j}$ est le nombre de leurs nœuds en commun, et $q = n_i + n_j - c_{i,j}$.

Finalement, un arbre est produit pour chaque ensemble d'arbres similaires moyennant un algorithme utilisant l'opérateur *fusion par greffe* proposé par Golfarelli et al. (1998) et Golfarelli et Rizzi (1999) et repris par Boussaid et al. (2006). L'algorithme utilise trois autres opérateurs nommés *fusion par inclusion*, *fusion par union des sous-arbres* et *fusion par union des nœuds* que nous avons définis dans Ben Messaoud et al. (2011). La fusion par greffe s'effectue lorsque les sous-arbres en commun n'ont pas la même structure de relations (i.e., relation entre les nœuds) dans les deux arbres en entrée ; les sommets non communs sont éliminés alors que les sommets identiques et leurs relations sont maintenus dans l'arbre fusionné. Ainsi lorsqu'un sommet est éliminé, ses descendants sont conservés dans l'arbre résultat. La fusion par inclusion est opérée quand l'un des deux arbres en entrée est inclus dans l'autre. La fusion par union des sous arbres est utile lorsque les nœuds communs des arbres en entrée ne partagent pas les mêmes nœuds fils. Dans ce cas, l'arbre résultat est composé de l'union des sous-arbres des arbres en entrée. Finalement, la fusion par union des nœuds est utile quand deux sous-arbres identiques possèdent des nœuds parents différents ; dans ce cas, l'arbre fusionné est caractérisé par le nœud spécifique *ou*. Ce nœud substitue les nœuds parents distincts des arbres en entrée et relie les nœuds en communs. Notons que lors de la fusion des arbres les cardinalités des nœuds sont prises en considération et sont traitées selon les règles présentées dans Hachaichi et al. (2010).

4.3 Validation des arbres unifiés

Dans cette étape, le décideur/concepteur peut intervenir pour valider les arbres unifiés en les ajustant à ses besoins. Il peut supprimer/renommer les nœuds des arbres. Les arbres unifiés résultats sont enregistrés dans le référentiel (cf. figure 2) que nous préparons pour aborder certains problèmes d'interrogation de l'EDoc distribués.

5 Modélisation multidimensionnelle des arbres unifiés

La modélisation multidimensionnelle vise à concevoir des schémas multidimensionnels reflétant des besoins OLAP. Elle permet de représenter les données tout en mettant en évidence le sujet d'analyse (fait et ses mesures) et les axes d'analyses.

Conformément à notre étude de l'état de l'art, et à notre connaissance, il existe deux types de modèle multidimensionnel. L'un est basé sur la dualité fait-dimension et l'autre est basé seulement sur le concept de dimension. Vu d'une part, les difficultés liées au premier modèle, et d'autre part, les avantages du deuxième modèle (cf. section 2), nous utilisons ce dernier pour la modélisation multidimensionnelle des documents de l'EDoc distribué. En fait, un modèle en galaxie est un regroupement de dimensions liées entre elles par un ou plusieurs nœuds centraux (cf. section 5.2.2, figure 6).

Nous décrivons, dans la suite de cet article, notre méthode de construction de schémas en galaxie à partir des arbres unifiés. Cette méthode s'articule autour des trois étapes suivantes :

- Prétraitement des arbres.
- Construction des schémas en galaxie.
- Validation des schémas en galaxie.

La figure 3 illustre l'enchaînement de ces étapes.

