Modèle conceptuel pour l'analyse multidimensionnelle de documents

Franck Ravat**, Olivier Teste*
Ronan Tournier*, Gilles Zurfluh**

*IRIT SIG/ED, UMR5505 – Université Toulouse 3, 118 rte. de Narbonne,
F31062 Toulouse CEDEX 9, France
** IRIT SIG/ED, UMR5505 – Université Toulouse 1, 2 rue du doyen G. Marty,
F31042 Toulouse CEDEX 9, France
{ravat, teste, tournier, zurfluh}@irit.fr

Résumé. OLAP et les entrepôts de données sont utilisés pour l'analyse de données transactionnelles. De nos jours, avec l'évolution d'Internet et le développement de formats d'échange de données semi-structurées comme par exemple XML, il est possible de considérer les documents comme source d'analyse. En conséquence, un environnement d'analyse multidimensionnel adapté à ce type de données est nécessaire. Dans cet article, nous introduisons un modèle conceptuel multidimensionnel adapté à l'analyse de données documentaires, reposant sur un unique concept : une dimension. Nous définissons aussi un ensemble d'opérateurs d'analyse multidimensionnelle adaptés.

1 Introduction

Les systèmes d'analyse en ligne OLAP (On-Line Analytical Processing) permettent aux analystes d'améliorer le processus de prise de décision. Ces systèmes facilitent la consultation et l'analyse de données économiques, statistiques ou scientifiques agrégées et historisées et reposent sur un outil de centralisation des données: un entrepôt de données (Kimball, 1996).

1.1 Contexte et problématique

Les systèmes d'aide à la décision, emploient des bases de données multidimensionnelles (BDM), qui permettent aux décideurs d'avoir une vision des performances d'une entreprise. Pour modéliser les BDM, des structures multidimensionnelles ont été définies permettant la représentation 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 d'attributs, agencés de manière hiérarchique, qui modélisent les différents niveaux de détails (granularité) des axes d'analyse.

L'analyse multidimensionnelle basée sur des BDM factuelles numériques est une tâche bien maîtrisée de nos jours (Sullivan, 2001). Ces BDM sont souvent construites sur des

données transactionnelles issues des systèmes d'information (SI) des entreprises. Cependant, seul 20% des données d'un SI sont des données transactionnelles et peuvent être traitées (Tseng et Chou, 2006). Les 80% restants, la « paperasserie », restent hors de portée de la technologie OLAP faute d'outils et de méthodes adaptées à la gestion de données textuelles. Afin de fournir des capacités d'analyse accrues, les systèmes d'aide à la décision devraient pouvoir exploiter de 100% des données des SI des entreprises. Les analystes devraient pouvoir intégrer les documents directement dans leur processus d'analyse. Ne pas prendre en compte ces données mène 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 ou erronées (Tseng et Chou, 2006).

Les systèmes OLAP fournissent de puissants outils mais dans un environnement rigide hérité des bases de données. Les données textuelles, faiblement structurées, ne sont pas facilement intégrées dans les entrepôts. 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 d'entrepôt de documents ou « document warehouse » (Sullivan, 2001) tels que Xyleme[2]. Désormais, les documents structurés ou semi-structurés représentent une source de données envisageable pour les processus OLAP.

Dans cet article, par *analyse multidimensionnelle de documents*, nous entendons l'analyse dans un environnement OLAP des sources de données textuelles. Pour des raisons de compatibilité avec les entrepôts, nous ne considérons que des documents structurés ou semi-structurés. Par exemple, des documents XML représentant les actes de conférences scientifiques, les diagnostics de dossiers patients, des rapports de contrôle qualité…

Dans les sources de données textuelles, les données internes sont exclusivement textuelles. Ce type de données étant non additif et non numérique, les fonctions d'agrégation traditionnelles (somme, moyenne…) sont inopérantes. (Park *et al.*, 2005) suggèrent l'emploi de fonctions d'agrégation adaptées. Dans cet article, en guise d'illustration, nous emploieront l'une d'elle : TOP_KEYWORDS qui sélectionne les principaux mots clés d'un texte.

1.2 Exemple d'application

En vue d'une évaluation, un décideur analyse les citations des travaux d'un institut de recherche. Cette tâche consiste à compter chaque fois qu'un auteur de l'institut est cité dans un article et afficher le résultat par auteur et par conférence. Dans la *TAB. 1* (a) l'auteur *Au1* a été cité trois fois par des auteurs de *DaWaK*. Le décideur analyse ensuite la portée des travaux cités en analysant le sujet des publications qui citent les travaux de l'institut. Comme cette analyse ne repose pas sur des données numériques mais sur des données factuelles textuelles (le sujet des publications), l'analyste utilisera la fonction TOP_KEYWORDS pour afficher les principaux mots clés des documents. Ces mots clés seront regroupés par conférence donnant ainsi une liste de sujets au lieu d'un nombre de publications. Dans la *TAB. 1* (b), les trois citations des travaux de *Au1* dans les conférences *DaWaK* concernent *XML* et les *Documents*. Quant à l'auteur *Au3*, il a toujours été cité dans le même contexte (*fouille de données* et *clustering*), ainsi la portée de ses travaux semble être moindre que celle des deux autres auteurs (*Au1* et *Au2*).

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

(a)

	Institut	Inst1		
	Auteur	Au1	Au2	Au3
Conference				
DaWaK		3	2	1
DEXA		2	-	-
CAiSE		1	1	2

(b)

	Institut	Inst1		
	Auteur	Au1	Au2	Au3
Conference				
DaWaK		XML, Docments	XML, Entrepôts de données	Fouille de données, Clustering
DEXA		XML, BD temporelles	-	-
CAiSE		Fouille de données	Fouille de données	Fouilles de données, Clustering

TAB. 1 – *Exemples d'analyses multidimensionnelles. (a) Analyse du nombre de fois qu'un auteur a été cité par conférence ; (b) La même analyse avec les principaux mots clés des publications contenant la citation.*

La combinaison de ces analyses seraient très complexes, voire impossible à exprimer en utilisant des modèles multidimensionnels classiques. Premièrement, l'analyse de données textuelles n'est pas prise en compte. Deuxièmement, l'analyste aurait besoin d'avoir à sa disposition un nombre non négligeable de magasins de données (Kimball, 1996). Et troisièmement, les approches de modélisation multidimensionnelle éclatent la structure du document en de nombreux sous éléments nécessitant des tâches d'adaptation lourdes et coûteuses pour l'administrateur de la BDM.

