Analyse multidimensionnelle de documents *via* des dimensions OLAP

Franck Ravat — Olivier Teste* — Ronan Tournier***

** IRIT, Institut de recherche en informatique, Université Toulouse 3*
118 route de Narbonne, F-31062 Toulouse cedex 9

*** IRIT, Institut de recherche en informatique, Université Toulouse 1*
2 rue du doyen G. Marty, F-31042 Toulouse cedex 9

{ravat, teste, tournier}@irit.fr

RÉSUMÉ. *Avec l'émergence de formats de données semi-structurés (tels que XML), le stockage de documents dans un entrepôt centralisé est apparu de façon naturelle comme une adaptation des entrepôts de données. De nos jours, les systèmes OLAP (On-Line Analytical Processing) font face à une part grandissante de données non numériques. Cet article présente un environnement pour l'analyse multidimensionnelle de données textuelles dans un environnement OLAP. La structure, les métadonnées et le contenu des documents orientés texte sont transposés en sujets d'analyse (faits) et en axes d'analyse (dimensions) au sein d'un schéma en étoile modifié. Ceci permet de plus amples possibilités d'analyses multidimensionnelles. Cet environnement permet à un utilisateur d'avoir une vision détaillée au sein d'une collection de documents.*

ABSTRACT. *With the emergence of semi-structured data format (such as XML), the storage of documents in centralized facilities has slowly appeared as a natural adaptation of data warehousing technology. Nowadays, OLAP (On-Line Analytical Processing) systems face growing non-numeric data. This paper presents a framework for the multidimensional analysis of textual data in an OLAP sense. Document structure, document meta-data and document contents are converted into subjects of analysis (facts) and analysis axes (dimensions) within an adapted star schema. This allows greater multidimensional analysis possibilities. This framework allows a user to gain insight within a collection of documents.*

MOTS-CLÉS : *document numérique, XML, OLAP, entrepôt de données, entrepôt de documents, analyse multidimensionnelle.*

KEYWORDS: *digital document, XML, OLAP, data warehouse, document warehouse, multidimensional analysis.*

DOI:10.3166/DN.10.2.85-104

DN – 10/2007. Entreposage de documents, pages 85 à 104

1. Introduction

L'essor des technologies de l'information a considérablement accru la quantité de données et de documents numériques disponibles. Le volume des informations a été démultiplié et de par sa taille, ce volume rend la masse d'informations difficilement appréhendable. Toutefois, les systèmes de traitement d'analyse en ligne OLAP (On-Line Analytical Processing) (Codd *et al.*, 1993) permettent l'analyse de grands volumes d'informations en générant une vision synthétique des données.

1.1. *Contexte et problématique : entrepôts de données et de documents*

Dans ce cadre, l'utilisation de bases de données multidimensionnelles (BDM) fournit un point de vue global sur les données d'une entité économique et permet aux décideurs d'avoir une vision détaillée des performances de cette entité grâce à un accès rapide et interactif aux données (Colliat, 1996). Les BDM, appelées magasins de données, agencent de manière multidimensionnelle les données issues de l'entrepôt afin de faciliter leur exploitation et analyse (cf. figure 1).

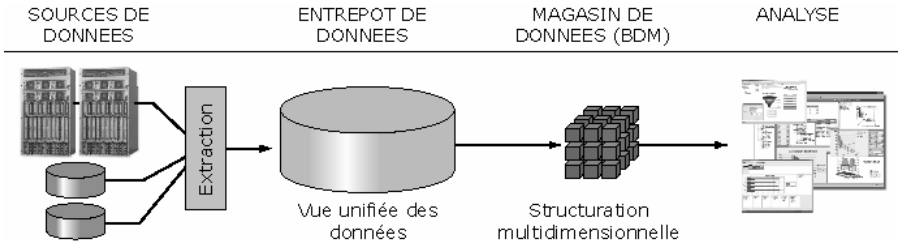

Figure 1. *Architecture d'un système d'aide à la décision*

La modélisation multidimensionnelle (Kimball, 1996) représente les données comme des points dans un espace multidimensionnel avec la métaphore du cube (ou hypercube). Cette approche orientée utilisateur intègre structures et données multidimensionnelles dans la représentation du cube. Par exemple, en figure 2, le *nombre de mots-clefs* utilisés dans une publication scientifique est analysé selon trois axes d'analyse : les *auteurs*, la *date*, et les *mots-clefs*. Une « tranche » (slice en anglais) est extraite du cube et présentée.

Pour concevoir les BDM, des structures multidimensionnelles ont été définies permettant la modélisation des concepts de sujets d'analyse, appelés *faits* et d'axes d'analyse, appelés *dimensions* (Kimball, 1996). Les faits sont des regroupements d'indicateurs d'analyse appelés *mesures*. Les dimensions sont composées de *paramètres*, agencés de manière hiérarchique, qui modélisent les différents niveaux de détails des axes d'analyse. Un paramètre peut être associé à des données complé-

mentaires représentées par des *attributs faibles* (par exemple, le nom du mois associé au numéro du mois). Un fait et ses dimensions associées composent un *schéma en étoile* (Kimball, 1996). La figure 3 illustre les structures multidimensionnelles de la représentation en cube de la figure 2. Les notations proviennent de (Ravat *et al.*, 2007b). Les concepts de faits et de dimension seront définis de manière plus précise dans la section 2.1.

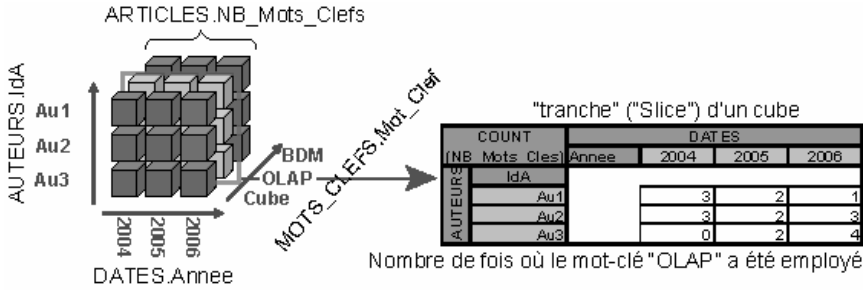

Figure 2. *Représentation en cube d'une base de données multidimensionnelle*

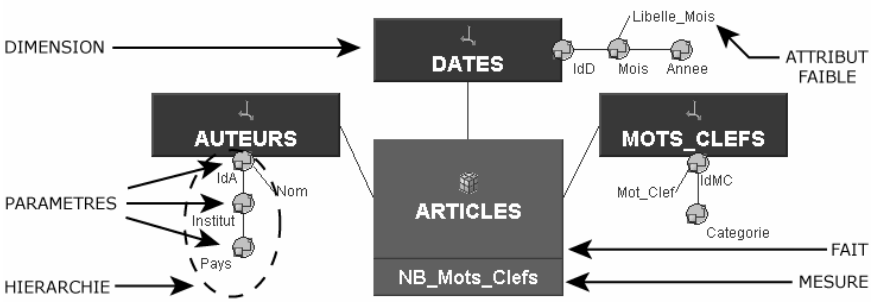

Figure 3. *Schéma en étoile d'une base de données multidimensionnelle*

Selon (Tseng *et al.*, 2006), les systèmes d'aide à la décision n'ont fait que survoler l'iceberg informationnel que représente l'aide à la décision. L'analyse multidimensionnelle basée sur des BDM factuelles numérique est une tâche bien maîtrisée de nos jours (Sullivan, 2001). Ces BDM sont construites sur des données transactionnelles extraites des systèmes d'information (SI) des entreprises ou du web. Cependant, seul 20 % des données d'un SI sont des données transactionnelles et peuvent être traitées par un système OLAP (Tseng *et al.*, 2006). Les 80 % restants, des documents, restent hors de portée de la technologie OLAP par absence d'outils ou de méthodes adaptées à la gestion de données textuelles.

