Modèle multidimensionnel en diamant dédié à l'OLAP sémantique de documents

Maha Azabou*, Kaïs Khrouf*, Jamel Feki*, Chantal Soulé-Dupuy**, Nathalie Vallès**

* Laboratoire MIR@CL, Université de Sfax
Route de l'Aérodrome km 4, B.P. 1088, 3018 Sfax, Tunisie
Azabou.Maha@yahoo.fr, Khrouf.Kais@isecs.rnu.tn, Jamel.Feki@fsegs.rnu.tn

** IRIT, Université Toulouse 1 Capitole,
2 rue du doyen Gabriel Marty, 31042 Toulouse Cedex 9, France
{Chantal.Soule-Dupuy, Nathalie.Valles-Parlangeau}@ut-capitole.fr

Résumé.
Le document électronique représente aujourd'hui un support d'information que les entreprises ne peuvent plus négliger si elles veulent être certaines d'identifier et de gérer toutes les données qui leur sont utiles au quotidien. Plusieurs travaux ont proposé l'application des techniques OLAP (« On-line Analytical Processing ») aux informations documentaires. Dans cet article, nous présentons un nouveau modèle multidimensionnel dédié à l'OLAP de documents. Ce modèle, dit *en diamant*, est organisé autour d'une dimension centrale qui traduit la *sémantique* du contenu textuel du document.

1 Introduction

Les données issues des processus métiers d'une organisation constituent un capital précieux pour la prise de décisions. La technologie des entrepôts de données a largement contribué à tirer profit de ces données opérationnelles via une réorganisation multidimensionnelle adaptée aux décideurs. Cependant, outre les données factuelles dont la modélisation et l'exploitation est maîtrisée par les outils d'entreposage de données (« data warehousing »), de nombreux travaux de recherche considèrent que les documents constituent également une source très riche de données textuelles dont l'utilité n'est pas moins importante que celle des données factuelles. Ainsi l'architecture d'un entrepôt de données classique s'avère inadaptée pour le stockage et la manipulation des documents.

Depuis plus d'une décennie, certains chercheurs recommandent d'entreposer les documents (McCabe et al., 2000) (Sullivan, 2001). Un entrepôt de documents doit permettre le stockage de documents hétérogènes, sélectionnés et filtrés, ainsi que leur classification structurelle et sémantique à des fins d'analyses multidimensionnelles. La préparation des données textuelles pour ce type d'analyse passe par une phase obligée de modélisation

multidimensionnelle spécifique basée sur les concepts multidimensionnels (*fait, dimension et hiérarchie*). Bon nombre de chercheurs ont constaté que les modèles multidimensionnels classiques (tels que les modèles en étoile) exprimant une série d'observations par le biais d'indicateurs purement numériques et d'axes d'analyses ne sont pas adaptés pour les traitements analytiques en ligne (OLAP :« On-line Analytical Processing ») des données textuelles issues des documents. C'est pour apporter une contribution à cette problématique que nous proposons un modèle multidimensionnel spécifique aux documents ; il permet d'exprimer l'aspect sémantique du contenu textuel.

Cet article est structuré comme suit. La section 2 étudie les travaux abordant l'exploitation OLAP des documents. La section 3 présente le *modèle en diamant* proposé avec ses différents composants. Enfin, dans la section 4, nous décrivons une approche pour la génération semi-automatique de schémas multidimensionnels en diamant.

2 Etat de l'art

Plusieurs travaux se sont intéressés à l'exploitation OLAP des documents ; la majorité de ces travaux adoptent le schéma en étoile initialement proposé dans le contexte des entrepôts de données. Par exemple, les auteurs de (Mothe et al., 2003) ont proposé une analyse multidimensionnelle avec l'environnement DocCube qui représente un fond documentaire avec des nuages de sphères où chaque sphère correspond à un ensemble de documents. (Khrouf, 2004) a défini une approche d'analyse multidimensionnelle des données documentaires ; néanmoins; cette approche est basée sur la structure logique des documents. Les auteurs de (Tseng et Chou, 2006) proposent d'analyser des documents (emails, articles, pages Web…) selon des dimensions construites à partir des métadonnées définies par le Dublin Core (Dublin Core, 2007).

