Rapport de recherche bibliographique mars 2003

DESS Ingénierie documentaire

L'entreposage et la fouille des données complexes

FARIZY Anne-Sophie

Sous la direction de Monsieur Jérôme DARMONT
Maître de conférences en informatique
Laboratoire ERIC
Université Lumière, Lyon2

enssib
école nationale supérieure des sciences de l'information et des bibliothèques

L'entreposage et la fouille des données complexes

Descripteurs: entreposage de données, fouille de données, données complexes.

Résumé: l'entreposage et la fouille de données sont deux techniques utilisées dans les systèmes d'aide à la décision qui ont vu le jour dans les années 1990. Actuellement, ces techniques sont relativement bien maîtrisées pour des données dites simples, en revanche, Internet et l'apparition de nouvelles formes de représentation de données plus complexes ont remis en cause les systèmes traditionnels existants. Le but de ce rapport est donc d'effectuer un état de l'art de la recherche scientifique accomplie pour l'entreposage de données complexes tout en exposant la méthodologie de recherche documentaire utilisée.

Data warehousing and data mining for complex data

Keywords: data warehousing, data mining, complex data.

Abstract: appeared in the nineties, data warehousing and data mining have usually been used in decision support systems. Nowadays, these techniques are mastered when they deal with mere data, however Internet and the emergence of more complex data representation methods question the traditional decision support tools. In this context, the purpose of this study is to make a data warehousing research overview for complex data and to present information retrieval methodology used.

Sommaire

Définition du sujet

1. Recherche du commanditaire

La définition de mon sujet de recherche s'est effectué en deux phases. Premièrement, mon objectif était de trouver un commanditaire dans le domaine qui m'intéressait, ensuite, de voir avec mon commanditaire l'orientation précise du sujet. J'ai donc choisi en relation avec ma formation initiale qu'est l'informatique de gestion, le domaine de l'ingénierie des connaissances. Mon choix était motivé par l'intérêt, la curiosité et l'idée que ce domaine est actuellement en plein développement et par conséquent très propice pour une recherche documentaire.

La recherche sur Internet des organismes susceptibles de correspondre à mes attentes sur Lyon m'a conduit à en sélectionner deux : le laboratoire LIRIS [1] et le laboratoire ERIC [2]. Une réponse favorable m'est parvenu du laboratoire ERIC par M. Jérome Darmont qui est responsable d'un pôle de recherche dans ce laboratoire.

2. Présentation du laboratoire ERIC

Le laboratoire ERIC a pour objectifs scientifiques le développement de méthodes et d'outils informatiques pour l'ingénierie des connaissances destinés, plus particulièrement, à l'Extraction automatique de Connaissances à partir de Données (ECD), ce qui englobe essentiellement les techniques d'entreposage et de fouille de données. Deux pôles de recherches majeurs divisent ce laboratoire en deux équipes : la première travaille sur le processus de fouille de données et la seconde sur les bases de données décisionnelles (BDD). En découvrant chacune des équipes, je me suis orientée vers le pôle BDD car certaines problématiques de leurs

[1] Laboratoire d'InfoRmatique en Images et Systèmes d'information, site Internet disponible sur : <http://liris.cnrs.fr/> (consulté le 28.11.2003).
[2] Equipe de Recherche en Ingénierie des Connaissances, site Internet disponible sur : <http://eric.univ-lyon2.fr/> (consulté le 28.11.2003).

recherches portaient sur la gestion des nouveaux modèles de représentation des données.

3. Attentes du laboratoire

Notre première rencontre à permis de déterminer trois points principaux de la recherche :

La définition du sujet

L'entreposage et l'analyse des données complexes.

Le but de la recherche

Effectuer une synthèse sur le sujet mais aussi réaliser une bibliographie qui devienne la bibliographie de base de l'équipe pour une alimentation future régulière. Ce besoin implique d'effectuer une bibliographie générale sur le concept d'entreposage de données en se restreignant à ne sélectionner que les documents phares, vu que ce travail est limité dans le temps.

Le type de la recherche

Se focaliser essentiellement sur l'état de la recherche pour le sujet précisé. Ceci évite certains aspects comme les documents de présentation de produits répondant à la problématique de base.

Méthodologie de recherche

1. Préparation de la recherche

1.1. Support de méthodologie pour ma recherche

Afin de mettre en place ma stratégie de recherche, je me suis appuyée sur différents supports de méthodologie orientés sur la recherche scientifique.

InfoSphère sciences et technologies

http://www.bibliotheques.uqam.ca/InfoSphere/sciences/index1.html

Ce site proposé par l'UQAM (Université du Québec à Montréal) nous a été conseillé par M. Lardy lors de son cours de recherche d'information sur Internet. Ce tutoriel décliné en deux versions, l'une pour les sciences humaines et sciences de la gestion, l'autre pour les sciences et la technologie, propose une méthodologie complète pour effectuer une recherche documentaire efficace. Je l'ai utilisé essentiellement pour la sélection des sources de recherche.

SAPRISTI!, Sentiers d'Accès et Piste de Recherche d'Information Scientifiques et Techniques sur Internet!

http://csidoc.insa-lyon.fr/sapristi/digest.html

Développé par Doc'INSA, ce site est orienté sur la recherche d'informations sur Internet essentiellement pour les sciences de l'ingénieur. Il propose une méthodologie de recherche par type de documents mais aussi conseille sur les outils de recherche sur Internet.

1.2. Sélection des sources

Voici les principaux types de documents qui me semblent appropriés pour faire un état des lieux de la recherche sur un sujet scientifique et pour chaque type de documents, les outils de repérage à utiliser :

- **Les actes de conférence** : ils abordent des thèmes très spécifiques et les derniers résultats de la recherche.
  - *Outils de repérage* : les catalogues de bibliothèques et le Web.
- **Les périodiques spécialisés** : ils diffusent des réflexions théoriques sur une discipline, des résultats de recherches originales et des expériences particulières.
  - *Outils de repérage* : les bases de données bibliographiques
- **Les publications préliminaires** : ils sont issus généralement de centres de recherches et concernent des résultats récents de recherche.
  - *Outils de repérage* : le Web
- **Les rapports scientifiques** : ils permettent une synthèse étayée sur des sujets très précis.
  - *Outils de repérage* : catalogues de bibliothèques, bases de données et le Web
- **Les thèses et mémoires** : ils ont pour but d'apporter des connaissances nouvelles et de "faire avancer la science".
  - Outils de repérage : catalogues de bibliothèques universitaires, sites de recherche spécialisés

Au vu de cette sélection, il m'a semblé primordial de m'appuyer essentiellement sur les bases de données et sur des outils de recherche Internet spécialisés sur la documentation scientifique.

1.3. Appropriation du sujet

Afin de cerner le sujet à traiter, j'ai tout d'abord commencé par feuilleter les livres traitant de l'entreposage de données directement accessibles grâce à la bibliothèque de l'ENSSIB et à la bibliothèque universitaire de campus Claude Bernard de Lyon1. Par ailleurs, j'ai aussi consulté le Web, ce qui m'a apporté différents éclairages sur le sujet, bien que la navigation ait été essentiellement aléatoire. Enfin, mon commanditaire m'a fait parvenir l'une de ses publications [3] me décrivant le sujet, et leurs différents axes de recherche. Ce document a été

[3] Voir bibliographie 4.1, référence [BOU03]

primordial par la suite, afin de délimiter les contours du sujet et d'étudier la pertinence des résultats.

2. Mise au point de la stratégie de recherche

Afin de faciliter ma recherche documentaire, j'ai décidé de classifier mes mots clés par groupes qui identifient une notion précise du sujet. Ensuite, j'ai décomposé mon sujet suivant différents axes de recherche auxquels j'ai associé les groupes de mots-clés correspondants.

2.1. Les mots clés

Six principales notions se distinguent de mon sujet de recherche

2.1.1 L'entreposage de données

Anglais	Français
data warehouse data warehousing	entrepôt de données entreposage de données
data mart	magasin de données
decision support system	système d'aide à la décision
knowledge discovery database	base de données décisionnelle

2.1.2 Les entrepôts de données spécialisés

Anglais	Français
web warehouse	entrepôt de données web
multimedia warehouse	entrepôt de données multimedia
XML warehouse	entrepôt de documents XML

2.1.3 Les données complexes

Anglais	Français
complex data	donnée complexe
multimedia	multimédia
web page web document web data	page web document web donnée web
image	image
audio	audio
video	vidéo
text	texte

2.1.4 La phase de préparation des données

Anglais	Français
ETL extract transform load	extraction transformation chargement
modelisation	modélisation
integration	intégration
transformation	transformation

2.1.5 L'extraction de caractéristiques

Anglais	Français
feature extraction	extraction de caractéristiques
feature construction	construction de caractéristiques
knowledge discovery	extraction de connaissances

2.1.6 La modélisation multidimensionnelle

Français	Anglais
multidimensional	multidimensionnel
OLAP	analyse en ligne
star schema snowflake schema	schéma en étoile schéma en flocon de neige
data cube	cube de données

2.1.7 Les processus de fouille de données

Anglais	Français
data mining	fouille de données
web mining	fouille de données web
XML mining	fouille de documents XML
multimedia mining	fouille de documents multimédias
text mining	fouille de textes
image mining	fouille d'images

2.2. Les axes de recherche

2.2.1 Généralités sur l'entreposage de données

Vu que l'une des attentes de mon commanditaire était de constituer une bibliographie de base pour une alimentation future régulière selon leurs travaux de recherche, il m'a donc été demander de rechercher les documents phares sur la notion d'entreposage de données.

2.2.1.1 Les groupes de mots clés

- L'entreposage de données

2.2.2 Le traitement des données complexes en analyse et fouille de données

Cet axe est en fait l'intitulé du sujet. Il permettra de recenser les documents les plus généraux par rapport au sujet

2.2.2.1 Les groupes de mots clés

- L'entreposage de données

- Les entrepôts de données spécialisés
- Les données complexes

2.2.3 La phase de préparation des données complexes

L'entreposage de données se découpe en deux phases principales, la première étant la phase de préparation des données qui se caractérise par l'extraction des données de leur source d'origine et leur transformation pour permettre leur intégration dans un entrepôt de données.

2.2.3.1 Les groupes de mots clés

- La phase de préparation des données
- Les données complexes

2.2.4 L'extraction de caractéristiques pour les données complexes

Cette étape intervient lors de la préparation des données. Il s'agit de dégager les principales caractéristiques des données à intégrer dans l'entrepôt de données. Cette phase est très importante car elle détermine la qualité des processus futurs de fouille de données. Notons que des outils de fouille de données peuvent servir à l'extraction de caractéristiques

2.2.4.1 Les groupes de mots clés

- L'extraction de caractéristiques
- Les données complexes
- Les processus de fouille de données

2.2.5 La modélisation multidimensionnelle des données complexes

Il s'agit de la seconde phase pour l'entreposage de données. Ce type de modélisation est typique aux entrepôts de données permettant les processus de fouille de données.

2.2.5.1 Les groupes de mots clés

- La modélisation multidimensionnelle
- Les données complexes

Le tableau suivant résume les différents axes de recherches, la numérotation des axes sera réutilisée lors du dépouillement des résultats

Numéro	Axes de recherche
1	L'entreposage de données
2	L'entreposage et la fouille des données complexes
3	La phase de préparation des données complexes
4	L'extraction de caractéristiques pour les données complexes
5	La modélisation multidimensionnelle des données complexes

Les différents axes de recherche

3. Dialog

3.1. Sélection des sources

Vu la foule de bases de données disponibles via Dialog, il était nécessaire de sélectionner les bases les plus pertinentes en fonction de mon sujet. J'ai donc utilisé le DialIndex.