Les systèmes OLAP fournissent de puissants outils mais dans un environnement rigide hérité des bases de données. Les données textuelles, faiblement structurées, ne peuvent être facilement intégrées. Récemment XML[1] a fourni un vaste environnement d'échange pour les documents au sein des SI des entreprises mais aussi sur le web. Ainsi peu à peu les documents semi-structurés ont commencé à être intégrés au sein des entrepôts de données (entrepôt de document ou document warehouse) (Sullivan, 2001), avec par exemple Xyleme[2]. Désormais, documents structurés ou semi-structurés représentent une source de données envisageable pour les processus OLAP. Le format d'échange XML permet le stockage de données dans un formalisme auto descriptif par l'intermédiaire d'une description de la structure appelée DTD (Document Type Definition) ou bien un XSchema[3].

De nos jours, l'environnement OLAP repose sur une analyse quantitative des données factuelles, par exemple, le nombre de produits vendus ou encore le nombre de fois qu'un mot-clef est employé dans un document. Nous souhaitons aller plus loin et fournir un environnement plus complet. Le système ne devrait pas être limité à l'analyse quantitative, mais devrait pouvoir inclure des analyses qualitatives. Toutefois, les données qualitatives doivent pouvoir être gérées par les environnements OLAP. Or ces données sont généralement non additives et non numériques, ainsi les fonctions d'agrégation traditionnelles, telles que SOMME ou MOYENNE, sont inopérantes. Dans (Park *et al.*, 2005), les auteurs suggèrent l'emploi de fonctions d'agrégation adaptées et dans (Ravat *et al.*, 2007a), nous avons proposé une telle fonction. Dans la suite, en guise d'illustration, nous emploieront une fonction qui sélectionne les principaux mots clés d'un texte : TOP_KEYWORDS.

Afin de fournir des capacités d'analyse plus exhaustives, les systèmes d'aide à la décision devraient pouvoir fournir l'exploitation de 100 % des données des SI des entreprises. Les documents ou des données issues du Web pourraient être directement intégrés dans les processus d'analyses. Ne pas prendre en compte ces données mènent inévitablement à l'omission d'informations pertinentes durant un processus de prise de décision important voire même l'inclusion de données non pertinentes générant ainsi des analyses approximatives voire erronées (Tseng *et al.*, 2006). En allant au-delà des propos de (Fankhauser *et al.*, 2003), nous pensons que la technologie XML permet d'envisager l'intégration de documents dans un environnement OLAP. Ainsi, notre problématique s'articule autour de l'analyse multidimensionnelle de documents. L'environnement OLAP actuel ne prend pas en compte l'analyse de données textuelles. De plus les données textuelles disposent d'une structure et d'un contenu inexploitable faute de moyens adaptés. Une analyse de documents textuels permet à un utilisateur d'obtenir une vision plus globale de l'ensemble d'une collection. Ainsi, l'utilisateur sera capable d'avoir une vision globale d'une collection de documents et de savoir si cette collection peut s'avérer intéressante ou non. En effet,

1. XML, Extended Markup Language de http://www.w3.org/XML/
2. Xyleme Server de http://www.xyleme.com/page/xml_server/
3 .XSchema, XML Schema de http://www.w3.org/XML/Schema

chercher une information dans des documents qui s'avèreraient au final non pertinents pour l'utilisateur représenterait une perte en terme de temps. Le contraire pourrait s'avérer crucial en termes de décisionnel.

1.2. *État de l'art*

L'état de l'art du domaine peut être divisé en trois catégories de travaux : 1) l'intégration de données XML dans un entrepôt de données ; 2) l'entreposage des données XML ; et 3) l'intégration de documents dans l'analyse OLAP.

La première catégorie présente l'intégration de données XML dans un entrepôt. (Golfarelli *et al.*, 2001 ; Pokorný, 2001) proposent d'intégrer des données XML à partir de leur DTD. (Vrdoljak *et al.*, 2003 ; Vrdoljak *et al.*, 2006) proposent de créer un schéma multidimensionnel à partir de la définition XSchema des documents XML. (Niemi *et al*., 2002) assemble à la volée des cubes de données XML à partir de requêtes utilisateur. (Zhang *et al*., 2003) propose une méthodologie pour constituer un entrepôt de données sur des données XML. Une alternative consiste à fédérer des données XML et un entrepôt classique si l'intégration des données XML au sein de l'entrepôt n'est pas envisageable. Ceci peut être dû à des contraintes légales (droit de détention des données) ou physiques (changement trop rapide des données sources pour permettre un processus de rafraîchissement efficace). Les auteurs de (Jensen *et al*., 2001), (Pedersen *et al.*, 2002 ; Yin *et al*., 2004) proposent une application qui répartit les requêtes entre les deux systèmes d'entreposage et fusionne les résultats à l'issue.

La seconde catégorie représente l'entreposage de données complexes au format XML. De part la nature complexe des données, la solution proposée est de les intégrer physiquement dans un entrepôt XML natif. Deux axes de recherches ont été développés. Le premier axe est centré sur l'adaptation du langage de requête XML, XQuery[4], pour faciliter l'expression de requêtes multidimensionnelles, notamment avec l'ajout d'un opérateur de regroupement (Beyer *et al*., 2005 ; Bordawerkar *et al*., 2005) ; l'adaptation de l'opérateur cube (Gray *et al*., 1996) aux données XML (Wiwatwattana *et al*., 2007) et l'agrégation de données XML par leur structure (Wang *et al*., 2005). Le second axe, quand à lui est centré sur la constitution d'entrepôts de données XML. Dans (Nassis *et al*., 2004), les auteurs proposent une structure xFACT permettant la définition d'un entrepôt de documents au format XML avec des données factuelles complexes. Se basant sur le xFACT, les auteurs de (Park *et al*., 2005) proposent un environnement multidimensionnel XML, intégrant des fonctions d'agrégation inspirées de la fouille de texte. Enfin, dans (Boussaid *et al*., 2006), les auteurs proposent aussi une méthode ainsi qu'un environnement pour un entrepôt de données XML. Ils proposent l'analyse multidimensionnelle de données complexes mais restent dans un cadre numérique. L'analyse de données

4. XQuery, XML Query Language de http://www.w3.org/XML/Query/

textuelles directement : dans (Khrouf *et al.*, 2004), les auteurs décrivent en entrepôt de documents (Sullivan, 2001), où les documents sont regroupés par structures similaires. Les utilisateurs peuvent effectuer des analyses multidimensionnelles, mais en restant toujours sur des données numériques.