Dans (Boussaid et al., 2006), les auteurs ont proposé une modélisation en flocon de neige des données multidimensionnelles XML avec des méthodes de fouille de données. Plus précisément, ils ont défini une approche appelée X-Warehousing qui permet de concevoir un entrepôt, de représenter son schéma conceptuel à l'aide de schémas XML et enfin d'alimenter la structure multidimensionnelle à l'aide de données initialement stockées dans des documents XML. Ce modèle implique beaucoup de redondance dans les données des dimensions puisque pour chaque mesure du fait, il faut indiquer les valeurs des dimensions correspondantes. Une telle redondance peut impliquer des difficultés de maintenance.

D'autres travaux ont proposé et utilisé le schéma en galaxie (Ravat et al., 2008) (Tournier, 2007) qui repose sur un seul concept, celui de *dimension*. Une fonction permet d'agréger des données textuelles afin d'obtenir une vision synthétique des informations issues de documents. Les travaux de (Ben Messaoud et al., 2012) proposent une démarche pour l'unification des structures des documents à des fins d'analyses multidimensionnelles, sans proposer de fonctionnalités spécifiques pour une analyse OLAP intégrant des aspects sémantiques.

En résumé des travaux relatifs à la modélisation multidimensionnelle des documents, nous avons noté que trois types de modèles multidimensionnels ont été utilisés : le modèle en étoile, le modèle en flocon de neige et le modèle en galaxie. Cependant, tous ces modèles ne tiennent pas compte de l'aspect sémantique des données textuelles. Pour pallier cet

inconvénient, certains auteurs ont proposé des approches ou fonctions pour l'analyse du contenu textuel (Tseng et Chou, 2006) (Ravat et al., 2008). L'objet de ce papier est alors de proposer un modèle multidimensionnel dédié à l'OLAP de documents et qui englobe les dimensions de données factuelles (comme *date, auteur, éditeur*), ainsi que la sémantique des données textuelles (comme *résumé, contenu, paragraphe*).

3 Modèle en diamant

La modélisation multidimensionnelle classique vise à représenter les données en fonction de l'analyse prévue par les décideurs. Elle représente l'information à analyser, comme un point dans un espace à plusieurs dimensions appelées axes d'analyses.

Dans cet article, nous proposons un nouveau modèle multidimensionnel dédié à l'OLAP de documents, que nous appelons *Modèle en diamant*. Ce modèle peut être généré à partir des structures logiques décrivant une collection de documents, à travers :

- les DTD ou Xschema des documents.
- les structures génériques existantes dans un entrepôt de documents : une structure générique est une structure commune à un ensemble de documents ; elle est obtenue par classification structurelle d'un ensemble de structures spécifiques (Khrouf et al., 2012).

Ce nouveau modèle en diamant se compose de :

- un fait qui correspond à une observation non forcément numérique sur les documents (par exemple, liste de parties de documents ou de documents se rapportant à un thème d'analyse, nombre de documents, …) ;
- un ensemble de dimensions : Une dimension *Sémantique*, une dimension *Version* et des dimensions *Classiques* :
    - la dimension *Sémantique* occupe un emplacement central ; elle se compose de la hiérarchie suivante : Concept → Ontologie. Le paramètre *Concept* sera relié aux éléments textuels (comme *Section*, *Paragraphe*) des documents afin d'utiliser ces concepts lors de l'analyse multidimensionnelle de ces éléments textuels. L'affectation de concepts au contenu textuel des documents se base sur un calcul de degré de similarité entre contenus textuels et concepts d'une ontologie comme présenté dans (Ben Mefteh et al., 2012).
    - la dimension *Version* de document concerne les différentes versions des documents, ainsi que les métadonnées associées, comme par exemple: *Propriété*, *Date de création* et *Description*. Il est à noter que chaque version de document sera reliée à une ontologie de domaine (Ben Mefteh et al., 2012).
    - les dimensions *Classiques* sont les axes d'analyse constitués des éléments du premier niveau de la structure logique de documents. Pour chaque élément, ses descendants constituent les paramètres (organisés sous forme de hiérarchies) et les attributs faibles. La détermination des paramètres, hiérarchies et attributs faibles fera l'objet de la section suivante.

