Equipe de Recherche en Ingénierie des Connaissances

Bilan – Projet

2002-2005

Université Lumière Lyon 2
5, avenue Pierre Mendès-France
69676 Bron, 69676
Tél. : (33) [0]4 78 77 44 92
Fax. : (33) [0]4 78 77 23 75
e-mail. nicolas.nicoloyannis@univ-lyon2.fr
Web http://eric.univ-lyon2.fr

1. Note de Synthèse

L'Equipe de Recherche en Ingénierie des Connaissances (ERIC) existe depuis 1995, d'abord comme Jeune Equipe (1995-1999), puis comme Equipe d'Accueil (depuis 1999). ERIC est actuellement composée de quatre professeurs (3 CNU 27 et 1 CNU 26), sept maîtres de conférences et d'une secrétaire à mi-temps. Le laboratoire accueille une vingtaine de doctorants et trois Attachés Temporaires d'Enseignement et de Recherche (ATER). Le laboratoire ERIC est localisé sur le campus Porte des Alpes à Bron et partage les locaux du Département d'Informatique et de Statistique de la Faculté de Sciences Economiques et de Gestion.

La recomposition des thématiques de recherche en informatique sur Lyon a conduit au départ du pôle image d'ERIC (un Professeur et deux Maîtres de Conférences) vers l'UMR CNRS LIRIS dont le rattachement principal est l'Université Lyon 1. Cette recomposition nous a amené à centrer nos recherches sur l'ECD en étendant le contexte des données tabulaires vers celui des données complexes.

Cette évolution s'est traduite par un élargissement de nos réflexions :

- vers les problèmes qui se situent sur l'ensemble du processus d'ECD : entreposage des données, représentation des données, construction d'attributs, séparabilité des classes, échantillonnage optimal, validation des modèles, … ;
- vers de nouveaux supports de données : les images, le texte et de manière générale les données complexes, distribuées et généralement non ou peu structurées, sont particulièrement abondantes. Développer une expertise pour extraire des connaissances à partir de ces données, dans leur forme naturelle, devient un enjeu stratégique.

Quantitativement, le bilan d'ERIC, pour les quatre dernières années universitaires, est :

- dix chercheurs ont pu effectuer ou achever leur doctorat au sein d'ERIC ;
- quatre collègues ont réalisé leur Habilitation à Diriger les Recherches ;
- ERIC continue d'attirer de jeunes doctorants titulaires d'allocations de recherche ministérielles et de bourses de l'industrie, avec près de vingt thèses en cours.

La qualité et la diversité des publications d'ERIC montre que l'équipe est active et présente à tous les niveaux : publications dans des revues internationales (18) ou nationales (13), communications dans des conférences internationales (79) ou nationales (59), chapitres (8) ou direction d'ouvrages (5), diffusion de logiciels, organisations de conférences majeures, contacts avec des universités étrangères … L'expertise développée par les chercheurs d'ERIC est reconnue comme en témoignent les nombreux contrats qui représentent près de 70% de ses ressources.

ERIC tient à maintenir un lien fort et explicite avec l'enseignement notamment à travers l'animation d'un Master d'Informatique.

Pour le futur, nous souhaitons :

- maintenir autant que possible une unité de lieu enseignement-recherche ;
- renforcer les thématiques sur lesquelles ERIC est reconnu, à savoir l'Extraction des Connaissances à partir des Données (ECD) ;
- maintenir une activité de recherche à trois niveaux : travaux à caractère théoriques, développement d'outils informatiques associés, recherche de terrains d'application en particulier dans les domaines des Sciences Humaines et Sociales et des Sciences Economiques et de Gestion ;
- renforcer la politique éditoriale et d'animation scientifique de l'équipe et notamment accroître le nombre de publications dans des journaux internationaux spécialisés ;
- développer nos relations avec la communauté scientifique locale, nationale et internationale autour à la fois d'une activité de recherche mais aussi d'enseignement telles que les co-tutelles de thèses, les doubles diplômes ou le master européen ;
- valoriser nos recherches sur le plan industriel par l'incubation de projets préindustriels.

2. Bilan Scientifique et Quantitatif

2.1. Place de ERIC à Lyon 2

Il n'est plus à démontrer que la modélisation informatique, et de manière générale l'informatique, est devenue plus qu'un outil mais un cadre méthodologique pour traiter les problèmes qui se posent au psychologue quand il cherche à comprendre les mécanismes cognitifs, au sociologue quand il analyse le comportement d'un groupe social, au géographe quand il souhaite reconstruire des reliefs en image de synthèse ou à l'archéologue quand il veut identifier et dater des vestiges du passé. Cette prise de conscience est quasi générale en recherche comme en enseignement et notamment dans les sciences humaines et sociales.

La création à l'Université Lyon 2 de filières d'enseignement à l'intersection des Sciences Humaines et Sociales, de Sciences Economiques et de Gestion, de l'Informatique et des Mathématiques appliquées, et plus récemment la mise en place d'un Master d'Informatique[1] avec trois spécialités professionnelles[2] et une spécialité recherche[3], sont une parfaite illustration de cette forte implication des collègues informaticiens et mathématiciens dans l'environnement spécifique de notre université.

[1] Le Master d'Informatique est commun aux universités Lyon 1 et Lyon 2, à l'Ecole Normale Supérieure (ENS Lyon), à l'Ecole Centrale de Lyon (ECL) et à l'Institut National des Sciences Appliquées (INSA – Lyon).

[2] Ingénierie Informatique pour la Décision et l'Evaluation Economiques (IIDEE) ; Statistique et Informatique Scio-Economique (SISE) et Organisation et Protection des Systèmes d'Information dans les Entreprises (OPSIE).

[3] Extraction des Connaissances à partir des Données (ECD).

2.2. POSITIONNEMENT SCIENTIFIQUE DE ERIC

Les travaux du laboratoire ERIC se situent dans le domaine de l'extraction des connaissances à partir des données (ECD). Par le terme ECD (*Knowledge Discovery in Databases - KDD*), on désigne tout le cycle de découverte de connaissances. Il regroupe donc la conception et les accès à de grandes bases de données ainsi que tous les traitements à effectuer pour extraire de la connaissance de ces données.

Les thématiques et activités du laboratoire ERIC se rattachent à un ou plusieurs aspects du processus d'ECD. Nous nous intéressons en particulier à la fouille de données (*data mining*) et aux entrepôts de données (*data warehouse*).

Le processus d'extraction de connaissances dans une application décisionnelle nécessite au préalable la constitution d'un entrepôt de données et son alimentation par un processus d'intégration depuis des sources souvent hétérogènes, l'analyse en ligne OLAP (*On-Line Analytical Processing*) et la transformation en vue de l'application d'une méthode de fouille de données. L'existence d'un entrepôt simplifie la tâche de fouille et permet donc d'optimiser le temps de développement d'un projet d'extraction de connaissances. En effet, il est beaucoup plus simple de trouver une information pertinente dans une structure organisée pour la recherche de connaissances.

L'émergence des entrepôts de données dans les années 90 a eu une répercussion importante aussi bien dans le monde industriel que dans la communauté de la recherche scientifique. Le processus décisionnel apporte une réponse au problème de la croissance continuelle des données et il supporte efficacement les processus d'analyse en ligne et de fouille de données. Que doit-on faire avec des données coûteuses à collecter et à conserver et dont on sait pertinemment qu'elles renferment des connaissances utiles ?

La fouille de données est également apparue au début des années 90, même si ses fondements méthodologiques sont antérieurs. Cette émergence n'est pas le fruit du hasard, mais le résultat de la combinaison de nombreux facteurs à la fois scientifiques, technologiques, économiques et même socio-politiques. On peut voir la fouille de données comme une nécessité imposée par le besoin de valoriser les données que toutes les organisations accumulent dans leurs bases.

La fouille de données est l'art d'extraire des connaissances à partir des données. Si elle comprend l'application d'algorithmes d'apprentissage en vue de construire des modèles, elle est en fait au centre d'un processus complet intégrant le pré-traitement des données (réduction

de la dimensionnalité, sélection des variables et des individus, nettoyage) et le post-traitement des connaissances (validation statistique, mise en forme des connaissances).

Les données peuvent être stockées dans des entrepôts, dans des bases de données distribuées ou sur Internet. La fouille de données ne se limite pas au traitement des données structurées sous forme de tables numériques ; elle offre des moyens pour aborder les corpus en langage naturel (*text mining*), les images (*image mining*), le son (*sound mining*) ou la vidéo.

Contrairement à d'autres processus fondés uniquement sur des logiciels, un processus décisionnel est un projet qui se construit. Il doit s'insérer dans un cadre pouvant contenir des données, des informations et des connaissances. L'entreposage de données propose une démarche d'intégration, d'organisation et de stockage de données en vue d'une analyse OLAP et la fouille de données propose une démarche et des outils pour répondre au besoin de valorisation. L'ECD a été employée avec beaucoup de succès dans de grands secteurs d'application : la gestion de la relation client, l'aide à la décision médicale, la recherche d'informations... Aucun domaine d'application n'est *a priori* exclu car dès que nous sommes en présence de données réelles, la fouille de données peut rendre de nombreux services.

L'ECD est un domaine à la fois scientifique et technologique récent qui a encore de nombreux défis à relever. La communauté des chercheurs dans ce domaine s'intéresse ainsi à des problèmes tels que la recherche de bons espaces de représentation ou l'agrégation de prédicteurs, etc. Grâce à Internet, une grande quantité de sites regroupant des logiciels, des données, des expertises, des cours, des communautés d'échanges et de la bibliographie sont à présent accessibles.

Nos travaux en ECD visent à intégrer de nouvelles dimensions pour mieux répondre à des problèmes réels. Cela passe par des évolutions et des adaptations des méthodes d'ECD pour qu'elles prennent mieux en compte :

- les données complexes : textes en langage naturel, images, données symboliques, numériques quantitatives ou qualitatives ;
- les données volumineuses qui peuvent se situer sur des supports structurés comme les entrepôts de données ou non structurés comme le Web ;
- les sources et les domaines d'application : médecine, archéologie, économie, chimie, enquêtes de satisfaction client, etc.

Les problèmes auxquels nous sommes confrontés en ECD couvrent un large spectre qui s'étend de l'organisation et du stockage de données dans les entrepôts de données à l'accès à ce large volume d'informations, en passant par les problèmes de représentation des données, de construction et de sélection d'attributs, d'échantillonnage, jusqu'aux méthodes de fouille de données, de validation des résultats et de mise en forme des connaissances produites.

Depuis le 1er janvier 2003, ERIC s'est réorganisé autour de deux axes de recherche principaux au sein de l'ECD :

- **Entreposage de données complexes**. Le premier axe s'intéresse aux problèmes liés à l'entreposage de données complexes dans le processus d'extraction de connaissances. En effet, les données disponibles sont non seulement volumineuses, mais également hétérogènes à plusieurs titres : natures et formats différents (données numériques et symboliques, textes, données multimédia…), représentations différentes (données similaires exprimées dans des langues différentes ou sous des formes différentes, comme des transcriptions textuelles de films), sources diverses (données de production, Web…), qui peuvent de surcroît être ou non structurées. Dans un tel contexte, il est important d'intégrer et de préparer ces données complexes avant de pouvoir leur appliquer des outils de fouille de données. Les travaux développés dans ce thème relèvent de l'intégration des données complexes dans un entrepôt de données et de leur analyse à travers un couplage OLAP et fouille de données. Les modèles de base de données relationnelles ou objets supportent difficilement les données semi-structurées et multimédia. Pour ces bases de données, la définition préalable d'un schéma, comme l'imposent les modèles relationnels et objets, n'est pas toujours possible. La force des modèles semi-structurés tels que XML est de ne plus imposer de structure a priori mais de la définir a posteriori de manière à maximiser les capacités d'intégration et de représentation.
- **Développement de nouvelles méthodes de fouille**. Le deuxième axe traite des méthodes et des algorithmes de fouille de données. Il s'agit de développer de nouvelles méthodes et de nouveaux outils destinés :
    - à l'apprentissage automatique,
    - aux stratégies de fouille de données en présence de données très volumineuses, plusieurs milliers d'attributs et plusieurs millions de n-uplets,
    - à évaluer la qualité des données traitées et des modèles produits, notamment ceux issus des algorithmes d'apprentissage et destinés à la prédiction, etc.

2.3. PUBLICATIONS

Le tableau ci-dessous résume le bilan scientifique quantitatif sur la période 2002-2005.