La troisième catégorie concerne l'ajout de documents dans une analyse multidimensionnelle. Cette catégorie peut être subdivisée en trois approches :

– premièrement en complétant une analyse multidimensionnelle. En associant l'analyse numérique et la recherche d'information, les auteurs de (Peréz *et al.*, 2007) proposent de retourner au décideur des documents pertinents vis-à-vis d'une analyse en cours. Ainsi, le décideur dispose d'un complément d'information pour l'analyse.

– deuxièmement, quatre propositions ont suggéré l'utilisation d'un environnement OLAP pour l'analyse du contenu d'une collection de documents. Dans (McCabe *et al.*, 2000 ; Mothe *et al.*, 2003), les auteurs suggèrent l'utilisation d'analyses multidimensionnelles pour avoir une vision globale au sein d'une collection de documents *via* l'analyse des mots-clefs utilisés dans les documents. Par cette approche, l'utilisateur ainsi spécifier des recherches d'informations plus efficaces en utilisant les mots-clefs de manière plus précise. Dans (Keith *et al.*, 2005 ; Tseng *et al.*, 2006), les auteurs suggèrent l'utilisation de l'environnement OLAP pour l'analyse de documents en vue de constituer des matrices de co-occurrences. Avec ces quatre propositions, les auteurs analysent le contenu des documents *via* des mots-clefs. Les données textuelles sont analysées *via* des dimensions qui modélisent des axes d'analyse et non des sujets. Les indicateurs d'analyse, les mesures, sont toujours numériques (le nombre d'occurrence d'un mot clé…). Ainsi seules des analyses quantitatives et non qualitatives peuvent être effectuées car l'environnement ne permet pas le support de l'analyse de données textuelles. Récemment, des solutions commerciales commencent à introduire l'analyse de documents, avec par exemple, Text OLAP de Megaputer[5].

– troisièmement, par l'analyse des données textuelles directement. Dans (Khrouf *et al.*, 2004), les auteurs décrivent un entrepôt de documents où les documents sont regroupés selon leur structure. Dans ce système, les utilisateurs peuvent effectuer des analyses multidimensionnelles en fonction des structures. Enfin dans (Park *et al.*, 2005), les auteurs introduisent le concept d'analyse multidimensionnelle de documents via des techniques de fouilles de texte. De manière complémentaire nous avons introduit une fonction permettant l'agrégation de mots-clefs, via une « pseudo-moyenne » entre mots-clefs et retournant des mots-clefs plus généraux moyennant une légère perte contrôlée de sémantique (Ravat *et al.*, 2007a).

Ces propositions avancées sont limitées : 1) les données textuelles ne sont généralement pas analysées et les systèmes utilisent principalement des mesures numériques pour contourner le problème ; 2) les analyses sont limitées aux mots-clefs tandis que structure et contenu du document sont ignorés ; 3) les structures des

5. Megaputer, Polyanalyst Suite de http://www.megaputer.com/com/products/pa

documents restent quasiment inexploitées ; et 4) il n'existe aucun cadre de spécification d'indicateurs non numériques.

1.3. *Objectifs et contributions*

L'objectif majeur de cette proposition est d'aller au-delà des capacités d'analyse fournies par des indicateurs numériques et de fournir un environnement OLAP enrichi associant l'analyse qualitative aux capacités actuelles d'analyse quantitative. Toutefois, l'analyse de données textuelles n'est pas aussi fiable que l'analyse de données numériques. Par *analyse multidimensionnelle de documents*, nous entendons l'analyse dans un environnement OLAP des sources de données textuelles. Pour être compatible avec l'environnement rigide hérité des entrepôts de données, nous considérons des documents structurés ou semi-structurés. Par exemple, des documents XML représentant les actes de conférences scientifiques, les diagnostiques de dossiers patients d'un système d'information hospitalier, des rapports de contrôle qualité…

Nous proposons une extension du modèle en constellation (Ravat *et al.*, 2007b) avec la spécification de mesures textuelles. Contrairement à (Park *et al*., 2005), nous proposons un cadre formel pour la spécification de ces mesures afin de faciliter la spécification d'analyses multidimensionnelles. Nous proposons également de réviser les types de dimensions proposées dans (Tseng *et al.*, 2006) pour permettre l'ajout de la dimension modélisant la structure des documents. Il s'agit de la première approche qui ajoute la structure des documents dans l'analyse multidimensionnelle. La modélisation de la structure permet une plus grande flexibilité quant à la spécification d'analyse de données textuelles.

Le présent document s'articule comme suit : la seconde section définit le modèle multidimensionnel, où nous introduisons le concept de mesure textuelle ainsi que la dimension structure ; la troisième section expose le niveau logique de la proposition, et la section 4 présente l'analyse de données textuelles ; enfin, la section 5 conclut l'article et liste quelques perspectives.

2. Modèle multidimensionnel conceptuel

Les modèles existants étant limités pour l'analyse de données textuelles, nous proposons de modifier un modèle en constellation pour permettre une modélisation conceptuelle d'une analyse multidimensionnelle d'une collection de documents. Par *collection*, nous entendons un ensemble de documents homogènes en structure correspondants à un besoin d'analyse. Ces collections sont supposées être accessibles au moyen d'entrepôt de documents (Sullivan, 2001). Nous renvoyons le lecteur aux états de l'art récents (Torlone, 2003 ; Abello *et al.*, 2006) pour un survol des modè-

les multidimensionnels. Nous utiliserons une extension d'un modèle en constellation (Ravat *et al.*, 2007b).

Dû à la rigidité héritée de l'environnement des entrepôts de données, nous supposons que les collections de documents sont composées de documents structurés ou semi-structurés : des articles scientifiques au format XML. En outre, la définition des dimensions nécessite une collection homogène en structure.

2.1. *Définition formelle*

Un schéma en constellation textuel est employé pour modéliser une analyse de contenus de documents où ce contenu est modélisé en tant que sujet d'analyse.

Un *schéma en constellation textuel CT* est défini par $CT = (F^{CT}, D^{CT}, Star^{CT})$ où $F^{CT} = \{F_1,\ldots,F_m\}$ est un ensemble de faits ; $D^{CT} = \{D_1,\ldots,D_n\}$ est un ensemble de dimensions et $Star^{CT} = F^{CT} \rightarrow 2^{D^{CT}}$ est une fonction associant chaque fait à ses dimensions associées[6]. Un *schéma en étoile textuel* est une constellation où F^{CT} est un singleton.

Un *fait F* est défini par $F = (M^F, I^F, IStar^F)$ où $M^F = \{M_1,\ldots,M_n\}$ est un ensemble de *mesures* ; $I^F = \{i^F_1,\ldots,i^F_q\}$ est un ensemble d'instances du fait et $IStar^F : I^F \rightarrow I^{D1} \times \ldots \times I^{Dn}$ est une fonction qui associe respectivement les instances du fait F aux instances des dimensions D_i liées.