Soit la structure logique *Article* de la Figure 1, le modèle en diamant correspondant est montré dans la Figure 2.

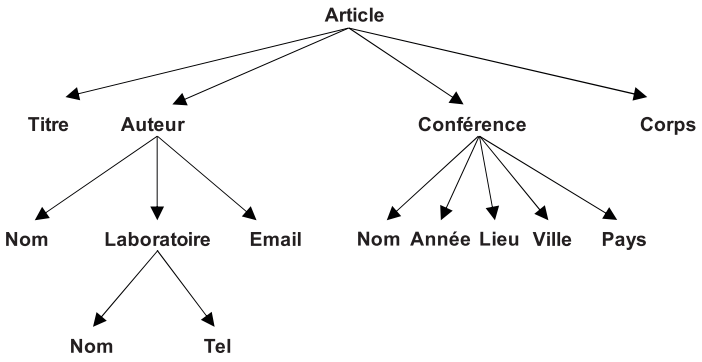

FIG. 1 – *Structure logique Article*

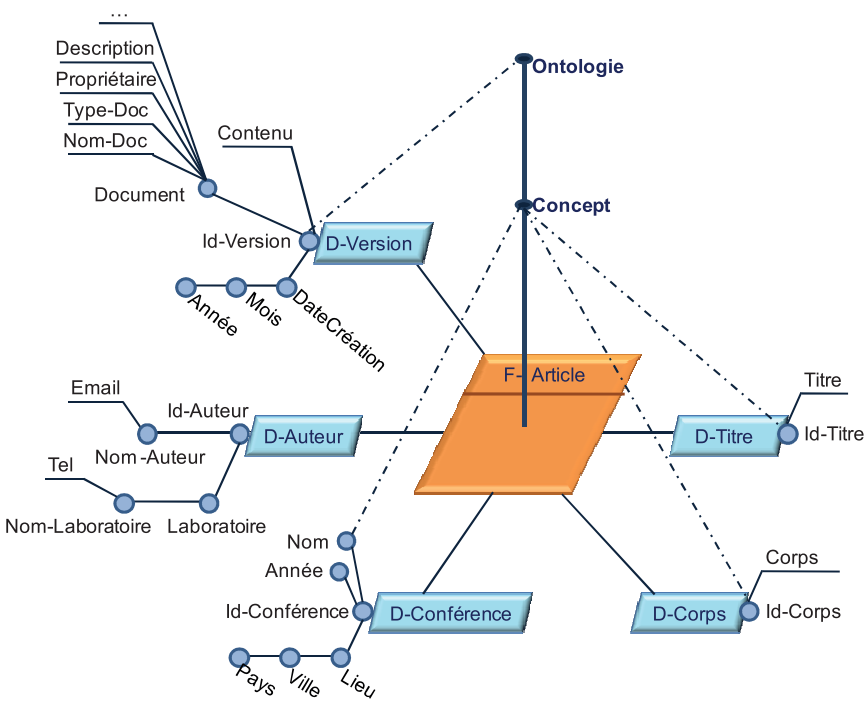

FIG. 2 – *Schéma en diamant obtenu pour la structure de la Figure 1*

4 Génération semi-automatique de modèles en diamant

Dans cette section, nous détaillons l'ensemble de règles que nous avons définies afin de générer un modèle en diamant à partir d'une structure logique décrivant une collection de documents. Ce modèle en diamant est composé d'un fait central, d'une dimension sémantique, d'une dimension version et de dimensions classiques.

Dans le reste de cet article, nous utilisons la structure logique *Article* de la Figure 1 pour illustrer les règles définies pour la génération d'un modèle en diamant.

