OLAP de documents

Modélisation et mise en oeuvre

Omar Khrouf[1], **Kaïs Khrouf**[1], **Jamel Feki**[2]

1. Laboratoire MIR@CL, Université de Sfax
Route de l'Aérodrome Km 4, B.P. 1088, 3018 Sfax, Tunisie
Omar.Khrouf@yahoo.fr, Khrouf.Kais@isecs.rnu.tn

2. Faculty of Computing and IT, University of Jeddah, Jeddah, Saudi Arabia
Jamel.Feki@gmail.com

RÉSUMÉ. *Les documents constituent une capitalisation des connaissances dans les systèmes d'information des organisations. La gestion du contenu de ces documents est, depuis plusieurs années, un besoin crucial permettant d'améliorer leurs processus de prise de décisions, de renforcer le succès des activités ainsi que sa pérennité. Pour les décideurs, l'analyse du contenu de ces documents représente un véritable défi. Dans ce contexte, nous proposons un nouveau modèle multidimensionnel générique étendu à base du modèle en galaxie appelé modèle en toile d'araignée dédié à l'OLAP (On-Line Analytical Processing) de documents XML (eXtensible Markup Language). Le modèle proposé se base sur une combinaison des différentes facettes standard extraites des documents XML afin d'augmenter l'utilité des requêtes analytiques et d'offrir une vision plus claire des données aux décideurs.*

ABSTRACT. *Documents constitute a capitalization of knowledge within organizations' Information Systems. Therefore, the management of the contents of these documents represents, during several years, a crucial need allowing organizations to improve their decision-making processes, in order to enhance the success of their activities and thus their sustainability. For decision-makers, analyzing the contents of documents represents a real challenge. In this context, we propose a new and generic multidimensional model called CobWeb; it is based on the galaxy model and dedicated to the OLAP (On-Line Analytical Processing) of XML documents (eXtensible Markup Language). The proposed model relies on a combination of different standard facets extracted from XML documents in order to provide more opportunities for the expression of analytical queries and an appropriate vision of data for decision-makers.*

MOTS-CLÉS : *OLAP, document XML, facette standard, modèle multidimensionnel, OROLAP.*

KEYWORDS: *OLAP, XML document, standard facet, multidimensional model, OROLAP.*

DOI:10.3166/ISI.21.1.11-37

1. Introduction

Les systèmes d'information des organisations accumulent au fil du temps une volumétrie importante d'informations qui peuvent être extraites directement à partir de sources opérationnelles (e.g., bases de données) ou encore de leurs documents manipulés et échangés. La modélisation, l'accès et l'analyse décisionnelle de ces informations documentaires deviennent une nécessité pour toute entreprise. Face aux évolutions rapides de données, la prise de décision constitue aujourd'hui une activité primordiale et un axe de recherche important dans le domaine des systèmes d'information ce qui a nécessité la mise en place de systèmes dédiés et efficaces. L'application des techniques d'analyse en ligne OLAP (*On-Line Analytical Processing*) sur les documents hétérogènes de point de vue structure et contenu est au cœur de notre problématique.

Plusieurs travaux de la littérature se sont intéressés aux analyses OLAP de documents. Pour ce faire, certains se sont basés sur les modèles multidimensionnels classiques : modèles en étoile, en flocon de neige, et en constellation (Hachaichi *et al*., 2010 et Feki *et al*., 2013) pour les documents XML centrés-données ; (Zhang *et al*., 2009) pour les documents XML centrés-documents. D'autres travaux ont proposé des modèles multidimensionnels spécifiques comme le modèle en galaxie (Ravat *et al*., 2008) et le modèle en diamant (Azabou *et al*., 2013). Cependant, l'inconvénient majeur de ces contributions est le fait qu'elles ne se focalisent pas sur l'hétérogénéité des structures et exigent la définition à l'avance des paramètres et des hiérarchies des dimensions.

D'autres travaux, comme (Hernandez *et al*., 2008) et (Charhad et Quénot, 2004), se sont intéressés à la *multi-représentation* des documents en s'appuyant sur un ensemble de facettes. Une facette permet de décrire un aspect, soit un point de vue, parmi plusieurs d'un document ; comme la représentation de la structure des documents, l'usage associé aux documents, etc. Ces travaux ont pour objectif la représentation de l'information selon divers points de vues des utilisateurs (dites facettes du document) ; cette représentation peut dépendre soit du profil de l'utilisateur, soit de son besoin.

L'objet de ce travail est de proposer une approche permettant la modélisation multidimensionnelle de documents XML à base de facettes. Plus précisément, nous proposons un nouveau modèle multidimensionnel appelé *modèle en toile d'araignée* étendu à base du modèle en galaxie (Tournier, 2007) et dédié à l'OLAP de documents XML. Ce modèle en toile d'araignée est basé sur un ensemble de facettes *standards* qui réunissent un ensemble de données complémentaires. L'entreposage envisagé dans nos travaux est le *stockage multidimensionnel* des documents XML, dans un entrepôt de documents.

La contribution de cette recherche est la combinaison de la *multi-représentation* avec l'*OLAP de documents* ; elle nous a permis de proposer un nouveau modèle multidimensionnel *générique*, c'est-à-dire regroupant toutes les données issues des

documents et de manière indépendante de tout domaine d'application ou de collections de documents à analyser. A partir d'un tel modèle, le décideur peut exprimer plus aisément ses besoins d'analyse.

Dans la modélisation multidimensionnelle, chaque dimension a une structure composée d'un ensemble d'attributs, appelés paramètres, hiérarchisé de la granularité la plus fine (attribut racine de la dimension) vers la granularité la plus générale (par exemple, la dimension temps est composée des attributs : Jour <Mois <Trimestre <Semestre <Année ; où le symbole < se lit plus fin que). La dimension peut être considéré comme un axe d'analyse ; un paramètre représente un niveau d'analyse et peut être associée à un ou plusieurs attributs descriptifs, appelés attributs faibles.

L'idée principale de ce modèle est la transformation des facettes en dimensions. Cependant, l'intégration des facettes dans un modèle OLAP nous a confronté à un ensemble de problèmes que les modèles classiques n'ont pas envisagés comme la récursivité d'un paramètre au moment de l'interrogation et la corrélation entre les dimensions (dans le modèle multidimensionnel les dimensions sont supposées orthogonales (Eder et Koncilia, 2001). Pour pallier ces problèmes, nous suggérons un ensemble d'extensions au modèle en toile d'araignée comme la contrainte d'exclusion inter-dimension, la contrainte de réflexivité de paramètres, les dimensions corrélées, etc.

Cet article est organisé ainsi : la section 2 détaille les travaux de la littérature relatifs à la représentation et l'exploitation des facettes des documents, la modélisation multidimensionnelle et l'analyse OLAP de documents. La section 3 décrit les facettes standards de documents. Dans la section 4, nous introduisons notre modèle multidimensionnel en toile d'araignée avec ses différentes spécificités. La section 5 présente le modèle logique ainsi que ses règles de dérivation. Dans la section 6, nous présentons le corpus académique utilisé dans la phase de l'interrogation multidimensionnelle ainsi que les différents résultats obtenus. Enfin, la section 7 dévoile les développements futurs de nos travaux.

2. État de l'art

Dans cette section, nous présentons les travaux touchant la représentation et l'exploitation des facettes extraites des documents puis nous continuons à étudier les travaux dédiés à l'analyse OLAP de documents.

2.1. Représentation multi-facette de documents

La notion de facette a été initialement introduite dans le domaine de la recherche d'information (Hernandez *et al.,* 2008 et Kumar *et al.,* 2012) dans le but de représenter les documents selon différents points de vue (facettes) afin de mieux satisfaire les besoins des utilisateurs.

Pour les documents textuels, Evéquoz *et al*. (Evéquoz *et al*., 2010) ont proposé un système de navigation par facette dans une collection personnelle nommé *Weena* qui permet à l'utilisateur de gérer ses informations personnelles avec plus de flexibilité et rapidité. Dans ce cadre, ils ont conçu une interface graphique composée de trois parties : *Facettes*, *Fil d'Ariane*, et *Panneau de résultat*. Les facettes sont définies pour faciliter l'accès aux documents. Le Fil d'Ariane permet à l'utilisateur de consulter le chemin parcouru, le nombre d'éléments obtenus pour chaque facette après les opérations de filtrage. Le panneau de résultat affiche sous forme tabulaire les documents correspondant à la recherche effectuée.

Dans le cadre du domaine de l'apprentissage en ligne, (Hernandez *et al.*, 2008) ont proposé un modèle basé sur une représentation multi-facette (scénarios pédagogiques, ontologie du domaine du thème, ontologie des théories éducatives, description *LOM* et *SCORM*) de documents pédagogiques en utilisant trois ontologies : ontologie des tâches, ontologie de thème, ontologie des théories pédagogiques afin d'éliminer les ambiguïtés des termes utilisés et d'enrichir la description des documents par l'indexation sémantique. Ils ont défini deux types de facettes : une facette qui représente la *sémantique* du contenu et les autres facettes (description des théories éducatives, description par des métadonnées, la structure du document, etc.) regroupent les différents paramètres qui doivent être pris en compte pour améliorer les résultats d'une recherche de document. Le modèle proposé peut couvrir le cycle de vie du document pédagogique à travers un ensemble d'ontologies depuis sa conception jusqu'à la recherche d'information sur une notion donnée.