1.3 État de l'art

L'état de l'art se subdivise en deux catégories avec la modélisation multidimensionnelle et l'analyse de documents avec les processus OLAP.

La catégorie concernant la modélisation multidimensionnelle et peut être subdivisée en deux approches. Premièrement avec la modélisation multidimensionnelle classique : un état de l'art peut être trouvé dans (Torlone, 2003) ou plus récemment avec (Abello *et al.*, 2006) et (Ravat *et al. 2007b*). La modélisation multidimensionnelle est basée sur les concepts de faits et de dimensions et principalement sur le très répandu schéma en étoile (Kimball, 1996). Tous ces modèles conceptuels ou logiques ont été conçus pour l'analyse de données numériques et ne peuvent traiter les documents nécessitant l'intégration de données non numériques. Deuxièmement, la modélisation multidimensionnelle permettant l'analyse de données complexes avec (Nassis *et al.*, 2004). Dans cette proposition, les auteurs définissent un xFACT, une structure hiérarchique complexe sur des données XML (tels que des documents), où les mesures appelées contextes peuvent être vues comme des données objets complexes. Les dimensions, appelées VDim, sont construites à partir de vues sur les données de l'entrepôt. Aucune opération n'est associée à ce modèle rendant son exploitation difficile.

Les propositions actuelles de modélisation multidimensionnelle sont incomplètes car elles ne répondent qu'à un besoin d'analyses numériques. A notre connaissance, il n'y a pas de proposition prenant en compte les spécificités des documents orientés documents.

La seconde catégorie concerne l'ajout de documents dans l'analyse multidimensionnelle. Ces documents, dans le contexte XML sont de deux types (Fuhr et Großjohann, 2001) :

- Les *Documents orientés données*, principalement employés par des applications pour l'échange de données. Par exemple : listes, logs de consultation de sites internet, factures, commandes, sorties d'applications e-commerce…
- Les *Documents orientés documents* sont les versions électroniques des documents papiers qui nous entourent. Par exemple : articles scientifiques, livres electroniques (e-books), pages web de sites internet…

L'analyse de documents orientés données a déjà été introduite dans plusieurs propositions tels que (Jensen *et al.*, 2001). Nous renvoyons le lecteurs aux travaux suivants : (Vrdoljak *et al.*, 2006), (Yin et Pedersen, 2004) et (Nassis *et al.*, 2004) pour des exemples et une liste plus complète. Récemment, (Boussaid *et al.*, 2006) propose l'intégration de données complexes XML et se ramène à des données transactionnelles dans l'environnement XML. Ces travaux considèrent les données textuelles au travers du XML, mais ces propositions traitent de documents orientés donnés ne prenant pas en compte les documents orientés documents.

En conséquence, le présent article est centré sur l'analyse multidimensionnelle de documents orientés documents. Cette catégorie peut être subdivisée en trois approches. Premièrement, en assistant une analyse multidimensionnelle classique en fournissant des documents complémentaires, les auteurs de (Peréz *et al.*, 2005) proposent de combiner l'analyse numérique traditionnelle et les techniques de recherche d'information. Ceci permet d'assister le décideur durant une analyse multidimensionnelle en lui fournissant des documents complémentaires jugés pertinents par rapport au contexte de l'analyse en cours. Deuxièmement, en fournissant une ébauche d'analyse multidimensionnelle de documents, les auteurs de (McCabe *et al.*, 2000), (Mothe *et al.*, 2003), (Keith *et al.*, 2005) et (Tseng et Chou, 2006), présentent des applications de l'analyse de documents orientés documents via un schéma en étoile. Ils proposent d'employer l'environnement OLAP pour permettre de compter des documents en fonction d'occurrences de mots-clefs ou de sujets. Mots-clefs et sujets sont organisés en dimensions et ces dernières permettent à l'utilisateur d'analyser le nombre de documents représentés par chaque mot-clefs en fonctions de divers autres axes d'analyse. De nos jours des solutions commerciales commencent à apparaître avec, par exemple, Text OLAP de Megaputer[3]. Troisièmement, en analysant des données textuelles directement. Dans (Khrouf et Soulé-Dupuy, 2004), les auteurs décrivent un entrepôt de documents où les documents sont regroupés par famille de structure. L'utilisateur peut effectuer des analyses multidimensionnelles en se basant sur des critères quantitatifs issues des données documentaires ou de leur structure (par ex. le nombre de document avec plus de trois sections). 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, dans (Ravat *et al.*, 2007) nous introduisons une fonction permettant l'agrégation de mots-clefs, générant ainsi un mot clé plus général.

Ces propositions montrent clairement les limites des modèles traditionnels pour l'analyse de documents : 1) les implantations suggérées ne préservent pas la structure des documents ; 2) ces structures restent inexploitées ; 3) les indicateurs non numériques ne sont pas gérés de manière aisée ; et 4) aucune flexibilité n'est fournie pour la sélection des sujets d'analyses.

Actuellement, à notre connaissance, il n'existe pas de proposition pour l'élaboration d'un modèle adapté pour l'analyse de documents orientés documents. Jusqu'à présent, les travaux de recherche se sont basés sur une analyse quantitative. Les données textuelles sont fournies pour l'analyse au moyen de dimensions qui modélisent des axes d'analyse et non des sujets d'analyse. Les indicateurs d'analyse (les mesures) sont systématiquement numériques.

1.4 Objectifs

Comme base vers un environnement plus complet pour l'intégration de documents dans un système OLAP, nous définissons un modèle conceptuel adapté pour l'analyse

[3] Megaputer, Polyanalyst Suite de http://www.megaputer/com/products/pa

multidimensionnelle de documents. Le but de cette proposition est de fournir à l'analyste une vue conceptuelle de haut niveau simple et adaptée (Golfarelli *et al.*, 2002), masquant les contraintes d'ordre logiques ou physiques. Afin de permettre la manipulation des concepts du modèle, les opérations de manipulation OLAP sont revisitées.