		2002	2003[1]	2004	2005	Total
Publications	Ouvrages	2	1	0	2	**5**
	Chapitres dans ouvrages	3	1	1	3	**8**
	Revues internationales	4	2	7	5	**18**
	Revues nationales	2	4	4	3	**13**
	Conférences internationales	23	21	14	21	**79**
	Conférences nationales	11	15	17	16	**59**
	Total	**45**	**44**	**43**	**50**	**182**
Effectifs	Chercheurs statutaires	14	11	11	11	
Thèses soutenues		3	2	3	0	**8**
HDR soutenues		1	0	1	0	**2**
Formation par la recherche	Effectifs DEA ECD	29	33	31	30	**123**

Tableau 1 : Bilan scientifique 2002-2005

2.4. RESSOURCES HUMAINES

2.4.1. Au 1er octobre 2005

[1] Au 1er Janvier 2003, départ du pôle Image

Enseignants-chercheurs statutaires

Nom, Prénom, Date de Naissance	Corps grade	Section CNU	Date d'arrivée
Bentayeb Fadila, 15 mai 1966	MCF	27	10/2001
Bousaïd Omar, 2 juin 1954	MCF	27	01/1995
Chauchat Jean-Hugues, 6 juillet 1946	PR2	27	06/1997
Darmont Jérôme, 15 janvier 1972	MCF	27	10/1999
Harbi Nouria, 27 août 1961	MCF	06	10/2005
Lallich Stéphane, 20 septembre 1947	PR2	27	06/1997
Loudcher Rabaséda Sabine, 27 octobre 1969	MCF	27	10/1998
Nicoloyannis Nicolas, 23 juin 1951	PR2	26	10/1999
Rakotomalala Ricco, 19 juillet 1967	MCF	27	10/1998
Viallefont Anne, 8 mars 1969	MCF	26	10/2000
Viallaneix Jacques, 6 juillet 1963	MCF	27	01/1995
Zighed Abdelkader, 12 mars 1955	PR1	27	01/1995

ATER rattachés au laboratoire

Nom, Prénom	Année universitaire
Effantin Dit Toussaint Brice	2004-2005
Legrand Gaëlle	2004-2005
Walid Erray	2003-2004 2004-2005

Post Doc rattaché au laboratoire

Nom, Prénom	Année universitaire
Jouve Pierre	2004-2005

HDR en cours

Nom, Prénom, Date de Naissance	Corps grade	Section CNU
Bousaïd Omar, 2 juin 1954	MCF	27
Darmont Jérôme, 15 janvier 1972	MCF	27

Thèses en cours

Nom, Prénom	Début	Directeur	Co directeur	Financement
Aouiche Kamel	2002	D. Zighed	J. Darmont	ACI
Ben Messaoud Riadh	2003	N. Nicoloyannis	O. Boussaid S. Loudcher	Bourse Fondation Vediorbis
Charbel Julien	2004	D. Zighed		Ressources propres
Clerc Frédéric	2003	N. Nicoloyannis	R. Rakotomalala	BDI CNRS
El Sayed Ahmad	2004	D. Zighed	F. Bentayeb	Ressources propres
Erray Walid	2001	D. Zighed		Contrat France Télécom
Fangseu Badjio Edwige	2002	D. Zighed	F. Poulet	Bourse ESIEA Laval
Favre Cécile	2003	N. Nicoloyannis	F. Bentayeb	Bourse CIFRE
Gaudin Rémi	2004	N. Nicoloyannis		Bourse MNERT Moniteur
Hacid Hakim	2004	D. Zighed		Bourse Région
Hupertan Vincent	2001	JH. Chauchat		Médecin des hôpitaux
Lazaar Naïma	2004	D. Zighed		Ressources propres
Mahboubi Hadj	2005	N. Nicoloyannis	J. Darmont	Ressources propres
Maïz Nora	2005	N. Nicoloyannis	F. Bentayeb	Ressources propres
Marcellin Simon	2004	D. Zighed		Bourse CIFRE
Mavrikas Efthimios	2002	N. Nicoloyannis S. Dascalopoulos (Co-tutelle avec l'Université de l'Egée)		Bourse de la Grèce
Prudhomme Elie	2005	S. Lallich		Bourse MNERT
Ralaivao Jean-Christian	2003	S. Lallich V. Manantsoa (Co-tutelle avec l'Université de Fianarantsao)	J. Darmont	Bourse de l'Ambassade de France

Stavrianou Anna	2005	N. Nicoloyannis		Bourse MNERT
Thomas Julien	2005	N. Nicoloyannis		Bourse CIFRE

Thèses soutenues

Nom, Prénom	Année	Directeur	Co directeur	Devenir
Baume Laurent	2004	N. Nicoloyannis	C. Mirodatos	Post Doc en Espagne
Clech Jérémy	2004	D. Zighed		Privé
Coeurjolly David	2002	S. Miguet		CR CNRS
Gavin Gérald	2001	D. Zighed		MCF Lyon 1
Jalam Radwan	2003	JH. Chauchat		MCF contractuel à l'ENSAR
Jouve Pierre	2003	N. Nicoloyannis		Post Doc
Legrand Gaëlle	2004	N. Nicoloyannis		ATER
Muhlenbach Fabrice	2002	D. Zighed	S. Lallich	MCF St Etienne
Scuturici Marian	2001	S. Miguet		MCF INSA
Scuturici Mihaela	2002	JM. Pinon	S. Miguet	ATER

HDR soutenues

Nom, Prénom	Année	Directeur	Devenir
Belkhiter Nadir	2001	D. Zighed	PR à Laval (Canada)
Chauchat Jean-Hugues	2001	D. Zighed	PR à Lyon 2
Lallich Stéphane	2002	D. Zighed	PR à Lyon 2
Poulet François	2004	D. Zighed	

Personnels administratifs

Nom, Prénom	Corps grade	Quotité recherche	Date d'arrivée
Gabrièle Valérie	IATOS	0,5	09/2000

Récapitulatif au 1er Octobre 2005

Catégorie	Effectif
Enseigants-chercheurs statutaires	12
ATER	3
Post Doc	1
Thèses en cours	20
HDR en cours	2
Thèses soutenues	10
HDR soutenues	4
Personnel administratif	1

2.4.2. Ayant terminé leur contrat ou quitté le laboratoire

Enseignants-chercheurs statutaires

Nom, Prénom	Corps grade	Section CNU	Date d'arrivée	Date de départ
Miguet Serge	PR1	27	09/1996	01/2003
Sarrut David	MCF	27	09/2001	01/2003
Tougne Laure	MCF	27	09/1998	01/2003

ATER

Nom, Prénom	Année universitaire
Clech Jérémy	2003-2004
El Dajani Rajai Mourid	2000-2001
Favetta Franck	2001-2002
Jalam Radwan	2000-2001 2001-2002
Muhlenbach Fabrice	2002-2003
Scuturici Marian	2003-2004
Scuturici Michaela	2002-2003 2003-2004
Sidhom Sahbi	2001-2002
Tweed Tiffany	2002-2003 2003-2004
Zellouf Yamina	2001-2002 2002-2003

Personnels admnistratifs

Nom, Prénom	Financement	Quotité recherche	Date d'arrivée	Date de départ
Varaine Astrid	Fonds propres	1	01/1998	04/2002
Delhomme Lydie	Fonds propres	1	10/2002	08/2004

3. Travaux Scientifiques

La production scientifique des chercheurs du laboratoire est riche et diversifiée, et par conséquent difficile à synthétiser en quelques pages. Nous avons préféré présenter succinctement quelques-uns des travaux que nous pensons être particulièrement significatifs en matière de contribution au domaine de l'ECD sur chacun des deux axes de recherche d'ERIC. Ces travaux s'articulent autour du traitement des données avec des préoccupations communes telles que la volumétrie, l'hétérogénéité des sources et des supports, la performance.

3.1. Contributions aux méthodologies de fouille de données

Nous regroupons ces contributions selon plusieurs thèmes.

Apprentissage automatique. Les graphes d'induction et plus particulièrement les arbres de décision continuent d'intéresser très fortement la communauté de l'apprentissage automatique à cause de leur simplicité de mise en œuvre et d'interprétation. Jusque là, les méthodes d'arbres proposées pour des variables à expliquer catégorielles sont ce que nous appelons les arbres de classification et quand la variable à expliquer est continue, on fait appel aux arbres de régression. Dans chacun des cas, le critère à optimiser est différent. Pour les arbres catégoriels, il s'agit généralement d'une mesure de gain informationnel et dans le second cas c'est une mesure d'inertie généralement basée sur la décomposition de la variance. Même si des liens sémantiques et même fonctionnels existent entre les notions de variance et d'entropie qui sont à la base des mesures de gain informationnel, dans la mise en œuvre pratique, il s'agit de méthodes différentes. Nous avons proposé une approche unifiée de ces méthodes quelle que soit la nature des variables à expliquer ou explicatives. Les règles issues de telles structures peuvent être à conclusions multiples et dans ce cas, il a fallu redéfinir les mesures d'estimation de l'erreur en apprentissage et sur l'échantillon test [AZRES05].

En matière d'apprentissage automatique non supervisé, nous avons développé des travaux originaux relatifs à la sélection de variables, à la mise au point des méthodes de classification non

supervisée pour données catégorielles, à l'évaluation de la qualité/validité des résultats d'un processus de classification non supervisée. Comme dénominateur commun à l'ensemble de ces travaux, nous proposons des concepts de comparaisons par paires, d'agrégation des préférences basés sur le critère de Condorcet [CJN03], [CJN03b], [CJN03d].

Mesure de séparabilité des classes. L'utilisation des graphes de proximité en fouille de données et plus particulièrement en apprentissage est l'un de nos principaux thèmes de recherche en apprentissage. Dans nos travaux antérieurs, nous avons introduit le concept de « réseaux de cooptation ». Celui-ci est formé de l'ensemble des points de l'échantillon d'apprentissage reliés entre eux par une propriété de voisinage. Parmi les modèles de voisinage que nous utilisons, on peut citer les graphes de Delaunay, les graphes de Gabriel, les graphes des voisins relatifs et l'arbre de recouvrement minimal. Tous ces modèles permettent d'engendrer un graphe connexe. Ce cadre nous a permis de déboucher sur de nombreux résultats significatifs et en particulier sur la construction d'une mesure de séparabilité des classes. Dans ce contexte nous avons mis au point un test statistique non paramétrique de séparabilité des classes [CZLM02, DZLM02b, AZLM05]. Le principe, consiste à étudier la statistique du nombre d'arêtes qu'il faut supprimer du graphe de voisinage afin d'obtenir des composantes connexes formées par des points qui appartiennent à la même classe. Ce test dépasse les limites inhérentes aux tests paramétriques telles que le lambda de Wilks qui repose sur des hypothèses rarement vérifiées (distribution multi-normale, égalité des paramètres d'échelle…). Des résultats établis dans le domaine de l'analyse statistique spatiale nous ont permis d'unifier nos travaux à ceux de l'analyse statistique spatiale et surtout d'ouvrir de nouvelles pistes comme, par exemple, la détection des individus atypiques [GL02, GM02, AMLZ04, CMLZ04b].

Réduction de la dimensionnalité. L'augmentation exponentielle du volume des données traitées en ECD rend nécessaire la réduction de la dimension des données, tant en termes d'individus que de variables. Lorsque la base de données comporte un très grand nombre d'individus, la simple lecture de la base est une opération très longue. Pour cette situation, nous avons conçu une méthode de sélection de variables qui ne nécessite qu'un seul parcours de la base de données [BRL02]. Notre méthode repose sur trois principes : (1) nous raisonnons sur les indicatrices associées aux paires d'individus, (2) nous utilisons le coefficient de corrélation linéaire entre ces indicatrices puisque les coefficients partiels se calculent à partir des coefficients simples, (3) la prévision se fait par un processus bayésien naïf. Il suffit alors de parcourir une fois la base pour former tous les tableaux croisés entre les descripteurs catégoriels et en déduire les

coefficients de corrélation entre variables indicatrices. Pour réduire le nombre d'individus, à côté des travaux antérieurs du laboratoire sur l'intégration des techniques d'échantillonnage dans les méthodes de fouille de données, nous avons proposé une méthode originale de construction d'un ensemble de prototypes qui repose sur le boosting. Chaque individu est considéré comme un classifieur de ses plus proches voisins. Le boosting permet de sélectionner les individus qui se complètent le mieux. Un critère original est utilisé pour contrôler le nombre d'itérations du boosting qui fixe la taille de l'ensemble de prototypes [ASNL02].

Validation. Les travaux concernant la validation des connaissances extraites ont porté aussi bien sur l'estimation de l'erreur de prédiction en général que sur des algorithmes d'extraction de connaissance particuliers. L'estimation de l'erreur de prédiction à l'aide de méthodes de ré-échantillonnage telles que la validation croisée est une technique bien connue pour évaluer les performances d'un modèle de prédiction. Cette évaluation repose sur un présupposé systématiquement passé sous silence : l'échantillon doit être constitué à partir d'un tirage au hasard dans la population de référence. Très souvent, cette hypothèse n'est pas respectée dans des applications réelles, les observations sont naturellement regroupées au sein d'entités plus larges. Nous avons pu montrer, dans le cadre d'une coopération avec le laboratoire Dynamique du Langage de l'Université Lyon 2, qu'il s'agit de classer des phrases selon la langue utilisée, un locuteur pouvant prononcer plusieurs phrases. Nous avons démontré que ne pas en tenir compte dans le mode opératoire de la validation croisée conduit à des résultats biaisés [CPCR02]. Une attention particulière a été accordée à la validation des règles d'associations extraites par des algorithmes du type Apriori. Les principales mesures de l'intérêt des règles d'association se référant à l'indépendance de l'antécédent A et du conséquent B de la règle, nous avons proposé une réécriture de ces mesures qui les fait dépendre d'un paramètre à la disposition de l'utilisateur [CLVL05]. Suivant la valeur choisie pour ce paramètre, on s'intéresse aussi bien aux règles qui s'écartent significativement de l'indépendance entre A et B, qu'aux règles de ciblage qui assurent par exemple qu'un événement B arrive 2 fois plus souvent lorsqu'un événement A est réalisé. Nous proposons alors des procédures de filtrage qui contrôlent le nombre de règles découvertes à tort [DLPT04].

3.2. Contributions à l'entreposage de données

Dans le cadre de l'extraction des connaissances à partir de données complexes, nous proposons une approche basée sur un processus complet d'entreposage et d'analyse des données complexes permettant de concevoir des systèmes décisionnels. La problématique d'intégration, de modélisation et d'analyse de données complexes nécessite une méthodologie et des outils génériques adaptés [BBBDR03]. Cette approche est complétée par une évaluation des performances des outils proposés.

Un entrepôt de données présente une modélisation dite « dimensionnelle » qui se compose classiquement d'une table de faits centrale et d'un ensemble de tables de dimension. Cette modélisation conceptuelle a pour objectif d'observer les faits, à travers des mesures (indicateurs), en fonction des dimensions qui représentent les axes d'analyse. Ce modèle est qualifié de modèle en étoile.

Intégration et modélisation des données complexes. L'intégration et la modélisation consistent à incorporer physiquement des données complexes dans une base de données jouant le rôle d'un sas à un entrepôt de données et à les préparer en vue de les analyser. Nous définissons dans la phase d'intégration des modèles conceptuel, logique, puis physique. Nous avons retenu XML comme formalisme pour décrire les modèles logiques et physiques, car il permet de stocker les données ainsi que leur description. Cette représentation se retrouve dans les entrepôts, qui stockent à la fois des données et des métadonnées. Le modèle conceptuel exprimé sous la forme d'un diagramme de classes UML permet de définir des objets complexes composés d'un ou plusieurs sous-documents. Le modèle conceptuel est traduit en un modèle logique sous la forme d'une grammaire XML exprimée à l'aide d'une DTD ou d'un schéma XML. Le modèle logique ainsi obtenu est instancié en modèle physique constitué de documents XML. Les documents XML générés sont valides et peuvent finalement être stockés soit dans une base de données native XML, soit dans une base de données relationnelle via un processus de mapping. Cette base de documents XML constitue un véritable ODS (*Operational Data Storage*) contenant des données de production avant d'alimenter un entrepôt de données [BDBBRZ03, CBBD03b].