Cabanac *et al.* (2010) ont proposé une approche permettant l'exploitation des données relatives aux documents et aux personnes de l'organisation selon différentes vues (*Thématique*, *Usage* et *Usager*) dont chacune est composée de quatre facettes. Le point de vue *Thématique* permet de classer les documents de l'organisation suivant leurs contenus. Le point de vue *Usage* s'intéresse à la manière dont sont utilisés les documents et offre une vue différente sur les documents de l'organisation. Le point de vue *Usager* permet de connaître les différentes thématiques partagées par une personne ou groupe de personnes. Ces différents points de vue d'exploitation des informations documentaires permettent aux usagers un accès complet aux documents en navigant d'une vue à une autre et d'une facette à une autre. Cette navigation permet à un utilisateur de rassembler des informations pertinentes à partir des documents de ses collègues.

Pour les documents vidéo, *EMIR*² « Extended Model for Image representation and Retrieval » (Mechkour, 1995) est un modèle basé sur les graphes conceptuels ; il permet de décrire une image fixe à travers un ensemble de facettes regroupées en deux niveaux de description : le *niveau physique* permet de définir les caractéristiques de l'image par une matrice de pixels, et le *niveau logique* qui englobe les facettes qui décrivent le contenu de l'image (facette structurelle, spatiale, perspective et symbolique). Afin d'étendre ce modèle aux documents vidéo, Charhad et Quénot (Charhad et Quénot, 2004) proposent la version *EMIR*² incluant un ensemble de facettes spécifiques aux aspects audiovisuels. Pour atteindre

leur objectif, les auteurs proposent de nouvelles facettes classées en deux catégories : 1) Les *facettes génériques* regroupent les facettes temporelles et événementielles qui permettent de décrire les caractéristiques communes du document vidéo comme par exemple la nature événementielle de la vidéo ; 2) Les *facettes spécifiques* offrent une description du contenu vidéo (image, audio ou texte), elles permettent une description orientée média à travers deux types de facettes qualifiés de sémantique et signal.

Pour les tweets, (Kumar *et al.,* 2012) ont suggéré un système de navigation par facette intitulé *Navigating Information Facets on Twitter (NIF-T)* basé sur le web dans le but de résoudre le problème de recherche et navigation de l'information dans Twitter. Dans ce cadre, ils ont conçu une interface graphique pour organiser les informations dans des *tweets* en trois facettes : la facette *Geo* indique les emplacements des tweets dans une carte. Cette facette offre les options de recadrage et de zoom pour détecter les régions importantes associées à un événement donné. La facette *Sujet* représente un nuage de mots indiquant les différentes thématiques générées à partir des tweets. Chaque mot est affiché avec une taille de police qui reflète sa fréquence d'apparition dans les tweets. Cette facette permet de comprendre le contexte des sujets extraits à partir des tweets. La facette *Temps* associe le nombre de tweets à une date donnée. Le système de navigation proposé par les auteurs peut être étendu à d'autres facettes comme la démographie des utilisateurs, les réseaux sociaux aigus, etc.

Le tableau 1 présente une synthèse des travaux étudiés précédemment et qui sont relatifs à la multi-représentation des documents. Les lignes du tableau représentent les travaux examinés et les colonnes sont les critères d'évaluation, à savoir :

- C1 : Types de documents.
- C2 : Nombre de facettes.
- C3 : Ressource sémantique utilisée pour construire les facettes.
- C4 : Mécanisme proposé pour l'interrogation des facettes.

Concernant la représentation et l'exploitation des facettes, notons que (Hernandez *et al.,* 2008) utilisent des facettes variables suivant le domaine d'application. Par contre, les autres travaux définissent des facettes spécifiques par domaine d'application. Contrairement à ces travaux, nous proposons des facettes *standard* non limitées à un ensemble de documents prédéfinis ou un domaine d'application spécifique et nous les intégrons dans le *modèle en toile d'araignée* (cf. section 4) puisque chaque facette peut être considérée comme un support pour l'expression des besoins des utilisateurs.

Tableau 1. Comparatif des travaux de multi-représentation des documents

Critères / Approches	C1	C2	C3	C4
Charhad et Quénot, 2004	Documents vidéo	4 facettes	Ontologie de domaine	Requêtes
Hernandez *et al.* 2008	Documents textuels	Variable (selon domaine d'application)	Ontologies de tâche, de thème et des théories pédagogiques	Requêtes
Evéquoz *et al.*, 2010	Documents textuels	7 facettes	Nd	Requêtes + Navigation
Cabanac *et al.*, 2010	Documents textuels	4 facettes	Nd	Navigation
Kumar *et al.*, 2012	Tweets	3 facettes	Nd	Requêtes + navigation

Nd : non défini.

2.2. Modélisation multidimensionnelle des documents

Pour la modélisation multidimensionnelle des données factuelles (issues de sources opérationnelles de l'organisation), trois principaux modèles ont vu le jour : le modèle en étoile, le modèle en flocon de neige, et le modèle en constellation (Kimball et Ross, 2013). Pour les documents, deux catégories de travaux peuvent être distinguées : travaux ayant repris ces trois modèles, et travaux ayant proposé des modèles spécifiques.

Dans ce qui suit, nous commençons par les travaux de la première catégorie.

Tseng et Chou (Tseng et Chou, 2006) proposent trois types de dimensions appelées : *Ordinaires, Métadonnées, et Catégories*. La dimension *Ordinaire* représente des données extraites à partir du contenu des documents (e.g., la dimension Mots-clés). La dimension *Métadonnées* représente les métadonnées extraites du standard Dublin Core (Dublin Core, 2012). La dimension *Catégories* regroupe un ensemble de données externes du document définies par les utilisateurs selon leur point de vue.

Boussaid *et al.* (Boussaid *et al.*, 2006) ont proposé une modélisation en flocon de neige des données multidimensionnelles XML avec des méthodes de fouille de données. Plus précisément, ils ont défini une approche appelée *X-Warehousing* qui permet de décrire le modèle logique d'un cube XML. Ce modèle implique beaucoup de redondances puisqu'il faut dupliquer les données des dimensions pour chaque mesure du fait (sujet d'analyse), ce qui peut entraîner des difficultés de mise à jour et de maintenance.

Zhang *et al.* (Zhang *et al.*, 2009) ont proposé un nouveau modèle basé sur un schéma en étoile intitulé *Topic Cube* qui permet d'étendre le cube de données traditionnel en intégrant une hiérarchie de thèmes *Topics* construite à partir des données reflétant une ontologie de domaine et adaptée aux préférences de l'analyste.

Oukid *et al.* (Oukid *et al.*, 2015) ont présenté un nouveau modèle appelé *CXT-Cube*, associé à des dimensions contextuelles. Chaque dimension est liée à un facteur contextuel correspondant à une définition de contexte dans l'entrepôt. Dans *CXT-Cube*, les dimensions contextuelles peuvent être soit sémantiques (où les attributs sont regroupés dans une hiérarchie de concepts, extraite à partir d'une ontologie de domaine utilisée comme une ressource externe), soit métadonnées (des informations externes concernant les documents, tels que: date, titre, auteur, etc.).

Aknouche *et al.* (Aknouche *et al.*, 2013) ont proposé une nouvelle architecture décisionnelle pour l'entreposage des données textuelles basé sur les techniques de recherche d'information (RI) et les tâches d'entreposage des données classiques. Cette architecture décisionnelle comporte un processus intitulé *ETLText* (Extract-Transform-Load-Text) pour intégrer les données textuelles dans un système décisionnel. Les auteurs ont proposé un nouveau modèle multidimensionnel intitulé *TWM* (*Text Warehousing Model*) pour prendre en compte la complexité des données textuelles.

D'autres travaux utilisent une dimension textuelle composée de mots-clés résumant le document, comme dans (Lin *et al.*, 2008) où les auteurs ont proposé un cube de textes nommée *Text Cube*, basé sur un schéma en étoile et dans lequel une dimension textuelle est représentée par une hiérarchie de termes. Récemment, Bautista *et al.* (Bautista *et al.*, 2013) ont proposé un modèle multidimensionnel qui prend en charge les informations textuelles en introduisant une dimension textuelle *AP-Dimension* obtenue à l'aide d'une structure sémantique appelée *AP-Structure*. Cette structure représente le sens caché derrière le texte au lieu d'un simple jeu de mots.

Pour les travaux de la deuxième catégorie, c'est-à-dire ceux qui ont proposé des modèles spécifiques, nous trouvons principalement les suivants.

Ravat *et al.* (Ravat *et al.*, 2008) ont proposé un modèle multidimensionnel intitulé *modèle en galaxie* adapté à l'analyse de documents XML orientés-documents. Ce modèle se caractérise par l'unique concept « Dimension » qui est susceptible de jouer à la fois le rôle d'axe et de sujet d'analyse. Une galaxie est un ensemble de dimensions liées entre elles par un ou plusieurs nœuds centraux ; chaque nœud modélise les dimensions compatibles c'est-à-dire qui peuvent être utilisées ensemble dans une même analyse.