Une *mesure M* est définie par $M = (m, F_{AGG})$ où m est la mesure et $F_{AGG} = \{f_1,\ldots f_x\}$ est un ensemble de fonctions d'agrégation compatibles avec l'additivité de la mesure, $f_i \in$ (SUM, AVG, MAX…). Les mesures peuvent être additives, semi-additives ou non-additives (Kimball, 1996 ; Horner *et al.*, 2004).

Une *dimension D* est définie par $D = (A^D, H^D, I^D)$ où $A^D = \{a^D_1,\ldots,a^D_u\}$ est un ensemble d'*attributs* (paramètres et attributs faibles) ; $H^D = \{H^D_1,\ldots,H^D_x\}$ est un ensemble de *hiérarchies* représentant l'agencement des attributs et $I^D = \{i^D_1,\ldots,i^D_p\}$ est un ensemble d'instances de la dimension.

Une *hiérarchie H* est définie par $H = (Param^H, Weak^H)$ où $Param^H = <p^H_1, p^H_2,\ldots, p^H_{np}, All>$ est un ensemble ordonné d'attributs, appelés *paramètres* (avec $\forall k \in [1..np]$, $p^H_k \in A^D$ et une racine commune $p^H_1 = a^D_1$, $\forall H \in H^D$) et $Weak^H : Param^H \rightarrow 2^{A^D - Param^H}$ est une application spécifiant l'association d'attributs faibles aux paramètres. Toutes les hiérarchies d'une dimension commencent par le même paramètre racine et se termine par le paramètre de plus haute granularité.

6. La notation est *2D* représente l'ensemble des parties de l'ensemble D.

2.2. *Différents types de mesures*

Pour répondre aux spécificités des collections de documents, nous définissons une extension du concept classique de mesure. Nous distinguons ainsi deux types de mesures : les mesures numériques et les mesures textuelles.

Une *mesure numérique* est exclusivement composée de données numériques. Elle est soit *additive* (toutes les fonctions d'agrégation traditionnelles peuvent être employées) ; soit *semi-additives* et représente des instantanés (des températures, des quantités de stock…). Avec des mesures semi-additives, les fonctions d'agrégation sont limitées. F_{AGG} permet la spécification des fonctions d'agrégations compatibles avec la nature additive ou semi-additive de la mesure.

Une *mesure textuelle* est une mesure dont les données textuelles sont à la fois non numériques et non additives. Le contenu d'une mesure textuelle peut représenter un mot, un paquet de mots, un paragraphe voire un document complet. Nous distinguons plus types de mesures textuelles :

– une *mesure textuelle brute* est une mesure dont le contenu correspond au contenu complet d'un document (par exemple le contenu d'un article au format XML privé des balises XML qui le structure).

– une *mesure textuelle élaborée* est une mesure dont le contenu est issu d'une mesure textuelle brute et ayant subit un certain nombre de prétraitements. Une *mesure textuelle mot-clef* est une mesure textuelle élaborée. Ce type de mesure est obtenu par exemple après application de traitements sur une mesure textuelle brute tel que le retrait des mots vides et le maintient des mots les plus significatifs vis-à-vis du contexte du document.

Avec une mesure non additive, seules les fonctions d'agrégation génériques peuvent être employées (telles que COUNT et LIST). Toutefois, dans (Park *et al*., 2005), les auteurs suggèrent l'emploi de fonctions d'agrégation inspirées de la fouille de texte, telles que TOP_KEYWORDS qui retourne les n principaux mots-clefs d'un texte et SUMMARY qui génère le résumé d'un fragment de texte. Plus récemment, nous avons proposé une fonction AVG_KW (Ravat *et al*., 2007a) qui opère exclusivement sur des mesures textuelles mot-clef et qui combine plusieurs mots-clefs en un mot-clef plus général selon une « pseudo-moyenne ».

2.3. *A données spéciales, dimensions spéciales*

Dans le cadre d'analyses OLAP à partir de documents, les auteurs de (Tseng *et al.*, 2006) distinguent 3 types de dimension :

1) *Dimensions ordinaires* : ces dimensions sont constituées à partir de données extraites du contenu des documents analysés. Ces données sont extraites puis organisées de manière hiérarchique Une dimension composée des principaux mots-clefs extraits des documents est une dimension ordinaire ;

2) *Dimensions métadonnées* : les données de ces dimensions sont composées par les métadonnées extraites des documents. Par exemple, les auteurs, l'éditeur, la date de publication… L'initiative des métadonnées Dublin Core[7] représente certaines de ces métadonnées ;

3) *Dimensions catégorie* : les données de ces dimension sont des données externes au document qui permette sa catégorisation. Par exemple, des catégories de documents, un ontologie telle que wordnet…

A ces trois types de dimensions, nous ajoutons deux autres :

4) *Dimension structure* : il s'agit d'une dimension qui modélise la structure commune des documents contenus dans une collection analysée. Une constellation textuelle contient un ensemble de dimensions structures, mais un fait ne peut être relié qu'à une unique dimension structure : $D^{STR}=\{D_{S1},\ldots,D_{Sn}\}$; $D^{STR}\subset D^{CT}$; $\|D^{STR}\|\leq\|F^{CT}\|$ et $\forall F\in F^{CT}$, $\nexists(D_{Si}\in D_{STR}\wedge D_{Sj}\in D^{STR})|(D_{Si}\in Star^{CT}(F)\wedge D_{Sj}\in Star^{CT}(F))$

5) *Dimensions complémentaires* : Ces dimensions sont constituées à partir de sources de données complémentaires. Par exemple, les données d'état civil correspondant aux auteurs d'articles.

Toutefois, il faut noter que les dimensions ordinaires ne sont guère employées dans notre modèle car le contenu des documents est modélisé par des mesures textuelles. Dans la même veine, les dimensions de catégories sont guère pratiques car comme présenté dans (Mothe *et al.*, 2003) ce type de dimension est délicat à mettre en œuvre, en plus d'être des dimensions *non strictes* (Malinowski *et al*., 2006). Ceci est du au fait qu'envisager le contenu de documents comme une mesure associée à un fait n'a jamais été pris en considération.

Les dimensions structure sont constituées à partir des structures extraites des documents (*via* la structure arborescente DTD ou XSchema des documents XML). Chaque paramètre de ces dimensions modélise les différents niveaux de granularité d'un même document. À savoir, la zone de données textuelles qui servira aux fonctions d'agrégation textuelles spécifiques. Les dimensions structures modélisent à la fois la structure générique (*section*, *sous-section*, *paragraphe*…) mais aussi la structure spécifique par des attributs *Type_section*, *Type_Paragraphe* (…). Par exemple, *introduction*, *conclusion* (…) sont des types de *sections* et *définition*, *théorème* (…) sont des types de *paragraphes*. La structure spécifique est extraite depuis le balisage qui délimite les éléments dans les documents XML.

De leur côté, les dimensions complémentaires représentent les dimensions que l'on retrouve de nos jours dans l'environnement OLAP (produit, client…).