- **Identification du fait**

Règle 1 : Le sommet *A* de toute structure logique contenant au moins une feuille constitue un fait *F-A*. (Hachaichi et al., 2010).

L'application de la *Règle 1* identifie le nœud *Article* comme un *fait* puisque la structure logique (cf. Figure 1), dont le sommet est *Article*, contient des descendants, ce fait est conventionnellement nommé *F-Article*.

- **Identification des dimensions classiques**

Règle 2 : Chaque nœud *d* descendant immédiat de la racine de la structure logique devient une dimension *D*.

Règle 3 : Chaque dimension *D* se verra affecter un identifiant *Id-D* (Ben Messaoud et al., 2011).

L'application de ces deux règles sur la structure logique *Article* identifie les quatre dimensions nommées *D-Titre*, *D-Auteur*, *D-Conférence* et *D-Corps avec les identifiants correspondants : Id-Titre*, *Id-Auteur*, *Id-Conférence* et *Id-Corps.*

Règle 4 : Chaque nœud, transformé en une dimension *D,* ne contenant pas de descendants aura un attribut faible ayant comme nom celui de la dimension *D* et directement relié à l'identifiant de *D* (*Id-D*).

Conformément à cette règle, la dimension *D-Titre* aura alors l'attribut faible *Titre* relié à l'identifiant *Id-Titre* (Idem pour *D-Corps*).

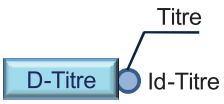

FIG. 3 – *Dimension D-Titre*

Pour les nœuds transformés en dimensions et possédant des descendants, nous définissons un ensemble de règles afin de dégager, à partir de ses descendants, les paramètres (organisés sous forme de hiérarchies) et leurs attributs faibles.

- **Identification des hiérarchies**

Pour l'identification des hiérarchies, nous devons consulter le contenu des documents afin de déterminer les *Dépendances Fonctionnelles* (DF[1]) entre les éléments de la structure logique.

Soient *X* et *Y* deux descendants immédiats d'un nœud-dimension *D* (nous appelons nœud-dimension un nœud transformé en une dimension).

Règle 5 : S'il existe une DF de *X* vers *Y* (notée X → Y) non symétrique (*i.e.*, sans avoir *Y*→X) alors *X* et *Y* constituent deux paramètres consécutifs de *D, i.e.,* de rang *i* et *i+1* respectivement (en allant du plus fin vers le moins fin).

Règle 6 : Eliminer les DF transitives. Soient les DF suivantes : *X*→*Y*, *Y*→*Z* et *X*→*Z* ; alors la DF *X*→*Z* est transitive et doit être éliminée. Nous obtenons alors la hiérarchie suivante : *X*→*Y*→*Z*.

Règle 7 : S'il n'y a aucune DF entre *X* et les autres descendants du nœud-dimension *D*, alors *X* constitue un paramètre de la dimension *D* de rang 2 (*X* relié à l'identifiant *Id-D* de *D*).

Reprenons le sous-arbre *Conférence* de la structure *Article* (cf. Figure 1) avec l'échantillon de contenu indiqué dans la Figure 4.

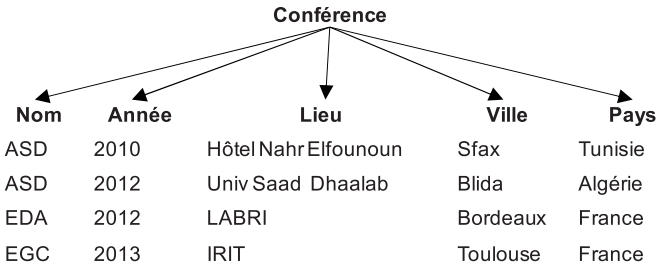

FIG. 4 – *Sous arbre Conférence*

Nous constatons à partir de ce contenu :

- L'existence des trois DF non symétriques (*Règle 5*) *Lieu*→*Ville*, *Lieu*→*Pays* et *Ville*→*Pays*. L'application de la *Règle 6* nous permet de déterminer la hiérarchie suivante : *Lieu*→*Ville*→*Pays*.