Dans notre démarche, nous recommandons une couche supplémentaire de modélisation multidimensionnelle des données complexes pour les préparer à l'analyse [CTBB04]. Les modèles en étoile étant les mieux adaptés, nous construisons un référentiel réunissant l'ensemble des données issues de la phase d'intégration. Il est complété par des informations sur les données indiquant leur origine, leur nature et le rôle qu'elles peuvent jouer. Les données complexes sont

décrites à la fois par des attributs de bas niveau et par des descripteurs sémantiques. Ces derniers peuvent être obtenus par diverses techniques de fouille de données, de statistique, de traitement d'images ou du signal. Une exploration des données du référentiel par une technique de fouille de données peut contribuer à l'identification des faits pertinents à analyser et permettre d'enrichir le référentiel par de nouveaux descripteurs sémantiques. Selon les objectifs d'analyse de l'utilisateur, les faits à observer sont identifiés et exprimés à l'aide de mesures et d'axes d'analyse (dimensions). Le cube de données construit correspond à une vue des données complexes et représente ainsi un espace d'analyse exploité par une analyse en ligne ou par des techniques de fouilles de données [CTBB05].

Analyse et fouille en ligne de données complexes. L'étude des interactions possibles entre le domaine des entrepôts de données et de l'analyse en ligne (OLAP) et le domaine de la fouille de données est nécessaire pour pallier les problèmes liés à la gestion et à l'analyse de ces gros volumes de données complexes. Nous étudions deux volets de recherche pour accomplir cette interaction. Le premier volet concerne le couplage entre OLAP et la fouille de données. Le deuxième volet étudie l'intégration des méthodes de fouille dans les systèmes de gestion de bases de données (SGBD).

Les opérateurs actuels d'analyse en ligne ne sont pas adaptés aux données complexes : on ne peut réaliser des opérations de somme ou de moyenne sur des textes ou des images. La complexité des données nécessite des opérateurs spécifiques d'agrégation, de navigation ou même d'extraction de connaissance. Nous pensons qu'un couplage entre l'analyse en ligne et la fouille de données est une solution à ce problème. Dans ce cadre, nous proposons un nouvel opérateur, baptisé OpAC (Opérateur d'Agrégation par Classification), d'analyse en ligne des données multidimensionnelles. OpAC consiste en l'agrégation des faits d'un cube de données, basée sur la classification hiérarchique ascendante (CAH) [CBBR04]. Il construit des classes de faits homogènes correspondant à des agrégats OLAP. L'extension des opérateurs OLAP est donc faisable à l'aide de combinaisons avec d'autres techniques de fouille de données. D'autre part, pour détecter les faits intéressants dans un cube, nous utilisons une technique d'analyse factorielle (Analyse des Correspondances Multiples). Cette approche permet de réarranger les faits dans le cube et de réduire son éparsité [CBBR05].

Le deuxième volet de cet axe de recherche vise à offrir à l'utilisateur des opérateurs de fouille en ligne sachant qu'OLAP constitue une première étape en matière d'intégration de processus d'analyse au sein des SGBD. Les algorithmes de fouille de données traditionnels fonctionnent sur

des tableaux attributs/valeurs chargés en mémoire centrale. De ce fait, ils se heurtent au problème de la limitation de la taille de la mémoire et par conséquent des bases à traiter. Dans ce contexte, nous proposons une nouvelle approche de fouille de grandes bases de données intégrée aux SGBD et ce avec des temps de réponse raisonnables. Elle consiste à réécrire et à implémenter les algorithmes de fouille de données (en particulier les méthodes de construction de graphes d'induction) en utilisant exclusivement les fonctionnalités et les outils du SGBD, comme le langage SQL et les vues relationnelles. Comme les temps de traitements restent élevés en raison des nombreuses lectures de la base, nous proposons de recourir à des résultats intermédiaires pré-calculés comme les tables de contingence ou encore d'utiliser les index bitmap [CFB05] pour optimiser les performances. Nos tests montrent que les temps de traitement de notre approche augmentent de façon linéaire avec la taille de la base. Mais contrairement aux méthodes de fouille opérant en mémoire, nous ne sommes pas limités par la taille des bases à traiter [CBDU04].

Performance. Les entrepôts de données classiques permettent d'analyser des activités représentées sous la forme de données numériques. Cependant, les données exploitées dans le cadre des processus décisionnels sont de plus en plus complexes, notamment depuis l'avènement du Web. Dans ce cadre, il est devenu primordial de réduire la fonction d'administration de base de données ou du moins de fournir des outils d'aide pour l'administrateur.

Une collaboration avec le groupe de recherche en bases de données OUDB de l'Université d'Oklahoma, initiée en 2001, nous a permis de proposer des solutions dans ce domaine. Le principe de nos propositions est d'exploiter une charge (ensemble de transactions) donnée à l'aide de techniques de fouille de données (recherche de motifs fréquents ou classification) pour générer automatiquement des structures physiques (index et vues matérialisées) permettant d'optimiser le temps d'accès aux données [AADG03, CADBB05, BADBB05b].

Nous avons pour objectif à court terme d'appliquer ces techniques d'auto-administration au sein de deux entrepôts de données complexes que nous avons développés, l'un dans le cadre d'une collaboration avec des collègues linguistes (CLAPI), l'autre dans le domaine de la médecine du sport (MAP). Ces deux entrepôts renferment des données complexes (qualitatives, numériques, textes, images...) et sont modélisés sous forme relationnelle.

Par ailleurs, afin d'évaluer la pertinence des solutions que nous proposons pour l'intégration, la modélisation et l'analyse des données complexes, il est nécessaire d'évaluer la performance des entrepôts de données mis en œuvre. Ces expérimentations sont habituellement menées à l'aide de bancs d'essais (*benchmarks*, dans la terminologie anglo-saxonne). Ce type d'outils permet à la fois aux utilisateurs de comparer les performances de différents systèmes et aux concepteurs de tester

les effets de divers choix techniques. Bien que les bancs d'essais décisionnels standards du TPC (*Transaction Performance Processing Council*) répondent au premier de ces usages, ils ne sont pas suffisamment paramétrables pour répondre au second et ne permettent pas de modéliser différents schémas d'entrepôts de données.

C'est pourquoi nous avons conçu un banc d'essais nommé DWEB (*Data Warehouse Engineering Benchmark*) qui permet de générer différents entrepôts de données et charges de requêtes décisionnelles synthétiques et personnalisés [CDBB05]. DWEB est totalement paramétrable, ce qui lui permet de répondre à tous les besoins d'ingénierie des entrepôts. DWEB permet actuellement la génération automatique d'entrepôts de données « simples ». Il doit désormais être étendu aux données complexes, et notamment à la génération d'entrepôts de données complexes modélisées sous forme de documents XML, de manière à fournir une plateforme de tests à nos travaux sur l'entreposage de données complexes.

3.3. Projets appliqués

Dans le cadre de partenariats avec le monde industriel et scientifique, le laboratoire ERIC est impliqué dans des projets d'envergure qui mobilisent plusieurs personnes au sein du laboratoire. En retour, ces projets mettent en avant la viabilité des solutions théoriques développées, ils servent également de support aux recherches que nous menons. Ces coopérations donnent lieu à des séries de publications dans des revues et des conférences.

Modèles topologiques pour l'indexation et la navigation dans les bases de données complexes. Dans le cadre de nos travaux sur les graphes de proximité mentionnés plus haut, nous avons testé avec succès leur utilisation dans le domaine de l'indexation des bases de données complexes soit dans un but de catégorisation soit dans un but de recherche d'informations. L'application visée dans ce contexte concerne le domaine médical et plus particulièrement les cancers du sein dans le cadre du projet SIRIUS avec le centre de lutte contre le cancer Léon Bérard (Lyon) soutenu par la Région Rhône-Alpes. Les images mammographiques sont dans un premier temps pré-traitées pour extraire les caractéristiques (vecteur d'attributs). Chaque image est ensuite considérée comme un point d'un espace euclidien à p dimensions. Nous construisons sur l'ensemble des points un graphe de proximité en utilisant l'algorithme des Voisins Relatifs. Nous obtenons un graphe connexe qui rend bien compte de la topologie des points. Grâce à un algorithme d'insertion locale de très faible complexité que nous avons mis au point, nous pouvons catégoriser de nouvelles images par agrégation des étiquettes des voisins. La différence

fondamentale par rapport aux k-plus proches voisins, est que la structure de voisinage est plus naturelle car elle ne repose pas exclusivement sur la distance qui sépare les points mais sur la présence ou non de points à proximité. Ces techniques étaient beaucoup employées dans le plan notamment avec les diagrammes de Voronoï mais pas dans des espaces de dimensions élevées à cause de leur complexité. La structure de voisinage sert également comme outil de navigation dans la base de données. Ainsi, dès qu'un point est inséré dans le graphe, grâce à la propriété de connexité, nous pouvons naviguer de façon 'naturelle' dans la base de proche en proche.

Action Concertée Incitative "Terrains, Techniques, Théories". Dans le cadre d'une collaboration avec le laboratoire ICAR (Interactions, Corpus, Apprentissages, Représentations) de l'université Lyon 2 et de l'École Normale Supérieure Lettres et Sciences Humaines de Lyon, sous forme d'une ACI TTT (Action Concertée Incitative "Terrains, Techniques, Théories"), nous avons contribué à la constitution, à la gestion, à la valorisation et à la mise en ligne de bases de données complexes (audio, vidéo, textes annotés) rassemblant des corpus de langues parlées en interaction (CLAPI). Du point de vue de nos collègues linguistes, ce projet a permis des avancées significatives en matière de gestion et d'exploitation automatique des corpus, tandis qu'il a constitué pour nous un terrain d'expérimentation notable pour nos recherches sur la structuration et la modélisation des données complexes.

Hybridation des algorithmes d'optimisation. Dans le cadre d'une collaboration avec l'Institut de la Recherche sur la Catalyse basée à Lyon, nous développons des méthodes spécifiques dans le cadre de la catalyse hétérogène. L'objectif est de produire le catalyseur – un produit synthétique composé de plusieurs substances – qui permet d'optimiser les performances d'une réaction chimique. Il est nécessaire de procéder à des tâtonnements, très coûteux en temps et en argent, pour définir la meilleure combinaison constituant le catalyseur pour une réaction donnée. Après un premier projet, le projet COMBICAT, en coopération avec d'autres laboratoires européens, a permis de dégager le financement d'une thèse qui a été soutenue avec succès en 2004. Un second projet, le projet OPTICAT, a été mis sur pied en 2003 avec le financement d'une nouvelle thèse. L'idée est d'utiliser les méthodes de fouille de données pour mieux guider le processus de tâtonnement destiné à produire un catalyseur optimal. Nous parlons alors d'hybridation des algorithmes d'optimisation [CCRF05]. Outre les résultats théoriques et expérimentaux qui crédibilisent l'approche, une plate-forme logicielle complète pour mener les expérimentations a été mise au point. Il s'agit d'un outil unique au monde qui permet de faire intervenir différentes techniques de fouille de données dans un processus d'optimisation.

Classement de protéines. Le classement de protéines est une activité importante pour le biologiste. Avec l'augmentation croissante de la taille des banques de données, il est devenu nécessaire de mettre en œuvre des stratégies informatiques pour automatiser cette activité. Nous utilisons le canevas de l'extraction de connaissances à partir de données pour classer automatiquement les protéines à partir de leur structure primaire. En effet, la description primaire d'une protéine s'appuie sur un alphabet de 20 caractères représentant les acides aminés, une protéine étant constituée d'une suite d'acides aminés. Dès lors nous avons pu transposer la démarche de la classification de textes à partir des n-grammes (suites de n caractères) dans le processus de classement de protéines, en développant des techniques *ad hoc* pour résoudre les problèmes spécifiques telles que la réduction de la dimensionnalité. Ce travail, en coopération avec des biologistes de l'Université de Tunis (URPAH), nous a permis de produire des résultats très encourageants [CRME05].

3.4. VALORISATION

Dans le cadre de la valorisation scientifique, les chercheurs du laboratoire ERIC sont à l'origine de la création en 2000 de la conférence nationale Extraction et Gestion des Connaissances (**EGC**) qui réunit plus de 250 chercheurs du domaine tous les ans. En juin 2003, à l'initiative du laboratoire ERIC, un groupe de travail national sur la fouille dans les données complexes (**GT FDC**) a été créé. Regroupant plus d'une centaine de chercheurs de nombreux laboratoires français, il se réunit trois fois par an, dont une fois à l'occasion de la conférence nationale Extraction et Gestion des Connaissances dans le cadre d'un atelier. Dans cette même dynamique, après avoir créé le groupe de travail sur la fouille de données complexes au sein de l'association EGC, les chercheurs d'ERIC viennent de lancer dans le cadre de *International Conference in Data Mining* de l'IEEE, un *Workshop Mining Complex Data* qui se tient pour la première fois à Huston (USA) à la fin du mois de novembre 2005.

Le laboratoire ERIC est à l'origine de la création, en 2002, de la collection d'ouvrage **RNTI** (Revue des nouvelles technologies de l'Information) publiée par Cepadues. Cette collection offre aux chercheurs un espace de publication de leurs travaux les plus avancés. Tous les articles publiés dans ses numéros spéciaux font l'objet d'une évaluation par trois rapporteurs indépendants. Trois à quatre numéros sont édités chaque année.

Finalement, l'intérêt pour les travaux liés aux entrepôts de données, à l'analyse en ligne (OLAP) et aux bases de données multidimensionnelles, n'a cessé de croître aussi bien au sein de la communauté des chercheurs que de celle des industriels et des utilisateurs de manière générale. Afin de créer et de pérenniser un cadre exclusivement dédié aux travaux dans ces domaines, nous avons initié les journées francophones sur les Entrepôts de Données et l'Analyse en ligne (**EDA**). Cette conférence favorise la rencontre de tous les chercheurs, industriels et utilisateurs francophones afin de discuter de l'avancement de la recherche ainsi que d'expériences de développement. Ces rencontres se doivent de devenir un rendez-vous national régulier. La première journée a eu lieu à Lyon le 10 juin 2005. La deuxième aura lieu à l'université de Versailles Saint-Quentin le 19 juin 2006.