Le modèle conceptuel doit faciliter la tâche de l'utilisateur et prendre en compte les caractéristiques des documents orientés documents. Premièrement, ces documents sont composés de données organisées de manière hiérarchique. Deuxièmement, un document peut pointer vers lui-même ou bien vers un autre document. Ces liens devraient être clairement indiqués afin de faciliter la compréhension et la navigation au sein des données au cours des analyses. Par exemple, lors de l'analyse des références d'une publication, l'analyste doit clairement voir et non deviner qu'une référence n'est autre qu'une publication. Et troisièmement, lors de l'analyse de documents, l'analyse de texte assistée par ordinateur ne donne pas nécessairement des résultats significatifs ou compréhensifs. C'est-à-dire que lors de l'analyse d'un sujet particulier, l'analyste peut se retrouver en face de résultats difficilement appréhendables. Ainsi, l'analyste doit pouvoir facilement réorienter son analyse en changeant de sujet et ne pas être restreint par des sujets prédéfinis. En conclusion, le modèle doit être capable de : 1) représenter les spécificité des données issues de documents orientés documents ; 2) faciliter la représentation et éviter de fournir à l'analyste des solutions d'analyse limitées ; enfin 3) permettre la manipulation des concepts via un ensemble d'opérations adaptées. En réponse à ces objectifs, nous proposons un modèle en Galaxie associé à un jeu d'opérations de manipulation.

Dans la suite, la section 2 définit le modèle conceptuel adapté et la section 3 présente un ensemble d'opérations de manipulations des concepts du modèle.

2 Modèle multidimensionnel

Le modèle défini dans la présente section s'inspire de la modélisation en constellation (Kimball, 1996). Notre approche consiste à utiliser uniquement le concept de dimension. Ces amas de dimensions constituent une « galaxie » et sont regroupés autour d'un ou plusieurs nœuds centraux, afin de percevoir les dimensions compatibles pour une même analyse.

2.1 Regroupement de dimensions en « Galaxies »

Un schéma *dimensionnel* est un regroupement de dimensions liées entre elles par un ou plusieurs nœuds centraux. Nous généralisons le concept de constellation en définissant celui de Galaxie. Chaque nœud modélise les dimensions compatibles pour une même analyse.

Définition: une *Galaxie G* est définie par (D^G, $Star^G$, Lk^G) où :
- $D^G = \{D_1,\ldots, D_n\}$ est un ensemble de *dimensions*,
- $Star^G : D^G \rightarrow 2^{D^G}$ est une fonction modélisant les *nœuds centraux*.[4] Elle associe chaque dimension $D_i \in D^G$ aux autres dimensions $D_j \in D^G$ ($D_j \neq D_i$) compatibles lors d'une analyse.
- $Lk^G = \{g_1,\ldots,g_u\}$ est un ensemble de fonctions appelées *liens récursifs* associant des *instances d'attributs* entre elles où $g^G : a_u^{D_i}\left(i_x^{D_i}\right) \rightarrow a_v^{D_j}\left(i_y^{D_j}\right)$ est l'association de l'instance i_x de a_x^{Di} avec l'instance i_y de a_y^{Dj}, $(D_i,D_j) \in D^G \mid D_j \in Star^G(D_i)$.

[4] La notation 2^D représente l'ensemble des parties de l'ensemble D.

Les liens Lk^G représentent des relations « correspond à » entre les valeurs des deux attributs. Ces liens sont employés au sein des expressions des opérations de manipulation.
Notations. Nous notons $D_j \in Star^G(D_i)$, le fait que D_i et D_j soient liées via un noeud central : $D_j \in D^G$, $D_i \in D^G \mid D_j \in Star^G(D_i)$.

2.2 Concept de dimension

Une dimension modélise de manière classique un axe d'analyse, mais aussi un sujet potentiel d'analyse. Une dimension est caractérisée par des attributs organisés de manière hiérarchique, chaque attribut étant une graduation de l'axe d'analyse, à savoir un niveau de détail (ou granularité).

Définition: une *dimension* D_i est définie par (A^{Di}, H^{Di}, I^{Di}, $IStar^{Di}$) où:
- $A^{Di} = \{a^{Di}_1, \ldots, a^{Di}_r\} \cup \{All\}$ est un ensemble d'*attributs*,
- $H^{Di} = \{H^{Di}_1, \ldots, H^{Di}_s\}$ est un ensemble de *hiérarchies*,
- $I^{Di} = \{i^{Di}_1, \ldots, i^{Di}_t\}$ est un ensemble d'*instances de dimension*, et chaque attribut a une valeur particulière $a^{Di}_u(i^{Di}_x)$ appelée *instance d'attribut*.
- $IStar^{D_i} : I^{D_i} \rightarrow I^{D_1} \times \ldots \times I^{D_m}$ est une fonction qui associe les instances de la dimension D_i aux instances des autres dimensions liées par un même noeud central ($\forall k \in [1..m]$, $D_k \in D^G$, $D_k \neq D_i$ et $D_k \in Star^G(D_i)$, c'est-à-dire D_k est associée à D_i).

Une hiérarchie modélise l'organisation des différents niveaux de granularité, à savoir, une vision particulière de la graduation de l'axe. Une hiérarchie H^{Dj}_i de D est une liste ordonnée d'attributs appelés paramètres. Chaque paramètre peut être associé à des attributs faibles qui représentent des informations complémentaires.

Définition: une *hiérarchie* H^{Dj}_i (ou H_i) est définie par ($Param^{Hi}$, $Weak^{Hi}$) où:
- $Param^{Hi} = <p^{Hi}_1, \ldots, p^{Hi}_{np}, All>$ est un ensemble ordonné d'attributs, appelés *paramètres* qui représentent les niveaux de granularité de la dimension, $\forall k \in [1..np]$, $p^{Hi}_k \in A^{Dj}$;
- $Weak^{H_i} : Param^{H_i} \rightarrow 2^{A^{Di} - Param^{Hi}}$ est une application associant éventuellement des *attributs faibles* aux paramètres complétant la sémantique de ces dernier.

Les attributs sont de deux types : un paramètre représente les données d'un niveau de détail particulier, par exemple un *institut de recherche* ou le *pays* de cet institut ; un attribut faible représente une donnée complémentaire comme le *nom* de l'*institut de recherche*.

Toutes les hiérarchies d'une dimension commencent par un paramètre racine commun ($\forall H_i \in H^{Dj}$, $p^{Hi}_1 = a^{Dj}_1$) et se terminent par un paramètre générique : *All* (Gray *et al.*, 1996).