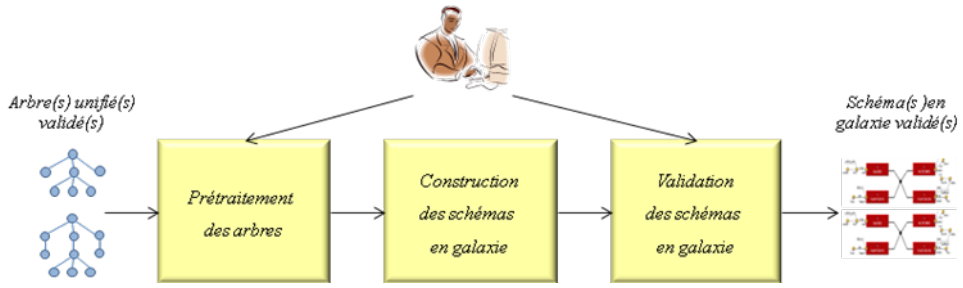

FIG. 3 – *Étapes de construction de schémas en galaxie.*

5.1 Prétraitement des arbres

L'étape de prétraitement est manuelle, elle vise à améliorer la sémantique des arbres en entrée (i.e., les arbres unifiés et validés). Elle permet d'ajouter des cardinalités pour chaque nœud père, autant de cardinalités que d'arcs sortants (à l'exception du nœud racine). Ces cardinalités doivent être soigneusement définies par le concepteur puisqu'elles influent sur la qualité du schéma multidimensionnel à générer. En effet, la construction des schémas en galaxie, et plus précisément l'identification des concepts multidimensionnels, à savoir, les dimensions et leurs hiérarchies exploitent ces cardinalités.

Par exemple, l'application de l'étape de prétraitement sur l'arbre *Article* (cf. figure 4), supposé obtenu par unification, produit l'arbre *Article'* illustré par la figure 5.

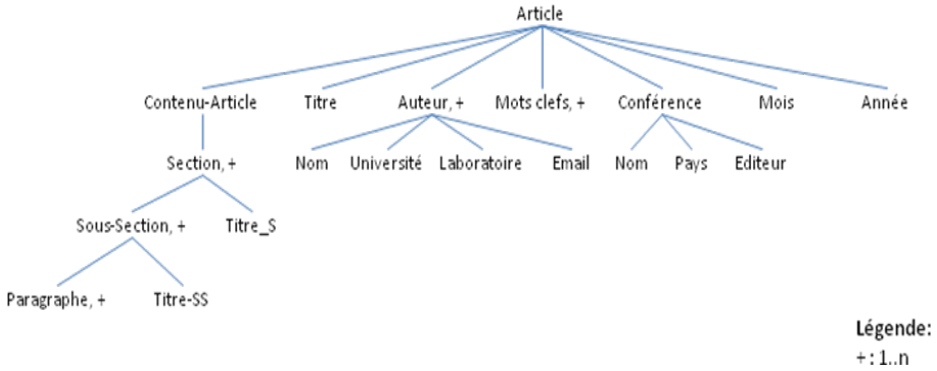

FIG. 4 – *Un exemple d'arbre unifié Article.*

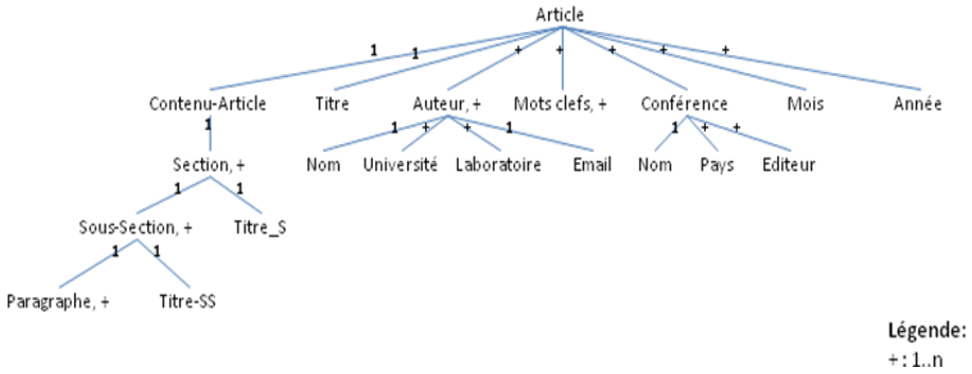

FIG. 5 – *Arbre Article' résultat du prétraitement de l'arbre Article.*

5.2 Construction des schémas en galaxie

Cette construction extrait les concepts multidimensionnels, à partir des arbres prétraités, et moyennant un ensemble de règles que nous détaillons dans les sous sections suivantes. Une fois ces concepts sont déterminés, nous obtenons des schémas en galaxie qui seront présentés au concepteur afin de les valider. Nous commençons par l'identification des dimensions et de leurs liens.