3.5. COLLABORATIONS

Les collaborations des membres d'ERIC (voir tableau récapitulatif paragraphe 4.2.) sont de différents ordres. On citera pour l'essentiel les collaborations avec :

- des équipes de recherche locales en Sciences Humaines et Sociales à l'Université Lyon 2 : DDL, CED, ICAR (cf. projets appliqués) ;
- des équipes nationales ou dans le cadre de projets nationaux ACI : ENST, LSIIT et LIV, IRC (cf. projets appliqués) ;
- l'industrie : par exemple, nous avons mené une collaboration avec France Telecom R&D autour de la problématique du *Churn* (départ des clients vers d'autres opérateurs). Nous avons à cette occasion réalisé un travail quasi-exhaustif sur le problème de la sélection de variables (feature selection) ;
- des équipes internationales, notamment avec les universités de Genève (Suisse), d'Oklahoma (USA), de la Mer Egée (Grèce) qui ont conduit à de nombreuses publications conjointes.

Notre coopération avec le laboratoire Dynamique du Langage de l'Université Lyon 2 (DDL) est le fruit d'un rapprochement autour de préoccupations communes en reconnaissance des formes. L'objectif est de rattacher automatiquement des phrases prononcées par un locuteur à la langue utilisée. L'utilisation du cadre de l'ECD a permis de systématiser l'approche et d'analyser finement les conséquences des choix réalisés à chaque étape du processus de classement automatique ; nous avons pu également mettre en œuvre des techniques de fouille très peu connues chez les linguistes [CPCR02].

Le Centre d'Etudes Démographiques de l'Université Lyon 2 (CED) dispose d'une base de données d'actes d'état-civil exhaustive sur les communes de la vallée de la Valserine pour les trois derniers siècles. Une collaboration originale a été entreprise entre ERIC et le CED pour gérer efficacement ces données, en déduire les listes d'ascendance ou de descendance dans la vallée et procéder à l'analyse statistique de ces listes dans une perspective socio-démographique [DBLB05].

A l'occasion de l'action spécifique GaFo-Qualité organisée en 2002 par le LIN, Nantes, une collaboration a été engagée avec l'UMR TAMCIC, Ecole Nationale Supérieure des Télécommunications Bretagne (ENST) sur le thème de la qualité des règles d'association. Cette recherche commune qui a donné lieu à de nombreuses publications (par exemple [CVLL04]) associe l'étude formelle des propriétés des mesures de l'intérêt des règles à l'analyse de leur comportement expérimental.

Une collaboration avec le Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection (LSIIT) et le Laboratoire Image et Ville (LIV) de l'Université de Strasbourg I, dans le cadre d'une Action Concertée Incitative "Masses de Données", a pour objectif de qualifier la végétation urbaine à partir de bases de données d'images. Pour cela sont proposés une méthode d'aide à l'interprétation à partir d'une masse de données images et un processus complet de fouille de données permettant une utilisation conjointe et complémentaire des différentes sources.

Des chercheurs du laboratoire ERIC entretiennent une relation suivie avec le département d'économétrie de l'université de Genève. Cette collaboration se situe à la fois sur le niveau pédagogique par des invitations de professeurs dans les deux sens et sur le niveau recherche par une production scientifique conséquente : plus d'une dizaine d'articles co-signés par les deux laboratoires.

En 2002 l'Université Lyon 2 et l'Université Egée en Grèce ont signé une convention de cotutelle de thèse. Dans ce cadre, une collaboration entre les deux établissements porte sur le thème de la construction d'un portail multimédia multilangues dans le domaine de l'héritage culturel. Ce projet de recherche a donné lieu à différentes publications [CMKN04, CMNK04]. De plus, en 2005, un projet Européen dans le cadre du 6ème PRC a été déposé.

3.6. DÉVELOPPEMENT DE LOGICIELS

Différents projets d'élaboration de logiciels dans le cadre des travaux de thèse ont été mis en place depuis la création du laboratoire. Ces projets ont pour objectif de créer des outils destinés au test et à l'évaluation des développements théoriques réalisés par les chercheurs. Ils sont également mis en oeuvre dans les contrats passés par le laboratoire.

Le projet **SIPINA_W** a été initié à l'été 1995, ce logiciel destiné à l'induction par graphes.

Une diffusion mondiale du logiciel a été assurée par sa mise à disposition gratuite sur Internet, sur les pages Web du laboratoire ainsi que sur les pages Web du forum KDD Nuggets (http://www.kdnuggets.com). Plusieurs centaines d'utilisateurs dans le monde sont recensés à ce jour, une dizaine d'entre eux ayant publiés des thèses ou des articles scientifiques en utilisant les résultats obtenus à l'aide du logiciel. Par ailleurs, SIPINA_W est utilisé en enseignement dans plusieurs universités : l'Université Lumière Lyon 2, l'Université Joseph Fourrier de Grenoble, l'école ESIEA (Ecole Supérieure d'Informatique Electronique Automatique) de Paris et l'Université de Chambéry.

TANAGRA est un logiciel gratuit de data mining destiné à l'enseignement et à la recherche, diffusé sur internet. Il implémente une série de méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'apprentissage automatique et des bases de données.

- Son premier objectif est d'offrir aux étudiants et aux chercheurs d'autres domaines (médecine, bio-informatique, marketing, etc.) une plate-forme facile d'accès, respectant les standards des logiciels actuels, notamment en matière d'interface et de mode de fonctionnement, il doit être possible d'utiliser le logiciel pour mener des études sur des données réelles.
- Le second objectif est de proposer aux enseignants une plate-forme entièrement fonctionnelle, le logiciel peut servir d'appui pédagogique pour les illustrations et le traitement des jeux de données en cours ou en TD.
- Enfin, le troisième objectif est de proposer aux chercheurs une architecture leur facilitant l'implémentation des techniques qu'ils veulent étudier, de comparer les performances de ces algorithmes. TANAGRA se comporte alors plus comme une plate-forme d'expérimentation. Point très important à nos yeux, la disponibilité du code source est un gage de crédibilité scientifique, elle assure la reproductibilité des expérimentations publiées, et surtout, elle permet la comparaison et la vérification des implémentations.

Le site de diffusion du logiciel (http://eric.univ-lyon2.fr/~ricco/tanagra) a été mis en ligne en janvier 2004, il compte en moyenne une vingtaine de visiteurs par jour. TANAGRA est également référencé par les principaux portails de l'ECD.

Pour mettre en œuvre l'intégration des données complexes, un système multiagents baptisé **SMAIDoC** a été développé. Il est basé sur une plate-forme d'agents génériques (http://bdd.univ-lyon2.fr/?page=logiciels). Le fonctionnement du système s'articule autour de cinq agents qui se chargent de l'intégration de données complexes dans une base de données relationnelle [CDRBB03, CDRBB03b]. Lorsque l'utilisateur choisit un site dans lequel se trouvent des données complexes, l'agent MenuAgent ordonne aux agents DataAgent et Wrap-perAgent de migrer. L'agent DataAgent collecte les données ainsi que les métadonnées et les transmet séquentiellement à l'agent WrapperAgent qui instancie progressivement la structure UML. Ce dernier transmet la structure UML créée à l'agent XMLCreator. XMLCreator traduit la structure UML en une DTD et génère des documents XML valides. Pour terminer, l'agent XMLCreator transmet les documents XML à l'agent XML2RDBAgent qui se charge du stockage des documents XML dans la base de données relationnelle. Le processus décrit ci-dessus se répète autant de fois que nécessaire [EBBDC03, FBBD03b]. Le modèle UML que nous avons proposé a été étendu avec une partie générique permettant de générer autant de classes et de liens entre elles que nécessaires pour représenter une sémantique spécifique un ensemble donné de données complexes [DTBB05].

Et enfin, les différents travaux sur la modélisation multidimensionnelle des données complexes, les nouveaux opérateurs OLAP, l'intégration de la fouille de données dans les SGBD, … ont également donné lieu au développement de prototypes (http://bdd.univ-lyon2.fr/?page=logiciels).

3.7. SYNERGIE ENSEIGNEMENT - RECHERCHE

Le laboratoire ERIC entretient des liens privilégiés avec le Département d'Informatique et de Statistique de la Faculté de Sciences Economiques et de Gestion de l'Université Lyon 2, dont la majorité de ses enseignants-chercheurs fait partie. Dans ce cadre, les membres d'ERIC animent un cycle de Master d'informatique complet, ainsi qu'une troisième année de Licence qui permet à divers étudiants en SHS de niveaux hétérogènes de se préparer au Master. Les enseignants-chercheurs d'ERIC sont également fortement impliqués dans de nombreuses autres formations,

notamment au sein de la Faculté de Sciences Economiques et de Gestion (Licence et Master), de l'IUT Lumière (département STID - Statistique et Traitement Informatique des Données, qui gère un DUT STID et une licence professionnelle « Chargé d'Etudes Statistiques ») et de la Faculté d'Anthropologie et de Sociologie (Licence MISASHS - Mathématiques, Informatique et Statistiques Appliquées aux Sciences Humaines et Sociales).

Au niveau du Master d'informatique, les membres d'ERIC animent la spécialité recherche ECD (Extraction des Connaissances à partir des Données). Cette formation multi-établissements (Nantes, Paris 11, Lyon 1 et Lyon 2) recrute environ vingt-cinq étudiants par an et sert par conséquent de pépinière pour la recherche du laboratoire. Cette formation offre une double originalité : c'est le seul Master en ECD de France et les enseignements peuvent être suivis à distance. Sur ce dernier point, tous les ans, des élèves-ingénieurs de l'Ecole Polytechnique de Bucarest (Roumanie) suivent la formation en visioconférence.

Les enseignants-chercheurs du laboratoire animent également des formations professionnalisantes (ex IUP et DESS) désormais intégrées dans la spécialité IDS (Informatique Décisionnelle et Statistique) du Master d'Informatique de Lyon 2. La proximité de ces formations avec ERIC permet des transferts de technologie rapides entre la recherche et l'industrie. En sens inverse, les contacts réguliers avec de nombreuses entreprises facilitent l'établissement de relations basées sur la recherche (contrats, financements CIFRE, etc.).

4. PROJETS DE RECHERCHE EN COURS

4.1. PROJETS DE RECHERCHE APPLIQUÉE

Amélioration de la prise en charge de patients cérébro-lésés et de patients ischémiques

Identification des partenaires	UMR 5823 LASS, Université Claude Bernard Lyon 1 UMR 5515 CREATIS, Université Claude Bernard Lyon 1 UMR 5015 ISC, Université Claude Bernard Lyon 1 TIMC, Université Joseph Fourier Grenoble
Objectifs recherchés	Constitution d'une banque de données anatomo-fonctionnelle aidant à la compréhension des mécanismes cérébraux et cardiaques chez l'homme. Création du premier centre distribué national de recueil de données médicales anatomiques et fonctionnelles de patients cérébro-lésés d'une part et de patients atteints d'affections du myocarde d'autre part.
Durée et financement	Durée : 3 ans. 2001-2003 Financement par la Région Rhône-Alpes : - une bourse de thèse de trois ans (Pierre-Emmanuel Jouve) - 30 000 €

Méthodes de ciblage de la clientèle tenant compte des bénéfices et coûts attendus

Identification du partenaire	Crédit Commercial de France

Objectifs recherchés	Méthodes d'optimisation du bénéfice global espéré Mise au point une représentation graphique de la liaison entre la taille de la cible optimisée et bénéfice global espéré. Propositions de plans d'expérience en vue de l'amélioration du ciblage
Durée et financement	Durée : 2001 Financement par le Crédit Commercial de France : 11 000 €

Prédiction des churns

Identification du partenaire	France Télécom R&D : Marc BOULLE
Objectifs recherchés	L'objectif de cette recherche concerne la sélection et la construction d'attributs pour la préparation des données appliquées aux grandes bases de données gestion-clients pour la réalisation du projet Customer Constructive Inductive Learning.
Durée et financement	Durée : 2001 – 2002 Financement par France Télécom : - 100 000 € - Une bourse de thèse de trois ans (Walid Erray)

Enquête trimestrielle des passagers d'un aéroport

Identification du partenaire	CCI de Lyon (Aéroport Lyon Saint-Exupéry)
Objectifs recherchés	Conception d'une enquête trimestrielle sur 4.000 passagers de l'aéroport pour connaître les caractéristiques et les origines-destinations des voyages (avec ou sans correspondances à Lyon) et mettre à jour de nouvelles lignes directes possibles au départ de Lyon. Réalisation de l'étude prototype (questionnaire, mode de contact, échantillonnage), des bases de données et des procédures d'exploitation.
Durée et financement	Durée : 2001 – 2002 Financement par la CCI de Lyon : 35 000 €

Enquête sur le devenir des apprentis de l'enseignements supérieur en Rhône-Alpes

Identification des partenaires	Région Rhône-Alpes Formasup Rhône-Alpes (IPRA) Rectorat des Académies de Lyon et Grenoble

Objectifs recherchés	Conception, réalisation, exploitation et présentation d'une enquête d'insertion de l'ensemble des apprentis de l'enseignement supérieur en Rhône-Alpes
Durée et financement	Financement par Formasup Rhône-Alpes : - 15 000 € en 2002-2003 - 3 600 les années suivantes

Modélisation des flux de transport de personnes

Identification du partenaire	Direction des transports de la Région Rhône-Alpes
Objectifs recherchés	Modélisation des flux de transport de personnes entre les villes de la région et des villes périphériques (Genève, Macon, …) Etude de la sensibilité à l'offre de trains et d'autoroutes, aux durées de transport, au cadencement des trains et aux tarifs. Recherche des données pertinantes, de leur qualité et construction du modèle économétrique et informatique.
Durée et financement	Durée : 2002 – 2003 Financement par la Région Rhône-Alpes : 21 500 €

Méthodes de fouille de données pour l'exploitation des bases de données CV

Identification du partenaire	Fondation Védior Bis
Objectifs recherchés	La Fondation VédiorBis Recherche (FVR) a pour vocation d'aider les laboratoires de recherche travaillant sur des thématiques pouvant aider à mieux caractériser l'offre et la demande d'emploi. Dans ce contexte, deux projets ont été financés sous forme de bourse de thèse sur une durée de deux années chacun. Les deux projets, le second dans le prolongement du premier ont pour but de déveopper des méthodes de fouille de données pour l'exploitation des bases de données CV. Le travail de Jérémie Clech, présenté dans sa thèse soutenue en mars 2004 a été dédié à la discrimination automatique des CV de cadre des autres. Le travail de Riadh BenMessaoud vise à approfondir ces questions.