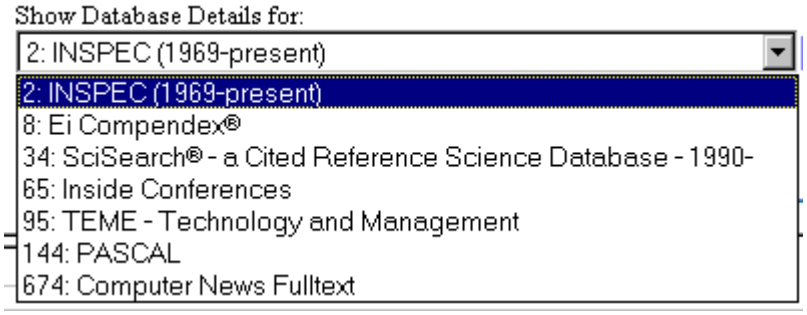
B 411
SF COMPSCI,ELECTRON,ENG,MINING,RESCENTE,SCITAPS,SCITECH
S data(w)warehous? OR data(w)mining
RF

J'ai sélectionné ensuite les bases de données qui retournaient plus de 1000 résultats, ce qui m'a fait un total de 38 bases. Cependant, les résultats obtenus sur ces bases faisaient apparaître un bruit non négligeable, provenant essentiellement de la description de produits informatiques spécialisés dans le domaine de l'entreposage et la fouille de données. Vu que mon sujet s'oriente essentiellement vers le domaine de la recherche scientifique, j'ai donc effectué une seconde sélection suivant les bluesheets [4] associées à chaque base de donnée afin de

[4] Description des bases de données sous Dialog en ligne, à l'adresse < http://library.dialog.com/bluesheets/>

supprimer celles à caractère commercial et de conserver celles susceptibles de traiter de la recherche scientifique.

Au total, 6 bases de données ont finalement été sélectionnées [5].



Bases de données retenues sous Dialog

3.2. La recherche

J'ai effectué mes recherches sur les six bases simultanément, en fonction de mes différents axes de recherche. Pour faciliter la recherche multibases, la commande set detail off pour visualiser les résultats pour l'ensemble des bases et la commande remove duplicates qui élimine les doublons provoqués par l'apparition d'un même résultat dans plusieurs bases, m'ont été très utiles.

L'accès au thésaurus des bases de données de Dialog n'étant pas permis, je me suis contentée de la recherche sur index grâce à la fonction expand, afin d'être sûre que les mots-clés utilisés renvoyait un ensemble significatif de résultats. Par ailleurs mes stratégies de recherche se sont largement appuyées sur le titre des références vu qu'il énonce généralement les principaux thèmes abordés par le document référencé. Cette limitation de recherche m'a permis ainsi de restreindre le nombre de résultats lorsque celui-ci s'avérait trop important.

[5] Voir description des bases de données sélectionnées sous Dialog en annexe II.1

Set	Term Searched	Items	
S1	FEATURE(W)EXTRACTION OR FEATURE(W)CONSTRUCTION OR KNOWLEDGE(W)DISCOVERY(1W)DATA	54125	Display
S2	TEXT? ?	306338	Display
S3	S1 AND S2	1626	Display
S4	S1/TI AND S2/TI	39	Display
S5	RD S4 (unique items)	29	Display

Exemple de restriction de résultats sur le titre sous Dialog

L'ensemble des recherches effectuées sous Dialog respectent la stratégie de recherche élaborée. Les principales requêtes effectuées sous Dialog sont présentées en annexe II.2

3.3. Mes impressions

Les principaux avantage de Dialog me semblent la quantité d'information mise à la disposition de l'utilisateur et les fonctionnalités de recherche qui sont très étendues. En revanche, bien qu'une technique d'utilisation des doublons soit disponible, il reste cependant de nombreuses références en double dans les résultats exploités. Enfin, j'ai eu beaucoup de difficultés à exploiter la fonction d'export des citations vers End Note. Mes seules tentatives réussies concernent les références de la base Inspec, cependant, les champs rapatriés étaient peu nombreux et parfois mal situés (éditeurs et auteurs tous les deux placés dans le champ auteur). J'ai donc choisi d'importer manuellement mes références vers End Note.

4. Science Direct

Science Direct m'a paru être un outil très intéressant vu qu'il permettait l'accès au texte intégral de plus de 1500 revues scientifiques d'Elsevier, revues très reconnues dans le domaine de la recherche scientifique. Par ailleurs j'ai consulté la base de donnée Science Direct Navigator qui offre une couverture complète des nouveaux développements dans l'ensemble des disciplines scientifiques.

4.1. Sélection des sources

L'interface de recherche de Science Direct se découpe suivant les différentes sources qu'il est possible de consulter (journaux, livres, bases de données..). En ce qui concerne les journaux, j'ai sélectionné ceux relatifs aux domaines de l'informatique, des sciences décisionnelles et enfin de l'ingénierie.

4.2. La recherche

L'annexe I présente les principales requêtes effectuées sous Science Direct.

4.3. Fonctionnalités de recherche

De nombreuses fonctionnalités facilitent la recherche sur Science Direct :

- L'interface de recherche est très conviviale et offre la possibilité d'écrire des requêtes complexes.
- La création d'un compte utilisateur afin de sauvegarder ses recherches et de créer des alertes m'a été très utile. J'ai ainsi pu modifier mes requêtes en fonction de l'évolution de mon sujet.
- En ce qui concerne les résultats, les articles « in Press » sont visualisables et pour chaque article, la fonction « Cited By » indique quels autres articles sur Science Direct citent cet article.
- Enfin, l'export des résultats vers le logiciel End Note pour la constitution de la bibliographie fonctionne très bien, ce qui, selon mon point de vue, est un avantage par rapport à Dialog.

5. La collection en ligne LNCS

Lecture Notes in Computer Science est une collection éditée par Springer dans le domaine de l'informatique et des technologies de l'information. A ce jour, 1500 volumes sont accessibles dont un tiers publiés à partir de 2003.

5.1. Pourquoi cette collection ?

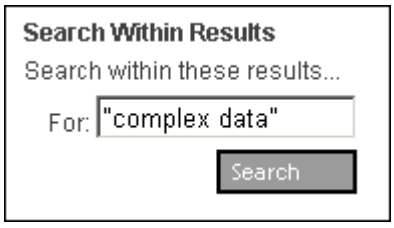
J'avais remarqué, lors de mes recherches sur différentes bases de données, que certains résultats provenaient de la collection LNCS. J'ai découvert, par la suite, lors de la journée de recherche documentaire spécialisée en sciences de l'ingénieur, que Doc'INSA possédait un abonnement pour l'accès au texte intégral de la collection LNCS. Vu que l'accès aux ressources de Doc'INSA était permis aux étudiants de l'université LyonI, j'ai ainsi pu procéder à une recherche documentaire sur cette collection.

5.2. La recherche

Une barre de recherche effectue une première sélection sur l'ensemble des documents de la collection.

1ère sélection

Ensuite, une seconde barre de recherche permet une seconde sélection à l'intérieur des résultats retournés



2nde sélection

En adaptant mes recherche aux fonctionnalités existantes, j'ai donc effectué une première sélection générale pour l'entreposage de données. Ensuite, pour les résultats sélectionnés, j'ai décliné mes différents mots-clés suivant mes différents axes de recherche.

6. La recherche sur Internet

Mes tentatives de recherche sur des outils généraux tels que les annuaires, méta-moteurs et moteurs de recherche se sont révélées peu satisfaisantes. D'une part, les interfaces de recherche ne sont généralement pas adaptées à l'écriture de requêtes complexes et d'autre part, la masse d'information est tellement importante et

hétéroclite qu'il m'a été très difficile d'obtenir un ensemble de résultats globalement pertinents, malgré l'utilisation de requêtes affinées. Mon choix s'est donc essentiellement porté sur des outils plus spécialisés dans le domaine de la recherche scientifique et dans le domaine de l'informatique.

6.1. Annuaires et moteurs de recherche

Les annuaires étant considérés comme les « outils humains » de la recherche d'information sur Internet, j'ai décidé de consulter les deux annuaires de recherche étudiés en cours [6] qui sont Yahoo et l'Open Directory. Google a été le moteur de recherche utilisé.

Yahoo

URL : http://dir.yahoo.com
Répertoire de recherche: racine
Requête : ("data warehousing" or "data warehouse" or "data warehouses" or "decision support system" or "decision support systems" or "knowledge discovery database" or "knowledge discovery databases" or "data mart" or "data marts") and ("complex data" or multimedia or web or image* or audio or video or text*)
26 résultats

Open directory project

URL: http://dmoz.org
Répertoire de recherche: Top : Computers : Software : Databases : Data warehousing
Requête : "complex data" or image* or text* or multimedia or web or audio or video
6 résultats

Sur l'ensemble des résultats retrouvés, aucun ne m'a semblé pertinent dans le cadre de ma recherche. En effet, les résultats retournent dans la plupart des cas des sites de présentation de sociétés ou des produits spécialisées dans l'entreposage et la fouille de données. En revanche, j'ai pu identifier les sites Internet ressources pour l'axe de recherche général sur l'entreposage de données.

Google

[6] Cours de recherche d'informations sur Internet, dirigé par Mme Defosse.

Suite à une utilisation régulière du moteur de recherche Google pour diverses recherches d'information, je pensais pouvoir l'exploiter pleinement dans le cadre de ma recherche. Cependant, force est de constater qu'il m'a été très difficile d'adapter ma stratégie de recherche à cet outil. Vu que les résultats renvoyés étaient généralement trop hétérogènes et nombreux, j'ai adopté les techniques suivantes dans le but de limiter les résultats, cependant sans grands résultats :

- Limitation des recherche au titre du document
- Limitation des recherches au fichier d'extension .pdf

Google ne m'a donc pas permis d'élaborer une stratégie de recherche pertinente, en revanche, il s'est révélé très efficace pour une recherche d'information ciblée. La plupart des recherches de références suivant une piste précise tel que l'auteur ou bien le titre m'a généralement permis de localiser assez rapidement cette référence.

6.2. Les outils de recherche spécialisés

6.2.1 SCIRUS

SCIRUS est un moteur de recherche spécialisé dans l'information scientifique. Il effectue sa recherche sur deux principales sources : les journaux et le Web. L'ensemble des résultats retournés ont d'une manière générale été très pertinents.

Searched for:	All of the words **(title:"web warehouse")** OR **(title:"multimedia warehouse")** OR **(title:"XML warehouse")**
Found:	**15 total** | 4 journal results | 11 Web results
Sort by:	**relevance** | date

Exemple de requête effectuée sur SCIRUS

6.2.2 L'archive ouverte ResearchIndex [7]

Cette archive ouverte de pré-publications et de post-publications d'articles à été créée en 1997 à l'initiative de S. Lawrence et C. Lee Giles de NEC Research Institute. Spécialisé dans le domaine de l'informatique, cet outil offre l'accès à une

[7] Disponible à l'adresse <http://citeseer.ist.psu.edu/> (consulté le 12.03.2004).

masse considérable d'information : sept millions de « pages » sont recensées dans la base ainsi que cinq millions de citations.

Les différentes fonctionnalités de recherche m'ont permis de récupérer un ensemble significatif de résultats :

- La recherche par mots-clés
- Recherche par l'utilisation d'un annuaire : le répertoire pour l'entreposage de données contient environ 120 documents.
- Un travail sur le contenu des articles. Pour un article donnée, la présentation de multiples liens vers d'autres articles au contenu similaire ou suivant une analyse des citations m'a permis de retrouver un ensemble de documents relatifs à un sujet précis.