7. Dublin Core Metadata initiative (DCMI) de http://dublincore.org/

2.4. *Exemple*

Afin d'analyser les activités d'un laboratoire de recherche, un utilisateur analyse le contenu d'une collection de documents composée d'articles scientifiques. Ces articles ont un en-tête avec une structure générique commune. De plus, ces articles contiennent un certain nombre de métadonnées : noms des auteurs, leur affiliation (institut, pays), la date de publication, une liste de mots-clefs…

Les articles sont aussi caractérisés par une organisation de leur contenu via la structure arborescente des données XML :

– une structure générique. Les articles scientifiques sont composés de paragraphes regroupés en sous-sections, elles-mêmes regroupées en sections. Les sections quant à elles sont regroupées en articles,

– une structure spécifique. Avec les types de sections : introduction, conclusion… et les types de paragraphes : définition, exemple, théorème...

Les éléments constitutifs de la structure générique et spécifique sont extraits des documents par analyse des balises XML qui découpent le contenu des documents de manière semi-automatique pour permettre la résolution éventuelle de conflits.

Le schéma en étoile textuel correspondant est présenté en figure 4. Ce schéma a l'avantage de compléter celui présenté en introduction (cf. figure 3). Le nouveau schéma a la possibilité d'analyser les métadonnées et les mots-clefs des documents comme dans les propositions de (McCabe *et al.*, 2000 ; Mothe *et al.*, 2003 ; Keith *et al.*, 2005 ; Tseng *et al.*, 2006). En outre, il permet aussi d'analyser le contenu de documents avec l'emploi de la structure. Comparé à l'exemple présenté en introduction, une mesure textuelle brute est ajoutée au fait *ARTICLES* et une dimension *STRUCTURE* est ajoutée. Notez que les trois autres dimensions sont de type métadonnées.

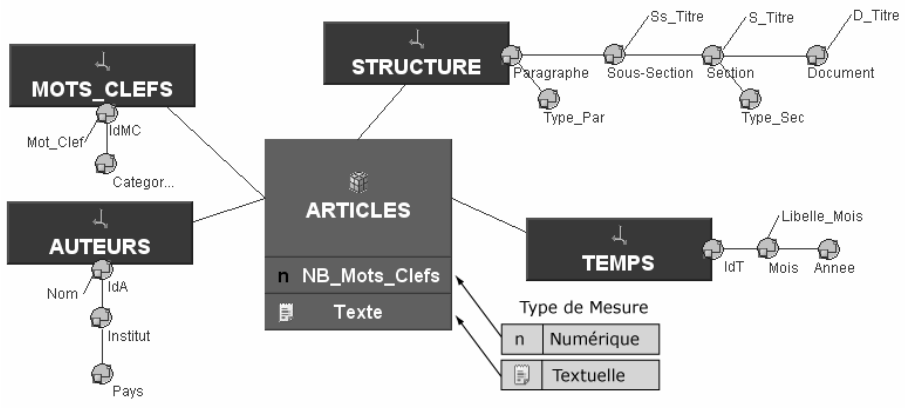

Figure 4. *Exemple de schéma pour l'analyse multidimensionnelle de documents*

La dimension *STRUCTURE* est composée de paramètres modélisant la structure générique et spécifique de la collection de documents. Chaque paramètre désigne une granularité spécifique de la mesure *Texte* (paragraphes, sous-section, section).

3. Modèle logique

3.1. *Modèle logique multidimensionnel*

Le niveau logique de l'environnement est basé sur une extension des technologies R-OLAP (OLAP relationnel) (Kimball, 1996). L'architecture est présentée en figure 5. Elle est composée de deux espaces de stockages : une base de données multidimensionnelle (1) et un espace de stockage de fichiers XML (2). Dans la base de donnée multidimensionnelle, chaque dimension est modélisée par une table relationnelle et la ou les tables des faits agissent comme pivot central hébergeant les clés étrangères vers les tables dimensions. Les documents sont stockés à part, dans des fichiers XML. Ils sont reliés aux données factuelles via une expression XPath. Dans notre exemple il s'agit d'un chemin Xpath pour chaque paragraphe des documents.

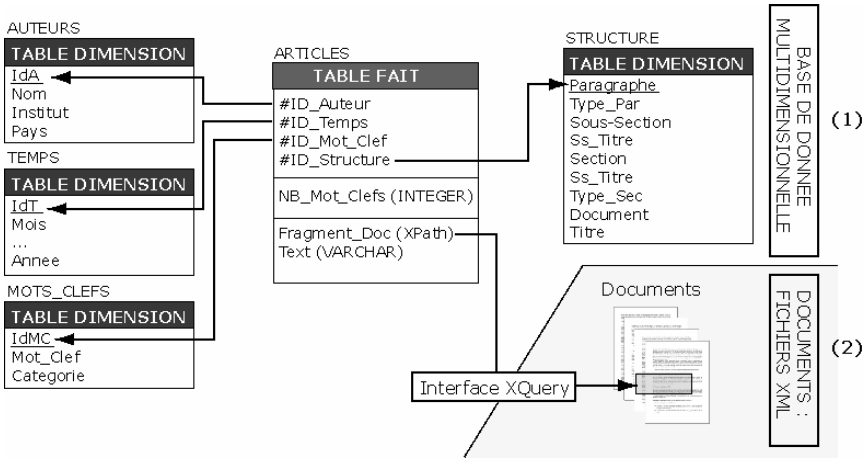

Figure 5. *Schéma en étoile, représentation du niveau logique*

Les tables sont constituées à partir de données extraites des documents (dimensions ordinaires et métadonnées) ou bien à partir de données complémentaires (dimensions catégories et complémentaires). La dimension structure est construite à partir de la structure interne des articles. Cette structure est extraite par le balayage de la DTD des fichiers XML représentant les documents. Chaque fragment de texte est lié à son fichier source via une expression XPath qui désigne la localisation de l'élément au sein du fichier XML. Le système utilise des expressions XQuery pour

sélectionner, retourner les fragments de documents. Par exemple, « Article/Section[@Id=1]/paragraphe » représente tous les paragraphes des premières sections des articles.

En cas de problème de stockage, il est possible de na pas dupliquer le contenu des documents dans la table de fait. Le système se contente alors du chemin XPath pour accéder aux données textuelles. Par contre, pour des raisons de performance, le système aura besoin de disposer de vues matérialisées afin d'optimiser les processus d'agrégation lors d'analyses multidimensionnelles.

Afin d'implanter les hiérarchies « *non-covering* » (Malinowski *et al*., 2006), le niveau logique utilise des valeurs virtuelles. Ainsi une section sans sous-section aura une sous-section virtuelle contenant les paragraphes de la section. Ceci permet aussi de gérer le cas particuliers des paragraphes introductifs de section qui sont avant la première sous-section.

3.2. *Exemple de données*

La figure 6 présente quelques données d'un schéma textuel en étoile (les données dimensionnelles ne sont pas toutes représentées).