Durée et financement	Financement par la Fondation Védior Bis : - bourse de thèse 1500 € par mois sur deux ans (Jérémy Clech) 2001-2003 - bourse de thèse 1500 € par mois sur de deux ans (Riadh Ben Messaoud) 2003-2005

Corpus de Langue Parlée en Interaction (CLAPI)

Identification des partenaires	Action Concertée Incitative "Terrains, Techniques, Théories" Laboratoire ICAR Université Lyon 2 et ENS Lettres et Sciences Humaines Lyon, Laboratoire RIM (Réseaux, Information, Multimédia), École Nationale Supérieure des Mines de Saint-Étienne
Objectifs recherchés	Assurer dans un délai de trois ans la constitution, la gestion, la valorisation et la mise en ligne de bases de données multimédia (audio, vidéo) rassemblant des corpus linguistiques oraux.
Durée et financement	Durée : 2002-2005 Financement par le Ministère de la Recherche (ACI) : - 36 000 € - bourse de thèse de trois ans (Kamel Aouiche)

Médecine d'anticipation personnalisée (MAP)

Identification du partenaire	Docteur Ferret, médecin du sport et porteur d'un projet de création d'entreprise accueilli au sein de l'incubateur CREALYS.
Objectifs recherchés	Etendre les résultats et avancées empiriques développés pour les sportifs de haut niveau à d'autres populations et de faire en sorte que les sujets analysés deviennent les gestionnaires de leur capital santé. Structurer, stocker et analyser un ensemble de données médicales complexes (qualitatives, numériques, textes, images...) concernant un grand ensemble de personnes.
Durée et financement	Durée : 2003-2004 Financement par la Région Rhône-Alpes et Lyon 2 : 29000 €

Entrepôt virtuel de données bancaires

Identification du partenaire	Crédit Lyonnais. Direction d'Exploitation Rhône-Alpes-Auvergne

Objectifs recherchés	L'objectif de ce projet est d'assurer dans un délai de trois ans la mise au point d'outils méthodologiques pour la gestion et l'analyse de données bancaires, qui se présentent sous forme de données hétérogènes. Du point de vue du Crédit Lyonnais, il s'agit de développer un système d'aide à la décision dans le domaine de ciblages clients. Du point de vue du laboratoire ERIC, il s'agit d'acquérir une expertise dans le domaine de l'entreposage virtuel de données hétérogènes. Il s'agit de construire des cubes de données à la volée en vue d'analyses (analyse en ligne (OLAP) et fouille de données), qui nécessite une intégration efficace de données.
Durée et financement	Durée : 2004-2007 Financement par le Crédit Lyonnais : - une bourse de thèse CIFRE de 3 ans (Cécile Favre)

Etude "Citoyens et usagers face aux évolutions des services publics marchands"

Identification du partenaire	Commissariat Général au Plan, service du Premier Ministre
Objectifs recherchés	Conception, réalisation et exploitation d'une enquête par sondage (1.000 enquétés, questions ouvertes et fermées). Fouille approfondie des résultats. Comprendre les attentes des Français vis à vis des services publics marchands (train, postes, transports urbains, électricité, gaz, …) dans une période de boulversements de leur organisation et selon leurs expériences concrètes (usages des services ; liens personnels avec ces entreprises ; etc.)
Durée et financement	Durée : 2004 Financement par le Commissariat Général au Plan : 54 000 €

DataMining pour la recherche pharmaceutique

Identification du partenaire	Laboratoires SERVIER

Objectifs recherchés	Fouille des données recueillies lors des tests de médicaments en phase IV (juste avant la demande d'Autorisation de Mise sur le Marché, AMM), en vue de découvrir les éventuels effets secondaires dangereux et leurs causes (la molécule testée en général, ou son association avec d'autres médicaments, ou des antécédents pathologiques de certains patients)
Durée et financement	Durée : 2004 Financement par les laboratoires SERVIER : 7 200 €

Fouille de Données Multistratégies (FoDoMuSt)

Identification des partenaires	Action Concertée Incitative "Masses de Données" Laboratoire LSIIT (Laboratoire des Sciences de l'Images, de l'Informatique et de la Télédétection), Université de Strasbourg I Laboratoire LIV (Laboratoire Image et Ville), Université de Strasbourg I
Objectifs recherchés	Les objectifs du projet, associés à l'imagerie spatiale, sont : d'une part, proposer une méthode d'aide à l'interprétation à partir d'une masse de données images et d'autre part, définir un processus complet de fouilles de données (structuration, construction des « objets », classification et interprétation de l'information) permettant une utilisation conjointe et complémentaire des différentes sources. Ce dernier aspect est rarement pris en compte dans les méthodes actuelles d'extraction. Le verrou principal réside dans la nécessité d'utiliser une multi-formalisation à plusieurs niveaux d'abstraction selon une approche multi-stratégie dans le processus de fouilles de données.
Durée et financement	Durée : 2004-2007 Financement par le Ministère de la Recherche (ACI) : 69 000 €

Système Intelligent pour la Recherche d'Information à l'Usage de la Santé (SIRIUS)

Identification des partenaires	Région Rhône-Alpes Centre anti cancéreux Léon Bérard Lyon

Objectifs recherchés	Le Système Intelligent pour la Recherche d'Information à l'Usage de la Santé (SIRIUS) sera développé et testé avec des usagers (Centre Léon Bérard). Le choix du secteur de la cancérologie résulte à la fois de la longue collaboration que nous entretenons avec le Centre Léon Bérard et de l'intérêt que porte la Région Rhône-Alpes à ce domaine, notamment à travers la création du cancéropôle.
Durée et financement	Durée : 2004-2007 Financement par la Région Rhône-Alpes : - 3000 € en fonctionnement - bourse de thèse de 1500€ par mois sur 3 ans (Hakim HACID)

Interdépendance des marchés immobiliers résidentiels

Identification du partenaire	Université de Genève
Objectifs recherchés	Etude sur l'interdépendance des marchés immobiliers résidentiels sur le bassin franco-valdo-genevois dans le cadre du programme européen INTERREG. Ce projet concerne l'analyse des marchés fonciers, des marchés résidentiels privés locatifs et des marchés résidentiels de vente d'appartements et de maisons individuelles simultanément sur les différentes zones du bassin. Son objectif est d'améliorer la connaissance du fonctionnement des marchés immobiliers résidentiels privés : - en visualisant l'évolution des prix des biens et des services sur une période de trente ans, - en observant la dynamique de ces marchés immobiliers simultanément dans les quatre zones du bassin, - en mettant en évidence les interdépendances existant entre les marchés immobiliers des différentes zones, - en créant des modèles économétriques permettant une analyse prospective.
Durée et financement	Durée : 2004 – 2006 Financement par le fond européen INTEREG : 55000 €

Positionnement relatif des législations du travail

Identification des partenaires	Bureau International du Travail Université de Genève Fondation RUIG.
Objectifs recherchés	Ce projet vise à développer des méthodes de fouille de texte visant à étudier et à positionner les législations du travail des différents pays. Le Bureau International du Travail (BIT) souhaite ensuite dresser des caratographies permettant aux représentants des différents pays de se positionner les uns par rapport aux autres. Le laboratoire ERIC assure la partie text mining du projet pour extraire les paramètres descriptifs des corpus juridiques. Il devra pour ce faire exploiter plusieurs centaines de textes relatifs à la legislation du travail.
Durée et financement	Durée : 2005 – 2007 Financement par le BIT : 12000 €

Méthodes et logiciels pour l'extraction de règles d'association

Identification du partenaire	Laboratoire d'Ingénierie des Connaissances de l'Université de Prague, République Tchèque.
Objectifs recherchés	Nous avons entamé une collaboration scientifique avec l'équipe d'ingénierie des connaissances de l'Université de Prague. L'objectif est de développer en commun des plates formes de data mining. ERIC apportant son expérience et son savoir faire à travers la plate forme SIPINA, l'équipe Tchèque a développé une plate forme pour l'extraction des règles d'association baptisée LispMiner.
Durée et financement	Durée : 2004 - 2006 Financement par le programme d'échange Franco-Tchèque Barrande : 6000 €

4.2. COLLABORATIONS INTERNATIONALES

Université Laval à Québec, Canada

Identification du partenaire	Professeurs Nadir Belkhiter et Guy Mineau,

Collaboration en enseignement	Nous entretenons depuis de longues années une collaboration régulière avec l'Université Laval à Québec. Le Professeur Nadir Belkhiter. Est régulièrement invité à l'Université Lyon 2 pour assurer des enseignements en DESS et DEA dans le domaine de la fouille des données et des interfaces de communication homme-machine.
Collaboration en recherche	Grâce aux compétences du Professeur Belkhiter dans le domaine des interfaces de communication homme-machine, nous développons une réflexion méthodologique sur les interfaces et le *data mining*. En effet, les utilisateurs de ces techniques sont potentiellement très nombreux mais, ces outils ne seront réellement utilisés que s'ils sont facile à appréhender. Cette recherche vise à étudier les modes d'interaction et les techniques de visualisation dans le domaine de l'ECD.

Université de Genève, Suisse

Identification du partenaire	Professeur Gilbert Ritschard
Collaboration en enseignement	Le Professeur G. Ritschard intervient depuis 1999 en tant que professeur invité dans un cours en DEA ECD sur les mesures d'association.
Collaboration en recherche	Nous travaillons avec le Professeur G. Ritschard depuis de nombreuses années et nous avons déjà de nombreuses publications communes.

Université de Prague, République Tchèque

Identification du partenaire	Professeurs Jan Rauch et Petr Berka
Collaboration en recherche	Développement d'outils communs pour la fouille de données.

Ecole Nationale d'Informatique de Tunis, Laboratoire RIADI-GDL, Tunisie

Identification du partenaire	Mme Hajer Bazaoui
Collaboration en recherche	Modélisation et analyse de data marts spatio-temporels.
Perspectives	Après avoir construit un modèle générique de data marts spatio-temporels, nous travaillons actuellement sur une démarche exploratoire incluant des analyses descriptives, de l'OLAP et de l'extraction des connaissances.

Université d'Oklahoma, Norman, USA

Identification du partenaire	Professeur Le Gruenwald
Collaboration en recherche	Un projet de recherche concernant l'utilisation de techniques de fouille de données pour l'auto-administration des entrepôts de données a abouti à plusieurs publications communes (concernant l'auto-indexation, principalement). Nous envoyons régulièrement des étudiants de Master en stage aux Etats-Unis depuis 2001.
Perspectives	Poursuivre et renforcer la collaboration sur le projet d'auto-administration. Un séjour de Mme Gruenwald au laboratoire est prévu en 2005. D'un point de vue scientifique, il s'agit d'une part d'étendre notre approche d'auto-indexation à d'autres techniques d'optimisation de performance (matérialisation de vues, notamment) et, d'autre part, de tester différentes technique de fouille dans ce cadre (motifs fréquents, motifs séquentiels, classification…) pour trouver la plus adaptée à chaque cas. Applications prévues aux données complexes

Ecole Nationale d'Informatique, Université de Fianarantsoa, Madagascar

Identification du partenaire	Victor Manantsoa
Collaboration en recherche	Performance des entrepôts de données complexes

Perspectives	Développer la collaboration entre ERIC et l'ENI. Les deux laboratoires ont des thématiques de recherche très proches et ont la volonté de développer des projets en commun. Le lien entre nos deux structures de recherche est actuellement assuré en grande partie par M. Ralaivao, dont le travail de thèse matérialise cette volonté de collaboration et se renforce à chacun de ses séjours au laboratoire ERIC.

Universités de Zagreb et de Ljubljana, Croatie et Slovénie

Identification du partenaire	Professeurs Dragan GAMBERGER, Ivan BRATKO, Zdenko SONICKI Avec l'aide du Ministère des Affaires Etrangères (programme EGIDE 2004-2005)
Collaboration en recherche	Méthodes de fouille des données médicales Organisation conjointe de « International Workshop on Intelligent Data Analysis and Data Mining, Application in Medicine » à Zagreb en juin 2004, avec 30 participants. Actes sur : http://lis.irb.hr/IDADM/ Dix jours de travaux coopératifs à Lyon en novembre 2004 pour fouiller de nouvelles données épidémiologiques et confronter les méthodes.
Perspectives	Un nouveau Workshop sera organisé en 2005, avec le soutien de EGIDE

HEC Montréal, Université de Montréal, Canada

Identification du partenaire	Professeur Denis LAROCQUE
Collaboration en recherche	Animation de séminaires pour les professeurs et les étudiants de Master du Département « Méthodes quantitatives de gestion » en 2002 et en 2004. Des collaboration de recherche sont engagées.
Perspectives	Le Professeur Denis LAROCQUE doit venir à ERIC pour une période sabbatique en 2005-2006

4.3. Sujets de thèse en cours

Kamel AOUICHE

Date de naissance : 10.10.1976

Directeur de thèse : J. Darmont

Financement : Allocation de recherche

Date de début de thèse : Octobre 2002

Date de soutenance prévue : Fin 2005

Titre : Auto-administration d'entrepôts de données à l'aide de techniques de fouille de données

Mots clés : Base de données, entrepôt de données, auto-administration, sélection d'index, sélection de vues matérialisées, fouille de données.

Résumé de thèse :

Avec le développement des bases de données en général et des entrepôts de données en particulier, il est devenu très important de réduire les tâches d'administration des systèmes de gestion de base de données. Les systèmes auto-administratifs ont pour objectif de s'administrer et de s'adapter eux-mêmes, automatiquement, sans perte ou même avec un gain de performance.

L'idée d'utiliser des techniques de fouille de données pour extraire des connaissances utiles à partir des données stockées pour leur administration est une approche très prometteuse, notamment dans le domaine des entrepôts de données, où les requêtes sont très hétérogènes et ne peuvent pas être interprétées facilement.