5.2.1 Identification des dimensions et des nœuds centraux inter dimensions

Les dimensions modélisent les perspectives d'analyse. Elles se caractérisent par des attributs organisés de manière hiérarchique, où chaque attribut modélise un niveau de granularité. Chaque dimension est caractérisée par son nom et possède un ensemble d'attributs.

Dans un arbre, les nœuds représentent généralement des groupements d'attributs, certains de ces attributs sont des éléments terminaux. Dans notre méthode, nous identifions les dimensions parmi les nœuds d'un arbre et les paramètres des hiérarchies à partir des éléments terminaux et des nœuds.

Identification des dimensions. Le modèle en galaxie se base sur le seul concept de dimension, nous définissons les trois règles *Rd1*, *Rd2* et *Rd3* pour identifier ces dimensions. Conventionnellement, une dimension construite sur un nœud de nom *N* sera nommée *D-N*.

Rd1 : Le sommet *A* de tout arbre constitue une dimension.

Nous identifions également des dimensions à partir d'autres nœuds par les deux règles suivantes.

Rd2 : Tous nœuds *N* et *M* non terminaux telle que l'arc *N-M* est de cardinalité + ou * des deux côtés, constituent deux dimensions.

Ces cardinalités multiples des deux côtés de l'arc reliant *N* et *M* dénotent l'existence d'une association pour laquelle les deux nœuds simulent des entités. Alors, nous considérons ces nœuds-entités comme des dimensions.

Généralement, la dimension temps figure dans tout entrepôt (Kimball, 1997). La règle suivante s'intéresse à cette dimension.

Rd3 : L'ensemble des nœuds d'un même niveau et décrivant un composant de date (e.g, jour, mois) constitue une dimension temporelle nommé *D-Date* où ces nœuds seront des paramètres.

Ces nœuds peuvent être déterminés et organisés en hiérarchies en se référant à un dictionnaire des termes temporels.

Notons que parfois une même dimension *D* peut être déterminée plus qu'une fois, c'est-à-dire, par plusieurs règles ; elle sera présentée une seule fois dans le schéma résultat.

L'application des règles *Rd1*, *Rd2* et *Rd3* sur l'arbre *Article'* identifie les quatre dimensions nommées *D-Article*, *D-Auteur*, *D-Conférence* et *D-Date*.

Détermination des nœuds centraux inter dimensions. Dans un schéma en galaxie, les nœuds centraux modélisent les dimensions compatibles pour une même analyse. Pour identifier ces nœuds, nous définissons la règle suivante :

Rn : Si *N* et *M* sont deux nœud-dimensions directement ou indirectement reliés alors les dimensions construites sur ces nœuds *N* et *M* sont des dimensions compatibles.

L'application de la règle *Rn* identifie un nœud central reliant les quatre dimensions : *D-Article*, *D-Conférence*, *D-Auteur* et *D-Date*.

5.2.2 Identification des hiérarchies

Les hiérarchies présentent les différentes perspectives d'analyses sur une même dimension. En fait, une hiérarchie organise les paramètres d'une dimension selon la relation *"est_plus_fin"* conformément à leur niveau de détail Teste (2000). Toutes les hiérarchies d'une dimension *D* partent nécessairement de l'identifiant de *D* qui est le paramètre le plus fin. Pour chaque dimension générée, nous définissons un identifiant de substitution (« Surro-

gate key ») codé *Id-D-DW*. Ainsi, les identifiants des dimensions *D-Article, D-Auteur, D-Conférence* et *D-Date* sont respectivement *Id-D-Article-DW, Id-D-Auteur-DW, Id-D-Conférence-DW et Id-D-Date-DW*. Notons que lorsqu'un nœud-dimension *D* possède un nœud fils de type identifiant, ce dernier sera associé comme un attribut faible à l'identifiant de substitution *Id-D-DW*.