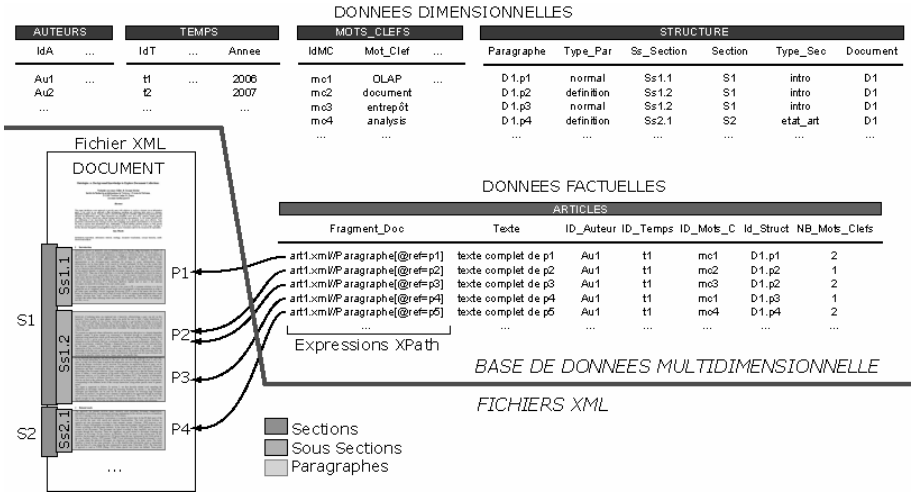

Figure 6. *Exemple de données*

4. Analyse multidimensionnelle de données textuelles

Les analyses multidimensionnelles OLAP présente les données des sujets d'analyse en fonction de différents niveaux de détails (granularité).Le processus agrège les données selon le niveau de détail désigné *via* des fonctions d'agrégation telles que SOMME, MIN, MAX, MOYENNE… Dans le tableau 1, un décideur analyse le nombre de théorèmes par mois et par auteur (1). Pour obtenir une vision plus globale, le décideur agrège les données mensuelles par année (2), projetant ainsi une valeur agrégée de théorèmes (le nombre total) pour chaque couple (auteur, année).

(1)

COUNT (Texte)	TEMPS					
	Annee	2005			2006	
	Mois	Sept.	Nov.	Dec.	Jan.	Fév.
AUTEURS / IdA						
Au1		3	4	2	2	2
Au2		2	2	3	6	7
STRUCTURE.Type_Par = "Theoreme"						

(2)

COUNT (Texte)	TEMPS		
	Annee	2005	2006
AUTEURS / IdA			
Au1		9	4
Au2		7	13
STRUCTURE.Type_Par = "Theoreme"			

Tableau 1. *Nombre de théorèmes par auteur et par mois (1) ; cumul par année (2)*

4.1. *Agrégation de données textuelle*

L'analyse de mesures textuelles requiert des moyens d'agrégation spécifiques. Les fonctions d'agrégations actuelles n'ont pas la capacité de prendre en entrée des données textuelles. Dans l'environnement standard OLAP, seules les fonctions d'agrégation COUNT et IDENTITÉ (aussi appelée LIST) peuvent être employées sur des données non numériques et non additives.

En plus des deux fonctions d'agrégation génériques (COUNT et LIST), nous proposons les fonctions d'agrégation suivantes :

– SUMMARY : fonction qui génère un résumé d'une mesure textuelle (Park *et al*., 2005) ;

– TOP_KEYWORD : fonction qui retourne les n principaux mots-clefs d'une mesure textuelle (Park *et al*., 2005) ;

– AVG_KW : fonction qui essaye de retourner un mot-clef « moyen », c'est-à-dire elle agrège des mots-clefs en un mot-clef plus général (avec une perte de sémantique contrôlée) (Ravat *et al*., 2007a).

En guise d'exemple dans la suite de cet article, nous emploierons la fonction d'agrégation TOP_KEYWORDS. Cette fonction prend en entrée un fragment de texte et retourne les n principaux mots-clefs du fragment. Dans nos exemples, nous nous limiterons aux deux premiers mots-clefs (*n=2*).

Contrairement à l'analyse de mesures numériques, l'analyse de mesures textuelles manque parfois cruellement de précision. En effet les fonctions d'agrégation opérant sur du texte ne sont pas aussi robustes que des fonctions basiques telles que

somme ou moyenne. L'intérêt de la dimension *STRUCTURE* est de permettre à l'utilisateur une flexibilité lors de la spécification d'analyses multidimensionnelles. Il peut changer le niveau de détails employé par la fonction d'agrégation, palliant ainsi le manque de précision de l'analyse textuelle.

4.2. *Exemple d'analyse*

Afin de faciliter la compréhension, nous utiliserons un pseudo langage de requête pour une spécification simple des analyses. Le langage désigne : un *sujet* d'analyse ; les éléments à être disposés dans les en-têtes en *lignes* et en *colonnes* ; le *niveau de granularité* (*via* l'emploi de la dimension *STRUCTURE*) ; et éventuellement une *restriction* sur le contenu par l'intermédiaire de la structure spécifique (*via* l'emploi des attributs *Type_Paragraphe* ou *Type_Section* de la dimension *STRUCTURE*). Pour visualiser les résultats, une table multidimensionnelle (mTable), adaptée pour l'affichage de données textuelles est employée (cf. tableau 1).

Dans l'exemple suivant, l'analyste analyse une collection d'articles scientifiques. L'analyse porte sur les publications des auteurs *Au1* et *Au2* durant *2005* et *2006*. L'analyse sélectionne les principaux mots-clefs de chaque ensemble de documents. Ces mots-clefs sont agrégés au niveau de granularité *section* et l'analyse est limitée au contenu des *introductions* de chaque document. Ainsi, la fonction liste les principaux mots-clefs des introductions de tous les articles. La figure 7 présente les résultats de cette analyse, spécifiée par l'expression suivante en pseudo langage :

```
Sujet: TOP_KEYWORDS(ARTICLES.Texte)
En lignes: AUTEURS.IdA="Au1" OR AUTEURS.IdA="Au2"
En colonnes: TEMPS.Annee="2005" OR TEMPS.Annee="2006"
Niveau de Granularité: STRUCTURE.Section
Restriction: STRUCTURE.Type_Sec="introduction"
```

L'interface de restitution est une table multidimensionnelle (mTable) (Gyssens *et al*., 1997 ; Ravat *et al.*, 2007b). Au sein de cette table, lors de l'analyse de mesures textuelles, chaque cellule contenant un résultat est en fait un lien hypertexte vers une page Web contenant la liste détaillée des éléments agrégés. Contrairement au système présenté dans (Tseng *et al.*, 2006) où l'interface retourne le texte complet des documents, la liste emploie des expressions XPath et ne retourne que les fragments correspondants à la granularité désignée (ici *Section*).

Définition. Une *table multidimensionnelle T* (*mTable*) est une matrice de *lignes*×*colonnes* cellules c_{ij}.Chaque cellule c_{ij}=(*R*, *Lk*) est composée d'un résultat agrégé *R* et d'un lien hypertexte *Lk*.