L'objectif de cette thèse est d'étudier les techniques d'auto-administration des entrepôts de données, principalement des techniques d'optimisation des performances, comme l'indexation et la matérialisation de vues, et de rechercher une manière d'extraire des données elles-mêmes des connaissances utilisables pour appliquer ces techniques. Nous avons réalisé un outil qui recommande une configuration d'index et de vues matérialisées permettant d'optimiser le temps d'accès aux données. Notre outil effectue une recherche de motifs fréquents fermés sur une charge donnée et une classification non supervisée des requêtes de la charge pour construire cette configuration d'index et de vues. Nous avons également couplé la sélection d'index et de vues matérialisées afin de partager efficacement l'espace de disque alloué pour stocker ces structures. Enfin, nous avons appliqué les principes développés dans le cadre relationnel aux entrepôts de données XML. Nous avons proposé une structure d'index précalculant les jointures entre les faits

et les dimensions XML et adapté notre stratégie de sélection de vues pour matérialiser des vues XML.

Mots clés : Bases de données, Entrepôts de données, Entrepôts de données XML, Indexation, Matérialisation de vues, Fouille de données, Recherche de motifs fréquents, Classification non supervisée, Modèles de coût, Performance.

Publications

[ADBF05] K. Aouiche, J. Darmont, O. Boussaid, F. Bentayeb, ``Automatic Selection of Bitmap Join Indexes in Data Warehouses", 7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK05), Copenhagen, Denmark, Août 2005, 64-73.

[DBRA05] J. Darmont, O. Boussaid, J. Ralaivao, K. Aouiche, ``An Architecture Framework for Complex Data Warehouses", 7th International Conference on Enterprise Information Systems (ICEIS 05), Miami, USA, May 2005.

[ABBBD03] K. Aouiche, F. Bentayeb, O. Boussaid, J. Darmont, ``Conception informatique d'une base de données multimédia de corpus linguistiques oraux : l'exemple de CLAPI 2, ``36ème Colloque International de la Societas Linguistica Europaea", Lyon, France, Septembre 2003, 11-12.

[ADG03a] K. Aouiche, J. Darmont, L. Gruenwald, ``Frequent itemsets mining for database auto-administration", 7th International Database Engineering and Application Symposium (IDEAS 03), Hong Kong, China, July 2003, 98-103.

[ADBB04] K. Aouiche, J. Darmont, O. Boussaid, F. Bentayeb, ``Auto-administration des entrepôts de données : une stratégie de sélection automatique d'index \textit{bitmap} de jointure", Revue de Nouvelles Technologies, Décembre 2004.

[ADG03b] K. Aouiche, J. Darmont, L. Gruenwald, ``Vers l'auto-administration des entrepôts de données", Revue des Nouvelles Technologies de l'Information, No. 1, 2003, 1-12.

R. Benmessaoud, K. Aouiche, C. Favre, ``Une approche de construction d'espaces de représentation multidimensionnels dédiés à la visualisation", 1ère journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA05), Lyon, France, Juin 2005, 34-50.

[ADB04] K. Aouiche, J. Darmont, O. Boussaid, ``Sélection automatique d'index dans les entrepôts de données", 1er atelier Fouille de Données Complexes dans un processus d'extraction des connaissances, EGC 04, Clermont-Ferrand, Janvier 2004, 91-102.

[ADG03c] K. Aouiche, J. Darmont, L. Gruenwald, ``Extraction de motifs fréquents pour l'auto-administration des bases de données", Journées francophones d'Extraction et de Gestion des Connaissances (EGC 03), Lyon, Janvier 2003; Revue des Sciences et Technologies de l'Information, Vol. 17, 547.

Riadh BEN MESSAOUD

Date de naissance : 28.02.1979

Directeur de thèse : Nicolas NICOLOYANNIS

Responsables scientifiques : Omar BOUSSAID et Sabine RABASEDA

Financement : La Fondation VediorBis pour la Recherche et l'Emploi

Date de début de thèse : Octobre 2003

Date de soutenance prévue : Fin 2006

Titre : Couplage OLAP et Data Mining pour l'exploration et l'explication de données complexes : application à la recherche d'emploi

Mots clés : Entrepôt de données,.analyse en ligne OLAP, fouille de données, extraction de connaissances, données complexes.

Résumé de thèse :

Le projet de cette thèse a pour but d'assurer la mise au point d'outils méthodologiques et logiciels pour et l'analyse en ligne de données complexes en vue d'améliorer les processus de prise de décision.

Par exemple, dans le domaine du recrutement, comme dans de nombreux domaines, nous assistons à l'heure actuelle à une prolifération de données dites complexes qui peuvent prendre la forme de texte (CV, lettres de motivation, profils de poste, …) d'images, son, vidéo (entretiens de recrutement filmés) ou encore de bases de données (contenant par exemple le contenu de formulaires, …). Sur le plan informatique, il s'agit de définir un processus qui permet de structurer et de modéliser des données complexes (multimédia, textes, bases de données, données provenant du Web, …) et ce dans une perspective d'analyser en ligne ces données pour l'aide à la décision.

Dans le cadre des recherches menées au sein du laboratoire ERIC, des éléments de solution sont d'ores et déjà avancés d'une part pour l'intégration et la modélisation des données complexes dans une même base et d'autre part pour la fouille dans les données complexes.

Au niveau de l'analyse des données complexes, l'objectif de cette thèse est de poursuivre ces travaux en définissant une nouvelle approche pour l'analyse des données complexes s'appuyant sur les principes de la fouille de données (Data Mining) et de l'analyse en ligne (OLAP). En effet, les technologies bien connues de fouille et d'analyse en ligne des données, ont largement fait leurs preuves sur des données dites simples (attributs-valeurs) mais elles sont à adapter quand il s'agit

de données complexes. Nous pensons possible d'enrichir l'analyse des données complexes en combinant les principes de ces deux technologies pour créer de nouveaux contextes d'analyse, d'exploration, d'explication et de prédiction des données complexes. Il devient nécessaire de définir des nouveaux moyens d'analyse plus élaborée s'appuyant sur des techniques de fouille de données.

La construction d'un nouvel opérateur d'analyse en ligne basé sur la classification ascendante hiérarchique (*OpAC*) a déjà donné des résultats prometteurs. La prise en compte de la nature multidimensionnelle des données stockées dans des cubes peut améliorer la recherche d'information pour les algorithmes d'extraction de connaissance. L'émergence de sous-ensembles de données pertinents peut être obtenue par les outils de fouille de données afin d'alimenter des systèmes d'analyse en ligne en données de qualité.

Publications :

R. BenMessaoud, S. Rabaséda, O. Boussaid, F. Bentayeb, « OpAC : Opérateur d'analyse en ligne basé sur une technique de fouille de données », *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04)*, Clermont-Ferrand, Janvier 2004; Revue des Nouvelles Technologies de l'Information, Vol. 2, 35-46.

R. BenMessaoud, S. Rabaséda, O. Boussaid, F. Bentayeb, « OpAC: A New OLAP Operator Based on a Data Mining Method », *Sixth International Baltic Conference on Databases and Information Systems (DB&IS 04)*, Riga, Latvia, June 2004.

R. BenMessaoud, O. Boussaid, S. Rabaséda, « A New OLAP Aggregation Based on the AHC Technique », *ACM 7th International Workshop on Data Warehousing and OLAP (DOLAP'04)*, Washington D.C., USA, November 2004.

Julien CHARBEL

Date de naissance : 31/08/1978 à Jdeidet Marjayoun (Liban)

Directeur de thèse : D.A. Zighed

Financement : Ressources propres

Date de début de thèse : Octobre 2004

Date de soutenance prévue : Décembre 2007

Titre : Réduction de la volumétrie pour la fouille dans les bases de données complexes

Mots clés : Fouille de données complexes, compression, recherche d'espace de représentation

Résumé de thèse :

Dans les bases de données complexes, nous trouvons des informations variées et volumineuses telles que les objets multimédia. La fouille des ces bases de données se heurte à de multiples facteurs comme l'hétérogénéité des données : images, son, descripteurs qualitatifs,… et la volumétrie. Dans cette thèse nous allons aborder plus spécifiquement la gestion de la volumétrie. Il s'agit de trouver des représentations de ces mêmes données qui soient plus économiques en taille. Dans un premier temps, nous allons travailler sur les données images et examiner ce problème sous l'angle de la compression. Dans un second temps, nous étudierons ce problème sous un autre angle, celui de la recherche d'un espace de représentation parcimonieux. Ce problème peut être traité comme un problème de codage et dans ce cas, on recherchera un codage optimal : le plus économique et avec un minimum de perte. Ces travaux seront ensuite étendus à d'autres modalités de données.

Dans ce travail, nous tenterons de mesurer l'impact de la réduction sur la qualité des résultats de la fouille. Autrement dit, pour une tâche d'apprentissage par exemple, existe-t-il un niveau de détérioration en dessous duquel les résultats se dégradent sensiblement ? Et si oui, peut on identifier ce seuil éventuellement selon les applications ?

Dans le contexte de la réduction par compression et notamment sur des images, on trouve dans la littérature deux types d'approches : compression sans perte ou non dégradante et compression avec perte ou dégradante. La compression avec perte apporte des gains de compression considérables qui peut dépasser les 90% mais ce type de compression peut détruire la qualité des objets images ou autres, ce qui n'est pas permis dans plusieurs domaines (médical, militaire,…). La compression sans perte conserve la qualité mais ce genre de compression aboutit à un gain de compression qui reste toujours faible. Il existe des approches mixtes qui

caractégorisent les objets présents sur un média et opèrent différemment selon la nature de la zone. Par exemple, dans le domaine des images radiographiques, le fond n'est pas porteur d'information, par conséquent, un pré-traitement initial permet d'isoler le fond qui sera codé selon une technique avec perte d'information et le reste sans perte.

Un premier travail bibliographique portera sur les techniques de compression d'image et les techniques de codage notamment par apprentissage automatique. Ces travaux seront testés dans le domaine de la fouille dans les bases de données médicales relatives au dépistage des cancers du sein qui contiennent des radiographies, des échographies, des comptes-rendus en lanagage naturel, etc.

Publications :

Charbel J. Traitement et compression des images numériques appliqués à la mammographies Mémoire de DEA ECD, Université Lyon 2,

Frédéric CLERC

Date de naissance : 2 juillet 1978 à ANNECY (Haute Savoie)

Directeurs de thèse : Nicolas Nicoloyannis, Claude Mirodatos, Ferdi Schüth

Co-directeurs de thèse : Ricco Rakotomalala, David Farrusseng

Financement : CDD CNRS

Date de début de thèse : Septembre 2003

Date de soutenance prévue : Printemps 2006

Titre de la thèse : Méthodes d'optimisation pour l'extraction de connaissances et apprentissage en catalyse hétérogène

Mots clés : optimisation combinatoire, catalyse hétérogène, algorithme génétique, benchmark, hybridation, extraction des connaissances

Résumé de thèse :

Les techniques d'optimisation combinatoire dans le domaine de la catalyse hétérogène ont ouvert un nouveau champ de recherches prometteur. L'application de méthodes informatiques issues de l'apprentissage automatique se prêtent particulièrement bien à la problématique.

La synthèse et les tests de catalyseurs fournissent une masse de données qu'il faut exploiter. L'objectif est, d'une part, de mettre en œuvre des techniques d'extraction de connaissances à partir de cette masse de données associées à des méthodes d'optimisation combinatoire afin de prédire quels catalyseurs sont susceptibles d'être performants pour une réaction donnée. D'autre part, l'ensemble des catalyseurs pour une réaction forme un vaste espace qui, exploré, permettrait de connaître le plus performant. L'exploration de cet espace passe par le positionnement pertinent d'individus : leur diversité doit être contrôlée.

L'extraction des connaissances à partir des données réelles fournies par la catalyse passe par la mise en œuvre d'algorithmes génétiques hybridés avec des systèmes à base de connaissances. L'ensemble des expérimentations se fait via la plateforme de traitement de données Opticat. Les résultats obtenus permettent de dire qu'il est tout à fait possible de fournir une solution au problème posé par la catalyse hétérogène combinatoire.

Plusieurs objectifs ont été atteints durant l'année qui s'est écoulée. Dans un premier temps, la plateforme Opticat est devenue réalité opérationnelle et est conçue de façon à pouvoir accueillir de nouveaux algorithmes de traitement de données rapidement et simplement. Un autre aspect important lié à la plateforme est le modèle de conception multi-tâche, qui permet de considérer des traitements cycliques, ce qu'aucune plateforme de traitement de données ne permet à l'heure

actuelle. Ceux-ci étaient indispensables afin de pouvoir concevoir des algorithmes génétiques à la carte, ce qui est un point crucial. L'utilisation en conditions réelles d'Opticat est validée. En effet, des expérimentations de catalyse ont été réalisées et ont conduit à l'élaboration de benchmarks à partir desquels de nombreux résultats utilisant des algorithmes génétiques classiques (sans système de prédiction) ont été obtenus via Opticat dans le cadre de collaborations européennes [1].

Dans un second temps, le système de prédiction de la valeur des individus KBS tel qu'il a été présenté dans le cadre d'une précédente thèse à été revu pour aboutir a IKBS, *Intelligent Knowledge Based System* puis rendu opérationnel comme un traitement à part entière d'Opticat. La validité de l'approche d'un point de vue optimisation a été démontrée récemment et plusieurs publications sont en cours de rédaction sur le sujet, visant principalement des revues scientifiques orientées sur les techniques d'optimisation combinatoire en catalyse. D'un point de vue data mining l'approche IKBS présente des similitudes avec l'approche de Michalski mais l'originalité de l'application entraîne la nécessité de comparaison avec des méthodes plus classiques (k plus proches voisins, arbres de régression), ce qui est un objectif à court terme. Dans une perspective plus lointaine, le lien entre les applications Stocat (base de données existante) et Opticat sera réalisé et dans le cadre d'un partenariat européen, l'application de ces techniques d'optimisation combinatoire via des systèmes de prédiction sera effectuée dans le champ de la catalyse homogène.