Pour déterminer la suite des paramètres d'une hiérarchie, nous continuons à extraire ceux qui suivent l'identifiant (i.e., de rang supérieur à 1). Pour ce faire, nous définissons quatre règles qui déterminent les paramètres d'une manière ordonnée dans chaque hiérarchie, et deux règles pour dégager les attributs faibles. Dans ces règles, le nom d'un paramètre construit sur un nœud de nom *N* sera conventionnellement nommé *P-N*.

Détermination des paramètres.

Rp1 : Tout nœud terminal *N* dont le père *M* est identifié comme nœud-dimension telle que l'arc *N-M* n'est pas de cardinalité 1 des deux cotés, constitue un paramètre de rang 2.

Rp2 : Tout nœud *N* terminal dont le père *M* est identifié comme un nœud-paramètre de rang *i* telles que l'arc *N-M* n'est pas de cardinalité 1 des deux côtés, constitue un paramètre de rang *i-1*. Les rangs des paramètres sont exprimés selon la relation « est-plus-fin ».

Naturellement, un nœud terminal *N* lié à un nœud-dimension *M* ou un nœud-paramètre *M* représente un nœud-paramètre. En effet, le nœud *M* regroupe des données élémentaires qui le décrivent. De plus, parmi ces nœuds *N* nous privilégions ceux de type caractère puisqu'un paramètre est généralement textuel. Nous écartons les nœuds terminaux tels que l'arc *N-M* est de cardinalités 1 des deux côtés du fait que ces nœuds représentent des données descriptives pour le nœud père *M*.

La règle *Rp1* identifie *P-Mots-Clefs* comme des paramètres de rang 2 pour la dimension *D-Article* ; et *P-Université* et *P-Laboratoire* comme des paramètres de rang 2 pour la dimension *D-Auteur* ; et *P-Pays* et *P-Editeur* comme des paramètres de rang 2 pour la dimension *D-Conférence*. L'application de la règle *Rp2* identifie le paramètre *P-Paragraphe de rang 2* pour la dimension *D-Article*.

Nous continuons à identifier les paramètres des hiérarchies à partir des nœuds non terminaux par le biais des deux règles *Rp3* et *Rp4*.

Rp3 : Si *N* est un nœud non terminal dont le nœud père *M* est un paramètre de rang *i* telle que l'arc *N-M* est de cardinalité + ou * du côté du nœud *N* et 1 du côté de *M*, alors *N* constitue un paramètre de rang *i-1*.

Cette règle identifie *P-Paragraphe, P-Sous-Section* et *P-Section* comme des paramètres de rang 2, 3 et 4 respectivement pour la dimension *D-Article*.

Rp4 : Si *N* est un nœud non terminal dont le nœud père *M* est une dimension telle que l'arc *N-M* est de cardinalité 1 des deux côtés, alors *N* constitue un paramètre qui s'ajoute à la fin de la hiérarchie.

La logique de cette règle s'apparente à un contexte relationnel où *M* est une table identifiée comme dimension et *N* est un attribut de cette table Feki et al. (2007). Naturellement, cet attribut joue le rôle de paramètre pour *M* puisqu'il est en dépendance fonctionnelle de l'identifiant de la dimension (lien 1-1 entre *N* et *M*).

L'application de cette règle sur notre exemple en cours dégage *P-Contenu-Article* comme paramètre de rang 5 pour la dimension *D-Article.*

Identification des attributs faibles. Certains paramètres d'une dimension peuvent être accompagnés de descripteurs appelés *attributs faibles*. Ces attributs faibles ont un rôle informationnel permettant de libeller les résultats des analyses. Les deux règles suivantes identifient les attributs faibles des paramètres.

Raf1 : Si *N* est un nœud terminal dont le nœud père *M* est une dimension telle que l'arc *N-M* est de cardinalité 1 des deux côtés, alors *N* constitue un attribut faible du paramètre identifiant de *D*.

Raf2 : Si *N* est un nœud terminal dont le nœud père *M* est un paramètre tel que l'arc *N-M* est de cardinalité 1 des deux côtés, alors *N* constitue un attribut faible du paramètre *M*.