Présentation des articles similaires à un article étudié

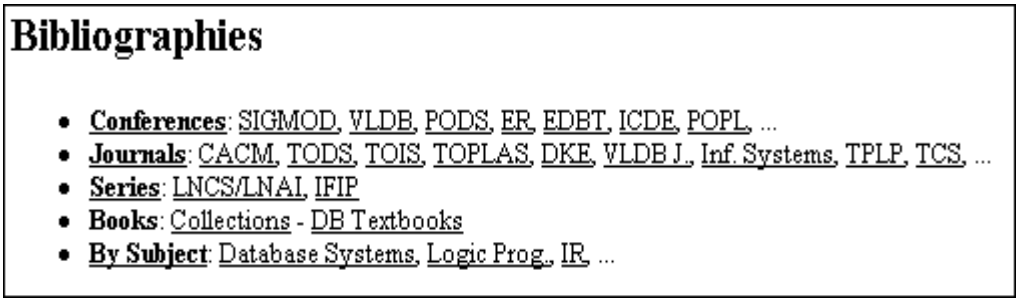
6.3. Les sites spécialisés

La recherche sur Internet m'a permis de localiser deux sites clés contenant une bibliographie plus ou moins spécialisée dans le domaine de l'informatique. Le premier site a été trouvé suite à mes différentes recherches sur Internet ou de nombreux résultats me renvoyaient sur ce site. Le second m'a été suggéré par mon commanditaire et se retrouve aussi dans les sites Web ressources en entreposage de données pour l'outil Open Directory Project.

DBLP computer science bibliography

http://www.informatik.uni-trier.de/~ley/db/index.html

Ce site proposé par l'université Trier en Allemagne propose des informations bibliographiques sur les principales ressources à consulter pour le domaine de l'informatique.



Sources bibliographiques disponibles sous DBLP bibliography

Cette ressource m'a essentiellement servi pour la constitution de la bibliographie sur le sujet de l'entreposage de données, afin de repérer les principales conférences et équipes de recherche concernées par ce domaine.

Data warehousing and OLAP research bibliography

http://www.ondelette.com/OLAP/dwbib.html

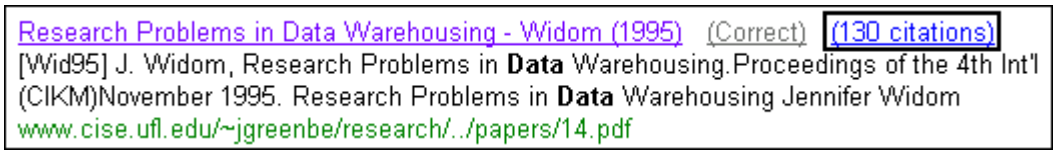
Plus spécialisé que le précédent, il recense les documents phares en entreposage de données.

7. Le dépouillement des résultats

7.1. Evaluation de la pertinence des résultats

J'ai évalué la pertinence des résultats suivants différentes méthodes:

- la lecture des résumés des documents afin d'effectuer un premier tri des différents résultats obtenus.
- la lecture du document quand mon choix était encore indécis suite à la lecture du résumé.
- l'analyse des références que permettent certains outils de recherche comme par exemple le Research Index [8]. Un article cité de nombreuses fois peut en effet être considéré comme un document de référence.



[8] Site Internet de recherche Research Index disponible à l'adresse : <http://citeseer.ist.psu.edu> (consulté le 10.03.2004).

Nombre de citations référençant l'article

- L'analyse de la bibliographie des documents pertinents. Cette analyse permet à la fois de récupérer des documents non encore trouvés et d'identifier les documents de référence.
- La redondance d'un même résultat suite à l'interrogation de différents outils de recherche et par différentes requêtes.

7.2. Quelques résultats de recherches

Les annexes I et II.2 rassemblent les principales requêtes effectuées sous Science Direct et Dialog. Pour chacune de ces requêtes, une analyse des résultats est présentée ci-dessous et montre le nombre de résultats obtenus, le nombre de résultats à première vue en adéquation avec le sujet, les résultats choisis, enfin le nombres de résultats finalement conservés suite à la suppression des doublons persistants et des documents similaires.

7.2.1 Dialog

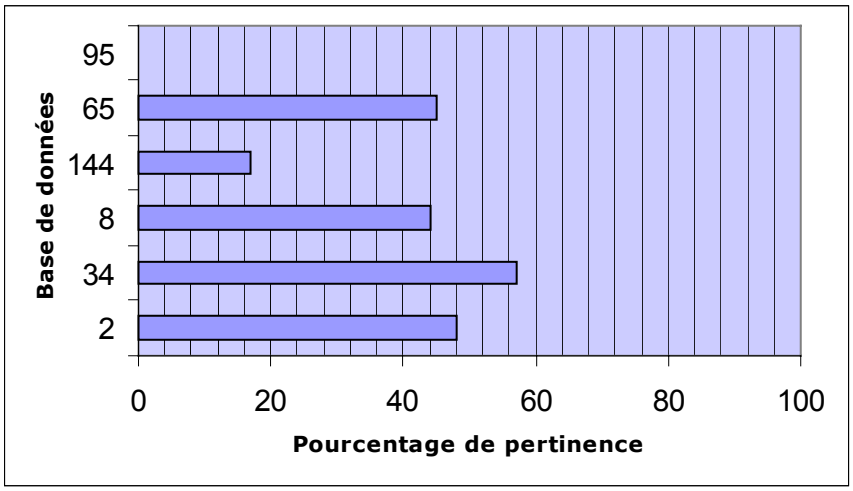
Numéro de requête	Axe de recherche	Références obtenues	Nb références pré-sélectionnées	Nb références retenues	Références retenues [9]
1	2	66	44	21	2.3 [ADI02], [ISH01], [ISH00], [KIM03a], [RIA99], [YOU02], [YOU01] 2.4 [NAC03], [RAM95] 4.1 [ZHU01] 4.3 [BHO03] 4.4 [RIA00]
2	3	28	19	10	2.3 [YOU02] 4.1 [AMO02], [BOU03], [KIM03c], [ZHU01] 5.1 [PED99]
3	4	79	30	14	2.3 [VEL03], [XIN02], [YOU02], [YOU01] 2.4 [ZAI01] 3.1 [FUK00] 3.2 [CAS01], [LOH03] 4.3 [TSE02]

[9] Une référence est identifiée par le numéro de bibliographie qui la contient 0.0, puis par son identifiant [AAA00]

4	5	49	34	22	2.3 [RAU02a], [RAU02b], [TCH01] 2.4 [NAC03], [RAU02c] 4.1 [ZHU01] 4.3 [JEN03] 4.4 [RIA00] 5.1 [PED99], [ZHU00] 5.2 [JEN01], [PED01a], [XIA01]
5	6	36	28	17	2.1 [KIM00] 2.3 [BLE01], [VRD03] 2.4 [ABI03], [TAN03] 4.1 [MIG00] 4.3 [BHO03] 5.1 [BLE00], [NGU03]

Analyse des résultats de Dialog

Le tableau ci-dessous nous présente le degré de pertinence des résultats suivant la base de donnée utilisée sous Dialog.



Pertinence des résultats suivant la base de données utilisée

Cette analyse se rapporte aux résultat de la requête n°1 mais elle me semble significative par rapport à l'ensemble des requêtes effectuées. Dans cet exemple, le faible niveau de pertinence la base PASCAL n°144 provient du fait que le champs titre contient à la fois le titre du document et le nom de la source d'ou est tiré le document. Vu que PASCAL référence les articles d'un périodique dénommé « Decision support systems » et que ce terme apparaissait dans la requête, de nombreux résultats ont été retournés avec un faible niveau de pertinence.

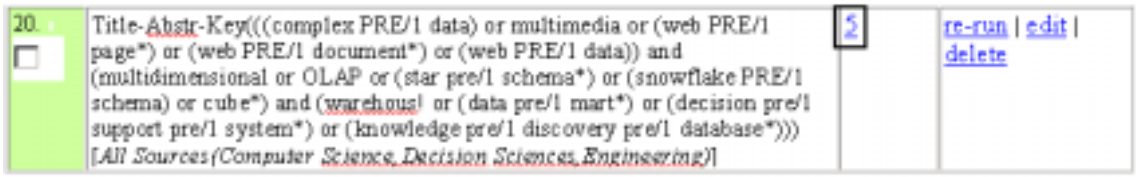
7.2.2 Science Direct

Numéro de requête	Axe de recherche	Références obtenues	Nb références pré-sélectionnées	Nb références retenues	Références retenues
1	4	8	6	4	2.4 [ZAI01] 3.2 [LOH03] 4.3 [TSE02]
2	2	71	32	12	2.3 [ADI02], [KIM03a], [QUA96] 2.4 [NAC03], [RAM95] 4.3 [BHO03], [CAO03], [JEN03] 5.2 [JEN01]
3	2	21	11	7	2.3 [ADI02], [KIM03a] 2.4 [NAC03], [RAM95] 3.2 [LOH03] 4.3 [BHO03]
4	6	36	24	12	2.3 [VRD03] 2.4 [ABI03], [ABI02], [TAN03], [ZAI01] 4.3 [BHO03], [CAO03], [DEL03], [PON03], [ROU03]
5	5	7	3	2	4.3 [JEN03] 5.2 [JEN01]

Analyse des résultats de Science Direct

7.2.3 Comparaison Dialog/Science-Direct

La consultation de Dialog est restée généralement plus prolifique que la consultation sur Science Direct. Voici d'ailleurs un exemple illustrant le nombre de références trouvées sur chacun de ces outils pour une même requête.



Références retrouvées sous Science Direct

S1	COMPLEX(W)DATA OR MULTIMEDIA OR WEB(W)PAGE? ? OR WEB(W)DOCUMENT? ? OR WEB(W)DATA	127940	Display
S2	MULTIDIMENSIONAL OR OLAP OR STAR(1W)SCHEMA? ? OR SNOWFLAKE(1W)SCHEMA OR CUBE? ?	108131	Display
S3	WAREHOUS? OR DECISION(W)SUPPORT(W)SYSTEM? ? OR DATA(W)MART? ? OR KNOWLEDGE(W)DISCOVERY(W)DATABASE? ?	70246	Display
S4	S1 AND S2 AND S3	79	Display
S5	RD S4 (unique items)	64	Display

Références retrouvées sous Dialog

J'ai donc plus souvent limité mes recherches au niveau du titre ou bien aux publications les plus récentes avec Dialog.

7.2.4 La collection LNCS

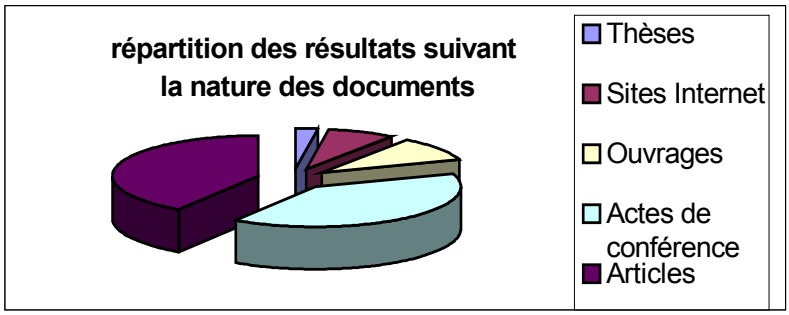
La recherche sur cette collection s'est aussi avérée très prolifique. Par exemple, pour la recherche suivante :

- Sélection d'un premier ensemble d'articles pour le mot-clé warehous* : 306 résultats
- Sélection à l'intérieur de ces résultats avec le mot-clé « complex data » : 38 résultats.

sur les 10 premiers résultats présentés, 5 se retrouvent dans les résultats finaux sélectionnés: [BEB02] [BIA02] et [TCH01] de la bibliographie 2.3, [BOU03] de la bibliographie 4.1, [GOP99] de la bibliographie 5.1.

7.3. Evaluation de la pertinence suivant la nature du document

Le diagramme suivant montre la répartition des résultats suivant la nature du document :



Les résultats ne semblent pas surprenant en considérant le sujet étudié. Les résultats de la recherche scientifique sont essentiellement diffusés par des actes de conférence ou bien par des périodiques spécialisés.

8. Constitution de la bibliographie

La bibliographie a été constituée une fois l'ensemble de mes recherches arrêtées et avant l'élaboration de la synthèse. J'ai choisi d'utiliser le logiciel End Note pour la gestion des références.

