DocWare : Vers l'entreposage et l'analyse multidimensionnelle de documents

Kaïs Khrouf — Chantal Soulé-Dupuy

Laboratoire IRIT, Equipe SIG/D2S2, Campus Université Paul Sabatier
118, route de Narbonne, F-31062 Toulouse Cedex 4

{khrouf, soule}@irit.fr

Université Toulouse I
Place France Anatole, F-31042 Toulouse Cedex

RÉSUMÉ. *L'augmentation du nombre de documents numériques gérés par les entreprises n'a fait qu'accroître les difficultés d'exploitation des informations textuelles. Ces difficultés sont en grande partie liées aux volumes à manipuler, mais également à l'hétérogénéité des sources et aux normes de structuration des informations documentaires. Il devient alors nécessaire, voire indispensable, de disposer d'outils d'intégration rendant les informations utiles accessibles, permettant de les manipuler et de les analyser. A cette fin, nous proposons le concept d'entrepôt de documents permettant d'intégrer et d'organiser des informations hétérogènes. Ces informations peuvent alors être analysées selon plusieurs points de vue (analyse multidimensionnelle) afin d'en déduire de nouvelles informations ou connaissances.*

ABSTRACT. *The increase in the number of numerical documents managed by the companies generated an increase in the exploitation difficulties of textual information. These difficulties are mainly related to volumes to be manipulated, but also to the heterogeneity of sources and the standards of documentary information. It then becomes necessary to have integration tools in order to render useful information accessible, making it possible to manipulate them and to analyze them. For this purpose, we propose the concept of document warehouse allowing to integrate and to organize heterogeneous information. This information can be then analyzed according to several viewpoints (multidimensional analysis) in order to deduce new information or knowledge.*

MOTS-CLÉS : *entrepôt, modèle générique, analyse multidimensionnelle.*

KEYWORDS: *warehouse, generic model, multidimensional analysis.*

1. Introduction

De nos jours, une entreprise est considérée non seulement comme une unité de production de biens ou de services, mais également comme une unité de production de connaissances à capitaliser au travers du concept de mémoire d'entreprise. Cette mémoire d'entreprise inclut non seulement une mémoire technique obtenue par capitalisation du savoir-faire de ses employés, mais également une mémoire documentaire permettant le stockage et l'exploitation des connaissances extraites des documents gérés par l'entreprise. Afin d'exploiter les informations textuelles, les utilisateurs jusqu'à nos jours ont généralement recours aux systèmes de recherche d'information traditionnels basés sur des index de termes. Ces systèmes ne permettent pas d'exploiter la structure logique des documents de façon appropriée ; la prise en compte de telle structure logique donne alors l'opportunité d'identifier et de se focaliser sur des parties ou « granules » de documents.

La définition de normes, telles que Structured Generalized Markup Language (SGML) puis eXtensible Markup Language (XML), et les travaux du World Wide Web Consortium (W3C) ont permis d'initier de nouvelles perspectives concernant l'exploitation des documents numériques et en particulier des informations textuelles semi-structurées. Ces perspectives se situent à la frontière entre les bases de données (Gardarin, 2002) et la recherche d'information (Kotsakis, 2002). C'est dans ce contexte de perspectives que nous proposons la constitution d'une mémoire documentaire sur la base de la spécification et de l'exploitation d'entrepôts de documents textuels comme une extension du concept d'entrepôts de données (Inmon et al., 1994). La principale caractéristique d'un entrepôt de documents est de permettre le stockage de documents hétérogènes, sélectionnés et filtrés, ainsi que leur classification selon des structures logiques génériques (structures communes à un ensemble de documents). Une telle organisation des entrepôts permet de faciliter l'exploitation structurelle et de contenu des informations documentaires au travers plusieurs techniques telles que la recherche, l'interrogation voire l'analyse.

L'originalité de notre approche basée sur les entrepôts de documents se situe principalement au niveau du modèle générique que nous proposons. Ce modèle permet l'intégration, à la différence des travaux proposés dans la littérature, de tout type de documents (structurés, semi-structurés ou non-structurés) et leur regroupement sous forme de collections en fonction de leurs structures sous-jacentes. De plus, le modèle générique proposé organise les informations documentaires intégrées dans l'entrepôt de manière à permettre leur analyse multidimensionnelle (technique habituellement appliquée aux données factuelles).

Ce papier se décompose comme suit. Après un panorama de travaux réalisés pour le stockage et la manipulation des documents (section 2), nous décrivons le modèle générique d'entrepôts de documents que nous proposons (section 3). Dans la section 4, nous détaillons notre approche d'analyse multidimensionnelle des informations documentaires. Enfin, nous présentons dans la section 5 les travaux d'expérimentations réalisés au travers de l'outil DocWare (Document Warehouse).

2. Etat de l'art

De nos jours, les travaux existants, visant au stockage et à la manipulation de documents numériques, le plus souvent au format XML, peuvent être classés en deux approches, à savoir : les SGBD natifs et les SGBD middlewares.