Les règles *Raf1* et *Raf2* identifient *Titre-S* comme attribut faible pour le paramètre *P-Section* ; *Titre-SS* comme attribut faible pour le paramètre *P-Sous-Section* ; *Titre* comme attribut faible pour l'identifiant de la dimension *D-Article* ; *Email* et *Nom* comme attribut faible pour l'identifiant de la dimension *D-Auteur*, et *Nom* comme attribut faible pour l'identifiant de la dimension *D-Conférence*.

La figure 6 montre le schéma en galaxie construit à partir de l'arbre *Article'* (cf. figure 5) suite à l'application de l'ensemble des règles d'extraction que nous avons définies.

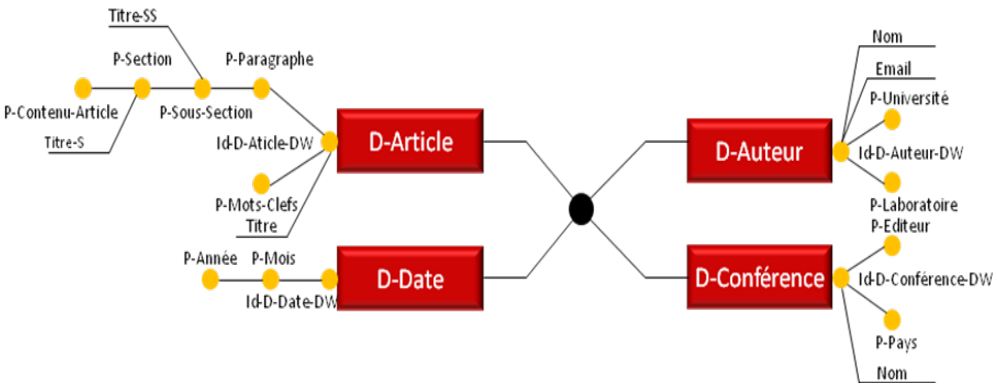

FIG. 6 – *Schéma en galaxie construit à partir de l'arbre de la figure 5.*

5.2.3 Validation des schémas en galaxie

Dans cette étape, le décideur valide les schémas en galaxie générés pendant l'étape précédente. Il peut supprimer et renommer les éléments multidimensionnels, réorganiser les paramètres d'une hiérarchie, et ajouter ou supprimer des nœuds centraux.

Par exemple, le schéma en galaxie illustré dans la figure 6 peut être ajusté pour dériver le schéma de la figure 7 en effectuant les opérations suivantes : suppression du paramètre *P-Paragraphe* et l'attribut faible *Titre-SS*, renommage de l'identifiant *Id-D-Article-DW* de la dimension *D-Article* en *id*.

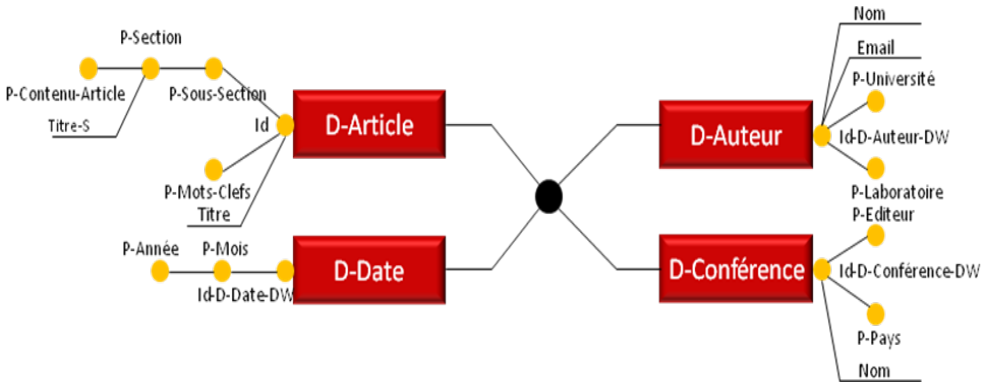

FIG. 7 – *Schéma en galaxie dérivé par ajustement du schéma en galaxie de la figure 6.*

Comme deuxième illustration de l'application des règles définies dans les sections 5.2.1 et 5.2.2 sur l'arbre de la figure 7, nous obtenons le schéma en galaxie de la figure 8.