8.1. End Note

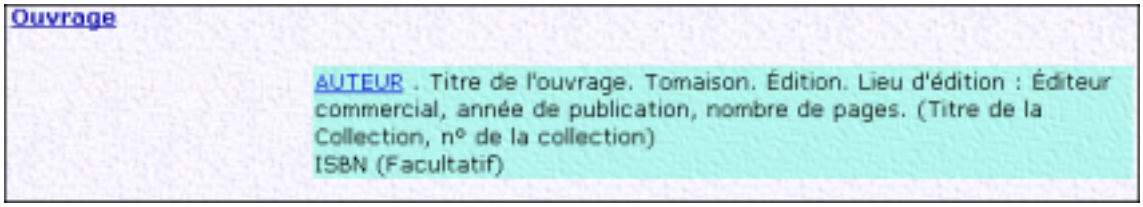
8.1.1 L'import sous End Note

Lors de la découverte de l'outil End Note [10], la possibilité d'import des références depuis les bases de données utilisées se révélait très avantageuse . Cependant, j'en retiens une certaine déception suite à mes différents essais vu que je n'ai réussi à importer mes références automatiquement vers End-Note que depuis une seule base de données. De plus l'import obtenu ne concernait que très peu de champs et une reprise manuelle des résultats étaient nécessaires. J'ai donc intégré manuellement la plupart de mes références dans l'outil End Note.

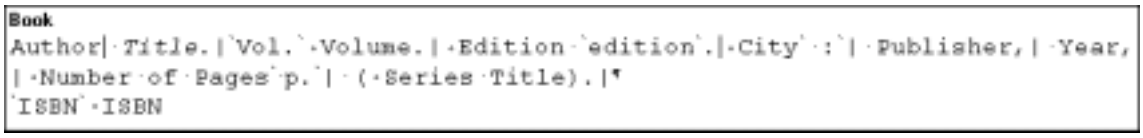
8.1.2 Constitution d'un style End Note

Je me suis servi du guide de rédaction des références bibliographiques [11] défini par le service de documentation de l'INSA afin de constituer ma bibliographie. Après avoir recherché un style End Note qui permette de constituer des références similaires à celles définies dans ce guide, j'ai décidé de créer mon propre style afin d'être en totale conformité avec la norme INSA.

[10] Cours de recherche documentaire informatisée, dirigé par Mme Baligand.
[11] Guide de rédaction de références bibliographique de Doc'INSA disponible sur : <http://csidoc.insa-lyon.fr/docs/refbibli.html> (consulté le 02.03.2004).



Présentation d'une référence de livre selon Doc'INSA



Constitution d'un style End Note pour la référence d'un livre

8.2. La bibliographie

8.2.1 Plan de constitution

J'ai choisi de découper ma bibliographie suivant mes différents axes de recherche. Ensuite, pour chaque axe de recherche, les documents sont classés suivant leur type. Enfin, les références, pour un axe de recherche et un type, sont classés par ordre alphabétique d'auteur. Il me semblait important d'effectuer un découpage de la bibliographie assez détaillée vu qu'elle est susceptible d'être enrichie par la suite.

8.2.2 Numérotation des références

Le choix d'une numérotation unique suivant les trois premières lettres du nom du premier auteur et les deux derniers chiffres de l'année de publication a été convenu avec mon commanditaire, bien que la bibliographie ne soit pas classée par ordre alphabétique.

Ce type d'identification permet ensuite un enrichissement de la bibliographie plus aisée qu'une numérotation numérique traditionnelle et le code utilisée est plus parlant.

8.2.3 Lien entre bibliographie et synthèse

La citation d'une référence s'effectue selon son identifiant ainsi que le numéro de bibliographie dans laquelle la référence est contenue.

9. Constitution de la synthèse

9.1. Les documents sélectionnés

Les principaux documents à étudier pour constituer la synthèse ont été définis avec mon commanditaire :

- **[2.3, BIA02]**
- **[2.3, KIM03a]**
- **[2.3, XIN02]**
- **[5.1, BLE00]**
- **[5.1, NGU03]**
- **[5.1, ZHU00]**
- **[5.2, JEN01]**

Le choix a été fait d'insister sur la partie modélisation multidimensionnelle des données complexes.

Bilan de la recherche

Je retiens de cette première expérience de recherche documentaire des débuts très douloureux mais un aboutissement personnellement très enrichissant. Suite à des débuts aléatoires, j'ai décidé d'adopter une stratégie de recherche rigoureuse consistant en la mise en place d'une structuration de mon sujet en différentes phases.

1. Estimation du temps passé

Recherche documentaire

- Préparation de la recherche : 7h
- Recherche sur les bases de données
  - Dialog : 10h
  - Science Direct : 8h
- Recherche sur Internet
  - Outils généraux : 5h
  - Outils spécialisés : 7h
- Recherches annexes : 10h
- Dépouillement des résultats : 12h

Réalisation

- Méthodologie de recherche : 13h
- Bibliographie : 5h
- Synthèse : 12h

Synthèse

1. Introduction

Face à l'évolution de l'origine et la présentation de l'information, les outils d'aide à la décision doivent constamment répondre à de nouvelles exigences. Le problème qui se pose alors est de savoir si les techniques traditionnelles d'aide à la décision, et plus spécialement l'entreposage de données, actuellement bien maîtrisées pour l'exploitation de données simples, sont capables ou non de s 'adapter à de nouveaux types de données de plus en plus complexes.

Suite à la définition des principales notions du sujet étudié, un état de la recherche pour l'entreposage de données complexes est présenté. Ensuite, l'accent sera mis sur le traitement des données complexes et nous terminerons par une étude de différents types de modélisation multidimensionnelle qui tiennent compte de la complexité de certaines nouvelles formes de données.

2. Définitions

2.1. L'entreposage de données

Un entrepôt de données peut-être défini comme "une structure informatique dans laquelle est centralisé un volume important de données consolidées à partir des différentes sources de renseignements d'une entreprise (notamment les bases de données internes) et qui est conçue de manière que les personnes intéressées aient accès rapidement à l'information stratégique dont elles ont besoin" [12]. Cette pièce maîtresse de l'informatique décisionnelle organise donc les données de manière à faciliter la prise de décision grâce à des outils d'analyse en ligne ou de fouille de données.

2.2. Les données complexes

La notion de donnée complexe n'a de sens que dans un contexte précis. Dans le cadre de ma recherche, une donnée est qualifiée de complexe quand [13] :

- elle est représentable par différents types de données (texte, images, sons, etc.)
- elle est multi-structurée (structurée, semi-structuré ou non structurée)
- elle provient de sources hétérogènes (base de données, Web, etc.)
- le phénomène qu'elle représente est décrit selon plusieurs point de vue
- elle est évolutive dans le temps

[12] Selon le grand dictionnaire terminologique de l'office québécois de la langue française. Disponible sur <http://www.granddictionnaire.com/btml/fra/r_motclef/index1024_1.asp> (consulté le 02.03.2004).

[13] Selon le pôle BDD du laboratoire ERIC. Disponible sur <http://bdd.univ-lyon2.fr/?page=donnees_complexes> (consulté le 02.04.2004).

3. Etat de la recherche sur l'entreposage des données complexes

3.1. Une remise en cause des outils traditionnels

3.1.1 L'apparition de nouvelles formes de représentation des données

L'explosion de l'Internet a constitué un nouvel enjeu pour l'entreposage et la fouille de données. Auparavant, les systèmes d'aide à la décision évoluaient dans un environnement clos ou les données provenaient essentiellement de systèmes internes. Puis Internet s'est imposé comme une source externe majeure d'information qu'il semble désormais vital d'intégrer dans les nouveaux systèmes d'aide à la décision **[2.3, XIN02]**.

L'amas d'information que constitue Internet se caractérise principalement par son hétérogénéité, sa multi-structuration et son instabilité temporelle.

3.1.2 Des systèmes de modélisation inadaptés

Par définition, la notion d'entreposage de données est en contradiction avec le fonctionnement du Web qui manipule des données multi-structurées **[5.1, ZHU00]**. En effet, l'un des principes de base de l'entrepôt de données est justement de proposer un modèle de données unique et structuré. La plupart des entrepôts reposent sur une modélisation et une algèbre relationnelles trop rigides et inadaptées à la représentation de données multimédias **[2.3, KIM03a]**. Il en est de même pour la modélisation orientée objet dont le système de typage empêche de tenir compte de l'hétérogénéité et la spécificité de chaque donnée multimédia.

3.2. Etat de la recherche scientifique

3.2.1 Les prémices

La recherche scientifique sur le sujet de l'entreposage de données complexes est d'une tendance générale assez récente. Un axe de recherche sous-jacent à ce sujet a précédemment été largement étudié, celui de la modélisation de données complexes pour les systèmes traditionnels de gestion de bases de données. Citons par exemple la création de bases de données multimédia **[4.3, GRO97]** ou bien le

projet LORE [14] visant à intégrer des données semi-structurées dans un système de gestion de bases de données.

3.2.2 Une orientation vers les données du Web

La définition d'une donnée complexe est très large et concerne de nombreux types de données. Les travaux de recherche recensés dans le cadre de cette étude traitent essentiellement de données issues du Web ou bien de données multimédias. Notons aussi qu'un format attire l'attention des chercheurs : XML. Pour des données de type image, audio, vidéo, les recherches sont plus orientées vers les techniques générales d'extraction de caractéristiques plutôt que d'entreposage.

3.3. Les principaux projets

Un entrepôt dynamique de données XML dénommé Xyleme **[2.4, ABI02]** à vu le jour en 2000 à l'instigation de différents centre de recherche dont l'INRIA. Ce projet ne s'est pas arrêté au stade de la recherche car il est actuellement exploité commercialement par la société Xyleme [15]. Pour des données issues du Web, un entrepôt de données dénommé Whoweda à vu le jour à l'université technologique de Nanyang à Singapour **[1.4, WHO04]** et de nouvelles recherches sont encore actuellement en cours sur ce projet. De même, le département informatique et ingénierie de l'université de Séoul à construit un entrepôt de données Web et multimédia **[2.3, KIM03a]**. Par ailleurs, la recherche concerne aussi des projets d'entrepôt de données plus spécialisés tels que des entrepôts d'images ou bien des entrepôts de textes **[5.1, BLE00]**.

---

[14] Site Internet de présentation du projet LORE disponible sur : <http://www-db.stanford.edu/lore/> (consulté le 02.03.2004).

[15] Site Internet de la société disponible sur : <http://www.xyleme.com/fr/> (consulté le 02.03.2004).

4. La gestion des données complexes

4.1. L'importance du contenu et de son interprétation

4.1.1 Un travail sur le contenu

Considérer et traiter le contenu d'une donnée multimédia est essentiel pour une recherche d'information par la suite efficace **[4.3, GRO97]**. En 2001, Tim Berners-Lee reprend cette idée au niveau d'Internet en évoquant le concept de Web sémantique [16]. Il imagine le Web non plus comme un outil seulement syntaxique, mais aussi sémantique. Cette nouvelle orientation se traduit actuellement par la mise en place de pierres angulaires telles que les ontologies [17] qui permettent une description des différents concepts et relations existants pour un domaine spécifique. Cette idée de classification suivant le sens de la donnée traitée se retrouve dans certains travaux d'entreposage soit par l'utilisation d'ontologies **[5.1, ZHU00]**, soit par l'utilisation de techniques similaires comme par exemple l'utilisation d'un arbre de catégories sémantiques pour hiérarchiser différents documents en fonction de son contenu dans un entrepôt de texte **[5.1, BLE00]**. Enfin, les cartes topiques XML [18] peuvent être une solution pour l'annotation du contenu de données multimédias dans le cadre de l'entreposage de ces données.