Publications :

S. R. M. Pereira, F. Clerc, D. Farrusseng, J. C. van der Waal, T. Maschmeyer, C. Mirodatos ; "*Effect of the Genetic Algorithm parameters on the optimisation of heterogeneous catalysts"*; QSAR & combinatorial science; Wiley; September 2004

D. Farrusseng, D. Tibiletti, C. Hoffman, A.S. Quiney, S.P. Teh, F. Clerc, M. Lengliz, L. Baumes, C. Mirodatos; "*Discovery of a WGS catalyst for intermediate temperatures by high throughput screening: which respective parts of chance and rationality?"* ; International Congres on Catalysis (ICC), 12-16 July, Paris

F. Clerc, "*Une approche modulaire de conception d'algorithmes génétiques hybrides*", mémoire de DEA, Université Lumière Lyon 2, Juin 2003. actionref dea03fc

O. Boussaid, F. Bentayeb, A. Duffoux, F. Clerc, "*Complex Data Integration based on a Multi-Agent System*", *1st International Conference on Industrial Applications of Holonic and Multi-Agent Systems (HoloMAS 03), Prague, Czech Republic,* September 2003; LNAI, Vol. 2744, 201-212.

actionref holomas03bbdc F. Clerc, A. Duffoux, C. Rose, F. Bentayeb, O. Boussaid, "*SMAIDoC : Un Système Multi-Agents pour l'Intégration des Données Complexes*", *Revue des Nouvelles Technologies de l'Information*, No. 1, 2003, 13-24.

actionref rnti03adgcdrbb F. Clerc, A. Duffoux, C. Rose, F. Bentayeb, O. Boussaid, "*SMAIDoC : Un Système Multi-Agents pour l'Intégration des Données Complexes*", *XXXVèmes Journées de Statistique, Session spéciale Entreposage et Fouille de Données, Lyon*, Juin 2003, 337-340. actionref sfds03cdrbb

Ahmad EL SAYED

Date de naissance : 11/03/1982 (Tripoli, Liban)

Directeur de thèse : D.A. Zighed

Co-directeur de thèse : F. Bentayeb

Financement : Ressources propres

Date de début de thèse : Octobre 2004

Date de soutenance prévue : Décembre 2007

Titre : Recherche d'information par le contenu dans les bases Multimédia

Mots clés : Fouille de données, bases de données complexes, recherche d'information, XML

Résumé de thèse :

Devant le flot des données multimédia, il est devenu nécessaire de mettre au point des modalités d'accès intelligent et rapide au contenu des bases de données complexes, et notamment celles contenant des données images, textuelles, graphiques, etc. En réalité, les méthodes de recherche classiques n'opèrent que sur un type de modalités, soit par une recherche sur les images, soit par une recherche sur les textes mais pas simultanément. Or, la recherche textuelle se heurte à un problème majeur, celui de l'ambiguïté qui peut résulter de la polysémie des mots. Quant à la recherche d'images par le contenu visuel, elle privilégie certaines caractéristiques de bas niveau comme la couleur, la texture, ou la forme. Cette technologie, basée sur la représentation vectorielle en attributs, rend la recherche d'images particulièrement difficile car les caractéristiques visuelles peuvent s'avérer insuffisantes, surtout quand il s'agit de rechercher une notion abstraite. D'où la nécessité de passer des pixels à la sémantique.

L'approche préconisée dans ce contexte, se base sur la construction d'ontologies pour décrire les concepts présents sur les images comme ceux présents dans les textes. Cette ontologie servira de modèle général unique d'indexation du contenu quel qu'il soit. Le domaine d'application principalement visé est celui du dossier médical.

Dans un premier temps, il conviendra d'étudier les modèles d'organisation du dossier médical et notamment la norme DICOM. Nous travaillerons sur les ontologies spécialisées que nous mettrons en place pour déduire le contenu conceptuel des images et des comptes rendus cliniques.

Publications :

El Sayed A., *Recherche Sémantique d'images*, Mémoire de DEA ECD, Université Lyon 2, Sept. 2004, rapport de recherche ERIC.

Walid ERRAY

Date de naissance : 07/04/1977

Directeur de thèse : D.A. Zighed

Financement : Contrat France Télécom (Oct. 2001 à Oct. 2003) ; ATER depuis Oct. 2003

Date de début de thèse : Oct. 2001

Date de soutenance prévue : Oct. 2005

Titre : Extensions et nouvelles approches en graphes d'induction. Application aux grandes bases de données

Mots clés : Apprentissage supervisé, Sélection de variables, Constructions de variables, Réduction d'une table de contingence, Graphes d'induction, Coût asymétrique en apprentissage.

Résumé de thèse :

Cette thèse aborde le problème de l'extraction des connaissances à partir des données volumineuses : grandes dimensions (plusieurs milliers d'attributs) et de nombreuses observations (plusieurs millions d'enregistrements). Ce travail a été mené en collaboration avec la société France Télécom R & D. La thèse qui est en cours de rédaction aborde trois questions fondamentales :

1°) L'amélioration de la qualité de l'espace de représentation à travers deux volets. Le premier a porté sur la réduction de la dimensionnalité de l'espace de représentation. Nous avons étudié et testé un grand nombre de méthodes de sélection d'attributs sur les bases de données d'UCI Irvine. Ces tests ont montré la capacité des méthodes basées sur un critère de consistance, comme ABB et LVF à retrouver l'ensemble optimal de variables dans le cas où les données utilisées ne sont pas bruitées. Nous avons ensuite appliqué ces mêmes techniques sur les données de France Télécom "prédiction des *churns*". Le second volet a porté sur la construction d'attributs où nous avons passé en revue l'ensemble des méthodes existantes.

2°) Les graphes d'induction comme les arbres de décision traitent essentiellement des variables à prédire de nature qualitative. Dans cette thèse nous avons proposé une nouvelle extension. Nous proposons une méthode qui vise à rechercher le regroupement optimal aussi bien sur les modalités (quantitatives ou qualitatives) des attributs prédictifs que de l'attribut à prédire. Nous obtenons ainsi une nouvelle espèce de graphe arborescent baptisé Arbogodaï [ZIG 03a, ZIG 03b, ZIG 04] capable de traiter des données prédictives et à prédire quantitatives et qualitatives. Nous avons également introduit des généralisations sur le calcul d'erreur en apprentissage et en

validation. Ce travail a été précédé d'un autre qui a porté sur la réduction optimale d'une table de contingence [ZIG 02]. Cette nouvelle approche est comparée aux méthodes classiques sur des benchmarks et sur des cas réels en médecine.

3°) En apprentissage supervisé classique, le coût de l'erreur est le même pour toutes les classes. Or dans de nombreux cas, notamment pour les prédiction de churners ou de maladie, l'erreur de prédiction n'a pas les mêmes conséquences. Pour cela, nous avons introduit dans la méthode Arbogodaï la possibilité d'affecter des coûts non symétriques lors de l'apprentissage.

Une plate forme logicielle a été réalisée, elle regroupe les méthodes de sélection dont nos propositions et de nombreux algorithmes d'apprentissage dont Arbogodaï.

Publications :

[ZIG 02], Zighed D.A, Ritschard G., Erray W., Bi dimensional partitioning with limited loss of information, in proceedings of Societa Italiana di Statistica,Milan,pp319-328(2002).

[ZIG 03a], Zighed D.A, Erray W., Scuturici V.M., Arbogodaï : Segmentation généralisée, "SETIT 2003: Sciences of Electronics, Technology of Information and Telecommunications, Mars 17-1, 2003- Sousse, Tunisia.

[ZIG 03b], Zighed D.A, Ritschard G., Erray W. and Scuturici V.M, Abogodaï, A New approach for Decision Trees, in Lavrac, N., D.Gamberger, L. Todorovski and H Blockeel (eds), Knowledge Discovery in Databases : PKDD 2003, LNAI 2838,Berlin:Springer,495--506(2003).

[ZIG 04], Zighed D.A, Ritschard G., Erray W., Scuturici Y-M., Decision tree with optimal join partitioning, International Journal of Intelligent Systems, Wiley (2004).

Edwige FANGSEU BADJIO

Date de naissance : 02 Septembre 1976

Directeur de thèse : Djamel Zighed

Co-directeur de thèse : François Poulet, ESIEA Pôle ECD

Financement : ESIEA

Date de début de thèse : Octobre 2002

Date de soutenance prévue : Fin 2005

Titre de la thèse : Visualisation et fouille de données

Mots clés : visualisation, fouille de données, analyse de la tâche, analyse des utilisateurs, aide aux utilisateurs, mesures statistiques, réduction d'attributs ou de dimensions, SMA, IHM, utilisabilité, utilité, acceptabilité.

Résumé de thèse :

L'extraction de connaissances à partir de données consiste schématiquement à extraire (rendre visibles, compréhensibles) des informations qui sont stockées de manière plus ou moins implicites dans les bases de données. L'ECD utilise la visualisation à la fois pour les données brutes et pour les résultats obtenus par extraction. Dans le premier cas, on s'intéresse plus particulièrement aux structures des données (donc aux capacités humaines dans le domaine de la reconnaissance de formes). Dans le second cas, on vise à rendre compréhensible par le plus grand nombre, les résultats des algorithmes de fouille de données (typiquement une classification des données).

Les domaines connexes à ces travaux de recherche sont aussi variés que : les interfaces homme-machine, la réalité virtuelle, les algorithmes de statistique, analyse de données, apprentissage automatique, intelligence artificielle, la psychologie cognitive (utilisation des résultats des études sur la perception humaine), les bases de données, …

Dans la plupart des cas, c'est un expert en statistique ou analyse de données qui est l'utilisateur du système. Une nouvelle approche du data mining est apparue il y a environ cinq ans, basée, entre autres choses, sur une interprétation graphique des données et appelée le *visual data mining*. La différence fondamentale avec les approches précédentes est le fait que l'utilisateur d'un tel système n'est pas un expert en analyse de données ou statistique, mais l'expert du domaine des données. Ce type d'approche présente notamment comme avantage la possibilité d'utiliser les connaissances du domaine des données lors du déroulement du processus de data-mining (par exemple pour guider ou restreindre la recherche, interpréter des résultats

intermédiaires, etc.). Notre objectif est de rendre cet utilisateur autonome, c'est-à-dire capable d'aboutir lui-même à des modèles et d'interpréter ces modèles. Concrètement, il s'agit d'optimiser sa condition de travail en développant des techniques de réduction de sa charge de travail (support aux différents choix à opérer sur l'environnement de fouille, méthode de visualisation de données, méthode d'analyse de données, réduction ou sélection de dimensions des données), de développer des guides pour la conception de logiciels de fouille visuelle de données de bonne qualité.

Pour ce faire, nous avons effectué une analyse de la tâche en fouille visuelle de données, l'objectif de cette analyse étant d'améliorer l'utilisabilité, l'utilité et l'acceptabilité de ces outils. L'idée tout au long de l'étude de l'utilisabilité est de détecter toutes les erreurs susceptibles de se produire après le développement d'un outil de fouille visuelle de données et de prévoir des approches de solution pour ces erreurs. A partir de cette analyse de la tâche, nous avons défini des critères à prendre en compte pour le développement de logiciels de ce type. Afin de valider ces différents critères et de les proposer comme recommandations, nous avons procédé à des tests utilisateurs des logiciels existants. Nous sommes actuellement en train de dépouiller les premiers résultats et nous affinons aussi les différentes techniques mises en oeuvre pour supporter la tâche de notre utilisateur, techniques s'appuyant sur l'intelligence artificielle en général et plus particulièrement les systèmes multi-agents et quelques mesures statistiques.

Publications :

E.Fangseu Badjio, F.Poulet, *Visual data mining tools: quality metrics definition and application*, to appear in in proc. of ICEIS'05, the 6th International Conference on Enterprise Information Systems, Miami, Florida, USA, May 2005.

E.Fangseu Badjio, F. Poulet, *Towards usable visual data mining environments*, to appear in proc. of HCII'05, the 11th International Conference on Human-Computer Interaction, Las Vegas, Nevada, USA, Jul 2005.

E.Fangseu Badjio, F. Poulet, *Feature Selection: CBR Retrieval Improvement for Knowledge Management*, in proc. of IMT Conference'04, The International Management and Technology Conference, Orlando, Florida, December 8 – 10, 2004.

E.Fangseu Badjio, F. Poulet, *A decision support system for data miners*, in proc. of AISTA'04, The International Conference on Advances in Intelligent Systems - Theory and Applications in cooperation with IEEE, Luxembourg-Kirchberg, Luxembourg, November 15 – 18 2004, ISBN 2-9599776-8-8.

E.Fangseu Badjio, F.Poulet, *Data Mining Algorithm Prediction*, in proc. of IFIP-AIAI'04, The Symposium on Professional Practice in AI, Toulouse, France, August 2004, 383-392, ISBN 2-907801-05-8.

E.Fangseu Badjio, F.Poulet, *Usability of Visual Data Mining Tools*, in proc. of ICEIS'04, the 6th International Conference on Enterprise Information Systems, Porto, Portugal, April 2004, Vol.5, 254-258, ISBN: 972-8865-00-7.

E.Fangseu Badjio, F. Poulet, *Qualité de prédiction des performances des algorithmes de classification de données*, SFC'04, Bordeaux, Sept.2004, pp189-192.

E.Fangseu Badjio, F.Poulet, *Qualité de l'Interaction Homme Machine en Fouille Visuelle de Données,* INFORSID'04, Biarritz, Mai 2004, 542-543, ISBN : 2-906855-20-0.

E.Fangseu Badjio, F.Poulet, *Utilisabilité d'un environnement de fouille de données*, SFC'03, Neuchatel, Suisse, Sept.2003, pp 117-120.

E.Fangseu Badjio, F.Poulet, Définition des spécificités de la fouille visuelle des données pour une évaluation de l'interaction homme machine, in proc. of 3e Atelier Visualisation et Extraction de Connaissances, EGC'05, Paris, 2005, pp 7-14.

E.Fangseu Badjio, F.Poulet, *Guidage des utilisateurs en fouille visuelle de données*, in proc. of 2e Atelier Visualisation et Extraction de Connaissances, EGC'04, Clermont-Ferrand, 2004, pp 13-18.

Cécile FAVRE

Date de naissance : 19/08/1980

Directeur de thèse : Nicolas Nicoloyannis

Responsable scientifique : Fadila Bentayeb

Financement : convention CIFRE en partenariat avec le Crédit Lyonnais

Date de début de thèse : 15 Janvier 2004

Date de soutenance prévue : Janvier 2007

Titre de la thèse : Entrepôt virtuel de données, application aux données bancaires

Mots-clés : intégration, extraction et capitalisation de connaissances, entreposage virtuel, systèmes coopératifs

Résumé de thèse :

Le laboratoire ERIC de l'Université Lumière Lyon 2 a débuté une collaboration avec le Crédit Lyonnais, via une convention CIFRE, qui devra déboucher sur la constitution, la gestion l'exploitation et la valorisation d'un entrepôt virtuel de données bancaires.