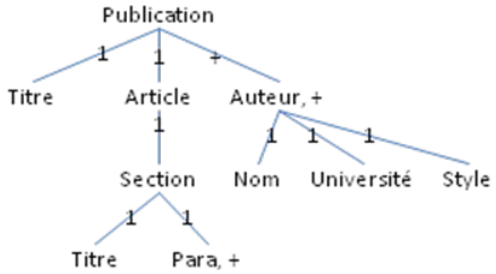

FIG. 7 – *Arbre Publication.*

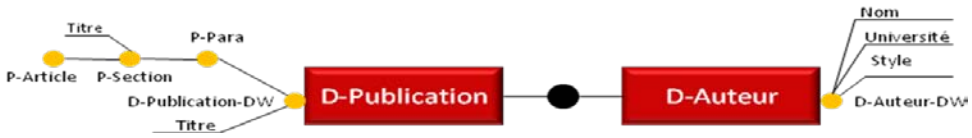

FIG. 8 – *Schéma en galaxie construit à partir de l'arbre de la figure 7.*

6 Conclusion

Le travail présenté dans cet article s'inscrit dans le contexte de la construction d'entrepôts de documents distribués. Plus précisément, nous avons élaboré et décrit notre méthode de construction de schémas multidimensionnels modélisant des documents XML structurellement hétérogènes. La méthode proposée est articulée autour de deux grandes étapes appelées : *Unification des structures des documents XML* et *Modélisation multidimensionnelle des arbres unifiés*.

L'étape d'unification nous a permis de définir une structure commune pour représenter les documents XML. Elle traduit les structures des documents XML en une représentation arborescente. Ensuite, elle génère des arbres unifiés à partir des arbres issus des différentes structures. Pour effectuer cette génération, nous avons défini une métrique de similarité qui évalue la ressemblance entre arbres, et nous avons utilisé une matrice de similarité qui facilite l'identification des arbres les plus prioritaires à fusionner. La fusion est réalisée moyennant un ensemble d'opérateurs.

Cet article a focalisé sur la deuxième étape de notre méthode, c'est-à-dire la modélisation multidimensionnelle qui vise à élaborer des schémas en galaxie à partir des arbres unifiés. Nous l'avons conduit en trois sous étapes : *Prétraitement des arbres*, *Construction des schémas en galaxie*, et *Validation des schémas en galaxie*. Le prétraitement améliore la lisibilité conceptuelle des arbres en les enrichissant par des cardinalités qui facilitent l'extraction des éléments multidimensionnels. La construction de schémas en galaxie identifie ces éléments, elle est fondée sur un ensemble de règles que nous avons définies. Finalement, les schémas multidimensionnels obtenus sont présentés au décideur/concepteur pour validation.

Actuellement, nous continuons à élargir l'évaluation de notre méthode sur d'autres cas. Egalement, nous sommes en cours de développement d'un prototype logiciel supportant les étapes de la méthode. Dans notre problématique, nous nous intéressons aussi à l'interrogation de l'entrepôt de documents répartis. Nous poursuivons donc nos investigations pour réunir toutes les métadonnées nécessaires à cette interrogation et nous comptons définir un langage graphique de requêtes OLAP.

Références

Ben Messaoud, I., J. Feki, et G. Zurfluh (2010). *Unification des structures des documents XML pour l'entreposage de documents.* Cinquième Atelier sur les Systèmes Décisionnels (ASD'10), pp. 1-12, ISBN 9973-9900-2-0, Sfax, Tunisie.

Ben Messaoud, I., J. Feki, K. Khrouk, et G. Zurfluh (2011). *Unification of XML document structures for Document Warehouse (DocW).* 13th International Conference on Entreprise Information Systems (ICEIS'11).

Boussaid, O., R. Ben Messaoud, R. Choquet, et S. Anthoard (2006). *Conception et construction d'entrepôts en XML*. 2ème journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA'06), Versailles, France.