Azabou *et al.* (Azabou *et al.*, 2015) proposent *un modèle en diamant* qui étend le modèle en galaxie par une dimension centrale traduisant la *sémantique* des contenus textuels d'un ensemble de documents ayant des structures identiques ou similaires. Le modèle en diamant est composé de deux types de dimensions dites *Sémantique* et

Classiques. La dimension *Sémantique* occupe un emplacement central ; elle est composée de la hiérarchie suivante : Concept <Taxonomie. Les paramètres de cette dimension seront reliés aux paramètres des autres dimensions. Les dimensions Classiques sont les axes d'analyse constitués des éléments du premier niveau de la structure générique des documents. Pour chaque élément, ses descendants constituent les paramètres (organisés sous forme de hiérarchies) et les attributs faibles.

Le tableau 2 présente une synthèse des travaux traitant la modélisation multidimensionnelle des documents, selon les critères suivants :

— C1 : La modélisation multidimensionnelle basée sur le modèle multidimensionnel de base (MMB) : modèle en étoile, en flocon de neige ou en constellation ; ou bien sur un nouveau modèle multidimensionnel (NMM).

— C2 : Format de documents.

— C3 : Schéma fixe par besoin d'analyse (toutes les dimensions et leurs paramètres sont connus à l'avance).

— C4 : Collections de documents ayant des structures similaires (même DTD ou schéma XML).

Tableau 2. Tableau comparatif des travaux de modélisation multidimensionnelle des documents

Critères / Approches	C1		C2	C3	C4
	MMB	NMM			
Tseng et Chou (2006)	Modèle en étoile		XML	Oui	Nd
Boussaid *et al.* (2006)	Modèle en flocon		XML	Oui	Similaire
Lin *et al*. (2008)	Modèle en étoile		Documents textuels	Oui	–
Ravat *et al.* (2008)		Modèle en galaxie	XML	Oui	Similaire
Zhang *et al*. (2009)	Modèle en étoile		Documents textuels	Oui	–
Bautista *et al.* (2013)	Modèle en étoile		Documents textuels	Oui	–
Aknouche *et al.* (2013)	Modèle en constellation		Documents textuels	Oui	–
Oukid *et al*. (2015)	Modèle en étoile		XML	Oui	Nd
Azabou *et al.*, 2015	Modèle en galaxie		XML	Oui	Similaire

Nd : non défini. (–) : les auteurs n'utilisent pas le critère de synthèse (C4) dans leur approche.

En conclusion, la majorité des travaux traitant de la modélisation multidimensionnelle de documents présentés ici s'intéressent essentiellement au contenu ; à l'exception de (Ravat *et al.*, 2008 et Azabou *et al.*, 2015) qui ont traité aussi l'aspect structurel. Cependant, ces deux travaux traitent les documents ayant des structures identiques ou similaires (même DTD ou schéma XML). Complémentairement à ces travaux, nous proposons une modélisation multidimensionnelle des documents *hétérogènes* de point de vue structure et contenu ; il s'agit d'un nouveau modèle multidimensionnel appelé *modèle en toile d'araignée.*

2.3. OLAP de documents

Dans cette section, nous nous intéressons aux travaux de la littérature traitant l'OLAP de documents.

Park *et al.* (Park *et al.*, 2005) ont proposé un nouveau cadre pour l'analyse multidimensionnelle des documents XML centrés-documents appelé *XML-OLAP*. Les auteurs ont défini un nouveau langage pour les expressions multidimensionnelles intitulé *XML-MDX*. Ils ont utilisé le langage *MDX* de Microsoft qu'ils ont adapté à XML pour construire des cubes XML, et le langage XQuery pour la définition des dimensions et le calcul des mesures. Ce travail suggère l'emploi des fonctions d'agrégation comme *TOPIC* pour l'extraction du sujet d'analyse, *CLUSTERING* pour répartir les textes en fonction de leur contenu, *TOP KEYWORD* pour extraire les mots-clés les plus importants, *SUMMARY* pour générer un résumé du texte, etc.

Oukid *et al.* (Oukid *et al.*, 2015) propose pour *CXT-Cube* une mesure d'analyse textuelle qui s'appuie à la fois sur un modèle vectoriel adapté à l'analyse OLAP et sur une technique de propagation de pertinence. Il est également associé à un nouvel opérateur d'agrégation appelé *ORank* (Olap-Rank) permettant d'agréger les données textuelles dans un environnement OLAP.

Ravat *et al.* (Ravat *et al.*, 2008) proposent pour le modèle en galaxie deux fonctions d'agrégation : *AVG_KW* qui permet de regrouper un ensemble de mots-clés en utilisant une ontologie de domaine et *TOP_KW* qui renvoie les *k* principaux mots-clés ayant les plus grands poids dans le document. La pondération de ces mots-clés est effectuée à l'aide la méthode *Tf-Idf.*

Zhang *et al.* (Zhang *et al* 2009) proposent deux mesures probabilistes pour le *Topic Cube* : la distribution d'un mot dans un thème *p(wi|topic),* et la couverture d'un thème par les documents notée *p(topic|dj)* qui est la probabilité qu'un document *dj* couvre le topic. Ainsi, le modèle est capable de prévoir quel est le sujet dominant dans l'ensemble de documents en agrégeant la couverture sur tous les documents.

Pour Lin *et al.* (Lin *et al.*, 2008), la dimension textuelle représentée par une hiérarchie de termes spécifie les relations sémantiques entre ces termes extraits des

documents, ce qui permet une navigation sémantique grâce à deux opérateurs associés : *pull-up* et *push-down*. Chaque terme extrait devient un nœud de bas niveau dans la hiérarchie ; un nœud parent se compose de tous les enfants à un niveau inférieur. Les auteurs définissent aussi dans leur cube deux mesures d'agrégation adaptées aux données textuelles : fréquence des termes *TF* et l'index inversés *IV*.

Azabou *et al.* (Azabou *et al.*, 2015) présentent de nouvelles fonctions permettant l'agrégation de données textuelles au sein d'un environnement OLAP. *List_Concept* agrège un ensemble de *n* concepts en un sous ensemble de *k* concepts les plus représentatifs. *G_Concept* et *S_Concept* extraient respectivement le ou les concepts les plus génériques ou spécifiques. Enfin, ils proposent la fonction *Top_Concept* qui regroupe les fonctions *List_Concept*, *G_Concept* et *S_Concept* dans le but d'afficher le premier concept obtenu avec chacune de ces fonctions.

Pour Bautista *et al.* (Bautista *et al.*, 2013), l'utilisation de l'*AP-Structure* (qui représente le sens caché derrière le texte) enrichit l'analyse des données textuelles de sorte que l'utilisateur peut introduire la sémantique des attributs textuels dans les requêtes.

Nous synthétisons les travaux de la littérature relatifs à l'OLAP de documents dans le tableau 3.

Tableau 3. Comparaison des travaux traitant l'OLAP de documents.

Critères / Approches	Nouveaux opérateurs OLAP : fonctions d'agrégation (FA) ou opérateur de navigation (ON)		Manipulation multi-dimensionnelle	Mesures textuelles
	FA	ON		
Park *et al.*, 2005	7 FA (Topic, Summary, etc.)	Non	Nd	Nd
Lin *et al.*, 2006	NON	2 ON (pull-up et push-down)	Sémantique	TF et IV
Ravat *et al.*, 2008	2 FA (AVG_KW et TOP_KW)	Non	Sémantique & structurelle	AVG_KW et TOP_KW
Zhang *et al.*, 2009	Non	Non	Sémantique	P (wi|topic) et P (topic|dj)
Bautista *et al.*, 2013	Non	Non	Sémantique	Non
Oukid *et al.*, 2015	Non	1 ON (ORank)	Sémantique	Nd
Azabou *et al.*, 2015	4 FA (Top_Concept, List_Concept, etc.)	Non	Sémantique & structurelle	Non

Nd : non défini.

Par ailleurs, le problème de visualisation des résultats OLAP de documents n'est pas encore suffisamment traité, nous proposons l'affichage du résultat d'une requête multidimensionnelle sous forme d'un nuage de mots pour mettre en évidence les termes ou concepts les plus importants : la taille d'un mot dans le nuage sera proportionnelle à sa fréquence d'apparition. De plus, nous suggérons un nouvel opérateur OLAP appelé *COR_DIM* permettant la corrélation entre les dimensions pour la même requête OLAP.

La suite de cet article sera consacrée à notre contribution dans la modélisation conceptuelle ; il présente un modèle permettant de décrire la vision du décideur indépendamment de toute technique d'implantation. Ensuite, nous décrivons la modélisation logique que nous proposons. Enfin, nous présentons la modélisation physique qui décrit les techniques utilisées pour l'implantation du modèle logique avec ses différentes spécificités (la contrainte d'exclusion entre les dimensions, la contrainte de paramètres réflexifs, etc.). Comme notre contribution est fondée sur le concept de facette, nous commençons par introduire ce concept.