4.1.2 Un travail sur l'interprétation

La représentation d'une donnée complexe peut aussi se fonder sur les impressions subjectives perçues par l'humain lors de l'analyse de cette donnée. L'idée qui provient de la « Kansei engineering community » du Japon (Kansei signifiant impression subjective) est donc d'associer les caractéristiques de bas niveau d'une donnée multimédia à l'impression subjective renvoyée par ces caractéristiques. Afin de capter l'interprétation des utilisateurs lors de l'assimilation d'un ensemble de données, un processus de rétroaction est utilisé pour la création de l'association entre les caractéristiques d'une donnée et l'interprétation qui en est faite. Ce

[16] site Internet de présentation du Web sémantique disponible sur: <http://www.w3.org/2001/sw/> (consulté le 02.03.2004)
[17] site Internet pour la définition d'une ontologie disponible sur: <http://www.semanticweb.org/knowmarkup.html#ontologies> (consulté le 02.03.2004).

processus peut encore être approfondi grâce à l'utilisation d'un méta schéma hiérarchique qui se représente sous la forme d'un arbre et qui permet d'affiner la description d'une caractéristique de l'image (par exemple la tonalité) suivant différents niveaux de hiérarchie **[2.3, BIA02]**. L'interprétation d'une donnée se fera donc suivant différents méta schémas hiérarchiques et suivant différents niveaux dans ce schéma, permettant ainsi de prendre en compte les impressions et descriptions multiples qui se dégagent d'une donnée.

4.2. La modélisation

Dans le cadre de l'entreposage de données, une donnée est extraite d'un système existant et doit être transformée afin de correspondre à une modélisation commune à l'ensemble des données prises en compte. En ce qui concerne le seul processus de modélisation, le standard MPEG-7 [19] est en cours de développement pour la description de données multimédias. La modélisation peut tenir compte de chaque type de donnée multimédia en proposant un modèle spécifique pour chaque type de données puis une structure homogène générale capable de modéliser un document suivant les différents types de données qu'il contient **[4.1, AMO02]**. XML peut être utilisé pour la représentation dynamique d'objets complexes, cependant, il semble nécessaire d'ajouter à cette technologie certains concepts orientés objet afin de pallier les lacunes de XML pour la gestion des relations entre objets **[2.3, KIM03a]**. Enfin, l'utilisation de graphes de relations entre objets peut permettre de spécifier des relations sémantiques entre différents objets et créer ainsi des objets complexes.

4.3. Les principales orientations

Un système intelligent d'aide à la décision capable d'analyser tout type d'information **[2.3, XIN02]** relève actuellement plus de l'utopie que de la réalité. Cependant, il est possible d'identifier deux orientations majeures :

---

[18] site Internet de spécification des cartes topiques XML disponible sur <http://www.topicmaps.org/xtm/1.0/> (consulté le 02.03.2004).

[19] site Internet de présentation du standard MPEG-7 disponible sur : <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm> (consulté le 02.03.2004).

4.3.1 Vers une structure dynamique des données complexes

Il est très difficile de représenter à l'avance la structure d'une donnée complexe. C'est pourquoi l'utilisation d'une structure dynamique permettant une modélisation évolutive semble appropriée afin de prendre en compte de nouveaux types de données et de contenus. En utilisant XML qui permet une auto-description des données, il est possible de profiter de l'absence du concept de classes et de la définition stricte de types pour utiliser une structuration dynamique des données **[2.3, KIM03a]**.

4.3.2 La norme XML

XML, eXtensible Markup Language [20] est un langage à balises utilisé pour la représentation des données et qui apparaît comme une solution à la modélisation de données complexes en vue de leur entreposage. En effet, la majorité des recherches effectuées dans ce domaine abordent cette technologie, comme par exemple la création d'un entrepôt entièrement construit à partir des fonctionnalités de XML **[2.3, KIM03a]**. Un format simple, auto-descriptif et permettant l'interopérabilité sont les principales justifications au choix de XML. De plus, il s'agit d'un format de plus en plus utilisé pour les échanges de données sur le Web, ce qui justifie l'intérêt à lui porter dans les systèmes d'aide à la décision **[5.2, JEN01]**.
En associant la structure et le contenu, le format XML se présente comme un puissant outil de modélisation pour les données complexes.

5. La modélisation multidimensionnelle des données complexes

5.1. Présentation

La modélisation multidimensionnelle est « une discipline de modélisation des données qui se positionne comme solution de rechange à la modélisation entité/relation » **[1.1, KIM02]**. Ce type de modélisation, qui caractérise des faits

à analyser par différentes dimensions organisées de façon hiérarchique, est particulièrement utilisé pour l'entreposage de données afin de supporter les techniques d'analyse en ligne et de fouille de données.

5.2. De l'utilisation de l'existant à la création de nouveaux concepts

Deux orientations majeures se distinguent pour la modélisation multidimensionnelle des données complexes. La première consiste à traiter les données complexes en fonction des outils existants alors que la seconde définit de nouveaux concepts directement adaptés à de nouveaux besoins.

5.2.1 L'utilisation de l'existant

Il est possible de trouver un compromis entre les outils OLAP traditionnels existants et l'intégration de données XML **[5.2, JEN01]**. La motivation essentielle de ce compromis est de pouvoir profiter des outils d'analyse en ligne existants, faute d'outils efficients pour l'analyse directe de données XML. L'inconvénient est que cette adaptation implique une série de méthodes de translation afin que les données représentées à la base selon un modèle spécifique, par exemple le modèle orienté objet, soient transférées dans un modèle multidimensionnel traditionnel **[5.1, ZHU00]**.

5.2.2 La création de nouveaux concepts

La remise en cause des outils existants implique la création de nouvelles techniques. Par exemple, le langage XML peut être utilisé pour la création d'un entrepôt de données multimédias, en enrichissant cependant ce langage de concepts orientés objets **[2.3, KIM03a]**. L'analyse des données s'effectue ensuite non pas par une modélisation multidimensionnelle classique, mais par la matérialisation de vues dans un environnement XML.

---

[20] Site Internet de présentation de XML disponible sur : <http://www.w3.org/XML> (consulté le 02.03.2004).

5.3. Différents modèles

Trois différents modèles sont exposés dans cette partie : un modèle conceptuel basé sur une représentation UML [21] **[5.2, JEN01]**, un modèle dynamique pour l'entreposage de textes **[5.1, BLE00]** et enfin un modèle multidimensionnel de métadonnées pour l'interopérabilité entre différents entrepôts **[5.1, NGU03]**.

5.3.1 Un modèle UML en flocon de neige

Ce modèle multidimensionnel basé sur UML, décrit à un niveau conceptuel, supporte à la fois la gestion des données XML et des données relationnelles. UML est considéré comme un outil capable de capturer un fort niveau sémantique pour des données XML, puis comme un standard facilement compréhensible pour les concepteurs et utilisateurs du système. Dans un premier temps, les documents XML sources, par le biais de leur structure logique, les DTDs [22], sont automatiquement modélisés sous forme de diagrammes UML. Ensuite, ce diagramme doit être transformé afin de supporter une modélisation multidimensionnelle constituée de classes UML. Cette transformation est effectuée de façon manuelle par le biais d'une interface utilisatrice et respecte des spécifications précises de structuration. Le modèle obtenu est un modèle en flocon de neige similaire au modèle relationnel traditionnel mais constitué de classes UML. La dernière étape consiste à transformer ce modèle en un modèle multidimensionnel relationnel afin de profiter des outils traditionnels d'analyse en ligne.

5.3.2 Un modèle dynamique pour l'entreposage de texte

Le modèle proposé est un modèle dynamique multidimensionnel en flocon de neige utilisé pour l'entreposage de textes. Les différentes dimensions du modèle représentent des catégories sémantiques qui caractérisent le contenu du texte analysé. Seulement, le nombre de catégories sémantiques qui caractérisent un texte ne peut être fixé à l'avance, ce qui implique de capturer dynamiquement le nombre

[21] Unified Modeling Language (UML) est un modèle conceptuel de modélisation orienté objet.
[22] Document Type Definition (DTD) représentation la structure logique d'un document XML.

de dimensions au niveau du modèle conceptuel. Cette tâche est réalisée grâce à l'ajout d'une table d'index au niveau du schéma multidimensionnel contenant les informations sur les dimensions utilisées pour chaque document. Le schéma contient cependant deux dimensions fixes qui sont les dimensions traditionnelles de temps et de lieu. Deux fonctions essentielles sont ensuite permises par cet entrepôt, la modification de la table d'index qui contient les catégories sémantiques puis la recherche de documents suivants ces différentes catégories sémantiques.

5.3.3 Le métacube XTM

Le protocole de métacube XTM permet l'interopérabilité entre différents entrepôts distribués de données Web. Le but est d'interroger de façon unique différents entrepôts quels que soientt le schéma de donnée et le modèle multidimensionnel utilisés par chaque entrepôt. Cette approche distribuée nécessite la mise en place d'un système centralisé destiné à gérer l'ensemble des entrepôts, d'ou le concept de métacube global et de métacubes locaux. Le rôle du premier est de gérer l'hétérogénéité entre les différents entrepôts et de proposer un système centralisé de recherche d'information, tandis que les seconds associés à un entrepôt de données décrivent le modèle multidimensionnel utilisé localement. Un métacube est conceptuellement modélisé par le langage UML et utilise la notion de carte topique permettant d'organiser les informations selon une vue unifiée du Web. Cette modélisation permet la représentation unique de tout type de schéma multidimensionnel basé sur les notions de fait et de dimension.

Bibliographie

1. L'entreposage de données

1.1. Ouvrages

[ANA97] ANAHORY S., MURRAY D. *Data warehousing in the real world: a practical guide for building decision support systems*. 1ère edition. Boston: Addison Wesley Professional, 1997, 368 p.
ISBN 0201175193

[INM02] INMON W.H. *Building the data warehouse*. 3ème edition. John Wiley & Sons, 2002, 356 p.
ISBN 0471081302

[JAR03] JARKE M., LENZERINI M., VASSILIOU Y., et al. *Fundamental of data warehouse*. 2nde édition. Berlin, Allemagne: Springer-Verlag, 2003, 207 p.
ISBN 3540420894

[KIM02] KIMBALL R., ROSS M. *The data warehouse toolkit: the complete guide to dimensional modeling*. 2nde édition. John Wiley & Sons, 2002, 416 p.

[SAN97] SANDOVAL V. *L'informatique décisionnelle*. Paris: Hermes, 1997, 126 p.
ISBN 2866016165

1.2. Ressources Internet

[CHU04] CHUO-HAN L. *Data warehousing* **[en ligne]**. Disponible sur: <http://www.1keydata.com/datawarehousing/datawarehouse.html>. (consulté le 02.02.2004).

[GRE04] GREENFIELD L. *The data warehousing information center* **[en ligne]**. Disponible sur: <http://www.dwinfocenter.org>. (consulté le 08.01.2004).

[LEM04] LEMIRE D. *Data warehousing and OLAP, a research-oriented bibliography* **[en ligne]**. Disponible sur: <http://www.ondelette.com/OLAP/dwbib.html>. (consulté le 02.02.2004).

[THE04] THE DATA WAREHOUSING INSTITUTE. *The data warehousing institute web site* **[en ligne]**. Disponible sur: <http://www.dw-institute.com>. (consulté le 02.02.2004).

[UNI04] UNIVERSITE TRIER. *DBLP computer science bibliography, database system**s*** **[en ligne]**. Disponible sur: <http://www.informatik.uni-trier.de/~ley/db/conf/index.html>. (consulté le 02.02.2004).

1.3. Conférences internationales

[DEX04] DEXA. *DAWAK, DAta WArehousing and knowledge discovery* **[en ligne]**. Disponible sur: <http://www.dexa.org> (consulté le 01.03.2004).

[DMD04] *DMDW, Design and Management of Data Warehouses* **[en ligne]**. Disponible sur: <http://sunsite.informatik.rwth-aachen.de/Societies/DMDW/> (consulté le 01.03.2004).