Pour répondre aux spécificités des structures des documents, les hiérarchies sont sémantiquement plus riches que dans un modèle traditionnel. Ceci permet d'obtenir une vision conceptuelle proche de la représentation des documents. Ainsi, les hiérarchies modélisant les documents peuvent être *non-strictes* (Malinowski et Zimányi, 2006).
Notations. $p_i \in H$ est une notation simplifiée de $p_i \in Param^H$. Lorsque cela est possible, si le contexte est évident, les notations H^D, $p^H{}_i \ldots$ seront simplifiées par H, $p_i \ldots$

2.3 Exemple

Afin d'observer les activités d'instituts de recherches, un décideur analyse des publications scientifiques ainsi que les projets effectués par ces instituts. En réponse à ce besoin, il utilise la galaxie G_1 qui représente dans sa partie supérieure : les articles publiés au

sein d'une conférence à une date donnée et écrits par des auteurs ; et dans la partie inférieure : les projets obtenus à une certaine date, pilotés par des instituts et employant des personnels scientifiques (qui sont aussi des auteurs d'articles). Au sein de cet exemple, deux liens récursifs peuvent être employés pour naviguer au sein des références des articles et des instituts des auteurs.

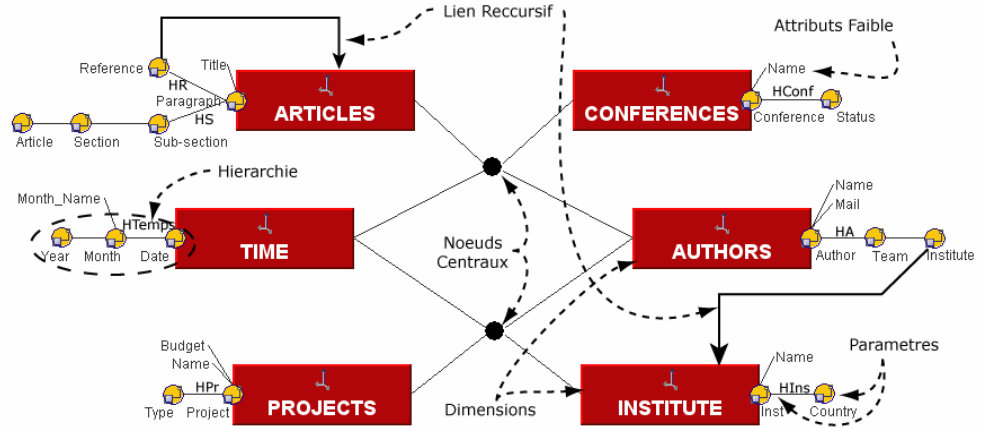

FIG. 1 – *Exemple de galaxie : analyse de projets et de publications scientifiques (G_1).*

3 Opérations Multidimensionnelles

Afin de manipuler les concepts représentés dans le modèle en galaxie, l'analyste a besoin de quatre opérations qui diffèrent légèrement des opérations OLAP traditionnelles (Rafanelli, 2003). Ces opérations sont basées sur les besoins suivants :

- Une opération de focalisation (Focus) est nécessaire pour mettre en avant le sujet d'une analyse, projetant les données du sujet sur plusieurs axes d'analyse.
- Pour restreindre la portée d'une analyse, une opération est nécessaire pour ne sélectionner qu'un sous-ensemble des données.
- Pour exploiter l'organisation hiérarchique des paramètres, deux opérations de forage sont nécessaires pour permettre de modifier le niveau de détail des données analysées : la première pour zoomer et explorer plus en détails les données ; la seconde pour « dézoomer » et permettre l'inverse, une exploration plus globale.
- Pour changer de critère d'analyse, une opération est nécessaire pour réorienter l'analyse, à savoir, effectuer un changement de sujet ou bien d'axe d'analyse.

Les auteurs de certains modèles ont souligné la nécessité du traitement symétrique des paramètres et des indicateurs d'analyse (mesures) pour faciliter la définition d'algèbres de requête ou de langage de calcul ainsi que pour introduire plus de flexibilité pour l'utilisateur (Agrawal *et al.* 1997), (Cabbibo et Torlone, 1997) et (Gyssens et Lakshmanan, 1997). Cependant certaines opérations spécifiques telles que les forages ne pouvaient opérer de manière symétrique entre tout type d'attributs. Avec notre modèle, ce problème est résolu.

Notations. Afin de faciliter la compréhension des définitions formelles qui suivent, nous introduisons les expressions suivantes. Les instances d'une galaxie G, composée de n dimensions, sont représentées par (1), où $dom(D_i)$ est le domaine de la dimension D_i, c'est-à-dire tout $i^{Di} \in I^{Di}$. Toutes les instances des attributs $a_j \in A^{Di}$ d'une dimension D_i sont

représentées par (2). Nous définissons une fonction d'agrégation f_{AGG} (3) où X^* représente un ensemble fini d'éléments de X et $dom(f_{AGG}(dom(D_i)))$ correspond au domaine des valeurs agrégées du domaine de la dimension D_i. Afin de pouvoir comparer le niveau des paramètres au sein d'une même hiérarchie H, nous introduisons la fonction *level* (4).

$$dom(D_1)\times\ldots\times dom(D_n)=\prod_{i=1}^{n} dom(D_i)=dom(G) \quad \textbf{(1)}$$

$$dom(D_i.a_1)\times\ldots\times dom(D_i.a_u)=\prod_{j=1}^{|A^{D_i}|} dom(D_i.a_j)=dom(D_i) \quad \textbf{(2)}$$

$$f_{AGG}:(dom(D_i.p_j))^* \rightarrow dom(f_{AGG}(dom(D_i.p_j))) \quad \textbf{(3)}$$
$$(x_1,\ldots,x_m)\mapsto f_{AGG}(x_1,\ldots,x_m)$$

$$\text{Etant donné } Param^H =< p_1,\ldots,p_{np},All >,\ level^H(p_1)=1,\ldots,level^H(p_n)=n \quad \textbf{(4)}$$
$$\text{et } \forall j\in[1..n],\ level^H(p_j)<level^H(All)$$

Toutes les opérations produisent des sorties compatibles permettant leur enchaînement, assurant la fermeture des opérations. L'opération de focalisation génère en sortie un sous-ensemble de la galaxie, noté s^G. Ce sous-ensemble est utilisé comme entrée pour toutes les autres opérations qui produisent à leur tour un sous-ensemble en sortie. La syntaxe générale des opérations est :

NOM_OPÉRATION(*entrée*, *paramètres_de_l_opération*) = *sortie*

3.1 Opérations de focalisation et de sélection

Cette sous-section présente les deux principales opérations qui permettent la spécification du jeu de données d'une analyse.