3. Notion de facette de documents

A ce que nous connaissons, la notion de facette n'a pas été utilisée dans la modélisation dimensionnelle. Nous l'exploitons pour proposer un ensemble de cinq facettes standards, c'est-à-dire indépendantes de tout domaine d'application spécifique (Khrouf *et al.,* 2013) dont chacune décrit un point de vue utile à l'interrogation des documents.

Ces facettes offrent au décideur la possibilité de consulter un ensemble de documents selon plusieurs points de vues (métadonnées, mots-clés, etc.) afin d'avoir un accès plus ciblé à l'information selon ses besoins. La figure 2 présente un exemple de ces différentes facettes pour un document XML.

– La facette *Contenu* : Cette facette permet l'accès au contenu proprement dit (texte, image, vidéo, etc.) du document. Elle a pour rôle de mettre en évidence l'information véhiculée par le document, en éliminant tout ce qui concerne la structuration, la présentation, les commentaires, etc.

– La facette *Structurelle* : Cette facette permet de définir une vue globale de la structure d'un document ou d'un ensemble de documents homogènes. Il s'agit plus précisément de l'arborescence du document. Cette facette a pour rôle de se focaliser sur des parties du document et non sur sa totalité.

– La facette *Métadonnée* : Elle représente un ensemble structuré des données décrivant un document. Dans nos travaux, nous utilisons les métadonnées définies par le *Dublin Core* (Dublin Core, 2012) comme le titre, l'éditeur, le format, les droits, etc.

– La facette *Sémantique* : Cette facette indique la sémantique véhiculée dans le contenu textuel d'un document XML. Pour la détermination de cette sémantique, nous ré-exploitons les travaux de (Ben Mefteh *et al.,* 2013) qui proposent une

approche d'extraction automatique de la structure sémantique d'un document XML. Il s'agit d'une structure superposée à la structure logique du document et dont les nœuds décrivent des concepts issus du contenu. La figure 1 présente un exemple de structure sémantique.

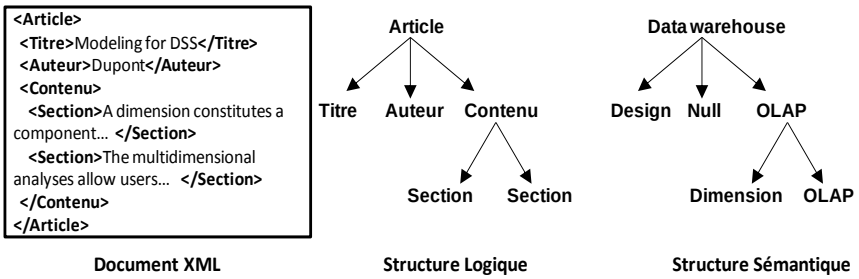

Figure 1. Exemple de document avec ses structures logique et sémantique

La structuration sémantique des documents se base sur quatre étapes principales, à savoir : 1) la détermination de la taxonomie[1] qui décrit la sémantique du document, 2) l'association d'un concept à chaque élément feuille de la structure logique du document. Il s'agit de chercher, dans la taxonomie retenue à la phase précédente, le concept le plus approprié à la description de la sémantique de l'élément feuille et ceci en tenant compte de ses mots-clés qui le décrivent, 3) l'inférence des concepts aux éléments non feuilles du document, et 4) les balises qui représentent des métadonnées (e.g., *Titre*, *Auteur*) dans la structure logique et qui seront remplacées par la sémantique qu'elles présentent ou par la valeur *NULL* si elles n'ont aucune sémantique.

– La facette *Mot-clé* : Cette facette représente les mots-clés les plus pertinents qui décrivent le contenu d'un ou de plusieurs documents. Ces mots-clés sont déterminés en utilisant les techniques d'indexation du domaine de la recherche d'information.

Les facettes que nous proposons sont standards, c'est-à-dire non limitées à un ensemble de documents prédéfinis ou de même structure, ce qui confère au modèle un potentiel analytique assez large. Ces facettes contiennent toutes les informations qui décrivent un document XML (mots-clés, métadonnées, sémantique) et qui sont utiles au décideur pour satisfaire ses besoins analytiques OLAP.

En se basant sur les facettes précédemment citées, nous présentons, dans ce qui suit, un nouveau modèle multidimensionnel dédié à l'OLAP de documents XML, que nous appelons modèle en toile d'araignée. Nous commençons par la

1. Une taxonomie est une ressource sémantique permettant la représentation hiérarchique de ses concepts.

modélisation conceptuelle. Ensuite, nous décrivons la modélisation logique. Enfin nous présentons la modélisation physique ainsi que le processus d'interrogation.

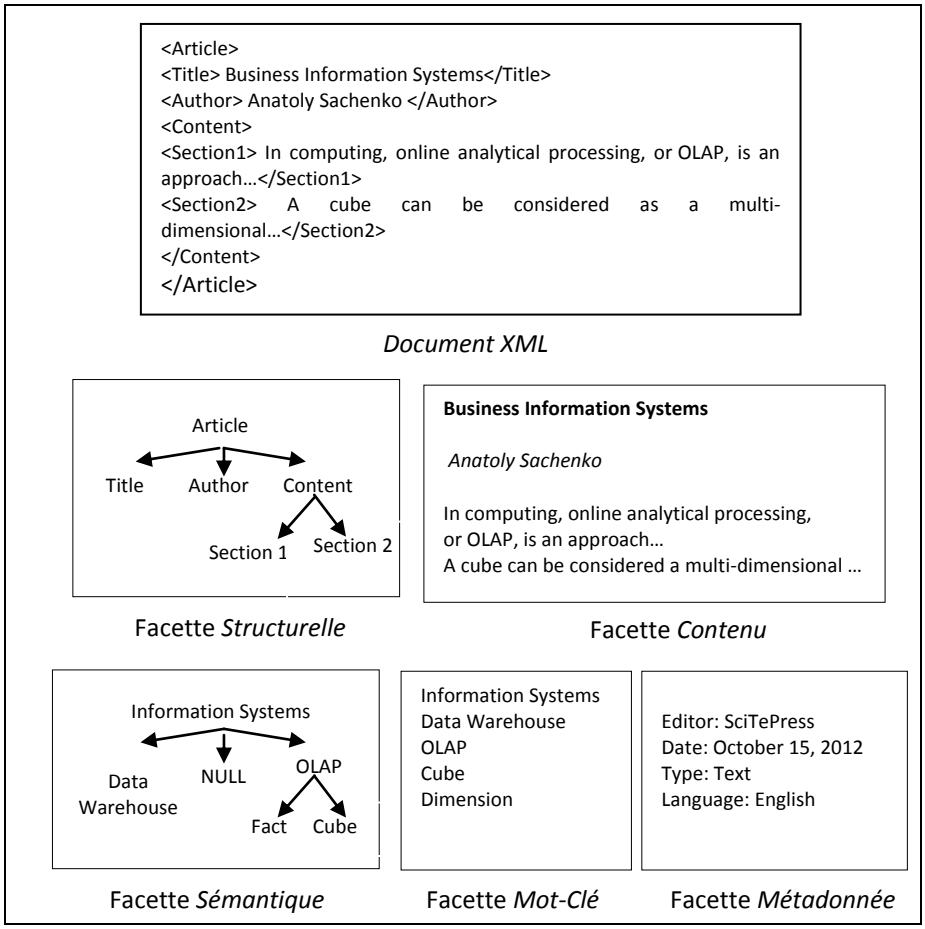

Figure 2. Les cinq facettes du document XML ci-dessus

4. Modélisation conceptuelle : modèle multidimensionnel en toile d'araignée

L'idée principale de notre modèle en toile d'araignée est de transformer chacune des facettes définies dans la section précédente en une dimension puisque ces facettes peuvent représenter un moyen d'expression de besoins pour les décideurs. En effet, elles regroupent différentes informations et métadonnées concernant les documents à analyser.

Nous considérons que le fait correspond à une observation sur les documents, il est décrit par des mots-clés, possède une sémantique, etc. Au moment de l'interrogation, une dimension pourrait jouer le rôle d'un fait en appliquant la fonction d'agrégation sur ses valeurs ; c'est pour cette raison que nous utilisons le modèle en galaxie qui repose sur l'unique concept de dimension ; la notion de fait est non explicitée. Une galaxie regroupe un ensemble de dimensions compatibles (qui peuvent être utilisées ensemble dans une même analyse) par un nœud central. La figure 3 présente un exemple d'un modèle en galaxie.