Les SGBD natifs sont conçus pour XML et permettent de stocker les documents sans les décomposer en éléments éclatés dans des tables. Ces systèmes utilisent des techniques d'indexation spécifiquement développées pour XML. Il s'agit de stocker les documents sous forme d'arbres avec des index arborescents ou hiérarchiques. Plusieurs travaux ont été proposés, parmi lesquels nous pouvons citer : Natix (Kanne et al., 2000), InfonyteDB (Huck et al., 2000) et Xylème (Abiteboul et al., 2002). L'interrogation de documents en utilisant ce type de SGBD se résume alors au calcul (généralement complexe) de chemins pour identifier les informations recherchées. Cependant, ces SGBD ne permettent pas une classification ou regroupement de documents qui permettrait de simplifier et d'améliorer les calculs de chemins.

Les SGBD middlewares permettent l'intégration des documents XML selon un modèle relationnel/objet. Il s'agit en fait de stocker les documents et de les « mapper » en tables ce qui permet de supporter des requêtes sur des collections de documents. Plusieurs travaux ont été proposés : XRel (Yoshikawa et al., 2001), e-XMLRepository (Gardarin et al., 2002) et (Darmont et al., 2003). Les approches basées sur les SGBD middlewares ne permettent pas de représenter complètement la flexibilité de XML. Il s'agit d'un langage semi-structuré et ses structures ne sont pas généralement fixes.

La plupart des travaux actuels (basés sur les SGBD natifs ou middlewares) traitent d'un type spécifique de documents, à savoir des documents au format XML (en s'appuyant sur l'hypothèse que XML deviendra le langage de description des données du Web de demain) définis selon un ensemble de balises. Par rapport à ces travaux, notre approche se veut générique dans le sens où aucune contrainte ne peut être imposée aux types de documents à manipuler, tout en se basant sur les concepts XML.

Dans ce sens, nous proposons un modèle générique permettant l'intégration et l'entreposage de documents hétérogènes, sans pour autant imposer une structure prédéfinie de bases de données. En outre, ce modèle générique permet de gérer la flexibilité des documents semi-structurés puisqu'il associe à chaque document intégré dans l'entrepôt une structure générique (schéma) ainsi qu'une structure spécifique (instance) qui sera reliée aux différentes informations de contenu.

Une des spécificités de nos travaux est de permettre d'analyser les documents de l'entrepôt sous n'importe quel angle ou perspective (axe). Les travaux existants ne traitent pas de l'aspect analyse qui peut en être fait au sens décisionnel. Dans ce papier, nous proposons notre approche d'analyse multidimensionnelle appliquée aux informations documentaires au travers de l'outil que nous avons réalisé DocWare.

3. Entreposage des documents

Dans le cadre de nos travaux, nous proposons le concept d'entrepôt de documents permettant d'intégrer et d'exploiter des documents jugés pertinents. Cependant, le principal problème à gérer dans le stockage et la manipulation des informations textuelles est celui de l'hétérogénéité, qu'elle soit structurelle ou sémantique (de contenu). En effet, ces informations peuvent être hétérogènes :

– sur le fond : elles peuvent concerner des domaines très divers. L'utilisateur doit fouiller et parcourir les documents pour trouver l'information utile. Elles peuvent également être écrites dans plusieurs langues,

– sur la forme : elles peuvent être de structures différentes, plus ou moins structurées, de formats standards ou non. L'utilisateur doit ainsi disposer d'outils adéquats pour visualiser et/ou exploiter l'information.

Afin de gérer cette hétérogénéité (structurelle, thématique voire linguistique), nous proposons un modèle générique (cf. figure 1) d'entrepôts de documents évolutif et indépendant de toute norme de représentation. Nous distinguons alors la structure logique générique de la structure logique spécifique :

– la structure logique générique : elle décrit les structures logiques communes à un ensemble de documents, elle regroupe ainsi toute une classe de documents ayant des structures logiques identiques ou similaires,

– la structure logique spécifique : elle correspond à une spécialisation de la structure logique générique. Elle est unique et correspond à un document et un seul.

REMARQUE. — Des exemples d'instanciation du modèle générique par différents types de documents ont été proposés dans (Khrouf, 2004).

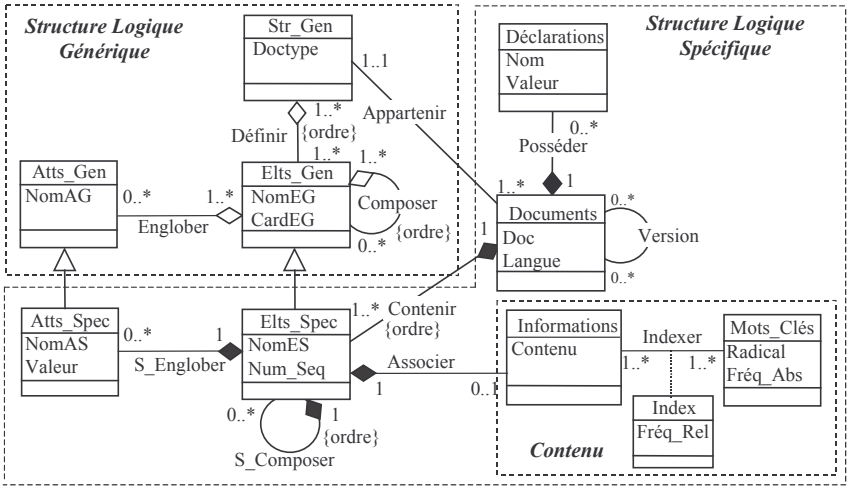


Figure 1. *Modèle générique d'entrepôts de documents*

L'apport de cette représentation réside dans le fait qu'une structure logique générique peut être considérée comme le schéma de tous les documents appartenant à cette structure alors que les structures logiques spécifiques de ces documents constituent les différentes instances du schéma. Ainsi, le modèle générique proposé (intégration de tout type de documents) peut être considéré comme un méta-modèle.