L'opération de focalisation est utilisée pour définir un sujet d'analyse et pour projeter les données orientées sujet sur plusieurs axes d'analyse. Concrètement, cette opération permet la spécification d'un sujet d'analyse (*DS*) agrégeant les données d'analyse au moyen d'une fonction d'agrégation (f_{AGG}) selon le niveau de détail sélectionné dans les axes d'analyse.

Syntaxe : $FOCUS(G, S, P)=s^G$ où G est l'entrée (une galaxie), $S=(f_{AGG}(DS.HS.p_i)$ est le sujet d'analyse « focalisé » sur le paramètre p_i de la hiérarchie HS de la dimension DS agrégé via la fonction f_{AGG} et $P=<(D_x.H_x, Param_x),(D_y.H_y, Param_y),\ldots>$ est l'ensemble (ordonné) des axes de projection où D_x est la dimension sélectionnée en tant que premier axe d'analyse, D_y le second,… H_x est la hiérarchie courante de l'axe représenté par D_x, H_y est la hiérarchie courante de D_y,… $Param_x=<p_{x\_min},\ldots,p_{x\_max}>$ est un ensemble ordonné de paramètre de H_x, où étant donné $Param^{Hx}=<p_1,\ldots,p_{np}, All>$, $level^{Hx}(p_{x\_min})\geq level^{Hx}(p_1)$ et $level^{Hx}(p_{x\_max})\leq level^{Hx}(p_{np})$. $Param_x$ représente les paramètres sélectionnés de D_x (il s'agit d'un sous-ensemble de $Param^{Hx}$). Idem pour $Param_y$…

Conditions : $\forall\ D_i\in P$, $D_i\in Star^G(DS)$, c'est-à-dire les dimensions sélectionnées en tant qu'axe d'analyse sont liées à la dimension sélectionnée en tant que sujet (*DS*). La fonction d'agrégation f_{AGG} doit être compatible avec les instances à agréger du paramètre p_i.

Mathématiquement : *FOCUS* (7) = *AGGREGATION* (6) ∘*PROJECTION* (5) où:

$$\prod_{i=1}^{n} dom(D_i) \xrightarrow{PROJECT} (dom(DS.p_i))^* \times \prod_{j=1}^{\|P\|} \left(\prod_{k=\min}^{\max} dom(D_j.p_k) \right) \quad (5)$$

$$(dom(DS.p_i))^* \times \prod_{j=1}^{\|P\|} \left(\prod_{k=j\_\min}^{j\_\max} dom(D_j.p_k) \right) \xrightarrow{AGGREGATE} dom(f_{AGG}(dom(DS.p_i))) \times \prod_{j=1}^{\|P\|} \left(\prod_{k=j\_\min}^{j\_\max} dom(D_j.p_k) \right) \quad (6)$$

$$\prod_{i=1}^{n} dom(D_i) \xrightarrow{FOCUS} dom(f_{AGG}(dom(DS.p_i))) \times \prod_{j=1}^{\|P\|} \left(\prod_{k=j\_\min}^{j\_\max} dom(D_j.p_k) \right) \quad (7)$$

Nous définissons aussi une notation simplifiée (8), où s^G représente un sous-ensemble de la galaxie avec une dimension désignée en tant que sujet (S_{AGG}) analysée (projetée et agrégée) selon les dimensions de l'ensemble de projection (P).

$$dom(G) \xrightarrow{FOCUS} dom(s^G) \text{ avec } dom(s^G) = dom(S_{AGG}) \times dom(P) \quad (8)$$

Exemple. Au sein de la galaxie G_1 représentée en *FIG. 1*, l'analyste focalise l'analyse sur les principaux mots-clefs des articles, les regroupant par auteur et par année. L'objectif est d'observer de manière grossière les recherches des différents auteurs. Nous supposerons que l'analyste emploie une table bidimensionnelle pour visualiser le résultat, (Gyssens et Lakshmanan, 1997) et (Ravat *et al.*, 2007b). En conséquence dans cet exemple, l'utilisateur focalise l'analyse sur une dimension (*DS*) et projette les données sur deux axes d'analyse. Dans l'exemple, la fonction d'agrégation TOP_KEYWORDS retournera seulement les deux principaux mots-clefs. L'instruction suivante produit la table représentée en *FIG. 2*.

FOCUS(G_1, TOP_KEYWORDS(*ARTICLES.HS.Section*),
((*TEMPS.HTemps*, <*Annee*>), (*AUTEURS.HA*, <*Auteur*>))) = $s^{G1}{}_1$

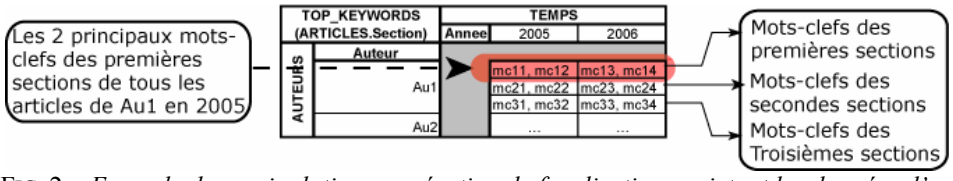

FIG. 2 – *Exemple de manipulations : opération de focalisation projetant les données d'un sujet d'analyse sur deux axes d'analyse.*

L'opération de sélection est employée pour réduire la quantité de données à analyser. En spécifiant un prédicat de restriction l'utilisateur peut effectuer une restriction sur un axe d'analyse ou bien sur le sujet d'analyse. Toutes les instances sélectionnées par un prédicat p sont maintenues dans la sélection courante, les autres instances étant retirées. Si cette opération est directement appliquée à la galaxie ceci permet le retrait d'instances avant le processus d'agrégation.