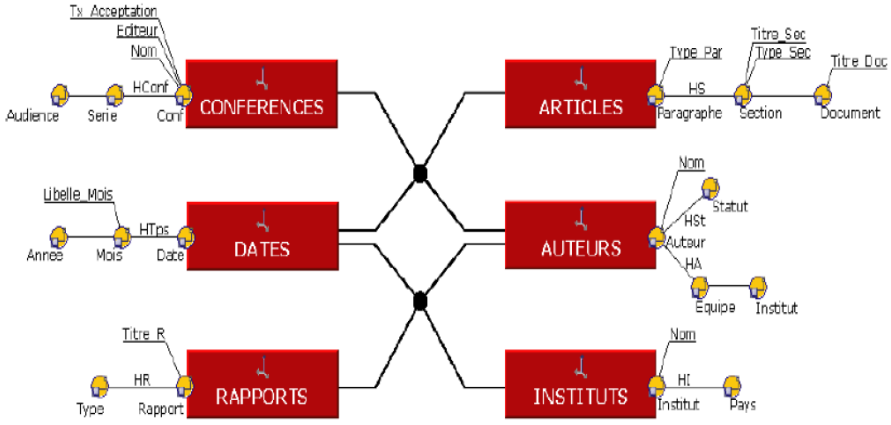

Figure 3. Exemple d'un modèle en galaxie (Tournier, 2007)

Tournier (Tournier, 2007) présente un modèle multidimensionnel par collection de documents structurellement homogènes; ce qui présente une restriction aux décideurs s'ils souhaitent analyser des données provenant de différentes collections. Pour remédier à cette limite, nous présentons un nouveau modèle *multidimensionnel* dite en *toile d'araignée* permettant l'analyse de documents hétérogènes (de point de vue structure et contenu) issues de plusieurs collections. Le modèle que nous proposons est *générique*, permettant ainsi de regrouper toutes les données (factuelles et textuelles) extraites des documents (Khrouf *et al.*, 2014).

La figure 4 présente notre modèle multidimensionnel en toile d'araignée, étendu à base du modèle en galaxie. Il est composé d'un seul nœud central reliant les six dimensions nommées :

- *D_Structurelle* : décrit les structures logiques des documents.
- *D_Contenu* : présente le contenu textuel des documents.
- *D_Métadonnée* : modélise les métadonnées définies dans le *Dublin Core* (*Dublin Core*, 2012). Nous citons dans la figure 4 un extrait de ces métadonnées.
- *D_MotClé* : contient les mots-clés les plus pertinents des documents.

- *D_Sémantique* : décrit la sémantique des documents par référence à des taxonomies.
- *D_Document* : Cette dimension relie les différentes informations issues des facettes (modélisées sous forme de dimensions) à leurs documents.

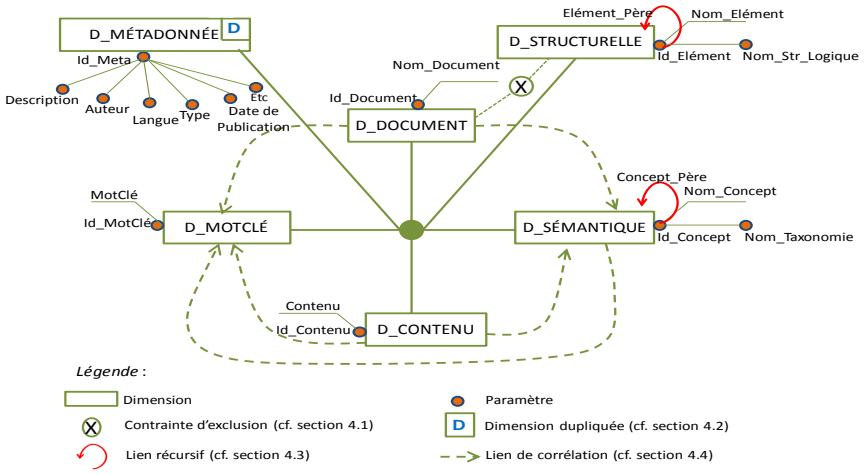

Figure 4. Modèle en toile d'araignée

La figure 5 est un exemple d'instanciation. Pour des raisons de clarté, nous instancions uniquement la balise *Title* du document XML de la figure 2.

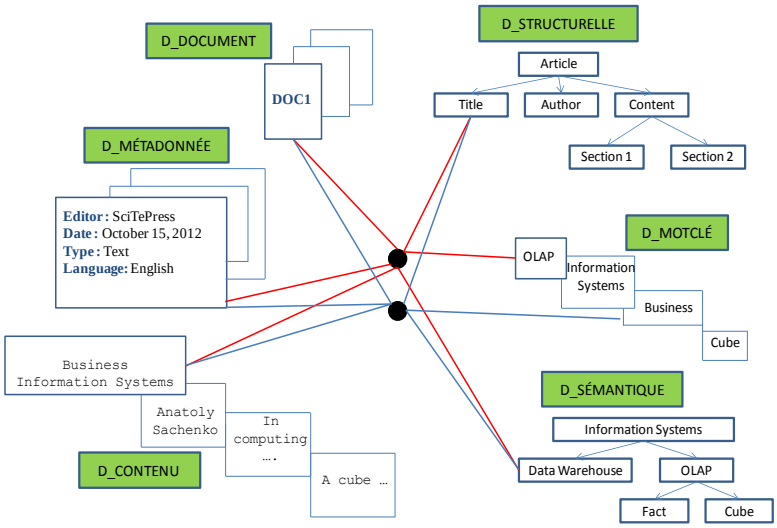

Figure 5. Exemple d'instanciation du modèle en toile d'araignée

REMARQUE.– Pour éviter la duplication du nœud central de la balise *Title* avec chacun des mots-clés correspondants, nous proposons une modélisation logique appropriée qui sera décrite dans la section 5.

Le modèle en toile d'araignée se caractérise par quatre contraintes qui lui sont spécifiques que nous décrivons ci-après.

4.1. Contrainte exclusion entre dimensions

La contrainte d'exclusion exprime la sémantique du fait que deux dimensions ne peuvent pas participer simultanément à une même analyse OLAP de documents.

Dans notre modèle multidimensionnel, la contrainte d'exclusion concerne les deux dimensions appelées *D_ Document* et *D_Structurelle*. Ces deux dimensions ne peuvent pas participer en même temps à une même analyse par document et par partie de ces documents (Titre, Section, Paragraphe, etc.).

Graphiquement, cette contrainte d'exclusion est montrée dans la figure 4 par un cercle contenant **X** relié aux dimensions concernées, à savoir : *Document* et *Structurelle*.

Dans la littérature, d'autres types de contraintes ont été proposées telles que des contraintes entre les hiérarchies d'une même dimension ou appartenant à des dimensions distinctes (Ghozzi, 2003). Cependant, ces contraintes ne convergent pas avec nos travaux.

4.2. Contrainte dimension dupliquée

Une dimension dupliquée est une dimension utilisée deux ou plusieurs fois dans la même analyse (requête multidimensionnelle). Ce genre de dimension est rarement utilisé dans la modélisation classique (telle que le modèle en étoile).

Comme exemple de dimension dupliquée, nous souhaitons analyser les produits achetés et vendues dans la même année. Ainsi, la dimension *Date* doit jouer le rôle de *Date d'achat* et de *Date de vente*. Une telle requête nécessite un niveau d'expertise élevé pour les décideurs, qui sont généralement non informaticiens.

Dans nos travaux, nous disposons d'une dimension spécifique *Métadonnée* ; elle se caractérise par un ensemble de paramètres complémentaires qui peuvent être utilisés à la fois dans la même requête, tels que : *Auteur*, *Langue*, *Date de publication*.

Pour mettre en évidence ce type de dimensions, nous proposons de les annoter par la lettre **D** (cf. figure 4).

4.3. Contrainte paramètre récursif

Dans les schémas conceptuels classiques des entrepôts de données, les hiérarchies des dimensions et leurs paramètres sont connus à l'avance et sont figés. Cependant, dans notre modèle multidimensionnel :

– La structure des documents peut différer d'une collection à une autre.

– La structure sémantique d'un document permet de décrire son contenu textuel ; elle est déduite à partir de son contenu en le projetant sur une taxonomie. Plus précisément, des concepts taxonomiques seront affectés aux différentes parties des documents. Notons que le nombre de concepts et de niveaux varient d'une structure sémantique à une autre.

Pour représenter les dimensions *Sémantique* et *Structurelle*, nous introduisons un nouveau type de paramètre, dit paramètre *Récursif* car les documents et les structures sémantiques utilisés dans notre modèle sont représentés d'une manière hiérarchique multi-niveaux.

— Pour la dimension structurelle, on peut passer ainsi d'un niveau à un autre, par exemple selon le chemin : Paragraphe <Sous-Section <Section <Contenu.

— Pour la dimension sémantique, on peut passer d'un concept spécifique vers un autre plus générique, par exemple selon le chemin : Cube <Entrepôt de données <Système d'information.

Graphiquement, un paramètre récursif est schématisé par une boucle sur le paramètre. Dans notre modèle multidimensionnel, nous disposons de deux paramètres récursifs, à savoir *IdConcept* et *IdElément* (cf. figure 4).

4.4. Contraintes Dimension corrélée

Les opérateurs multidimensionnels « *Drill-Down* » et « *Roll-Up* » permettant respectivement de détailler ou d'agréger les niveaux d'analyse OLAP et s'appliquent à des paramètres appartenant à une même hiérarchie. Cependant, dans l'OLAP de documents, il serait intéressant par exemple de passer de la facette *Contenu* aux concepts de la facette *Sémantique* ou bien du *Contenu* aux *Mots-clés*.