Le modèle générique (cf. figure 1) comporte les composants suivants :

– la description des éléments des structures logiques génériques. Une structure logique générique est définie par un ensemble d'éléments génériques pouvant être composés d'autres éléments génériques et/ou décrits par des attributs génériques,

– la description des éléments décrivant une structure logique spécifique. Une structure logique spécifique est définie par un ensemble d'éléments spécifiques pouvant englober des attributs spécifiques et/ou des informations,

– la description des éléments décrivant le contenu textuel de chaque élément de la structure logique spécifique. Il s'agit d'extraire les termes descriptifs en se basant sur des techniques issues de l'indexation automatique de texte (Soulé-Dupuy, 2001).

L'intérêt d'un tel modèle générique réside dans :

– l'intégration de tous types de documents caractérisés par une absence totale ou partielle de structure. Dans (Khrouf et al., 2003a), nous avons défini une technique d'extraction de structures et de contenus pour chaque type de documents : (1) pour les documents structurés, nous avons déterminé un ensemble de règles d'extraction permettant de gérer les liens hiérarchiques existants entre les balises, (2) pour les documents semi-structurés, des règles de réécritures ont été proposées et utilisées permettant de supprimer les balises de présentation et de mettre en évidence les balises structurelles et (3) pour les documents non structurés, nous avons utilisé les techniques de segmentation permettant de déterminer des unités documentaires fines et cohérentes formant des blocs sémantiques,

– le regroupement des documents selon des structures communes identiques ou approchantes. Dans (Khrouf et al., 2003b), nous avons défini une approche basée sur le calcul de similarité (ressemblance) d'arborescences hétérogènes d'éléments ordonnés et étiquetés. Il s'agit de comparer la structure logique du document à intégrer avec celles de l'entrepôt. Deux cas se présentent : si une structure identique ou approchante existe (degré de similarité > seuil de similarité fixé par des expérimentations), le système fusionne les deux structures en une structure logique générique et rattache le document à cette structure. Dans le cas contraire, le système crée une nouvelle structure logique générique afin d'y rattacher le document,

– la persistance des documents jugés pertinents. Les informations issues du Web par exemple sont très volatiles, c'est à dire qu'elles évoluent rapidement et peuvent disparaître, être modifiées ou encore changer de localisation fréquemment. Le modèle générique proposé permet de garder les différentes versions d'un même document afin de détecter les changements structurels et de contenu,

– la mise en évidence de la granularité des documents. Ce modèle générique permet de représenter les index des différentes parties et non la totalité d'un document (partie « contenu » dans figure 1). Pour cela, nous avons adapté les formules de pondération de (Sparck Jones, 1972) afin de tenir compte de la granularisation des documents. Ceci permet une grande précision de la recherche puisque nous calculons la similarité entre les requêtes et les granules des documents,

– l'application des techniques d'analyse multidimensionnelle aux informations documentaires. Cette partie est détaillée dans ce qui suit.

4. Analyse multidimensionnelle

La modélisation multidimensionnelle consiste à considérer un sujet analysé comme un point dans un espace à plusieurs dimensions (Kimball et al., 2002). A cette fin, nous proposons de dériver des magasins de documents (extraits de l'entrepôt) supportant les processus d'analyses décisionnelles. Un magasin de documents est dédié à un type d'utilisateurs et il doit répondre à un objectif décisionnel précis ou un besoin spécifique.

Le processus proposé, pour analyser d'une manière multidimensionnelle les informations contenues dans l'entrepôt, peut être schématisé comme l'indique la figure 2. Ce processus se compose de trois phases.

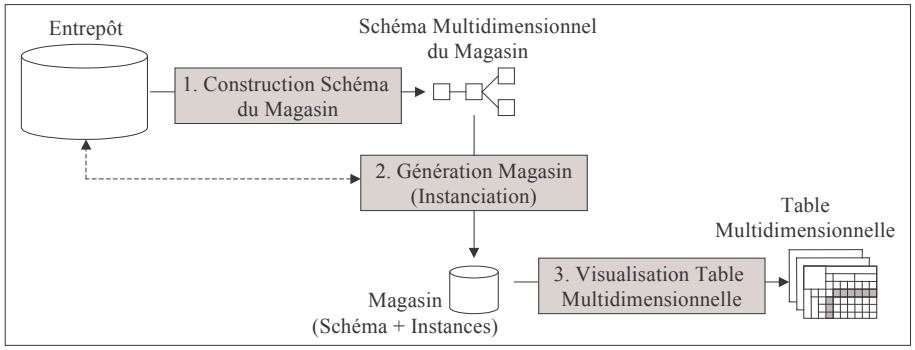


Figure 2. *Processus d'analyse multidimensionnelle*

4.1. *Construction des schémas des magasins*

La première phase du processus d'analyse multidimensionnelle consiste à générer, à partir de l'entrepôt, le schéma du magasin de documents désiré. Cette phase se compose de quatre étapes, à savoir : (1) choix du type d'analyse, (2) sélection des composants d'analyse, (3) filtrage et (4) visualisation du schéma.