Feki, J., (2004). *Vers une conception automatisé des entrepôts de données : Modélisation des besoins OLAP et génération de schémas multidimensionnels.* 8th Maghrebian Conference on

Software Engineering and Artificial Intelligence (MCSEAI'04), pp. 473-485, ISBN 9973-37-193-3, Sousse, Tunisie.

Feki, J., et Y. Hachaichi (2007). *Conception assistée de MD : Une démarche et un outil*. Journal of Decision Systems (JDS). Ed. Lavoisier, vol. 16 – No.3/2007, pp. 303-333, ISSN 1246-0125, ISBN 978-2-7462-1976-2.

Golfarelli, M., D. Maio, et S. Rizzi (1998). *Conceptual Design of Data Warehouses from E/R Schema*. Proceedings of the 31st Annual Hawaii International Conference on System Sciences (HICSS'98), IEEE Computer Society, pp. 334–343, Washington, DC, USA.

Golfarelli, M., et S. Rizzi (1999). *Designing the Data Warehouse: Key Steps and Crucial Issues*. Journal of Computer Science and Information Management 2(3), pp. 88–100.

Hachaichi, Y., J. Feki, et H. Ben-Abdallah (2010). *Modélisation multidimensionnelle de documents XML centrés-données.* Journal of Decison Systems, vol 19/3, pp. 313-345.

Júnior, C. A. S. et R. S. Mello (2008). *An ontology-driven process for unification of XML instances.* Brazilian Symposium on Multimedia and the Web, Vila Velha, Brazil, 242-249.

Kimball, R. (1997). *The Data Warehouse Toolkit*, John Wiley and Sons, Inc.

Lee, M. L., L. H. Yang, W. Hsu, et X. Yang (2002). *XClust: clustering XML schemas for effective integration*. Proc. of the ACM International Conference on Information and Knowledge Management (CIKM'02), pp. 292–299, McLean, Virginia,

McCabe, M. C., J. Lee, A. Chowdhury, D. Grossman, et O. Frieder (2000). *On the design and evaluation of a multi-dimensional approach to information retrieval*. Proceedings of the 23th Annual International ACM SIGIR Conference, pp. 363–365.

Pérez-Martinez, J. M, R. Berlanga-Llavori, M. J. Aramburu-Cabo, et T. B. Pederson (2008). *Contextualizing data warehouses with documents*. Decision Support System (DSS), Elsevier, pp. 77-94.

Ravat, F., O. Teste, T. Ronan, et G. Zurfluh (2007). *Modèle conceptuel pour l'analyse multidimensionnelle de documents*. 3ème journées francophones sur les Entrepôts de Données et Analyse en ligne (EDA'O7), Revue des Nouvelles Technologies de l'Information (RNTI), pp. 161–175, 2007.

Ronan, T. (2007). *Analyse en ligne (OLAP) de documents*. Thèse de doctorat en informatique, Université Paul Sabatier, Toulouse (France).

Sullivan, D. (2001). *Document Warehousing and Text Mining: Techniques for Improving Business Operations*. Marketing and Sales. John Wiley & Sons, Inc.

Teste, O. (2000). *Modélisation et manipulation d'entrepôts de données complexes et historisées*. Thèse de doctorat en informatique, Université Paul Sabatier, Toulouse.

Tseng, F. S. C., et A. Y. H. Chou (2006). *The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence*. Decision Support Systems (DSS), vol 42, Elsevier, pp. 727– 744.

W3C-XML (2008). *Extensible Markup Language (XML) 1.0*, http://www.w3.org/xml/.

Yoo, C. S., S. M. Woo, et Y. S. Kim (2005). *Unification of XML DTD for xml Documents with Similar Structure*. Computational Science and its Applications – ICCSA, LNCS 3482, pp. 954-963.

Summary

Nowadays, and with the large accessibility of organizations to Internet, documents constitute an interesting source for decisional analyses; they help decision makers to better understanding the evolution of their business processes. Generally, these documents exist in XML format, are geographically distributed and described by heterogeneous structures. This paper proposes a method for distributed document warehouse construction that is composed of two stages: *i) Unification of the structures of XML documents*, and *ii) Multidimensional Modeling of unified trees*. More precisely, this paper focuses on the multidimensional modeling stage.