- Aucune DF entre *Nom* et les autres descendants de *Conférence* (*Année*, *Lieu*, *Ville* et *Pays*), *Nom* est alors un paramètre de rang 2 relié à *Id-Conférence* selon la *Règle 7*. (Idem pour *Année*).

[1] Une Dépendance Fonctionnelle (DF) de l'attribut *X* vers l'attribut *Y*, notée *X*→*Y*, exprime qu'à une valeur de *X* est associée une seule valeur de *Y* ; l'inverse d'une DF n'est pas forcément une DF.

La dimension *D-Conférence* sera définie de la manière suivante :

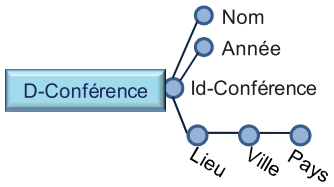

FIG. 5 – *Dimension D-Conférence*

- **Identification des attributs faibles**

Certains paramètres de dimension peuvent être accompagnés de descripteurs appelés attributs faibles.

> *Règle 8* : Etant donnés deux attributs *X* et *Y*, s'il existe deux dépendances fonctionnelles symétriques $X \rightarrow Y$ et $Y \rightarrow X$ et si *Y* n'a pas de descendants, alors *Y* constitue un attribut faible pour *X*.

Soit le sous arbre *Auteur* (cf. Figure 6) issu de la structure *Article* de la Figure 1.

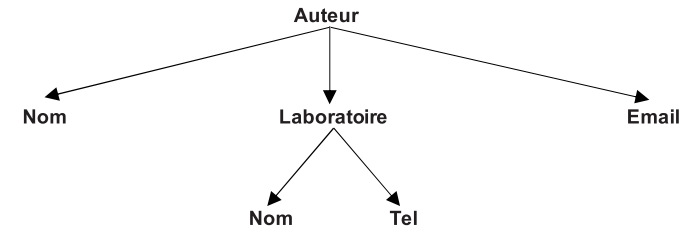

FIG. 6 – *Sous arbre Auteur*

Concernant la *Règle 8*, on remarque l'existence des deux dépendances fonctionnelles symétriques $Nom \rightarrow Email$ et $Email \rightarrow Nom$, *alors Email* peut jouer le rôle d'un attribut faible pour le paramètre *Nom*.

Remarque : Soient *X* et *Y* deux descendants de la structure logique tels que $X \rightarrow Y$ et $Y \rightarrow X$, nous considérons arbitrairement que l'un d'entre eux est un attribut faible de l'autre attribut assimilé lui à un paramètre. C'est le concepteur qui vérifiera et validera ce choix par rapport à la sémantique des deux descendants.

Il convient d'itérer les règles 5, 6, 7 et 8 pour chaque élément ayant des descendants afin de déterminer tous ses paramètres à tous les niveaux et leur(s) attribut(s) faible(s).

Le paramètre *Laboratoire* possède des descendants (*Nom* et *Téléphone*) ; il s'agit d'un nouveau niveau. Selon la règle 8, il existe deux DF symétriques *Nom→Téléphone* et *Téléphone→Nom,* alors nous choisissons arbitrairement *Téléphone* comme attribut faible pour le paramètre *Nom*.

La dimension *D-Auteur* sera définie de la manière suivante :

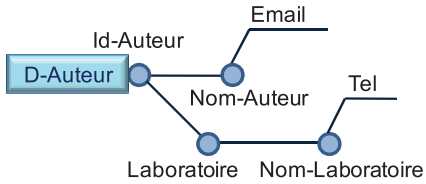

FIG. 7 – *Dimension D-Auteur*

Remarque : Si une dimension contient des paramètres ayant la même appellation (Exemple *Nom* dans *D-Auteur*), nous modifions le nom du paramètre par l'ajout de l'intitulé de son élément-père. Ainsi, dans notre exemple, nous aurons : Nom-Auteur et Nom-Laboratoire.