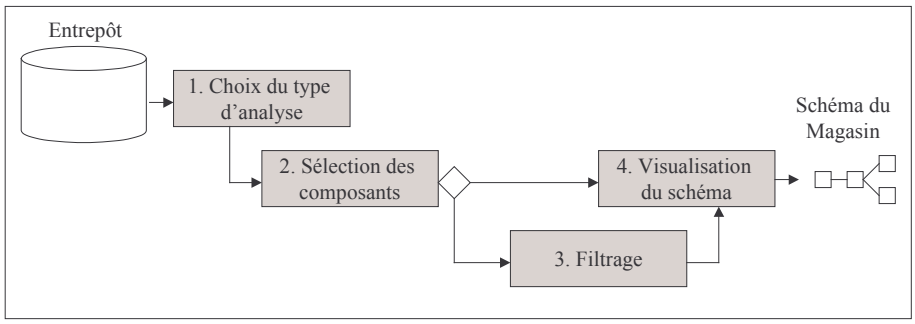


Figure 3. *Phase de construction des schémas des magasins*

1. La première étape doit permettre à l'utilisateur de choisir un type d'analyse. Il s'agit de décider de travailler sur des documents ayant des structures similaires ou différentes ou même sur un seul document. Nous distinguons ainsi trois types d'analyse, à savoir :

– par collection : cela consiste à analyser les documents appartenant à la même structure logique générique (ayant le même schéma),

– par composants : cela consiste à analyser des documents appartenant à plusieurs structures logiques génériques (ayant des schémas identiques ou différents mais des éléments communs),

– par document : cela consiste à analyser le contenu d'un et d'un seul document (ayant un seul schéma).

Ces différents types d'analyse permettront à l'utilisateur de se focaliser sur une structure précise, sur un domaine bien défini ou même sur un document, selon ses besoins. Ceci va tout à fait dans le sens des objectifs de l'analyse multidimensionnelle, à savoir se focaliser sur un sujet ou un thème bien particulier.

2. La deuxième étape doit permettre à l'utilisateur de sélectionner les composants d'analyse, à savoir :

– un fait qui modélise un sujet de l'analyse, il constitue un centre d'intérêt décisionnel. Il regroupe un ensemble d'attributs représentant les mesures d'activité. Une mesure est un indicateur d'analyse de type numérique et cumulable.

EXEMPLE. — Considérons le fait *Apparition* d'une maison d'édition pouvant être constitué de la mesure d'activités suivante : *Nombre* d'ouvrages apparus.

– ses dimensions qui modélisent des axes d'analyse selon lesquels sont visualisées les mesures d'activité d'un sujet d'analyse. En d'autres termes, ce sont les critères sur lesquels nous souhaitons évaluer, quantifier et qualifier les faits.

EXEMPLE. — Le fait *Apparition* peut être analysé selon les dimensions suivantes : *Domaine*, *Année* et *Type* (livre, revue, etc.).

L'utilisateur doit indiquer aussi l'ordre des dimensions et la fonction d'agrégation pour la mesure (indicateur d'analyse) du fait (Compte, Somme, Maximum, Minimum, Moyenne).

3. La troisième étape, celle de filtrage, doit permettre à l'utilisateur de sélectionner des valeurs précises afin d'affiner ses analyses. Nous distinguons deux types de filtrage :

– pour une dimension, nous choisissons, parmi toutes ses valeurs, celles que nous voulons intégrer dans le magasin,

– pour le fait qui est toujours sous forme numérique, nous proposons un filtrage plus fin qui nous permet de fixer des critères de sélection, en utilisant des opérateurs classiques de comparaison (<, >, =, <>, <=, >=).

4. La dernière étape, celle de visualisation, consiste à restituer à l'utilisateur le schéma du magasin de documents selon une représentation graphique facilitant les analyses décisionnelles.

4.2. *Génération automatique des magasins*

A ce niveau, la tâche de l'utilisateur est terminée. La phase suivante consiste à générer le magasin d'une manière automatique afin de récupérer les informations de l'entrepôt. Cette génération se fait en deux étapes : (1) génération d'une vue pour chaque composant d'analyse et (2) jointure et groupement des différentes vues générées.

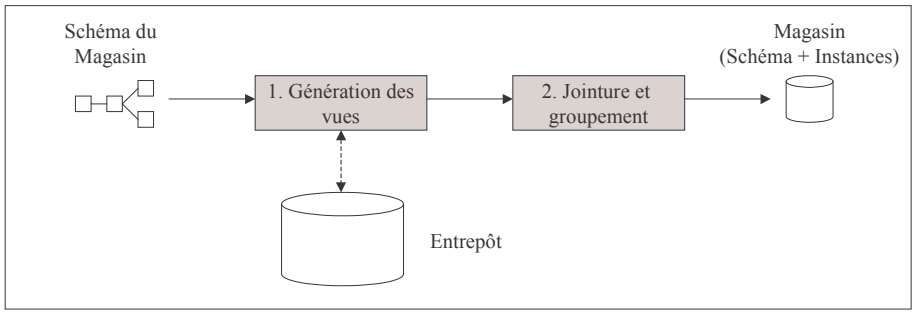


Figure 4. *Phase de génération automatique des magasins*

4.2.1. *Génération d'une vue pour chaque composant d'analyse*

Pour chaque objet (élément ou attribut) constituant un composant d'analyse (dimension ou fait), le système doit générer une vue qui englobe trois attributs *Doc*, *Anc* et *Inf* :

– le premier attribut *Doc* de la vue correspond aux numéros des documents extraits de l'entrepôt concernant tous les éléments ou attributs spécifiques qui héritent respectivement de l'élément ou de l'attribut générique jouant le rôle d'un composant d'analyse (fait ou dimension),

– le deuxième attribut *Anc* de la vue correspond aux numéros des éléments ou des attributs spécifiques qui héritent du premier ancêtre commun de tous les éléments d'analyse,

– le dernier attribut *Inf* de la vue correspond à l'information contenue dans l'élément ou l'attribut spécifique qui hérite respectivement de l'élément ou de l'attribut générique correspondant.