[IEE04] IEEE COMPUTER SOCIETY. *ICDE, International Conference on Data Engineering* **[en ligne]**. Disponible sur: <http://dblp.uni-trier.de/db/conf/icde/> (consulté le 01.03.2004).

[INT04] *International workshop on data warehousing and OLAP* **[en ligne]**. Disponible sur: <http://www.cis.drexel.edu/faculty/song/dolap.html> (consulté le 01.03.2004).

[IGI04] ACM. *SIGIR, Special Interest Group on Information Retrieval* **[en ligne]**. Disponible sur: <http://www.acm.org/sigir/> (consulté le 01.03.2004).

[KDD04] ACM. *SIGKDD, Special Interest Group on Knowledge Discovery and Data mining* **[en ligne]**. Disponible sur: <http://www.acm.org/sigkdd/> (consulté le 01.03.2004).

[MOD04] ACM. *SIGMOD, Special Interest Group on Management of Data* **[en ligne**]. Disponible sur: <http://www.acm.org/sigmod/index.html> (consulté le 01.03.2004).

[VLD04] *VLDB, Very Large Databases* **[en ligne]**. Disponible sur: <http://www.vldb.org/> (consulté le 01.03.2004).

1.4. Les laboratoires de recherche

[AAL04] *Data warehousing at Aalborg university* **[en ligne]**. Disponible sur: <http://www.cs.auc.dk/~strategy/Warehousing/index.html> (consulté le 01.03.2004).

[ALB04] *Database systems research group, department of computing science, university of Alberta* **[en ligne]**. Disponible sur : <http://www.cs.ualberta.ca/research/labs/database/> (consulté le 01.03.2004).

[COL04] *Database research group, computer science department, university of Columbia* **[en ligne]**. Disponible sur : <http://www.cs.columbia.edu/database> (consulté le 01.03.2004).

[IBM04] *IBM research knowledge discovery and data mining* **[en ligne]**. Disponible sur: <http://www.research.ibm.com/compsci/kdd/index.html> (consulté le 01.03.2004).

[INR04] *INRIA, Institut National de Recherche en Informatique et en Automatique* **[en ligne]**. Disponible sur <http://www.inria.fr> (consulté le 01.03.2004).

[KNO04] *Knowledge and database systems laboratory, national technical university of Athens* **[en ligne]**. Disponible sur: <http://www.dbnet.ece.ntua.gr/> (consulté le 01.03.2004).

[PRI04] *PRISM, laboratoire de recherche en informatique* **[en ligne]**. Disponible sur: <http://www.prism.uvsq.fr/> (consulté le 01.03.2004).

[STA04] *Data warehousing at Stanford* **[en ligne]**. Disponible sur: <http://www-db.stanford.edu/warehousing/links.html> (consulté le 01.03.2004).

[WHO04] *Whoweda, the Web warehousing and mining group* **[en ligne]**. Disponible sur: <http://mandolin.cais.ntu.edu.sg/~whoweda/index.htm> (consulté le 01.03.2004).

2. Le traitement des données complexes en analyse et fouille de données

2.1. Ouvrages

[KIM00] KIMBALL R., MERZ R. *The data webhouse toolkit: building the Web-enabled data warehouse*. 1ère édition. John Wiley & Sons, 2000, 416 p.
ISBN 0471376809

[THU01] THURAISINGHAM B. *Managing and mining multimedia databases*. CRC presse, 2001, 352 p.
ISBN 0849300371

2.2. Chapitres d'ouvrage

[DAR03] DARMONT, J., BOUSSAID, F., BENTAYEB, S. *Web multiform data structuring for warehousing*. **In:** DJERABA, C. Multimedia mining: a highway to intelligent multimedia documents. Boston, Etats-Unis: Kluwer Academic Publishers, 2003, pp. 179-194. (Multimedia systems and applications, n° 22)

2.3. Communications dans une conférence

[ADI02] ADIBA M., ZECHINELLI-MARTINI J.L. *Building spatio-temporal presentations warehouses from heterogeneous multimedia web servers*. **In:** BANKS PIDDUCK A., MYLOPOULOS J., WOO C.C., et al. Eds. Advanced information systems engineering, 14th international conference, CAiSE 2002, 27-31 mai 2002, Toronto, Canada. Heidelberg, Berlin, Allemagne: Springer-Verlag, 2002, pp 692-696. (Lecture notes in computer science, n° 2348)

[BEB02] BEBEL B., KROLIKOWSKI Z., WREMBEL R. *On method's materialization in object-relational data warehouse*. **In:** YAKHSNO T. Ed. Advances in information systems: second international conference, ADVIS 2002, 23-25 octobre 2002, Izmir, Turquie. Heidelberg, Berlin, Allemagne: Springer Verlag, 2002, pp 425-434. (Lecture notes in computer science, n° 2457)

[BIA02] BIANCHI-BERTHOUSE N., HAYASHI T. *Subjective interpretation of complex data: requirements for supporting kansei mining process*. **In:** SIMEON J.S., DJERABA C., ZAIANE O.R. Eds. MDM/KDD'2002 third international workshop on multimedia data mining, 23 juillet 2002, Edmonton, Alberta, Canada **[en ligne]**. University of Alberta, 2002, pp 93-99. Disponible sur : <http://www-staff.it.uts.edu.au/~simeon/mdm_kdd2002/KDD02-mdmkdd.pdf> (consulté le 22.01.2004).

[BLE01] BLEYBERG M.Z., PARANJAPE P.S. *A content delivery strategy for text warehouses*. **In:** 2001 IEEE international conference on systems, man and cybernetics, e-systems and e-man for cybernetics in cyberspace, 7-10 octobre 2001, Tucson, AZ, Etats-Unis. Piscataway, NJ, USA: IEEE, 2001, pp 2322-2325.
ISBN 0780370872

[FAN03] FANKHAUSER P., KLEMENT T. *XML for data warehousing chances and challenges*. **In:** Data Warehousing and Knowledge discovery, 5th international conference, DaWaK 2003, 3-5 septembre 2003, Prague, République Tchèque. Heidelberg, Berlin, Allemagne: Springer-Verlag, 2003, pp 1-3. (Lecture notes in computer science, n° 2737).

[ISH01] ISHIKAWA H., OHTA M., KATO K. *Document warehousing: a document-intensive application of a multimedia database*. **In:** ABERER K., LIU L Eds. IEEE 11th workshop on research issues in data engineering, 1-2 avril 2001, Heidelberg, Berlin, Allemagne. Los Alamitos, CA, USA: IEEE computer society, 2001, pp 25-31.
ISBN 0769509576

[ISH00] ISHIKAWA H., OHTA M., KATO K*. A multimedia database support for document warehousing*. **In:** FURHT B. Ed. Proceedings of 2000 conference on internet and multimedia systems and applications, 19-23 novembre 2000, Las Vegas, NV, Etats-Unis. Anaheim, CA, Etats-Unis: IASTED, 2000, pp 418-423.

[KIM03a] KIM H.H., PARK S.S. *Building a web-enabled multimedia data warehouse*. **In:** The second international human.society@internet conference, 18-20 juin 2003, Seoul, Corée. Heidelberg, Berlin, Allemagne: Springer-Verlag, 2003, pp 594-600. (Lecture notes in computer science, n0 2713).

[KIM03b] KIM H.H., PARK S.S. *A semantics-based versioning scheme for multimedia data*. **In:** DAtabase Systems For Advanced Applications, 9th international conference, DASFAA 2004, 17-19 mars 2003, Jeju Island, Corée. Heidelberg, Berlin, Allemagne: Springer-Verlag, 2003, pp 277-288. (Lecture notes in computer science, n° 2973).
ISBN 3540210474

[QUA96] QUASS D., WIDOM J., GOLDMAN R. et al. *LORE: A Lightweight Object REpository for semistructured data*. **In:** JAGADISH H.V., MUMICK I.S. 1996 ACM SIGMOD international conference on management of data, 04-06 juin 1996, Montreal, Quebec, Canada **[en ligne]**. ACM press, 1996, pp 549. Disponible sur : <http://www-db.stanford.edu/lore/pubs/lore-demo.pdf> (consulté le 16.02.2004).

[RAU02a] RAUBER A., ASCHENBRENNER A., WITVOET O. *Austrian online archive processing: analyzing archives of the World Wide Web*. **In:** AGOSTI M., THANOS C. Eds. Research and advances technology for digital technology , 6th european conference, ECDL 2002, 16-18 septembre 2002, Rome, Italie.

Heidelberg, Berlin, Allemagne: Springer-Verlag, 2002, pp 16-31. (Lecture notes in computer science, n° 2458).
ISBN 0302-9743

[RAU02b] RAUBER A., WITVOET O., ASCHENBRENNER A. *Putting the World Wide Web into a data warehouse: a DWH-based approach to Web analysis*. **In:** TJOA A.M., WAGNER R.R. Eds. Proceedings 13th international workshop on Database and Expert Systems Applications, DEXA, 2-6 septembre 2002, Aix en Provence, France **[en ligne]**. Los Alamitos, CA, Etats-Unis: IEEE computer society, 2002, pp 822-826. Disponible sur : <http://citeseer.nj.nec.com/cache/papers/cs/26752/http:zSzzSzwww.ifs.tuwien.ac.atzSzifszSzresearchzSzpub_pszSzrau_vldwh02.pdf/rauber02putting.pdf> (consulté le 17.02.2004).
ISBN 0769516688

[RIA99] RIAHI F., MOTHE J. *Dataweb : construction automatique d'un entrepôt de données à partir de documents issus du Web*. **In:** Ecrit et multimedia, 9 septembre 1999, Tours, France. 1999, p. 37-38.

[TCH01] TCHOUNIKINE A., MIQUEL M., FLORY A. *Information warehouse for medical research*. **In:** KAMBAYASHI Y., WINIWARTER W., ARIKAWA M. Eds. Data Warehousing and Knowledge discovery : third international conference, DaWaK 2001, 5-7 septembre 2001, Munich, Allemagne. Heidelberg, Berlin, Germany: Springer-Verlag, 2001, pp 208-218. (Lecture Notes in Computer Science, n° 2114)
ISBN 3540425535

[VEL03] VELASQUEZ J.D., YASUDA H., AOKI T. et al. *A generic data mart architecture to support Web mining*. **In:** ZANASI A. Data mining, 4th international conference, 2003, Rio de Janeiro, Brésil. Southampton, Royaume-Uni: WIT press, 2004, pp 389-400. (Management information systems, n° 4)
ISBN 1853128066

[VRD03] VRDOLJAK B., BANEK M., RIZZI S. *Designing web warehouses from XML schemas*. **In:** DAta WArehousing and Knowledge discovery, 5th international conference, DaWaK 2003, 3-5 septembre 2003, Prague,

République Tchèque. Heidelberg, Berlin, Allemagne: Springer-Verlag, 2003, pp 89-98. (Lecture notes in computer science, n° 2737).

[XIN02] XIN-ZHONG C., YAN L., BING-RU Y. *New intelligent decision support systems based on information mining*. **In:** 2002 international conference on machine learning and cybernetics, 4-5 novembre 2002, Beijing, Chine **[en ligne]**. Piscataway, NJ, Etats-Unis: IEEE, 2002, pp 962-967. Disponible sur : <http://web.it.kmutt.ac.th/~chakarida/dmpaper/paper9.pdf> (consulté le 05.03.2004).
ISBN 0780375084

[YOU02] YOU J., LIU J., LI L. et al. *On data mining and data warehousing for multimedia information retrieval*. **In:** ISHI N. ACI'02: IASTED international conference on artificial computational intelligence, 25-27 septembre 2002, Tokyo, Japon. Anaheim, CA, Etats-Unis: ACTA Press, 2002, pp130-135.
ISBN 088986358X

[YOU01] YOU J., DILLON T., LIU J*. An integration of data mining and data warehousing for hierarchical multimedia information retrieval*. **In:** Proceedings of 2001 International Symposium on Intelligent Multimedia, video and speech Processing, ISIMP 2001, 2-4 mai 2001, Hong Kong, Chine. Piscataway, NJ, USA: IEEE, 2001, pp 373-376.
ISBN 9628576623

[ZHU99] ZHU Y. *A framework for warehousing Web contents*. **In:** HUI L.C.K., LEE D.L. ICSC'99: 5th international computer science conference, 13-15 décembre 1999, Hong Kong, Chine. Heidelberg, Berlin, Allemagne: Springer-Verlag, 1999, pp 83-92. (Lecture notes in computer science, n° 1749).
ISBN 3540669035

2.4. Articles de périodiques

[ABI03] ABITEBOUL S. *Managing an XML Warehouse in a P2P Context*. Lecture notes in computer science, 2003, vol. 2681, pp 4-13.