Syntaxe : *SELECT*(G, p)= $s^{G'}$ ou *SELECT*(s^G, p)= $s^{G'}$ où G (ou s^G) est l'entrée et p est un prédicat restrictif sur un attribut a_j d'une dimension.

Conditions : $a_j \in D_i$ et $D_i \in Star^G(DS)$.

Mathématiquement :

$$dom(G) \xrightarrow{SELECT(p)} dom(G) - dom(\neg p) \text{ ou } dom(s^G) \xrightarrow{SELECT(p)} dom(s^G) - dom(\neg p) \quad (9)$$

La notation $dom(\neg p)$ représente l'ensemble du domaine ne respectant pas le prédicat p. L'opération inverse *UNSELECT*, supprime tous les prédicats restrictifs.

Exemple. Afin de réduire la portée de l'analyse, l'analyste décide de se restreindre aux seules publications de *Au1* et d'analyser les principaux mots-clefs uniquement dans les introductions (la première section). En utilisant le sous-ensemble précédemment défini (s^G), l'instruction suivante produit la table définie en *FIG. 3* :

SELECT(*SELECT*($s^{G1}{}_1$, *ARTICLE.Section*='*Introduction*'), *AUTEURS.Auteur*='*Au1*') = $s^{G1}{}_2$

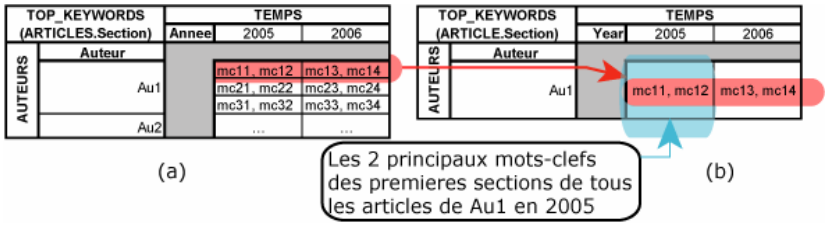

FIG. 3 – *Exemple de manipulations: application de deux restrictions.*

3.2 Opérations de forage

Une fois une analyse spécifiée (définition de s^G), l'utilisateur peut vouloir changer le niveau de détail selon lequel les données d'analyse sont projetées.

En employant une opération de forage vers le bas (Drill-Down) l'analyste peut « zoomer » et obtenir des données plus détaillées. Cette opération consiste à ajouter au sein de la liste des paramètres de l'un des axes de projection ($Param_i$), un paramètre p_{new} de la hiérarchie courante dont le niveau est inférieur au niveau du paramètre de plus fine granularité actuellement sélectionné (p_{min}).

Syntaxe : *DRILLDOWN*(s^G, D_i, p_{new})=$s^G{}_1$ où s^G est l'entrée, D_i est une dimension de l'ensemble de projection P de s^G, c'est à dire $\exists(D_i.H_i, Param_i)\in P$ et $p_{new}\in H_i$.

Condition : Le paramètre doit être d'un niveau inférieur au paramètre de plus fine granularité déjà sélectionné : $level^{Hi}(p_{new})<level^{Hi}(p_{min})$

Mathématiquement :

$$DrillDown: dom(s^G) \xrightarrow{DRILLDOWN} dom(S_{AGG}) \times dom(P) \times dom(D_i.p_{new})$$

$$\text{où} \quad dom(P) = \prod_{\substack{j=1 \\ j\neq i}}^{\|P\|} \left(\prod_{k=j\_\min}^{j\_\max} dom(D_j.p_k) \right) \times \prod_{k'=i\_\min}^{i\_\max} dom(D_i.p_{k'}) \qquad \textbf{(10)}$$

Il est à noter que $dom(P)$ représente le domaine des paramètres sélectionnés des dimensions ne prenant pas part au forage ($\forall D_j \mid \exists(D_j.H_j, Param_j)\in P$ et j≠i), mais aussi le domaine de la dimension prenant part au forage (D_i). Nous rappelons aussi que $Param_j$=<p_{j_min},…,p_{j_max}>.

L'opération inverse, le forage vers le haut (Roll-Up), est employé pour obtenir une vision plus globale des données analysées. Cette opération est utilisée pour « dézoomer » une vision détaillée des données d'analyse. L'opération consiste à retirer tous les paramètres de la liste des paramètres sélectionnés ($Param_i$) dont les niveaux sont inférieurs à un paramètre spécifié. Ce dernier paramètre est ajouté dans la liste s'il n'y est pas déjà.

Syntaxe : *ROLLUP*(s^G, D_i, p_{sup})= $s^G{}_1$ où s^G est l'entrée, D_i est une dimension de l'ensemble de projection P de s^G, c'est-à-dire $\exists(D_i.H_i, Param_i)\in P$ et $p_{sup}\in H_i$.

Condition : Le paramètre p_{sup} doit être d'un niveau supérieur au paramètre de granularité la plus fine déjà sélectionné : $level^{Hi}(p_{sup})>level^{Hi}(p_{min})$.
Mathématiquement : dans l'équation suivante nous définissons $sup=level^{Hi}(p_{sup})$.

$$RollUp: dom(s^G) \xrightarrow{ROLLUP} dom(S_{AGG}) \times \prod_{\substack{j=1 \\ j\neq i}}^{\|P\|} \left(\prod_{k=j\_\min}^{j\_\max} dom(D_j.p_k) \right) \times \prod_{k'=\sup}^{i\_\max} dom(D_i.p_{k'}) \quad (11)$$

Ici, $\prod_{k'=\sup}^{i\_\max} dom(D_i.p_{k'})$ est le domaine des paramètres de la dimension prenant part à l'opération de forage (D_i). Le domaine des paramètres dont les niveaux sont inférieurs à p_{sup} sont retirés (ainsi la borne inférieure de k' est $level^{Hi}(p_{sup})=sup$).
Exemple. Comme dans les modèles traditionnels, l'opération de forage vers le bas peut être employée pour représenter les mots-clefs par mois plutôt que par années. Cependant dans notre modèle, cette opération peut aussi être employée sur la hiérarchie courante de la dimension focalisée. Cette fonctionnalité est cruciale lorsque l'agrégation textuelle produit un résultat manquant de sens car elle permet à l'utilisateur d'avoir un aperçu du processus d'agrégation. Dans l'exemple suivant, plutôt que d'analyser des mots-clefs par section, l'analyste décide de les analyser par sous-section. L'instruction suivante produit la table présentée en *FIG. 4* :