Dans la modélisation multidimensionnelle classique, ces opérations ne sont pas réalisables en raison de la contrainte d'orthogonalité des dimensions (absence de relations inter-dimensionnelles). Face à cette problématique, nous proposons le concept *Dimension corrélée* pour indiquer les dimensions en relation et permettre ainsi le passage d'une dimension à une autre pour la même requête OLAP.

Graphiquement, la corrélation entre les dimensions de notre modèle multidimensionnel est schématisée par des flèches entre les dimensions (cf. figure 4). Le passage d'une dimension à une autre n'est accepté que si on respecte le sens de la flèche.

5. Modélisation logique

Pour notre *modèle en toile d'araignée* nous avons opté pour la modélisation logique *OROLAP* que nous présentons dans un premier temps, ensuite nous décrivons les règles de passage du modèle conceptuel multidimensionnel au modèle logique *OROLAP*. Enfin, nous présentons le modèle logique obtenu.

5.1. Modèle OROLAP (Object-Relational OLAP)

Dans la littérature, trois modèles logiques pour les systèmes OLAP ont été proposés : Les modèles *ROLAP* (Relational OLAP) qui se basent sur un *SGBD* relationnel en transformant les concepts dimensionnels de base (fait, dimension) en des tables relationnelles. Les modèles *MOLAP* (Multidimensional OLAP) utilisant un *SGBD* Multidimensionnel (SGBDM) capable de stocker et traiter des données multidimensionnelles et les représenter sous forme des cubes de données, des matrices ou des vecteurs à *n* dimensions. Ces modèles optimisent les temps d'accès aux données et réduisent les temps de réponse des requêtes. Les modèles *HOLAP* (Hybrid OLAP) qui englobent un modèle *ROLAP* et un modèle *MOLAP*. Plus précisément, il utilise un *SGBD* Relationnel pour stocker, gérer les données détaillées et un *SGBD* Multidimensionnel pour stocker, gérer les données agrégées afin de gérer de très grandes quantités de données et optimiser les temps de réponse des analyses OLAP. Notons que le modèle *ROLAP* est le plus utilisé en décisionnel puisqu'il repose sur un nombre limité de tables.

Cependant, les modèles *ROLAP* exigent que le fait soit relié à chacune de ses dimensions par une clé étrangère. Plus précisément, à chaque valeur d'une clé étrangère du fait correspond une seule ligne de la dimension associée. Dans nos travaux, le nœud central peut concerner plusieurs mots-clés ; il sera dupliqué ainsi autant de fois que de mots-clés (cf. figure 5). Pour éviter ce problème, nous proposons l'*OROLAP* (*Object-Relational* OLAP). Ce modèle se caractérise par l'existence des liens monovalués (référencent une seule ligne) et des liens multivalués (pour référencer plusieurs lignes) de la table du nœud vers la table de dimension. A titre d'exemple, la figure 6 schématise un lien monovalué de la table *NŒUD* vers la table *DDOCUMENT* et un lien multivalué de la table *NŒUD* vers la table *DMOTCLE*.

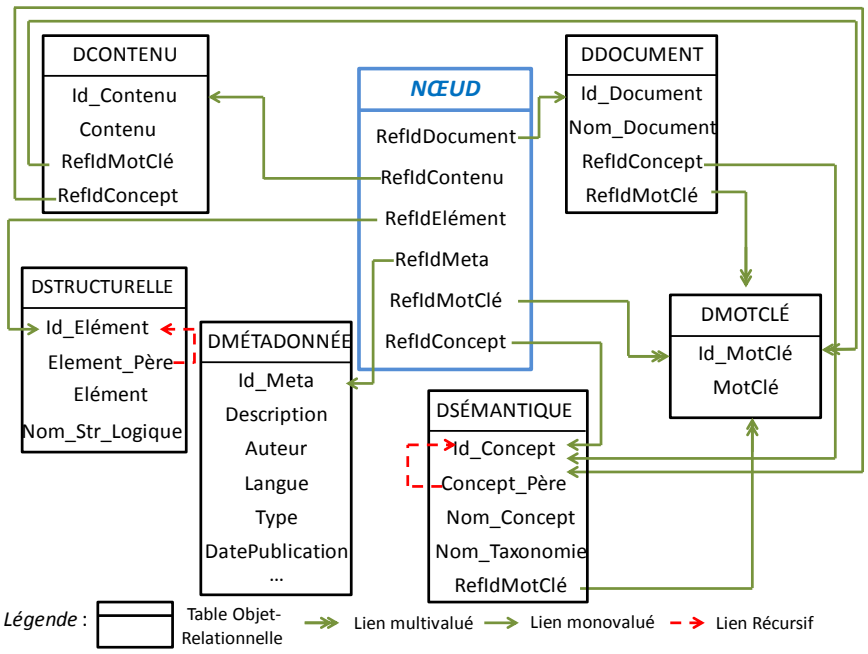

Figure 6. Modèle logique OROLAP

5.2. Règles de transformation

Pour les besoins de mise en œuvre, le modèle conceptuel multidimensionnel en toile d'araignée se transforme en modèle logique *OROLAP*. Pour ce faire, nous définissons les quatre règles suivantes :

– Chaque dimension *d* se transforme en une table objet-relationnelle composée d'un ensemble d'attributs obtenu par l'union des paramètres et les attributs faibles de toutes les hiérarchies de *d*. La clé primaire de la table construite à partir de *d* est l'attribut appartenant au niveau de granularité le plus fin.

– Chaque contrainte de paramètre récursif se transforme en un lien monovalué récursif sur l'attribut représentant ce paramètre.

– Chaque corrélation entre deux dimensions se transforme en un lien multivalué si on référence la dimension Mots-clés et en un lien monovalué pour les autres dimensions (*Document, Sémantique, Contenu,*).

– Le nœud *n* se transforme en une table objet-relationnelle composée d'un lien multivalué vers la dimension Mots-clés et d'un lien monovalué vers chaque dimension (*Document, Sémantique, Structurelle, Contenu, Métadonnée*). La clé primaire de *n* est formée par la concaténation des attributs représentant les liens monovalués et multivalués.

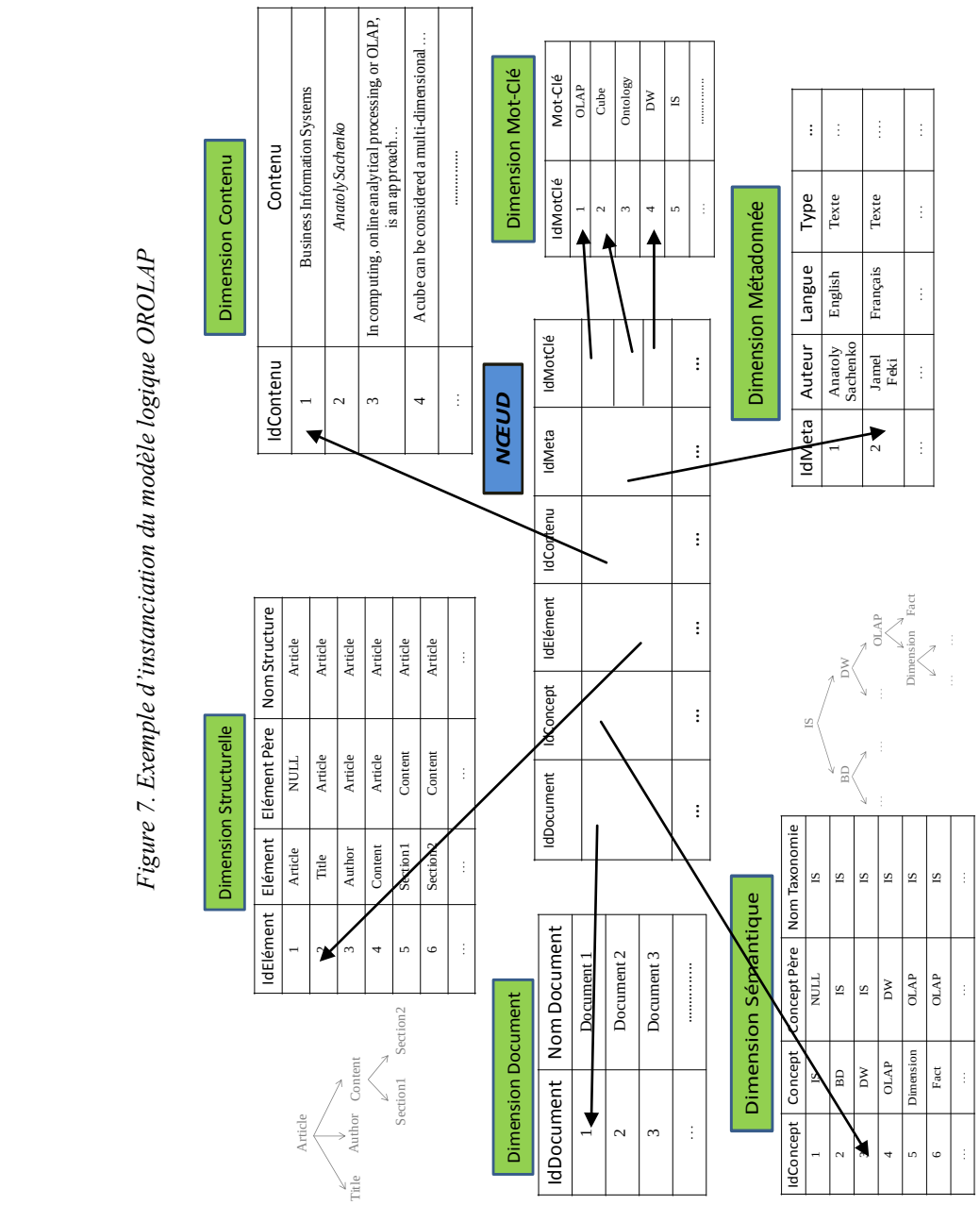

Figure 7. Exemple d'instanciation du modèle logique OROLAP

5.3. Modèle logique OROLAP

Le résultat de transformation du *modèle en toile d'araignée* en un modèle logique conformément à ces règles est présenté dans la figure 6. Le formalisme utilisé est celui de (Soutou *et al.*, 1999).