Le lien hypertexte *Lk* mène à une page Web contenant le résultat agrégé de la cellule liée ainsi qu'une liste des éléments qui ont servi pour générer l'agrégation. Chaque élément source est listé et un chemin XPath le lie au fragment correspondant. Par exemple (cf. figure 7), en *2006*, *Au1* a publié au moins deux articles (*ARTICLE12* et *ARTICLE23*) et les principaux mots-clefs des introductions sont

OLAP et *Requêtes* : $c_{Au1,2006}$=(*R*={*OLAP*, *Requêtes*}, $Lk_{Au1,2006}$).. Remarquez que les liens XPath sont spécifiés selon le niveau de granularité sélectionné au sein de la dimension STRUCTURE. Par exemple, l'introduction d'un document est l'ensemble des paragraphes qui le compose.

Un autre exemple est la liste de tous les théorèmes écrits par un auteur au cours d'une année. Cette analyse est exprimée par la requête en pseudo langage suivante :

```
Sujet: LIST(ARTICLES.Texte)
En lignes: AUTEURS.IdA= "Au1" OR AUTEURS.IdA="Au2"
En colonnes: TEMPS.Annee="2005" OR TEMPS.Annee="2006"
Niveau de Granularité: STRUCTURE.Document
Restriction: STRUCTURE.Type_Par="theoreme"
```

Il est possible de constater que les théorèmes sont désignés par une contrainte sur le type de paragraphe (paramètre *Type_Par*). La granularité d'un document complet est désignée par le paramètre *Document* de la dimension *STRUCTURE*. Ainsi la fonction d'agrégation va regrouper les éléments à agréger par article. Si l'utilisateur avait sélectionné la fonction d'agrégation COUNT, il aurait obtenu pour chaque cellule de la table multidimensionnelle un ensemble de nombres correspondant aux nombres de théorèmes par article (cf. tableau 1 (2)).

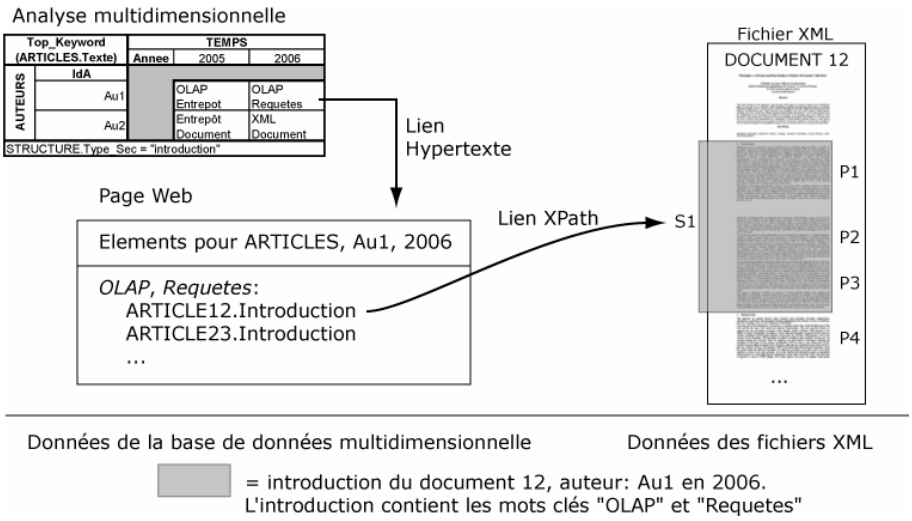

Figure 7. *Interface de restitution avec un exemple d'analyse*

5. Conclusion

Dans cet article, nous avons proposé d'étendre de précédents travaux sur l'analyse multidimensionnelle de documents afin d'obtenir des capacités d'analyse étendues au sein de l'environnement OLAP.

Premièrement, nous avons adapté le modèle en constellation (Kimball, 1996) et (Ravat *et al.*, 2007b), avec l'ajout de mesures textuelles ainsi qu'une dimension spécifique représentant la structure des documents. Cette dimension fournit une plus grande souplesse à l'utilisateur afin d'obtenir une meilleur fiabilité lors de l'agrégation de données textuelles. Elle permet aussi la spécification de requêtes complexes portant sur des éléments particuliers dans une collection de documents. En outre, le modèle supporte de multiples analyses avec un schéma en constellation. Enfin, afin de retourner à l'utilisateur des résultats compréhensibles, nous avons adapté une table multidimensionnelle (mTable). Cette table permet d'afficher à la demande de l'utilisateur les fragments de documents correspondant aux sources de données à l'origine des agrégations présentées dans la table.

Cette proposition a pour but de fournir à l'analyste des capacités d'analyse accrues. Comme la majorité des données qui transitent dans un système d'information est composé de texte, ces données ne peuvent être gérées par les systèmes actuels. Nous pensons que cet environnement OLAP adapté fournira de nouvelles perspectives d'analyse pour le processus d'aide à prise de décision.

Nous implantons actuellement un prototype basé sur le SGBD Oracle 10*g*2 et une interface client Java (1.5). Les documents sont stockés sous la forme de fichiers XML, mais l'emploi d'un entrepôt de document tel que Xylème est envisageable pour la gestion des données XML. L'expression des requêtes est effectuée via une manipulation graphique des éléments conceptuels représentés dans un schéma conceptuel d'une constellation ou d'une étoile textuelle.

Nous envisageons plusieurs perspectives. Premièrement, au sein de l'article, nous avons employé une table multidimensionnelle (mTable) comme interface de restitution des résultats d'analyses textuelles. Cette structure tabulaire bidimensionnelle est bien adaptée pour la restitution de données numériques, mais lorsqu'il s'agit de données non numériques telles que du texte, l'affichage se retrouve rapidement surchargé. Ainsi nous envisageons d'employer des interfaces de restitution mieux adaptées à la restitution de données textuelles. Deuxièmement, comme les documents ne contiennent pas uniquement du texte, mais aussi des graphes, des figures ou encore des références, nous envisageons aussi d'adapter cet environnement pour tout type de contenus, ne limitant pas le système uniquement à l'analyse numérique ou textuelle.

6. Bibliographie

Abelló A., Samos J., Saltor F., "YAM²: A Multidimensional conceptual model extending UML", *Information Systems* (*IS*), vol. 31, (6), Elsevier, 2006, p. 541-567.

Beyer K.S., Chamberlin D.D., Colby L.S., Ozcan F., Pirahesh H., Xu Y., "Extending XQuery for Analytics", *ACM SIGMOD Int. Conf. on Management of Data* (*SIGMOD*), ACM Press, 2005, p. 503-514.

Bordawekar R., Lang C. A., "Analytical processing of XML documents: opportunities and challenges", *SIGMOD Record*, vol. 34, (2), ACM Press, 2005, p. 27-32.

Boussaid O., Messaoud R.B., Choquet R., Anthoard S., "X-Warehousing: An XML-Based Approach for Warehousing Complex Data", *10th East European Conf. on Advances in Databases and Information Systems* (*ADBIS 2006*), LNCS 4152, Springer, 2006, p. 39-54.

Codd E.F., Codd S.B., Salley C.T., Providing OLAP (On Line Analytical Processing) to user analyst: an IT mandate, rapport technique, E.F. Codd and associates, (white paper de Hyperion Solutions Corporation), 1993.