$DrillDown(s^{G1}{}_2, ARTICLE, SousSection) = s^{G1}{}_3$

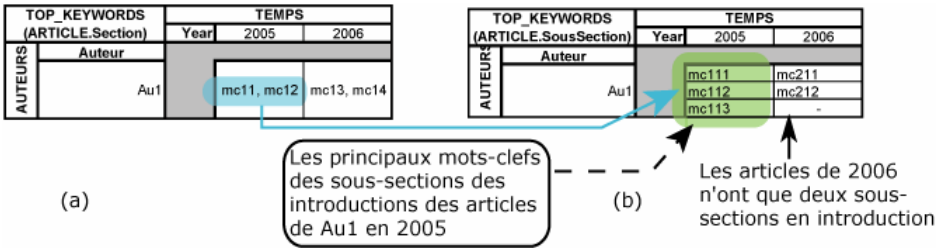

FIG. 4 – *Exemple de manipulations: forage sur la dimension focalisée.*

3.3 Opération de réorganisation d'analyse

Dans certains cas, l'utilisateur pourrait vouloir réorganiser les données analysées. A cette fin, il peut employer l'opération qui changera les éléments structurels du sous-ensemble de la galaxie s^G.

L'opération de rotation remplace par une nouvelle dimension l'une des dimensions de s^G. Le sujet d'analyse (*DS*) ou bien l'un des axes d'analyse, c'est-à-dire une des dimensions de l'ensemble de projection ($D_i \in P$).
Syntaxe : $ROTATE(s^G, D_{old}, D_{new}.H_{new}, A)=s^G{}_1$ où s^G est l'entrée, D_{old} est la dimension à remplacer, D_{new} est la nouvelle dimension, H_{new} est sa hiérarchie courante actuelle et A dépends de D_{old}. Si $D_{old}=DS$ alors $A=f'_{AGG}(p_{new})$, sinon si $D_{old}\in P$ alors $A = Param_{new} = <p_{new\_min},\ldots, p_{new\_max}>$ (un sous-ensemble de $Param^{Hnew}$).
Conditions : si $D_{old}=DS$ alors $\forall D_k \in P$, $D_k \in Star^G(D_{new})$ et $p_{new} \in H_{new}$. Si $D_{old}\in P$ alors $D_{new}\in Star^G(DS)$ $Param_{new} \subseteq Param^{Hnew}$ et $level^{Hnew}(p_{new\_min})<\ldots<level^{Hnew}(p_{new\_min})$
Mathématiquement : si $D_{old}=DS$ alors l'opération correspond à (12), sinon si $D_{old}\in P$, alors l'opération correspond à (13).

$$Rotate : dom(s^G) \xrightarrow{ROTATE} dom(f'_{AGG}(dom(D_{new}.p_{new}))) \times dom(P) \quad (12)$$

$$Rotate : dom(s^G) \xrightarrow{ROTATE} dom(S_{AGG}) \times \prod_{\substack{j=1 \\ j \neq old}}^{\|P\|} \left(\prod_{k=j\_\min}^{j\_\max} dom(D_j.p_k) \right) \times \prod_{k'=new\_\min}^{new\_\max} dom(D_{new}.p_{k'}) \quad (13)$$

Remarquez que si $D_{old} = D_{new}$, ceci permet le changement d'une des hiérarchies courantes (*HS*, H_x, H_y…). Remarquez aussi que la rotation du sujet d'analyse est l'équivalent des opérations *FRotate* (Ravat *et al.*, 2006) ou *DrillAcross* (Abelló *et al.*, 2003).

3.4 Utilisation des liens récursifs

Les liens récursifs au sein de la galaxie peuvent être employés pour accéder à un ensemble particulier de données. Ceci permet une plus grande flexibilité lors de la désignation de sous-éléments de documents et simplifie la spécification de requêtes.

Par exemple, la séquence d'opérations suivante utilise les liens entre *Reference* et *ARTCLE* (cf. *FIG. 1*). L'opération centre l'analyse sur les principaux mots-clefs des articles qui sont cités par *Au1*, c'est-à-dire les articles dans les sections « *Références* » des publications d'*Au1*.

SELECT (*SELECT* (*FOCUS* (*TOP_KEYWORDS*(*ARTICLES.Reference.HS.Section*),
((*TEMPS.HTemps*, <*Annee*>), (*ARTICLE.Reference.AUTEURS.HA*, <*Auteur*>))),
AUTEURS.Auteur='*Au1*'), *ARTICLE.Reference.TEMPS.Annee* > *2005*)

Où *ARTICLE.Reference.AUTEURS* sont les auteurs des articles cités par les publications d'*Au1*, *ARTICLE.Reference.TEMPS.Annee* sont les années de publication des articles cités par *Au1* alors que *TEMPS.Annee* sont les années de publication des articles d'*Au1*.

Autre exemple, la table présentée en *TAB. 1* (b), est obtenue avec la séquence d'opérations suivantes :

SELECT (*FOCUS* (*TOP_KEYWORDS*(*ARTICLES.HS.Article*),
((*ARTICLES.Reference.AUTEURS.HA*, <*Auteur, Institut*>),
(*CONFERENCES.HConf*, <*Nom*>))),
ARTICLE.Reference.AUTEURS.Institut = *'Inst1'*)

Où *ARTICLES.Reference.AUTEURS* sont les auteurs des articles cités dans les conférences *CONFERENCES.Nom* dans les articles dont le contenu est spécifié par *ARTICLES.Article*. Il est à noter que les hiérarchies ne sont spécifiées que dans l'instruction focus afin de permettre par la suite l'emploi d'opérations de forages qui suivent l'agencement hiérarchique des paramètres.

Ces liens assurent une flexibilité quand à l'expression de requêtes sur des données fortement interconnectées et permettent une exploration plus complète des jeux de données.

4 Conclusion et perspectives

Dans ce document nous avons fourni un modèle conceptuel multidimensionnel adapté pour l'analyse de documents orientés documents. Ce modèle est basé sur l'unique concept de dimension. Le modèle est associé à un ensemble d'opérations de manipulations permettant l'analyse multidimensionnelle OLAP.