REMARQUE. — Les attributs *Doc* et *Anc* constituent les clés primaires pour les vues générées par le système. Ces attributs seront utilisés par la suite pour effectuer la jointure entre ces différentes vues.

La vue d'une dimension, correspondant à un élément générique dans le cas d'une analyse par collection, aura la forme suivante.

```
CREATE VIEW Dim_n (Doc, Anc, Inf) AS
SELECT e.s_composer… .s_composer.sondoc.numdoc,
       e.s_composer… .s_composer.numes,
       e.contenu
FROM   Elts_Spec e
WHERE  e.herite.nomeg = "Inf"
AND e.s_composer… .s_composer.sondoc.appartient.doctype = "SG"
AND   (e.contenu = "V1" OR … OR e.contenu = "Vn") ;
```

4.2.2. *Jointure et groupement des différentes vues générées*

Le système doit ensuite, à partir des vues générées, établir une nouvelle vue par une jointure sur les deux premiers attributs *Doc* et *Anc* de toutes les vues. La nouvelle vue aura alors la forme suivante.

```
CREATE VIEW Jointure (Inf_1, Inf_2… , Inf_n, Inf) AS
SELECT d1.Inf, d2.Inf… , dn.Inf, f.Inf
FROM   Dim_1 d1, Dim_2 d2… , Dim_n dn, Fact f
WHERE  d1.Doc = d2.Doc     AND    d2.Doc = d3.Doc
…
AND    dn-1.Doc = dn.Doc   AND    dn.Doc = f.Doc
AND    d1.Anc = d2.Anc     AND    d2.Anc = d2.Anc
…
AND    dn-1.Anc = dn.Anc   AND    dn.Anc = f.Anc ;
```

Le système doit effectuer à ce niveau une opération de groupement en utilisant la fonction d'agrégation choisie par l'utilisateur, pour générer une dernière vue qui représente le contenu du magasin. Cette vue aura la forme suivante.

```
CREATE VIEW Vue (j.Inf_1, j.Inf_2… , j.Inf_n) AS
SELECT      j.Inf_1, j.Inf_2… , j.Inf_n, Function(j.Inf)
FROM        Jointure j
GROUP BY    j.Inf_1, j.Inf_2… , j.Inf_n ;
```

4.3. *Visualisation*

Une fois que le magasin de documents a été généré, la dernière phase, celle de visualisation, est déclenchée. Elle consiste à afficher le contenu de la dernière vue générée par le système sous forme de tables multidimensionnelles assez simples à manipuler et à interpréter. En effet, ces tables permettent de mieux apprécier le contenu des magasins de documents. Elles organisent les données en les classant suivant les dimensions, déjà choisies par l'utilisateur. Ainsi, les colonnes représentent la première dimension, les lignes représentent la deuxième dimension et les plans représentent la troisième dimension. Alors que les valeurs des mesures des faits sont représentées à l'intérieur des tables sous formes d'interrelation entre les différentes valeurs des dimensions.

Le passage de la dernière vue générée par le système en une table multidimensionnelle se fait de la manière suivante : Etant donné que chaque plan de la table multidimensionnelle correspond à une seule valeur de la troisième dimension, le système génère une vue en effectuant une sélection sur une valeur précise. Cette nouvelle vue contient trois colonnes : (1) la première dimension, (2) la deuxième dimension et (3) le fait.

A partir de cette vue, le système doit :

– récupérer toutes les valeurs possibles de la première dimension, ces valeurs seront affichées dans les colonnes du plan correspondant,

– récupérer toutes les valeurs possibles de la deuxième dimension, ces valeurs seront affichées dans les lignes du plan correspondant,

– restituer pour chaque couple (une colonne *i* et une ligne *j*) la mesure à partir de la troisième colonne de la vue (le fait). Cette mesure sera affichée dans la case correspondante (intersection entre *i* et *j*).

Nom table			Dimension 1						
			P1	$ValP1_1$			$ValPar1_2$		
			P2	$ValP2_1$	$ValP2_2$	…	$ValP2_i$	$ValP2_{i+1}$	…
Dimension 2	P3	Par4	(M1,M2)						
	$ValP3_1$	$ValP4_1$		(X,Y)	…	…	…	…	…
		$ValP4_2$		(*,*)	…	…	…	…	…
		…		…	…	…	…	…	…
	$ValP3_2$	$ValP4_j$		…	…	…	…	…	…
		$ValP4_{j+1}$		…	…	…	…	…	…
		…		…	…	…	…	…	…

Figure 5. *Représentation graphique d'une table multidimensionnelle*

5. Implantation et expérimentations

Afin de valider les propositions présentées, nous avons réalisé un outil d'aide à l'intégration et à l'analyse de documents textuels, intitulé DocWare (Document Warehouse). Plus précisément, DocWare assure les fonctionnalités suivantes : (1) supporter une construction incrémentale des entrepôts de documents à partir de documents filtrés et sélectionnés récupérés de sources hétérogènes et (2) assister l'administrateur (ou l'utilisateur) dans l'élaboration des magasins de documents.