- **Ajout de la dimension *Version***

A ce niveau, nous ajoutons la dimension *D-Version* composée d'un identifiant *Id-Version*, de son *Contenu* et de deux hiérarchies : *i)* une hiérarchie temporelle se rapportant à la date de création du document organisée comme suit *DateCréation→Mois→Année* ; *ii)* une hiérarchie décrivant un ensemble de métadonnées (*Nom* physique du document, *Type* d'extension, *Propriétaire* créateur du document, *Description* sommaire du document, etc.).

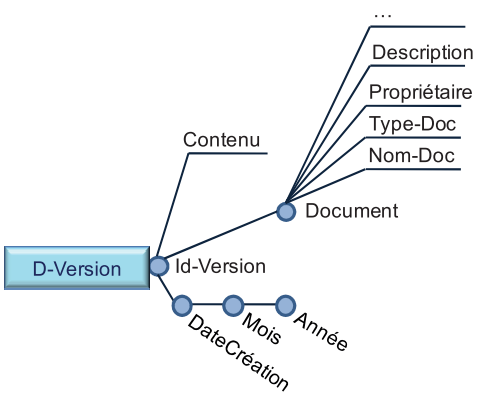

FIG. 8 – *Dimension D-Version*

- **Ajout de la dimension sémantique**

Les données textuelles véhiculent une sémantique. Elle est exprimée, dans le modèle en diamant, à travers la *dimension sémantique* faisant référence à un ensemble d'ontologies et de concepts décrivant les documents associés à la structure logique correspondante.

La détermination de la dimension sémantique se fait conformément à l'approche décrite dans (Ben Mefteh et al., 2012) et selon les étapes suivantes :

- Extraction des termes. Il s'agit d'extraire les mots-clés significatifs des éléments feuilles du document (fragments textuels associés aux éléments feuilles) selon un processus d'indexation classique tel que défini en recherche d'information (Baeza-Yates et al., 1999).
- Choix de l'ontologie. L'objectif est de déterminer, parmi un ensemble d'ontologies de domaines, celle qui convient le mieux pour décrire la sémantique du document, et ce à partir des mots-clés du langage d'indexation générés lors de l'étape précédente.
- Association de concepts aux éléments feuilles. Consiste à rechercher, dans l'ontologie de domaine retenue à la phase précédente, le concept le plus approprié à la description de la sémantique de l'élément feuille compte tenu des mots-clés qui le décrivent.
- Inférence de concepts aux éléments non-feuilles. Les concepts feuilles d'un nœud servent ensuite à inférer un concept à ce nœud à partir de l'ontologie sélectionnée.

La représentation de la dimension sémantique dans le modèle en diamant se fait de la manière suivante :

- relier le paramètre *Version* au paramètre *Ontologie* de la dimension sémantique puisque dans nos travaux, nous associons toute version d'un document de l'entrepôt à une ontologie de domaine.
- relier les paramètres textuels des dimensions classiques, « porteurs » d'un point de vue sémantique, au paramètre *Concept* de la dimension sémantique. Ces paramètres sont les éléments correspondants à la structure logique et dont les contenus sont reliés aux concepts des différentes ontologies (Ben Mefteh et al., 2012). A titre d'exemple, l'attribut *Nom* de la dimension *D-Conférence* sera relié au paramètre *Concept*, alors que l'attribut *Nom-Auteur* de la dimension *D-Auteur* ne sera pas relié. Ainsi, nous pouvons faire l'analyse soit par le nom de la conférence (ASD, par exemple), soit par le ou les concepts associés (Système décisionnel, par exemple) ; dans ce cas, les autres conférences appartenant aussi à ce concept (e.g., EDA) seront intégrées dans l'analyse.

En utilisant ces règles, le modèle en diamant obtenu à partir de la structure *Article* est celui présenté dans la Figure 2.