Contrairement aux précédents modèles multidimensionnels, cette proposition a l'avantage de préserver la structure des documents, mais aussi les liens entre ces structures. Les opérations de manipulation permettent à l'utilisateur de naviguer en exploitant les liens et de faciliter l'expression des requêtes. Ces requêtes seraient très complexes dans d'autres environnements. L'absence d'entité factuelle ne restreint pas l'utilisateur à des sujets d'analyse prédéfinis qui pourraient le contraindre et le mener à des analyses manquant de sens sur des données textuelles. Les opérations de manipulation associées permettent des permutations aisées du sujet d'analyse. Ainsi, l'utilisateur peut bénéficier d'une plus grande flexibilité dans cet environnement OLAP. Ceci permet de pallier de manière non négligeable au manque de précision généralement imputé à l'analyse textuelle. Toutefois, le modèle proposé est plus générique que les modèles multidimensionnels classique car il permet aussi de modéliser les faits par des dimensions simplifiées (un attribut pour une mesure).

Nous implantons actuellement un prototype basé sur un SGBD Oracle *10g*, des fichiers XML et un outil Java. Dans notre implantation, afin de maintenir un certain niveau de performance chaque dimension est reliée, au niveau logique, à l'ensemble des instances des autres dimensions permettant ainsi des rotations aisées et rapides entre les sujets d'analyse. Dans une architecture ROLAP, ceci se traduit par une clé étrangère dans chaque table dimension pour chaque autre dimension liée (les autres dimensions liées au nœud central).

Ce modèle conceptuel est une étape vers un environnement plus complet. Au sein de ce document, nous avons suggéré l'emploi d'une fonction d'agrégation simple (TOP_KEYWORDS). Nous envisageons d'aller plus loin et nous avons proposé une nouvelle fonction d'agrégation (Ravat *et al.*, 2007) pour faciliter l'analyse de documents orientés documents. En parallèle, le but de ce modèle et de ces opérations associées étant de simplifier l'expression d'analyses multidimensionnelles, nous adaptons et développons un langage graphique de requêtes multidimensionnelles OLAP.

Références

Abelló, A., J. Samos, et F. Saltor (2003). Implementing operations to navigate semantic star schemas. *6th Int. Workshop on Data Warehousing and OLAP* (*DOLAP*), ACM Press, p.56–62.

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

Agrawal, R., A. Gupta, et S. Sarawagi (1997). Modeling Multidimensional Databases. *13th Int. Conf. on Data Engineering* (*ICDE*), IEEE Computer Society, p.232–243.

Boussaid O., R.B. Messaoud, R. Choquet, et S. Anthoard (2006). Conception et construction d'entrepôts XML. *2ème journée francophone sur les Entrepôts de Données et l'Analyse en ligne* (*EDA*), RNTI, p. 3–21.

Cabibbo, L., et R. Torlone (1997). A Systematic Approach to Multidimensional Databases. 5th *National* (*Italy*) *Conference on Advanced Database Systems* (*SEBD*), p.361–377.

Fuhr, N., et K. Großjohann (2001). XIRQL: A Query Language for Information Retrieval in XML Documents. *4th Int. ACM conf. on research and development in Information Retrieval* (*SIGIR*), ACM Press, p. 172–180.

Golfarelli, M., D. Maio, et S. Rizzi (1998). The Dimensional Fact Model: a Conceptual Model for Data Warehouses. *Int. Journal of Cooperative Information Systems* (*IJCIS*), vol.7(2&3), p. 215–247.

Golfarelli, M., S. Rizzi, et E. Saltarelli (2002). WAND: A CASE Tool for Workload-Based Design of a Data Mart. 10th *National* (*Italy*) *Conference on Advanced Database Systems* (*SEBD*), p.422–426.

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

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

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

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

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

Khrouf K., et C. Soulé-Dupuy (2004). A Textual Warehouse Approach: A Web Data Repository. *Intelligent Agents for Data Mining and Information Retrieval*, Idea Group, p. 101–124.

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

McCabe C., J. Lee, A. Chowdhury, D.A. Grossman, et O. Frieder (2000). 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, p. 363–365.

Mothe J., C. Chrisment, B. Dousset, et J. Alau (2003). 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, p. 650–659.

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

Park, B.K., H. Han, et I-Y. Song (2005). XML-OLAP: A Multidimensional Analysis Framework for XML Warehouses. *7th Int. Conf. on Data Warehousing and Knowledge Discovery* (*DaWaK*), LNCS 3589, Springer, p.32–42.

Pérez, J.M., R.B. Llavori, M.J. Aramburu, et T.B. Pedersen (2005). A relevance-extended multi-dimensional model for a data warehouse contextualized with documents. *8th ACM Int. Workshop on Data Warehousing and OLAP* (*DOLAP*), ACM Press, p.19–28.

Rafanelli, M. (2003). Operators for Multidimensional Aggregate Data. Chap. V de *Multidimensional Databases: Problems and Solutions*, Idea Group, p. 116–165.

Ravat F., O. Teste, et G. Zurfluh (2006). Algèbre OLAP et langage graphique. *XXIV^e Congrès Informatique des organisations et systèmes d'information et de decision* (*INFORSID*), Inforsid, p. 1039–1054.

Ravat F., O. Teste, et R. Tournier (2007). OLAP Aggregation Function for Textual Data Warehouse. *9^th Int. Conf. on Enterprise Information Systems* (*ICEIS*), INSTICC Press, Juin 2007 *(à paraître)*.

Ravat, F., O. Teste, R. Tournier, et G. Zurfluh (2007). Algebraic and graphic languages for OLAP manipulations. *Int. Journal of Data Warehousing and Mining* (*ijDWM*), IDEA Group Publishing (*à paraître*).

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

Torlone, R. (2003). Conceptual Multidimensional Models", Chap. III de *Multidimensional Databases: Problems and Solutions*, Idea Group, p.69–90.

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

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

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

Summary

Data warehousing and OLAP are mainly used for the analysis of transactional data. Nowadays, with the evolution of Internet, and the development of semi structured data exchange format (such as XML), it is possible to consider documents as analysis sources. As a consequence, a multidimensional analysis framework for such data needs to be provided. In this document, we introduce an OLAP multidimensional conceptual model adapted for the analysis of textual data that rests on a unique concept: a dimension. We also define a set of manipulation operations for multidimensional analysis.