Ainsi, nous obtenons une ligne dans la table *NŒUD* par balise (une ligne pour *Title*, une ligne pour *Author*, etc.). La figure 7 décrit la ligne *Title* pour le document XML (cf. figure 2), elle correspond au document *Document1*, à la structure *Article*, au concept *DW* (*Data Warehouse*), et elle contient les métadonnées (*Auteur*, *Langue* et *Type*). Cette ligne de fait contient un lien multivalué vers les trois Mots-clés *OLAP, Cube et DW*.

Ce modèle logique regroupe toutes les données issues des documents à analyser. A ce niveau, le décideur peut exprimer sa requête multidimensionnelle en spécifiant le fait (une des six dimensions de notre modèle) et les axes d'analyse (parmi les autres dimensions). Deux exemples de requêtes multidimensionnelles seront traités dans la section suivante.

6. Expérimentations

Pour valider nos propos, nous proposons d'évaluer notre approche par un ensemble d'expérimentations sur un corpus académique composé de 250 documents XML. Nous avons construit manuellement ce corpus afin de couvrir et tester les différentes extensions que nous avons définies dans le modèle en toile d'araignée.

Ravat *et al.* (2010) distinguent deux types de documents : *les documents XML orienté-données* et les *documents XML orienté-documents*.

– Les *documents XML orienté-données* sont constitués d'un ensemble d'éléments généralement courts et précis et sont similaires aux données relationnelles.

– Les *documents XML orienté-documents* sont riches en texte et constituent des versions électroniques des documents papier (e.g., articles scientifiques, rapports internes).

La figure 8 montre un extrait du corpus reconstitué de documents XML orienté-documents.

```
<?xml version="1.0" encoding="UTF-8"?>
<Article>
   <Title>CobWeb Multidimensional Model: From Modeling to
   Querying</Title>
   <Authors>
      <Author>Omar Khrouf</Author>
      <Author>Kais Khrouf</Author>
      <Author>Jamel Feki</Author>
   </Authors>
```

```
   <Conference>
      <Name> International Conference on Model & Data Engineering
      (MEDI) </Name>
      <Date>2014</Date>
   </Conference>
   <Abstract>Nowadays, the information can be derived from their
operational sources… </Abstract>
   <Keywords>XML documents, Facets, OROLAP </Keywords>
</Article>
```

Figure 8. Exemple de document XML extrait du corpus académique

Nous avons implanté le modèle logique *OROLAP* sous Oracle 10g en utilisant les attributs de type *REF* pour les liens monovalués et les tables imbriquées *(nested tables)* pour les liens multivalués.

Par la suite, nous avons instancié le modèle physique et testé un ensemble de requêtes multidimensionnelles afin de vérifier et de valider notre modèle en toile d'araignée.

Pour l'interrogation multidimensionnelle, le décideur spécifie sa requête en indiquant le fait avec la ou les mesures et les différentes dimensions. Suite à cela, le système génère automatiquement les requêtes intermédiaires nécessaires. Nous distinguons deux types de requêtes :

– Les requêtes ne nécessitant que les liens monovalués (sans la dimension Mots-clés). Nous utilisons ainsi les jointures implicites en utilisant les points (.).

Exemple : Supposons que nous souhaitons analyser l'année de la dernière publication par auteur, langue et thématique (dimension sémantique). Le système génère la requête suivante à partir des données stockées dans les différentes tables de dimensions.

```
SELECT      N.RefIdMeta.Createur, N.RefIdMeta.Langue,
            N.RefIdConcept.NomConcept,
            Max (TO_CHAR(N.RefIdMeta.DatePublication,
            'YYYY'))
FROM        Nœud N;
GROUP BY    F.RefIdMeta.Createur, F.RefIdMeta.Langue,
            F.IdConcept.NomConcept;
```

– Les requêtes utilisant aussi les liens multivalués (intégration de la dimension Mots-clés). Pour ces requêtes, nous utilisons l'opérateur THE[2].

2. Nous avons utilisé la version Oracle 10g.

Exemple : Supposons que nous souhaitons sélectionner les mots-clés par auteur, par année de publication et par éditeur.

```
SELECT     N.RefIdMeta.Createur, N.RefIdMeta.Date,
N.RefIdmeta.Editeur, Nt.RefMct.Motcle
FROM       Nœud N, THE(SELECT M.IdMotCle
FROM Nœud M
WHERE M.RefIdContenu.IdContenu=
N.RefIdContenu.IdContenu)Nt;
```

Nous avons développé un outil logiciel appelé *IMDF « Interrogation Multidimensionnelle de Documents à base de Facettes »* ; il comprend trois modules :

– Module d'extraction : c'est un analyseur syntaxique (parseur) permettant l'extraction des données issues de documents XML qui représentent les différentes facettes.

– Module d'intégration : il permet de générer automatiquement les scripts SQL pour le stockage multidimensionnel des données extraites.

– Module de visualisation : il s'agit d'une interface graphique conviviale pour l'interrogation multidimensionnelle.

Dans le volet gauche de l'interface (cf. figure 9), sont spécifiées les dimensions et le fait d'analyse. Le volet droit est consacré au résultat de la requête présentée sous forme d'une table multidimensionnelle.

La figure 9 montre le résultat de la requête précédente dont les colonnes et les lignes représentent respectivement les dimensions *Créateur* et *Date*, et le plan (fiches superposées) assimile la troisième dimension *Editeur*. Les mesures, à savoir les mots-clés, sont placées à l'intersection des lignes et des colonnes pour un plan donné. Le symbole asterisk (*) indique l'absence de valeur de mesure.

Dans le souci d'aider le décideur à porter son attention sur les données les plus pertinentes des tables multidimensionnelles, nous avons emprunté au domaine *Big Data* la *technique* de visualisation des résultats sous forme de nuage de valeurs. Ainsi, nous représentons le résultat d'une requête par un nuage de mots où chaque mot possède une taille de police proportionnelle à sa fréquence d'apparition dans la table multidimensionnelle. Ce nuage représente un moyen de classification graphique du résultat dans le but d'attirer l'attention de l'analyste-décideur sur les valeurs les plus fréquentes, de la table multidimensionnelle résultat, qui méritent d'être soigneusement examinées (Khrouf *et al.*, 2015).

La figure 10 visualise un nuage de mots issus de la requête précédente (cf. figure 9). Dans la figure 10 nous montrons que chaque mot-clés parmi *Data*, *Modeling* et *OLAP* a une fréquence plus élevée que les autres mots-clés ; c'est-à-dire que les

documents analysés traitent beaucoup plus ces concepts que d'autres concepts considérés moins importants.

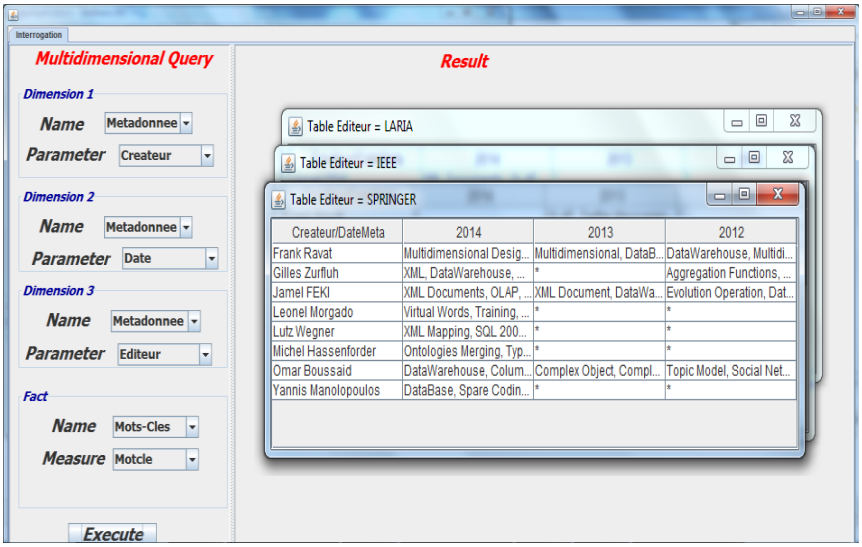

Figure 9. Interface IMDF d'interrogation multidimensionnelle

Figure 10. Nuage des mots correspondant à la requête OLAP de la figure 9

7. Conclusion

Dans cet article nous avons présenté le modèle multidimensionnel en toile d'araignée à base de facettes. Ce modèle est dédié à l'OLAP de documents XML. Ce modèle est générique puisqu'il s'appuie sur des facettes indépendantes de tout domaine d'application. Il se distingue des autres modèles de la littérature par les contraintes suivantes : 1) *l'exclusion sémantique* entre deux dimensions qui signifie que deux dimensions ne peuvent pas participer simultanément à une même analyse OLAP des documents lorsque la sémantique est violée ; 2) la contrainte *dimension dupliquée* c'est-à-dire qu'une dimension peut participer deux fois, voire trois fois, dans la même analyse ; 3) les *paramètres récursifs* qui sont utilisés pour les dimensions *Sémantique* et *Structurelle* où les documents et les taxonomies sont représentés d'une manière hiérarchique multi-niveaux ; et 4) les *dimensions corrélées* pour indiquer les dimensions en relation et permettant ainsi le passage d'une dimension à une autre pour la même requête OLAP. Ce modèle se distingue également par sa capacité à traiter des données textuelles et ceci en plus des mesures numériques.