Nous proposons dans cette section une validation de la démarche présentée concernant le module d'analyse multidimensionnelle. Les documents ayant servi à nos expérimentations ont été extraits de sources diverses : documents XML issus de sites Web et de CD-ROM fournis dans le cadre de benchmarks (Reuters, …) ou de bases de tests (TREC, …). Ces documents ne sont pas associés à un domaine particulier ; leurs caractéristiques sont représentées dans le tableau suivant.

Nombre de structures logiques génériques	23
Nombre de documents	1675
Nombre d'éléments génériques	292
Nombre d'attributs génériques	52
Nombre d'éléments spécifiques	61299
Nombre d'attributs spécifiques	19000

Tableau 1. *Caractéristiques du contenu de l'entrepôt*

5.1. *Cas d'une analyse par collection*

Nous souhaitons analyser les films sortis lors des dernières années (2001, 2002, 2003) par genre et par pays. Le nombre de films vérifiant les trois critères (Année, Genre et Pays) doit être supérieur à 4.

Contexte

En observant le contenu de l'entrepôt, nous déduirons que les documents décrivant les films sont regroupés selon une seule structure logique générique, intitulée *Film*. Chaque film est décrit par l'année de sa sortie, son titre, son genre, le pays où il a été tourné et ses principaux acteurs. Cette structure logique générique contient alors tous les éléments nécessaires pour réaliser cette analyse.

Démarche

1. Choix du type d'analyse

La première étape de la construction du magasin consiste à sélectionner le type d'analyse (dans notre exemple, c'est par collection). Ainsi, le système affiche la liste

de toutes les structures existantes dans l'entrepôt. Parmi ces structures, nous choisissons la structure logique générique *Film* qui sera par la suite visualisée d'une manière automatique sous forme d'une arborescence.

2. Sélection des composants

A ce niveau, nous sélectionnons et définissons les composants d'analyse, il s'agit de préciser les dimensions et le fait. L'affectation de ces rôles se fait au travers des menus contextuels. Pour cela, nous pointons l'élément désiré et avec un clic sur le bouton droit nous fixons notre choix (Fait ou Dimension) ainsi que les attributs, à savoir : l'ordre pour les dimensions et la fonction d'agrégation pour le fait (Compte, Somme, Maximum, Minimum, Moyenne). Dans notre exemple, la première dimension est l'année, la deuxième dimension est le genre et la troisième dimension est le pays. La mesure du fait est le calcul du nombre de titres.

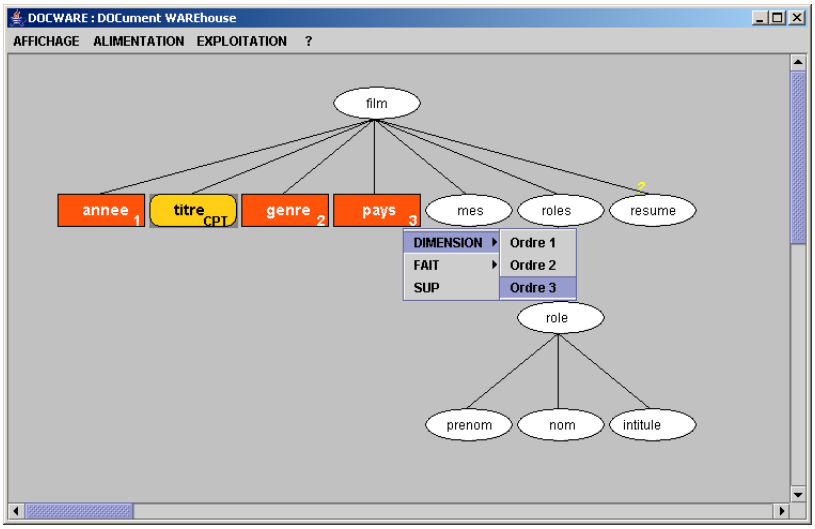


Figure 6. *Cas n°1 — Sélection des composants*

3. Filtrage

Nous souhaitons analyser les films pour les années suivantes : 2001, 2002 et 2003. Nous appliquons ainsi un filtre sur la première dimension. Pour cela, le système affiche alors toutes les valeurs de l'élément *Année*. Parmi ces valeurs, nous choisissons les années correspondantes, à savoir : 2001, 2002 et 2003.

Une deuxième contrainte indique que le nombre de films par année, genre et pays doit être supérieur à 4. Nous appliquons alors un filtre sur la mesure du fait. Le système affiche une boite de dialogue pour nous permettre de spécifier le critère de sélection.

4. Visualisation du schéma

A ce niveau, nous pouvons consulter le schéma multidimensionnel du magasin d'une manière graphique (cf. figure 7).