[ABI02] ABITEBOUL S., CLUET S., FERRAN G., et al. *The Xyleme project*. Computer networks, 2002, vol. 39, n° 3, pp 225-238.

[COM03] COMAI S., MARRARA S., TANCA L. *Representing and querying summarized XML data*. Lecture notes in computer science, 2003, vol. 2736, pp 171-181.

[NAC03] NACHOUKI G., CHASTANG M.P. *On-line analysis of a web data warehouse*. Lecture notes in computer science, 2003, vol. 2822, pp 112-121.

[PED01b] PEDERSEN T.B., JENSEN C.S., DYRESON C.E. *A foundation for capturing and querying complex multidimensional data*. Information systems, 2001, vol. 26, n° 5, pp 383-423.

[RAM95] RAMESH B., SENGUPTA K. *Multimedia in a design rationale decision support system*. Decision support systems, 1995, vol. 15, n° 3, pp 181-196.

[RAU02c] RAUBER A., ASCHENBRENNER A., WITVOET O., et al. *Uncovering information hidden in Web archives* **[en ligne]**. D-Lib magazine, 2002, vol. 8, n° 12. Disponible sur: <http://www.dlib.org/dlib/december02/rauber/12rauber.html> (consulté le 04.02.2004).

[TAN03] TAN X., YEN D.C., FANG X. *Web warehousing: Web technology meets data warehousing*. Technology in society, 2003, vol. 25, n° 1, pp 131-148.

[ZAI01] ZAIANE O.R. *Building virtual web views*. Data & knowledge engineering, 2001, vol. 39, n° 2, pp 143-163.

2.5. Thèses

[TES00] TESTE O. *Modélisation et manipulation d'entrepôts de données complexes et historisées*. Thèse doctorale. Toulouse, France: Université de Toulouse 3, Toulouse, France, 2000, 214 p.

3. L'extraction de caractéristiques pour les données complexes

3.1. Les actes de conférences

[FUK00] FUKUDA F.H., PASSOS E.L.P., PACHECO M.A., et al. *Web text mining using a hybrid intelligent system based on KDT, expert system and neural network*. **In:** EBECKEN N., BREBBIA C.A. Data Mining II, second international conference on data mining, juillet 2000, Cambridge, Royaume-Uni. Southampton, Royaume Uni: WIT press, 2000, pp 363-372.
ISBN 185312821X

[JIZ01] JI Z., WYNNE H., MONG LI L. *An information-driven framework for image mining*. **In:** MAYR H.C., LAZANSKY J., QUIRCHMAYR G. et al. Eds. Database and EXpert systems Applications, 12th international conference, DEXA 2001, 3-5 septembre 2001, Munich, Allemagne. Heidelberg, Berlin, Allemagne: Springer-Verlag, 2001, pp 232-242. (Lecture notes in computer science, n° 2113).
ISBN 3540425276

3.2. Les articles de périodiques

[HSU02] HSU W., LEE M.L., ZHANG L. *Image mining: Trends and developments*. Journal of intelligent information systems, 2002, vol. 19, n° 1, pp 7-23.

[KES01] KESONG H., YONGCHENG W. *Text mining, data mining vs. knowledge management: the intelligent information processing in the 21st century*. Journal of the China society for scientific and technical information, 2001, vol. 20, n° 1, pp 100-104.

[LEI02] LEI Z., AIBING R., AIDONG Z. *Advanced feature extraction for keyblock-based image retrieval*. Information systems, 2002, vol. 27, n° 8, p. 537-557.

[LOH03] LOH S., DE OLIVEIRA J.P.M., GAMEIRO M.A. *Knowledge discovery in texts for constructing decision support systems*. Applied intelligence, 2003, vol. 18, n° 3, pp 357-366.

[LOS00] LOSIEWICZ P., OARD DOUGLAS W., KOSTOFF RONALD N. *Textual data mining to support science and technology management*. Journal of intelligent information systems, 2000, vol. 15, n° 2, pp 99-119.

4. Le processus de préparation des données complexes

4.1. Communications dans une conférence

[AMO02] AMOUS I., JEDIDI A., SEDES F. *A contribution to multimedia document modeling and organizing*. **In:** BELLAHSENE Z., PATEL D., ROLLAND C. Eds. Object-oriented information systems : 8th international conference, OOIS 2002, 2-5 septembre 2002, Montpellier, France. Heidelberg, Berlin, Allemagne: Springer-Verlag, 2002, pp 434-444. (Lecture notes in computer science, n° 2425)

[BOU03] BOUSSAID O., BENTAYEB F., DUFFOUX A., et al. *Complex data integration based on a multi-agent system*. **In:** 1st international conference on industrial applications of Holonic and Multi-Agent Systems, HoloMAS 03, septembre 2003, Prague, République Tchèque. Heidelberg, Berlin, Allemagne: Springer-Verlag, 2003, pp 201-212. (Lecture notes in computer science, n° 2744).
ISBN 3540407510

[KIM03c] KIM H.H., PARK S.S. *Mediaviews: a layered view mechanism for integrating multimedia data*. **In:** The 9th international conference on object-oriented information systems, 2-5 septembre 2003, Geneve, Suisse. Heidelberg, Berlin, Allemagne: Springer-Verlag, 2003, pp 250-261. (Lecture notes in computer science, n° 2817)
ISBN 3540408606

[MIG00] MIGNET L., PREDA M., ABITEBOUL S. *Acquiring XML pages for a webhouse*. **In:** BDA'2000, 16èmes journées bases de données avancées, 24-27 octobre 2000, Blois, France **[en ligne]**. 2000, pp 241-263. Disponible sur : <http://www1.cs.columbia.edu/~amelie/papers/bda2000-acq.pdf> (consulté le 16.01.2004).

[ZHU01] ZHU Y., BORNHOVD C., BUCHMANN A.P. *Data transformation for warehousing web data*. **In:** Third international workshop on advanced issues of e-commerce and Web-based information systems, WECWIS 2001, 21-22 juin 2001, San Juan, CA, Etats-Unis **[en ligne]**. Los Alamitos, CA, Etats-Unis: IEEE computer society, 2001, pp 74-85. Disponible sur : <http://www.dvs1.informatik.tu-darmstadt.de/publications/pdf/wecwis01-zbb.pdf> (consulté le 17.02.2004)
ISBN 0769512240

4.2. Sites Internet

[GAR04] IBM The Garlic project **[en ligne]**. Disponible sur : <www.almaden.ibm.com/cs/garlic> (consulté le 26.01.2004).

4.3. Articles de périodiques

[ABI99] ABITEBOUL S., CLUET S., MILO T. et al. *Tools for data translation and integration* **[en ligne]**. IEEE data engineering bulletin, 1999, vol. 22, n° 1, pp 3-8. Disponible sur : <ftp://ftp.research.microsoft.com/pub/debull/99MAR-CD.pdf> (consulté le 16.02.2004)

[BHO03] BHOWMICK S.S., SOURAV S., MADRIA S., et al. *Representation of web data in a web warehouse*. Computer journal, 2003, vol. 46, n° 3, pp 229-262.

[CAO03] CAO Y., LIM E.P., NG W.K. *Data model for warehousing historical Web information*. Information and software technology, 2003, vol. 45, n° 6, pp 315-334.

[DEL03] DELOBEL C., REYNAUD C., ROUSSET M.C., et al. *Semantic integration in Xyleme: a uniform tree-based approach*. Data & knowledge engineering, 2003, vol. 44, n° 3, p. 267-298.

[GRO97] GROSKY W.I. *Managing multimedia information in database systems* **[en ligne]**. Communications of the ACM, 1997, vol. 40, n° 12, pp 72-80. Disponible sur : <http://citeseer.nj.nec.com/cache/papers/cs/20363/http:zSzzSzzeus.cs.wayne.eduzSz~groskyzSzPaperszSz97CACM.pdf/grosky97managing.pdf> (consulté le 17.02.2004).

[HUA02] HUANG Z., CHEN L.D., FROLICK M.N. *Integrating web-based data into a datawarehouse*. Information systems management, 2002, vol. 19, n° 1, pp 23-34.

[JEN03] JENSEN M.R., MOLLER T.H., PEDERSEN T.B. *Converting XML DTDs to UML diagrams for conceptual data integration* **[en ligne]**. Data & knowledge engineering, 2003, vol. 44, n° 3, pp 323-346.Disponible sur : <http://www.cs.auc.dk/~mrj/publications/diweb.pdf> (consulté le 17.02.2004).

[MIL01] MILLER R., HERNANDEZ M., HASS L., et al. *The Clio project: managing heterogeneity* **[en ligne]**. SIGMOD record, 2001, vol. 30, n° 1, pp 78-83. Disponible sur : <www.acm.org/sigmod/record/issues/0103/JP-Sys.pdf> (consulté le 26.01.2004).

[PON03] PONTIERI L., URSINO D., ZUMPANO E. *An approach for the extensional integration of data sources with heterogeneous representation formats*. Data & knowledge engineering, 2003, vol. 43, n° 3, pp 291-331.

[ROU03] ROUSSET M.C., REYNAUD C. *Knowledge representation for information integration* **[en ligne]**. Information systems international journal, 2003, vol. 29, n° 1, pp 3-22. Disponible sur : <http://www.lri.fr/~mcr/publis/is.pdf> (consulté le 16.02.2004).

[SOU02] SOURAV S., BHOWMICK S.S., NG W.K., et al. *Anatomy of the coupling query in a web warehouse*. Information and software technology, 2002, vol. 44, n° 9, pp 513-539.

[TSE02] TSENG F., HWUNG W.J. *An automatic load/extract scheme for XML documents through object-relational repositories*. Journal of systems and software, 2002, vol. 64, n° 3, pp 207-218.

[WHI00] WHITE C. *First analysis [data warehouses]* **[en ligne]**. Intelligent enterprise, 2000, vol. 3, n° 9, pp 50-52,54-55. Disponible sur <http://www.intelligententerprise.com/000605/feat3.jhtml?_requestid=460837> (consulté le 17.02.2004).

[WOE86] WOELK D., KIM W., LUTHER W. *An object-oriented approach to multimedia databases*. SIGMOD record, 1986, vol. 15, n° 2, pp 311-325.

4.4. Thèses

[RIA00] RIAHI F. *Mécanismes pour l'élaboration automatique d'un entrepôt d'informations à partir de documents semi-structures issus du Web*. Thèse doctorale. Toulouse: Université de Toulouse 3, Toulouse, France, 2000, 148 p.

5. Modélisation multidimensionnelle des données complexes

5.1. Communications dans une conférence

[BLE00] BLEYBERG M.G., GANESH K. *Dynamic multi-dimensional models for text warehouses*. **In:** 2000 IEEE international conference on systems, man and cybernetics, `cybernetics evolving to systems, humans, organizations, and their complex interactions', 8-11 octobre 2000, Nashville, TN, Etats-Unis **[en ligne]**. Piscataway, NJ, Etats-Unis: IEEE, 2000, pp 2045-2050. (Proceedings of IEEE international conference on systems, man, and cybernetics, n° 3). Disponible sur : <http://www.cis.ksu.edu/~maria/PAPERS/nldw.pdf> (consulté le 05.03.2004).