Colliat G., "OLAP, Relational, and Multidimensional Database Systems", *SIGMOD Record*, vol. 25, (3), ACM Press, 1996, p. 64-69.

Fankhauser P., Klement T., "XML for Data Warehousing Chances and Challenges", (Extended Abstract), *5th Int. Conf. on Data Warehousing and Knowledge Discovery* (*DaWaK 2003*), LNCS 2737, Springer, 2003, p. 1-3.

Golfarelli M., Rizzi S., Vrdoljak B., "Data Warehouse Design from XML Sources", *4th ACM Int. Workshop on Data Warehousing and OLAP* (*DOLAP 2001*), ACM Press, 2001, p. 40-47.

Gray J., Bosworth A., Layman A., Pirahesh H., "Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total", *12th Int. Conf. on Data Engineering* (*ICDE*), IEEE Computer Society, 1996, p. 152-159.

Gyssens M., Lakshmanan L.V.S., "A Foundation for Multi-dimensional Databases", *23rd Int. Conf. on Very Large Data Bases* (*VLDB*), Morgan Kaufmann, 1997, p. 106-115.

Horner J., Song I-Y., Chen P.P., "An analysis of additivity in OLAP systems", *7th ACM Int. Workshop on Data Warehousing and OLAP* (*DOLAP 2004*), ACM Press, p. 83-91, 2004.

Jensen M.R., Møller T.H., Pedersen T.B., "Specifying OLAP Cubes On XML Data", *13th Int. Conf. on Scientific and Statistical Database Management* (*SSDBM*), IEEE Computer Society, 2001, p. 101-112.

Keith S., Kaser O., Lemire D., "Analyzing Large Collections of Electronic Text Using OLAP", *APICS 29th Conf. in Mathematics, Statistics and Computer Science*, Acadia University, 2005, p. 17-26.

Khrouf K., Soulé-Dupuy C., "A Textual Warehouse Approach: A Web Data Repository", in *Intelligent Agents for Data Mining and Information Retrieval*, Masoud Mohammadian (Eds.), Idea Publishing Group (IGP), ISBN : 1-59140-277-8, 2004, p. 101-124.

Kimball R., *The data warehouse toolkit*, Ed. John Wiley and Sons, 1996, 2nd ed. 2003.

Malinowski E., Zimányi E., "Hierarchies in a multidimensional model: From conceptual modeling to logical representation", *Journal of Data & Knowledge Engineering* (*DKE*), vol. 59, (2), 2006, p. 348-377.

McCabe C., Lee J., Chowdhury A., Grossman D. A., Frieder O., "On the design and evaluation of a multi-dimensional approach to information retrieval", *23rd Int. ACM Conf. on Research and Development in Information Retrieval* (*SIGIR*), ACM Press, 2000, p. 363-365.

Mothe J., Chrisment C., Dousset B., Alau J., "DocCube: Multi-dimensional visualisation and exploration of large document sets", *Journal of the American Society for Information Science and Technology* (*JASIST*), vol. 54, (7), Wiley Periodicals, 2003, p. 650-659.

Nassis V., Rajugan R., Dillon T.S., Rahayu J.W., "Conceptual Design of XML Document Warehouses", *6th int. Conf. on Data Warehousing and Knowledge Discovery* (*DaWaK 2004*), LNCS 3181, Springer, 2004, p. 1-14.

Niemi T., Niinimäki M., Nummenmaa J., Thanisch P., "Constructing an OLAP cube from distributed XML data", *5th ACM Int. Workshop on Data Warehousing and OLAP* (*DOLAP*), ACM Press, 2002, p. 22-27.

Park B-K., Han H., Song I-Y., "XML-OLAP: A Multidimensional Analysis Framework for XML Warehouses", *6th Int. Conf. on Data Warehousing and Knowledge Discovery* (*DaWaK*), LNCS 3589, Springer, 2005, p. 32-42.

Pedersen D., Riis K., Pedersen T.B., "XML-Extended OLAP Querying", *14th Int. Conf. on Scientific and Statistical Database Management* (*SSDBM*), IEEE Computer Society, 2002, p. 195-206.

Pérez-Martinez J.M., Berlanga-Llavori R.B., Aramburu-Cabo M.J., Pedersen T.B., "Contextualizing data warehouses with documents", *Decision Support Systems* (*DSS*), Elsevier, (In Press) disponible en ligne : http://dx.doi.org/10.1016/j.dss.2006.12.005

Pokorný J., "Modelling Stars Using XML", *4th ACM Int. Workshop on Data Warehousing and OLAP* (*DOLAP*), ACM Press, 2001, p. 24-31.

Ravat F., Teste O., Tournier R., "OLAP Aggregation Function for Textual Data Warehouse", *9th International Conference on Enterprise Information Systems* (*ICEIS 2007*), vol. DISI, INSTICC Press, 2007, p. 151-156.

Ravat F., Teste O., Tournier R., Zurfluh G., "Querying Multidimensional Databases", *11th East-European Conference on Advances in Databases and Information Systems* (*ADBIS 2007*), LNCS 4690, Springer, 2007, p. 298-313.

Sullivan D., *Document Warehousing and Text Mining*, Wiley John & Sons, 2001.

Torlone R., "Conceptual Multidimensional Models", Capitre 3 de *Multidimensional Databases: Problems and Solutions*, M. Rafanelli (ed.), Idea Publishing Group (IGP), 2003, p. 69-90.

Tseng F.S.C., Chou A.Y.H, "The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence", *Decision Support Systems* (*DSS*), vol. 42, (2), Elsevier, 2006, p. 727-744.

Vrdoljak B., Banek M., Rizzi S., "Designing Web Warehouses from XML Schemas", *5th Int. Conf. on Data Warehousing and Knowledge Discovery* (*DaWaK*), LNCS 2737, Springer, 2003, p. 89-98.

Vrdoljak B., Banek M., Skocir Z., "Integrating XML Sources into a Data Warehouse", *2nd Int. Workshop on Data Engineering Issues in E-Commerce and Services* (*DEECS 2006*), LNCS 4055, Springer, 2006, p. 133-142.

Wang H., Li J., He Z., Gao H., "OLAP for XML Data", *5th Int. Conf. on Computer and Information Technology* (*CIT*), IEEE Computer Society, 2005, p. 233-237.

Wiwatwattana N., Jagadish H.V., Lakshmanan L.V.S., Srivastava D., "X^3: A Cube Operator for XML OLAP", *23rd Int. Conf. on Data Engineering* (*ICDE*), IEEE Computer Society, 2007, p. 916-925.

Yin X., Pedersen T.B., "Evaluating XML-extended OLAP queries based on a physical algebra", *7th ACM Int. Workshop on Data Warehousing and OLAP* (*DOLAP*), ACM Press, 2004, p. 73-82.

Zhang J., Ling T.W., Bruckner R.M., Tjoa A.M., "Building XML Data Warehouse Based on Frequent Patterns in User Queries", *5th Int. Conf. on Data Warehousing and Knowledge Discovery* (*DaWaK*), LNCS 2737, Springer, 2003, p. 99-108.