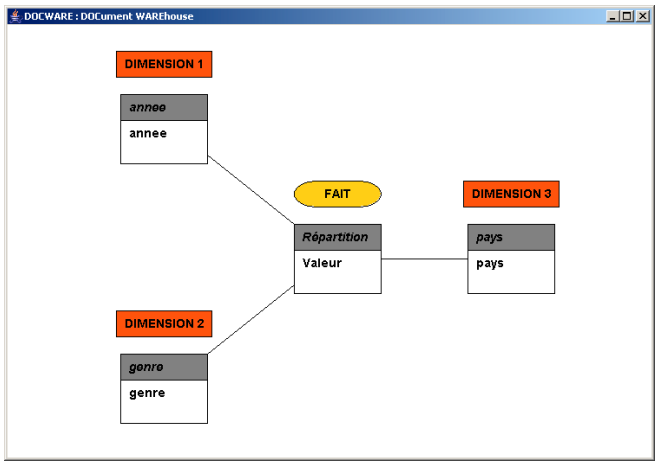
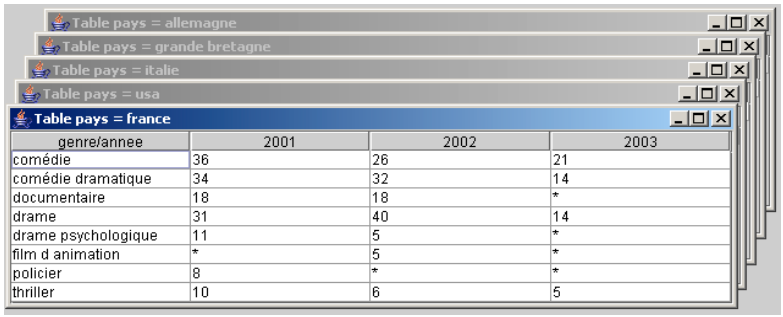


Figure 7. *Cas n°1 — Visualisation du schéma multidimensionnel*

Résultat

Pour visualiser le résultat, le système crée les vues selon la démarche décrite dans la section 4.2 et affiche la table multidimensionnelle suivante.



genre/annee	2001	2002	2003
comédie	36	26	21
comédie dramatique	34	32	14
documentaire	18	18	*
drame	31	40	14
drame psychologique	11	5	*
film d animation	*	5	*
policier	8	*	*
thriller	10	6	5

Figure 8. *Cas n°1 — Table multidimensionnelle*

5.2. *Cas d'une analyse par composants*

Nous souhaitons analyser les articles de notre laboratoire de recherche par type de publication, année d'apparition et thème de recherche.

Contexte

Les documents, qui décrivent les publications de notre laboratoire de recherche, sont regroupés dans l'entrepôt selon deux structures logiques génériques *Pub-conf* et *Pub-revue*.

Démarche

1. Choix du type d'analyse

Les documents nécessaires pour cette analyse sont répartis sur deux structures logiques génériques, l'analyse par collection c'est à dire une seule structure logique générique n'est pas alors envisageable. Nous utilisons ainsi l'analyse par composants.

2. Sélection des composants

Pour ce type d'analyse, le système affiche une fenêtre dans laquelle nous précisons les éléments nécessaires pour l'analyse. Dans notre exemple, nous affectons l'élément *Type* à la première dimension, l'élément *Année* à la deuxième dimension, l'élément *Thème* à la troisième dimension, ainsi que l'élément *Titre* au fait en précisant sa fonction d'agrégation *Compte*.

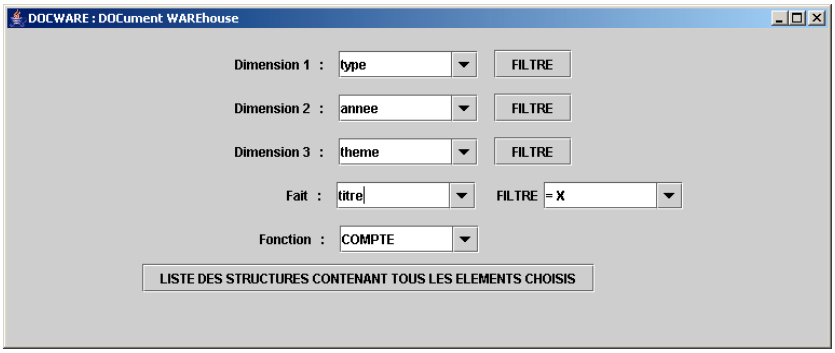


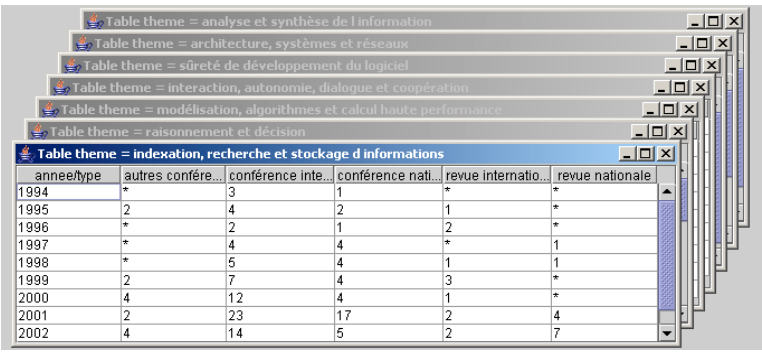
Figure 9. *Cas n°2 — Sélection des composants*

3. Filtrage

Le système affiche par la suite la liste de toutes les structures logiques génériques de l'entrepôt contenant tous les éléments sélectionnés. Parmi ces structures, nous choisissons celles qui nous intéressent. Dans notre exemple, nous sélectionnons les deux structures logiques génériques *Pub-conf* et *Pub-revue*.