Au niveau logique, nous avons opté pour la modélisation *OROLAP* (Object-relationnel OLAP) pour permettre à une ligne de fait de référencer plus qu'une seule ligne de dimensions. Une telle mise en œuvre est impossible avec *ROLAP*. Afin de montrer la faisabilité et la pertinence de nos propositions, nous avons développé un outil logiciel nommé *IMDF (Interrogation Multidimensionnelle de Documents à base de Facettes)* pour l'interrogation multidimensionnelle de documents et nous avons appliqué l'OLAP sur un corpus académique défini manuellement composé de 250 documents XML. Le résultat d'une requête OLAP s'affiche sous forme d'une table multidimensionnelle et d'un nuage de mots ; l'affichage sous forme de nuage de mots explicite visuellement les concepts les plus traités par les documents analysés, et permettrais une meilleure interprétation des résultats.

Plusieurs perspectives sont envisageables pour ce travail. Dans l'immédiat, il est important d'offrir à l'analyste-décideur plus de solutions pour effectuer des analyses plus approfondies à partir de la sélection d'un ou plusieurs mots affichés dans le nuage. A long terme, nous souhaitons partager les analyses OLAP entre différents utilisateurs d'une même organisation, par l'introduction de l'aspect collaboratif comme par exemple, un système de recommandation de requêtes et de collaboration entre des utilisateurs ayant des intérêts communs.

Bibliographie

Aknouche R., Asfari O., Bentayeb F., Boussaid O. (2013). Decisional architecture for text warehousing: ETL-text process and multidimensional model TWM. *Proceedings of the 19th International Conference on Management of Data*, p. 101-104.

Azabou M., Khrouf K., Feki J., Vallès N., Soulé-Dupuy C. (2015). Diamond multidimensional model and aggregation operators for document OLAP. *IEEE 9th*

International Conference on Research Challenges in Information Science, Athens, Greece, p. 363-373.

Bautista M., Molina C., Tejeda E., Vila A. (2013). A new multidimensional model with text dimensions: definition and implementation. *International Journal of Computational Intelligence Systems*, vol. 6, n° 1, p. 137-155.

Ben Mefteh S., Khrouf K., Feki J., Soulé-Dupuy C. (2013). Semantic Structure for XML Documents: Structuring and Pruning. *Journal of Information Organization*, vol. 3, n° 1, p. 37-46.

Boussaid O., Ben Messaoud R., Choquet R., Anthoard S. (2006). Conception et construction d'entrepôts XML. *Journée francophone sur les Entrepôts de Données et l'Analyse en ligne*, Versailles, France, p. 3-22.

Cabanac G., Chevalier M., Chrisment C., Julien C. (2010). Organization of digital resources as an original facet for exploring the quiescent information capital of a community. *International Journal on Digital Libraries*, Vol. 11, n° 4, p. 239-261.

Charhad M., Quénot G. (2004). Semantic Video Content Indexing and Retrieval using Conceptual Graphs. *IEEE Conference on Information and Communication Technologies: From Theory to Applications*, Damascus, Syria, p.19-23.

Dublin Core. (2012). The Dublin Core Metadata Element Set de http://dublincore.org/, Version 1.1.

Eder J., Koncilia C. (2001). Changes of dimension data in temporal data warehouses. *Proceedings of the International Conference on Data Warehousing and Knowledge Discovery (DaWaK'01)*, Munich, Germany, p. 284-293.

Evéquoz F., Thomet J., Lalanne D. (2010). Gérer son information personnelle au moyen de la navigation par facettes. *Proceedings of the 22nd Conference on l'nteraction Homme-Machine*, Luxembourg, p. 41-48.

Feki J., Ben Messaoud I., Zurfluh G. (2013). Building an XML Document Warehouse. *Journal of Decision Systems (JDS)*, Taylor & Francis, vol. 22, n° 2, p. 122-148.

Ghozzi F., Ravat F., Teste O., Zurfluh O. (2003). Modèle multidimensionnel à contraintes. *Extraction et Gestion des Connaissances*, Lyon, France, p. 43-55.

Hachaichi Y., Feki J. (2013). An Automatic Method for the Design of Multidimensional Schemas from Object Oriented Databases. *International Journal of Information Technology and Decision Making*, vol. 12, n° 6, p. 1223-1260.

Hernandez N., Mothe J., Ralalason B., Ramamonjisoa B., Stolf P. (2008). A Model to Represent the Facets of Learning Objects. *Interdisciplinary Journal of e-Learning and Learning Objects*, vol. 4, p. 65-82.

Khrouf O., Khrouf K., Feki J. (2013). A new multidimensional model for the OLAP of documents based on facets. *The International Arab Conference on Information Technology*, Khartoum, Soudan.

Khrouf O., Khrouf K., Feki J. (2014). Modèle multidimensionnel en toile d'araignée : Modélisation conceptuelle et logique. *Conférence sur les Avancées des Systèmes Décisionnels*, Hammamet, Tunisie, p. 146-156.

Khrouf O., Khrouf K., Altalhi A., Feki J. (2015). CobWeb Multidimensional Model: Filtering Documents using Semantic Structures and OLAP. *The Tenth International Conference on Internet and Web Applications and Services*, Brussels, Belgium, p. 92-98.

Kimball R., Ross M. (2013). The Data Warehouse Toolki: The Definitive Guide to Dimensional Modeling, *3rd edition. John Wiley & Sons*, New York.

Kumar S., Morstatter F., Marshall G., Liu H., Nambiar U. (2012). Navigating Information Facets on Twitter (NIF-T). *Proceedings of the 18th ACM SIGKDD International conference on Knowledge discovery and data mining*, Beijing, China, p. 1548-1551.

Lin C. X., Ding, B., Han J., Zhu F., Zhao B. (2008). Text cube: Computing in measures for multidimensional text database analysis. *Proceedings of the 8th IEEE International Conference on Data Mining*, Pisa, p. 905-910.

Mechkour M. (1995). A Multifacet Formal Image Model for Information Retrieval. *Proceedings of the Final WorkShop on Multimedia Information Retrieval*, Glasgow, UK, p. 18-20.

Oukid L., Benblidia N., Bentayeb F., Asfari O., Boussaid O. (2015). Contextualized Text OLAP Based on Information Retrieval. *International Journal of Data Warehousing and Mining IJDWM*, vol. 11, n° 2, p.1-21.

Park B.-K., Han H., Song I.-Y. (2005). XML-OLAP: A Multidimensional Analysis Framework for XML Warehouses. *Proceedings of the 7th international conference on Data Warehousing and Knowledge Discovery*, Copenhagen, Denmark, p. 32-42.

Ravat F., Teste O., Tournier R., Zurfluh G. (2008). Designing and Implementing OLAP Systems from XML Documents. *Proc. Annals of Information Systems, Springer, Special issue New Trends in Data Warehousing and Data Analysis*, vol. 3, p. 1-21.

Ravat F., Teste O., Tournier R., Zurfluh G. (2010). Finding an Application-Appropriate Model for XML Data Warehouses. *Information Systems*, vol. 35, n° 6, p. 662-687.

Soutou C. (1999). Relational-objet sous oracle 8 : Modélisation avec UML. Edition Eyrolles.

Tournier R. (2007). *Analyse en ligne (OLAP) des documents.* Thèse de doctorat en Informatique, Université Toulouse III, Paul Sabatier, Toulouse, France.

Tseng F.S.C., Chou A.Y. (2006). The concept of document warehousing for multidimensional modeling of textual-based business intelligence. *Journal Decision Support System (DSS)*, vol. 42, n° 2, p. 727-744.

Zhang D., Zhai C., Han J. (2009). Topic cube: Topic modeling for olap on multidimensional text databases. *Proceedings of the SIAM International Conference on Data Mining*, NV, USA, p. 1124-1135.