ISBN 0780365836

[GOP99] GOPALKRISHNAN V., LI Q., KARLAPAREM K. *Star/snow-flake schema driven object-relational data warehouse design and query processing strategies*. **In:** MOHANIA M., TJOA A.M Eds. DAta WArehousing and Knowledge discovery, first international conference, DaWaK'99, 30 août au 1er septembre 1999, Florence, Italie. Heidelberg, Berlin, Allemagne: Springer-Verlag, 1999, pp 11-22. (Lecture notes in computer science, n° 1676)

[NGU03] NGUYEN T.B., TJOA A.M., MANGISENGI O. *MetaCube XTM: a multidimensional metadata approach for semantic web warehousing systems.* **In:** DAta WArehousing and Knowledge discovery, 5th international conference, DaWaK 2003, 3-5 septembre 2003, Prague, République Tchèque. Heidelberg, Berlin, Allemagne: Springer-Verlag, 2003, pp 76-88. (Lecture notes in computer science, n° 2737).

[PED99] PEDERSEN T.B., JENSEN C.S. *Multidimensional data modeling for complex data*. **In:** KITSUREGAWA M., MACIASZEK L., PAPAZOGLOU M. et al. Eds. 1999 15th International Conference on Data Engineering, ICDE-99, 23-26 mars 1999, Sydney, Australie **[en ligne]**. IEEE computer society press, 1999, pp 336-345. Disponible sur : <http://www.cs.auc.dk/research/DP/tdb/TimeCenter/TimeCenterPublications/TR-37.pdf> (consulté le 01.03.2004).

[ZHU00] ZHU Y., BORNHOVD C., SAUTNER D., et al. *Materializing Web data for OLAP and DSS*. **In:** LU H., ZHU A. Eds. Web-Age Information Management, first international conference, WAIM 2000, 21-23 juin 2000, Shanghai, Chine. Heidelberg, Berlin, Allemagne: Springer Verlag, 2000, p. 201-214. (Lecture notes in computer science, n° 2846).

5.2. Articles de périodiques

[HAN98] HAN J., NISHIO S., KAWANO H. et al. *Generalization-based data mining in object-oriented databases using an object cube model*. Data & knowledge engineering, 1998, vol. 25, n° 1-2, pp 55-97.

[JEN01] JENSEN M.R., MOLLER T.H., PEDERSEN T.B. *Specifying OLAP cubes on XML data* **[en ligne]**. Journal of intelligent information system: integrating artificial intelligence and database technologies, 2001, vol. 17, n° 2-3, pp 255-280. Disponible sur : <http://www.cs.auc.dk/~tbp/articles/R015003.pdf> (consulté le 01.03.2004).

[LIJ00] LI J.Z., GAO H. *Multidimensional data modeling for data warehouses*. Journal of software, 2000, vol. 11, n° 7, pp 908-917.

[PED01a] PEDERSEN T.B., JENSEN C.S., DYRESON C.E. *A foundation for capturing and querying complex multidimensional data* **[en ligne]**. Information systems, 2001, vol. 26, n° 5, pp 383-423. Disponible sur : <http://citeseer.nj.nec.com/cache/papers/cs/24899/http:zSzzSzwww.cs.auc.dkzSz~tbpzSzTeachingzSzDAT5E01zSztbpmodel.pdf/bachpedersen01foundation.pdf> (consulté le 17.02.2004).

[XIA01] XIAOLING W., YISHENG D. *XML based data cube and X-OLAP*. Journal of southeast university, édition anglaise, 2001, vol. 17, n° 2, pp 5-9.

Table des annexes

Annexe I: Science Direct

Les principales requêtes

		Results
5.	Title-Abstr-Key((multidimensional or OLAP or (star pre/2 schema*) or (snowflake PRE/2 schema) or cube*) AND ((complex PRE/1 data) or multimedia or image* or (web PRE/1 page*) or (web PRE/1 document*) or (web PRE/1 data) or audio or video or text*) and (warehous! or (data pre/1 mart*) or (decision pre/1 support pre/1 system*) or (knowledge pre/1 discovery pre/1 database*) or (data PRE/1 mart*))) *[All Sources(Computer Science,Decision Sciences,Engineering)]*	7
4.	(multimedia PRE/1 warehouse*) or (web PRE/1 warehouse*) or webhous* or (XML PRE/1 warehouse*) *[All Sources(Computer Science,Decision Sciences,Engineering)]*	36
3.	Title(((complex PRE/1 data) or multimedia or (web PRE/1 page*) or (web PRE/1 document*) or (web PRE/1 data) or image* or audio or video or text*) and (warehous! or (data pre/1 mart*) or (decision pre/1 support pre/1 system*) or (knowledge pre/1 discovery pre/1 database*) or (data PRE/1 mart*))) *[All Sources(Computer Science,Decision Sciences,Engineering)]*	21
2.	Title-Abstr-Key(((complex PRE/1 data) or multimedia or (web PRE/1 page*) or (web PRE/1 document*) or (web PRE/1 data)) and (warehous! or (data pre/1 mart*) or (decision pre/1 support pre/1 system*) or (knowledge pre/1 discovery pre/1 database*) or (data PRE/1 mart*))) *[All Sources(Computer Science,Decision Sciences,Engineering)]*	71
1.	Title-Abstr-Key(((((complex PRE/1 data) or multimedia or (web PRE/1 page*) or (web PRE/1 document*) or (web PRE/1 data) or text*) and ((feature pre/1 extraction) or (feature PRE/1 construction) or (knowledge PRE/1 discovery PRE/2 data))) or ((text PRE/1 mining) or (web PRE/1 mining) or (multimedia PRE/1 mining) or (image PRE/1 mining) or (audio PRE/1 mining) or (video PRE/1 mining))) and (warehous! or (data pre/1 mart*) or (decision pre/1 support pre/1 system*) or (knowledge pre/1 discovery pre/1 database*))) *[All Sources(Computer Science,Decision Sciences,Engineering)]*	8

Annexe II : Dialog

Description des bases de données utilisées [23]

Base n°2:
INSPEC

Base de données bibliographique produite par IEE

Principaux domaines couverts : physique, électronique et informatique.

Période couverte : de 1969 à actuellement

Ressources : 4 300 revues, 2 000 actes de conférences, rapports, thèses…

Langue d'interrogation : anglais

Base n°8:
EI. Compendex

Base de données bibliographique produite par Engineering Information inc.

Principaux domaines couverts : sciences de l'ingénieur, technologie

Ressources : 4 500 revues (63 % de la base), des livres, des rapports, des comptes rendus de conférences (25 %).

Langue d'interrogation : anglais

Base n°34:
SciSearch

[23] Description des bases de données sous Dialog en ligne, à l'adresse < http://library.dialog.com/bluesheets/> (consulté le 02.03.2004).

Base de données bibliographique produite par l'ISI (Institute for Scientific Information) qui sélectionne les périodiques indexés selon une technique d'analyse de citations

Principaux domaines couverts : science, technologie, bio-médecine

Ressources : 4500 prestigieuses revues scientifiques et techniques

Base n°65:
Inside
conferences

Base de données bibliographique produite par la British Library

Principaux domaines couverts : les domaines sont divers

Période couverte : de 1993 à actuellement

Ressources : rapports de 16000 conférences reçues au British Library Document Supply Centre

Base n°95:
TEME

Base de données bibliographique Allemande

Principaux domaines couverts : technologie et gestion

Langue d'interrogation : anglais ou allemand

Base n°144:
PASCAL

Base de données bibliographique produite par l'INIST

Principaux domaines couverts : science, technologie, médecine.

Période couverte : de 1973 à actuellement

Ressources : articles de périodiques (93 %), thèses françaises, rapports, comptes rendus de congrès...

Base n°674:
Computer News
Fulltext

Base de données plein-texte

Principaux domaines couverts : informatique

Ressources : 120 périodiques relatifs à l'informatique produits par IDG communications

Les principales requêtes

Requête n° 1

S1	COMPLEX(W)DATA OR MULTIMEDIA OR WEB(W)PAGE? ? OR WEB(W)DOCUMENT? ? OR WEB(W)DATA	127940	Display
S3	WAREHOUS? OR DECISION(W)SUPPORT(W)SYSTEM? ? OR DATA(W)MART? ? OR KNOWLEDGE(W)DISCOVERY(W)DATABASE? ?	70246	Display
S6	S1/TI	45532	Display
S7	S3/TI	16378	Display
S8	S6 AND S7	93	Display
S9	RD S8 (unique items)	66	Display

Requête n° 2

Set	Term Searched	Items	
S1	ETL	2498	Display
S2	WAREHOUS? OR DECISION(W)SUPPORT(W)SYSTEM? ? OR KNOWLEDGE(W)DISCOVERY(W)DATABASE? ? OR DATA(W)MART? ?	73912	Display
S3	COMPLEX(W)DATA OR MULTIMEDIA OR WEB(W)DATA OR WEB(W)DOCUMENT? ? OR WEB(W)PAGE? ? OR IMAGE? ? OR TEXT? ? OR AUDIO OR VIDEO	1801411	Display
S4	S1 AND S2 AND S3	6	Display
S5	RD S4 (unique items)	5	Display
S6	MODELI? OR INTEGRAT? OR DESCIPT? OR TRANSFORM?	3643434	Display
S7	S6/TI AND S3/TI AND S2	38	Display
S8	RD S7 (unique items)	28	Display

Requête n° 3

Set	Term Searched	Items	
S1	FEATURE(W)EXTRACTION OR FEATURE(W)CONSTRUCTION OR KNOWLEDGE(W)DISCOVERY(1W)DATA	54125	Display
S2	WAREHOUS? OR DECISION(W)SUPPORT(W)SYSTEM? ? OR KNOWLEDGE(W)DISCOVERY(W)DATABASE? ? OR DATA(W)MART? ?	73912	Display
S3	WEB(W)PAGE? ? OR WEB(W)DOCUMENT? ? OR WEB(W)DATA OR COMPLEX(W)DATA OR MULTIMEDIA OR TEXT? ? OR AUDIO OR VIDEO	676449	Display
S4	WEB(W)MINING OR TEXT(W)MINING OR MULTIMEDIA(W)MINING OR IMAGE(W)MINING OR AUDIO(W)MINING OR VIDEO(W)MINING	1300	Display
S11	((S1 AND S3) OR S4) AND S2	106	Display

S12	RD S11 (unique items)	79	Display

Requête n° 4

Set	Term Searched	Items	
S1	COMPLEX(W)DATA OR MULTIMEDIA OR WEB(W)PAGE? ? OR WEB(W)DOCUMENT? ? OR WEB(W)DATA	135591	Display
S2	MULTIDIMENSIONAL OR OLAP OR STAR(1W)SCHEMA? ? OR SNOWFLAKE(1W)SCHEMA OR CUBE? ?	108949	Display
S3	WAREHOUS? OR DECISION(W)SUPPORT(W)SYSTEM? ? OR DATA(W)MART? ? OR KNOWLEDGE(W)DISCOVERY(W)DATABASE? ?	73917	Display
S4	S1 AND S2 AND S3	101	Display
S5	RD S4 (unique items)	85	Display
S6	S4 AND PY>1999	49	Display

Requête n° 5

Set	Term Searched	Items	
S1	WEBHOUS? OR WEB(W)WAREHOUS? OR XML(W)WAREHOUS? OR IMAGE(W)WAREHOUS? OR MULTIMEDIA(W)WAREHOUS? OR TEXT(W)WAREHOUS? OR IMAGE(W)WAREHOUS? OR AUDIO(W)WAREHOUS? OR VIDEO(W)WAREHOUS?	153	Display
S2	RD S1 (unique items)	84	Display
S3	S2/TI	36	Display