Résultat

Le résultat de cette analyse est la table multidimensionnelle suivante comme l'indique la figure 10.



annee/type	autres confére...	conférence inte...	conférence nati...	revue internatio...	revue nationale
1994	*	3	1	*	*
1995	2	4	2	1	*
1996	*	2	1	2	*
1997	*	4	4	*	1
1998	*	5	4	1	1
1999	2	7	4	3	*
2000	4	12	4	1	*
2001	2	23	17	2	4
2002	4	14	5	2	7

Figure 10. *Cas n°2 — Table multidimensionnelle*

6. Conclusion

Une mémoire d'entreprise doit pouvoir servir de base à des processus de veille scientifique et technique. Cependant, les informations utiles à ces processus ne se trouvent pas uniquement dans les bases opérationnelles des entreprises mais également dans la masse d'informations textuelles et de documents échangés dans le cadre de leurs activités. Il apparaît clairement qu'un entrepôt de documents semble être le meilleur outil informatique pour la gestion et l'exploitation aisée des informations documentaires.

Dans ce papier, nous avons présenté un modèle générique d'entrepôts de documents permettant d'intégrer des documents issus de sources disséminées et hétérogènes. Nous avons présenté aussi une démarche de construction de magasins de documents permettant l'utilisation des techniques d'analyse multidimensionnelle aux informations documentaires contenues dans l'entrepôt. L'originalité de ce type d'analyse réside dans : (1) une possibilité de visualisation du contenu documentaire de l'entrepôt selon plusieurs axes d'analyse ou dimensions et (2) une manipulation et une spécification faciles des magasins de documents car la démarche proposée ne nécessite l'apprentissage d'aucun langage. Cette démarche est réalisée sous forme graphique.

Les travaux menés jusqu'à aujourd'hui dans le cadre des entrepôts de documents textuels nous encouragent à approfondir les approches proposées. Plusieurs perspectives à ces travaux sont envisagées : (1) une analyse multidimensionnelle basée sur le contenu des documents en plus de la structure s'avère intéressante, (2) l'adaptation du modèle générique pour intégrer des documents multimédia permettra plus de fonctionnalités à l'entrepôt (manipulation des images, vidéo, etc.) et (3) une expérimentation « à plus grande échelle » afin de procéder à des évaluations quantitative et qualitative plus approfondies de l'outil réalisé DocWare.

7. Bibliographie

Abiteboul S., Cluet S., Ferran G., Rousset M.C., « The Xyleme Project », *Computer Networks*, (3): 225-238, 2002.

Darmont J., Boussaid O., Bentayeb F., Rabaseda S., Zellouf Y., « Web Multiform Data Structuring for Warehousing », *Multimedia Systems and Applications*, Vol. 22, p. 179-194, Kluwer Academic Publishers, 2003.

Gardarin G., Mensch A., Tomasic A., « An Introduction to the e-XML Data Integration Suite », *Conference on Extending Database Technology (EDBT'02)*, p. 297-306, Prague, Czech Republic, March 2002.

Gardarin G., *XML : Des bases de données aux services Web*, Edition Dunod, Novembre 2002.

Huck G., Macherius I., Fankhauser P., « PDOM: Lightweigt Persistency Support for the Document Object Model », *Succeeding with Object Databases*, p. 107-118, John Wiley, 2000.

Inmon B., Hackathorn R.D., *Using the Data Warehouse*, Wiley-QED Publication, 1994.

Kanne C.C., Moerkotte G., « Efficient Storage of XML Data *», International Conference on Data Engineering*, San Diego, California, USA, 2000.

Kimball R., Ross M., *The Data Warehouse Toolkit*, John Wiley & Sons, New York, second edition, 2002.

Khrouf K., Soulé-Dupuy C., « Vers une mémoire d'entreprise via les entrepôts de document : Extraction de structures logiques », *Extraction et Gestion des Connaissances (EGC'03)*, p. 201-206, Lyon, France, Janvier 2003.

Khrouf K., Ravat F., Soulé-Dupuy C., « Comparaison et fusion de structures logiques de documents semi-structurés », *Ingénierie des Systèmes d'Information (ISI)*, Edition Hermès, Vol. 8, N°5-6/2003, p. 127-151, 2003.

Khrouf K., Entrepôts de documents : De l'alimentation à l'exploitation, Thèse de doctorat, Université Paul Sabatier, Toulouse III, Juillet 2004.

Kotsakis E., « Structured Information Retrieval in XML Documents », *ACM Symposium on Applied Computing (SAC'02)*, p.663-667, Madrid, Spain, 2002.

Soulé-Dupuy C., Bases d'informations textuelles : Des modèles aux applications, Mémoire d'HDR, Université Paul Sabatier, Toulouse III, Décembre 2001.

Sparck Jones K., « A Statistical Interpretation of Term Specificity and its Application in Retrieval », *Journal of Documentation*, Vol. 28, N°1, p. 11-20, 1972.

Yoshikawa M., Amagasa T., Shimura T., Uemura S., « XRel: A Path-Based Approach to Storage and Retrieval of XML Documents Using Relational Databases », *ACM Transactions on Internet Technology*, 1(1):110-141, 2001.