Notons que toutes les règles proposées ci-dessus sont automatisables. Une fois que la construction du modèle en diamant est terminée, c'est le concepteur (assisté du décideur) qui vérifie et valide le modèle en diamant obtenu. Il peut renommer, supprimer les éléments multidimensionnels et les liens entre les dimensions, réorganiser les paramètres d'une hiérarchie si nécessaire, …

Exemple : Supposons que le nom d'auteur *Olivier Dupond* est relié au concept *Oliviers* de l'ontologie *Agriculture*, le système relie alors le paramètre *Nom-Auteur* de la dimension *D-Auteur* au paramètre *Concept* de la dimension sémantique. Le concepteur doit intervenir pour supprimer ce lien.

5 Conclusion

Cet article présente un nouveau modèle multidimensionnel dédié à l'OLAP de documents, à savoir le *modèle en diamant*. Ce modèle se compose d'un fait, de dimensions *classiques* (construites à partir des structures génériques de l'entrepôt, DTD ou Xschema), la dimension version de document et la dimension sémantique. Cette dimension sémantique a pour rôle de passer du simple texte à un niveau plus sémantique, que sont les concepts associés. Nous avons défini aussi dans cet article un ensemble de règles en vue de la construction semi-automatique de modèles en diamant.

Plusieurs perspectives sont envisageables. Un prototype logiciel est en cours de développement supportant les différentes règles présentées dans cet article et la visualisation du modèle en diamant en 3D. En l'absence de Benchmarck de tests, nous comptons valider ces règles sur des exemples pris de la littérature, qui construisent des schémas multidimensionnels de documents. Il serait intéressant aussi de définir un processus d'instanciation de ces modèles en diamant à partir du contenu des documents et la visualisation des résultats sous forme de cubes ou tables multidimensionnelles.

Références

Baeza-Yates R. et Ribero-Neto B. (1999), *Modern Information Retrieval*, Addison Wesley, 1999.

Ben Mefteh S., Khrouf K., Feki J., Ben Kraiem M. et Soulé-Dupuy C. (2012). *Une approche d'extraction automatique de structures sémantiques de documents XML*, INFormatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2012), p. 523-538, Montpellier, France.

Ben Messaoud I., Feki J. et Zurfluh G. (2011). *Modélisation multidimensionnelle des documents XML.* Journées francophones sur l'Entrepôts de Données et l'Analyse en ligne (EDA 2011), p. 55-70, Clermont-Ferrand, France.

Ben Messaoud I., Feki J. et Zurfluh G. (2012). *A First Step for Building a Document Warehouse: Unification of XML Documents* (S). International Conference on Research Challenges in Information Science (RCIS), Valencia, Spain, 2012.

Boussaid O., Ben Messaoud R., Choquet R. et Anthoard S. (2006). *Conception et construction d'entrepôts XML.* Journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA'2006), p. 3–21, Versailles, France.

Dublin Core Metadata Initiative (DCMI) (2007): Dublin Core Metadata Element Set, Version 1.1, ISO Standard 15836, from http://dublincore.org/documents/dces/.

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

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

Khrouf K., Azabou M., Feki J. et Soulé-Dupuy C. (2012). *Towards a Multi-user Document Warehouse*. International Conference on Web Information Systems and Technologies, p. 149-154, Porto, Portugal.

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

Mothe J., Chrisment C., Dousset B. et Alau J. (2003). *DocCube: Multidimensional visualization 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.

Sullivan D. (2001). *Document Warehousing and Text Mining.* Wiley John & Sons, ISBN: 0471399590, 2001.

Ravat F., Teste O., Tournier R. et Zurfluh G. (2008). *Top_Keyword: an Aggregation Function for Textual Document OLAP*. International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2008), p. 55-64, Turin, Italy.

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

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

Summary

Today, the electronic document represents an important support for information, therefore organizations cannot neglect these documents if they want to identify and manage all data useful for the decision support system. Several works have addressed how to apply On-line Analytical Processing (OLAP) techniques on documents. In this paper, we present a new multidimensional model called *diamond model;* it is for documents modeling and OLAP analyses. This model takes into account the semantics of the textual content of documents by means of a new central *semantic* dimension.