L'une des premières étapes de cette thèse a consisté à effectuer une analyse de l'existant sur les demandes de ciblage[1]. Deux principaux constats ont été faits. D'une part, il existe différents types de ciblage et de nouveaux types sont appelés à émerger. D'autre part, la demande et le contexte dans lequel celle-ci est émise sont amenés à évoluer, et il est nécessaire de conserver la trace de ces évolutions. Ceci nous a amené à étudier les travaux de recherche relatifs à la modélisation et au versionnement dans les bases de données.

Nous avons donc, dans un premier temps, défini un modèle des demandes de ciblage qui tienne compte à la fois de l'aspect générique et de l'aspect « version ».

La deuxième étape consiste à analyser (fouille de données, analyse en ligne OLAP) les données bancaires provenant de sources hétérogènes (base des demandes de ciblage, systèmes d'information client et des logiciels dédiés). Dans ce cadre, nous proposons une architecture d'entrepôt virtuel de données, qui permet de construire des cubes de données à la volée. Par ailleurs, étant donné le caractère hétérogène des données sources, nous proposons un système coopératif dans le but d'intégrer ces données dans l'entrepôt virtuel.

La troisième étape a pour but de capitaliser la connaissance extraite à partir des données et de pouvoir ainsi réutiliser celles-ci dans le processus d'aide à la décision.

[1] Les demandes de ciblage permettent de formuler des critères déterminant des populations de clients

Du point de vue du laboratoire ERIC, il s'agit d'acquérir une expertise dans le domaine de l'entreposage virtuel de données hétérogènes, de l'intégration des données basée sur les systèmes coopératifs, à l'analyse, en passant par la construction des cubes de données. Afin de valider nos propositions théoriques, une plateforme logicielle sera développée. Cette plateforme logicielle assurera la gestion et l'exploitation de l'entrepôt de données bancaires et la mise à disposition des outils d'analyse et d'évaluation.

Ainsi, la plateforme logicielle aura un double objectif : d'une part, elle aura pour but de répondre aux besoins exprimés par le Crédit Lyonnais ; d'autre part elle nous permettra d'étudier la performance de notre architecture d'entrepôt virtuel sur des données réelles.

Publication :

C. Favre et F. Bentayeb, "Bitmap index-based decision trees", 15th International Symposium on Methodologies for Intelligent Systems (ISMIS 05), New York, Mai 2005.

C. Favre et F. Bentayeb, "Intégration efficace des arbres de décision dans les SGBD : utilisation des index bitmap", 5èmes Journées d'Extraction et de Gestion des Connaissances (EGC 05), Paris, Janvier 2005; Revue des Nouvelles Technologies de l'Information.

C. Favre, F. Bentayeb, O. Boussaid et N. Nicoloyannis, "Entreposage Virtuel de demandes marketing : de l'acquisition des objets complexes à la capitalisation des connaissances", 2ème atelier Fouille de Données Complexes, EGC 05, Paris, Janvier 2005.

Riadh Ben Messaoud, Kamel Aouiche et Cécile Favre, "Une approche de construction d'espaces de représentation multidimensionnels dédiés à la visualisation", 1ère journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA'05), Lyon, France, Juin 2005, Revue des Nouvelles Technologies de l'Information, Vol. B-1, 34-50.

Rémy GAUDIN

Date de naissance : 01/06/1981

Directeur de thèse : N. Nicoloyannis

Financement : Bourse MENRT

Date de début de thèse : Oct. 2004

Date de soutenance prévue : Oct. 2007

Titre : Contribution méthodologique à l'apprentissage et la fouille de données complexes évolutives

Mots clés : Séries temporelles, Séries multidimensionnelles et hétérogènes, clustering et classification automatique.

Résumé de thèse :

En matière de fouille de données nous sommes très souvent confrontés à des données complexes évoluant dans le temps.

L'objectif de ce travail de thèse est la représentation des différentes problématiques associées à la découverte des structures stables dans des données complexes de type trajectoire, ainsi que la conception d'une nouvelle approche méthodologique d'extraction des connaissances à partir de ce type des données.

La construction des mesures de similarité adaptées aux objets complexes sera l'objet de la première partie de cette recherche. En effet, pour traiter efficacement ce type des données nous devons faire appel aux mesures de similarité ou de distance adaptées, capables à prendre en compte la complexité des objets à étudier et ne pas se contenter aux mesures de distances classiques, efficaces pour le traitement des objets simples.

Après un tour d'horizon de différentes approches existantes et l'étude de l'état de l'art en matière de recherche de patterns stables dans les données complexe nous allons nous intéresser à la conception d'un système capable de traiter efficacement les données évolutives.

L'approche méthodologique que nous adopterons initialement pour la découverte des structures stables est celle de la théorie de l'apprentissage non supervisé et plus particulièrement nous allons s'intéresser à l'approche neuronale basée sur les cartes de Kohonen (SOM : Self Organizing Maps). Une extension de l'approche aux objets complexes « séries temporelles » est envisagée.

Publications :

R. Gaudin et N. Nicoloyannis. Apprentissage non supervisé de séries temporelles à l'aide des k-means et d'une nouvelle méthode d'agrégation de séries. In Proc. of the 5èmes Journées d'Extraction et de Gestion des Connaissances (EGC'05), Paris, France, pages 201–212, 2005.

R. Gaudin, "Apprentissage non supervisé sur les séries temporelles et les objets complexes évolutifs", Mémoire de DEA, Laboratoire ERIC Université Lyon 2, Juin 2004.

Hakim HACID

Date de naissance : 05/03/1979

Directeur de thèse : D.A. Zighed

Co-directeur de thèse :

Financement : Bourse Région Rhône-Alpes

Date de début de thèse : Oct. 2004

Date de soutenance prévue : Oct. 2007

Titre : Un environnement informatique pour l'interrogation et l'accès intelligent aux bases de données complexes

Mots clés : Données complexes, Recherche d'informations, Espace de representation, Voisinage, interrogation par le contenu, données hétérogénes.

Résumé de thèse :

La numérisation du dossier patient est quasi générale. Du cabinet du généraliste à l'hôpital, tous les acteurs de la santé collectent, stockent et échangent de l'information numérique. La consommation médicales restera en croissance pour différentes raisons : vieillissement de la population, allongement de la durée de vie, accroissement du niveau socioculturel, etc. De ce fait, le nombre d'actes médicaux ne peut que croître contribuant ainsi à une augmentation du volume des données collectées.

De nos jours, l'industrie informatique offre des solutions techniquement satisfaisantes et économiquement supportables pour assurer une conservation et un partage de l'information médicale. La baisse des coûts du stockage et de transmission des données va encore amplifier le volume des données disponibles en ligne.

Le défi de la prochaine décennie est la valorisation des données médicales collectées. Accéder aux connaissances cachées dans ce colossal amas de données hétérogènes, distribuées et peu structurées constitue un enjeu à la fois scientifique et technologique majeur. Les retombées de ce défi concerneront les connaissances médicales qui pourraient se développer plus vite et les technologies à très forte valeur ajoutée qui pourraient être mises sur le marché.

Si la technologie de l'Extraction des Connaissances à partir des Données est relativement mûre pour exploiter les bases de données classiques où les données sont sous forme tabulaire (attribut-valeur), elle reste malheureusement démunie face aux structures de données complexes telles que celles du dossier patient.

La problématique de l'accès aux connaissances cachées dans des données complexes n'est pas propre au domaine de la santé. L'environnement, l'industrie ou l'économie regorgent de problèmes similaires. Notre projet se situe dans le domaine de l'extraction des connaissances à partir des données complexes et vise plus particulièrement la mise au point d'un Système Intelligent pour la Recherche d'Information à l'Usage de la Santé (**SIRIUS**), mettant en œuvre les principaux résultats obtenus. Cette plate forme servirait à la fois de terrain d'expérimentation pour tester différents algorithmes de fouille dans les données complexes mais également de cadre de validation des concepts informatiques que nous allons mettre à la disposition des usagers.

En effet, il n'existe aucun système de gestion de base de données, même chez les grands constructeurs comme Oracle ou IBM, qui offre une technologie intelligente d'accès à l'information notamment hétérogènes. Les logiciels disponibles qui offrent un accès à l'information par le contenu sont également limités à un seul type d'information. Par exemple, les logiciels de fouille de texte, ne prennent pas en compte les données images et vice versa. Or, un dossier médical contient des données de nature différentes qu'il convient d'intégrer.

Dans le cadre de cette thèse, nous proposons :

- D'élaborer et de tester des stratégies d'intégration des données multiformes, muti-sources, multi-supports dans le cadre du format XML.
- L'interrogation, la visualisation et la navigation dans les bases de données complexes. Cela suppose des recherches à caractère théorique, notamment sur les mesures de similarités entre objets complexes.
- De mettre au point des algorithmes de fouille de données capables de passer à l'échelle et possédant des mécanismes de réduction de la dimensionnalité.
- De mettre au point un prototype logiciel intégrant toutes ces notions.
- Tester et valider les concepts dans le cadre du dossier médical

Publications :

H. Hacid and A. D. Zighed. An effective method for locally neighborhood graphs updating. Database and Expert Systems Applications, (16): LNCS 3588, pp. 930-939.

H. Hacid and A. D. Zighed. An Incremental Algorithm for neighborhood graphs construction. In the 3rd world conference on Computational Statistics & Data Analysis. Cyprus, October, 2005. (à paraitre).

Hakim Hacid, Djamel Zighed. Neighborhood Graphs for Image Databases Indexing and Content-Based Retrieval. In the First IEEE International Workshop on mining Complex Data (IEEE MCD'05). 2005. Texas, USA. (à paraitre)

Vincent Pisetta , Hakim Hacid , Djamel Zighed. Automatic juridical texts classification and relevance feedback. In the First IEEE International Workshop on mining Complex Data (IEEE MCD'05). 2005. Texas, USA. (à paraitre).

Vincent HUPERTAN

Date de naissance :

Directeur de thèse : Jean-Hugues Chauchat

Financement : par son emploi de chirurgien des hopitaux de Paris

Date de début de thèse : octobre 2001

Date de soutenance prévue : courrant 2005

Titre de la thèse : Extraction de connaissances sur la tolérance à partir des données inter-essais cliniques, dans la recherche thérapeutique

Mots clés : data mining, essais cliniques, tolérance thérapeutique, Overall Safety Assement

Résumé de thèse :

Cette thèse est réalisée en liaison avec les travaux de fin d'étude de nouveaux médicaments dans un grand laboratoire pharmaceutique. Il s'agit de mettre au point une démarche de fouille des données recueillies lors de plusieurs essais thérapeutiques, dits de « Phase IV », c'est-à-dire sur des patients atteints de la maladie visée. Cette phase ultime est lancée après l'étude sur des volontaires sains. La fouille des données vise ici à repérer les éventuels effets secondaires indésirables et les conditions dans lesquels ils se produisent (association avec d'autres médicaments, etc.).

Publications :

1. Hupertan, V., N. Deltour, J.H. Chauchat*, et al.*, *Knowledge discovery on clinical trials to explore the overall safety of the medicinal products. A case study*. International Workshop on Intelligent Data Analysis and Data Mining,. 2004: Zagreb, Croatia. Actes sur : http://lis.irb.hr/IDADM/

2. Hupertan, V., N. Deltour, M. Coste*, et al. Data Mining on clinical trials to explore the safety in a context of an Integrated Safety Sumary (ISS)*. in *International Conference on Statistics in Health Sciences*. 2004. Nantes, France. Actes en préparation

3. Hupertan, V., J.F. Poisson, M. Roupret*, et al. Etude de validation du nomogramme postopératoire de Kattan dans le cancer de rein*. in *Congres A.F.U. Association Française d'Urologie,* 2004. Paris, France.

4. Messas, A., V. Hupertan, J. Ghossein*, et al. The length of the prostate core biopsy involved as a predictor of extra-capsular extension in prostate cancer*. in *EAU 2004*. 2004. Vienne, Autriche

5. Hupertan, V., M. Roupret, J.H. Chauchat*, et al.*, *Value of bootstrapping for small series of patients: application to survival analysis for 26 patients followed for bilateral sporadic renal cell carcinoma.* Prog Urol, 2003. **13**(6): p. 1307-10.

6. Hupertan, V., V. Ravery, J.-H. Chauchat*, et al. Réseaux de neurones, apprentissage automatique ou statistiques dans la construction des nomogrammes*. in *97e Congres Français d'Urologie*. 2003. Paris.

7. Messas, A., V. Hopirtean, V. Ravery, et al. La longueur de tissu biopsique envahie pour prédire l'extension extra prostatique. in 97e Congres Français d'Urologie. 2003. Paris.

Naïma LAZAR

Directeur de thèse : D.A. Zighed

Financement : Ressources propres

Date de début de thèse : Oct. 2004

Date de soutenance prévue : Oct. 2007

Titre : Utilisation des méthodes basées sur les SVM et les noyaux dans le contexte des modèles topologiques : Application au domaine de l'imagerie et l'aide au diagnostique

Mots clés : SVM - modèles topologiques - Graphes de voisinage

Résumé de thèse :

Depuis quelques années, l'apprentissage artificiel a connu une grande révolution, que ce soit à l'échelle des applications mises en œuvre ou des concepts utilisés. Sur le plan des applications, on s'est tourné vers la fouille de très grandes bases de données, avec tous les défis qui lui sont liés, notamment, en termes de complexité du calcul et de gestion de l'espace, de nécessité d'un prétraitement des données et de l'adéquation de la présentation des résultats aux utilisateurs finaux. Sur le plan conceptuel, de nouvelles méthodes ont été mises en place. Les SVM ou les machines à vecteurs de support sont une des méthodes qui se sont très largement répandues.

Nous visons, dans le cadre de cette thèse, l'adaptation de ces méthodes basées sur les SVM et les noyaux aux méthodes topologiques, notamment les graphes de voisinage. Le but étant de rechercher le lien entre la dimension de Vapnik , l'apprenabilité et le test de séparabilité mis au point par le laboratoire ERIC. Le domaine d'application auquel on s'intéressera est le domaine de l'imagerie et l'aide au diagnostic.

Hadj MAHBOUBI

Date de naissance : 17/03/1981
Directeur de thèse : Nicolas Nicoloyannis
Co-directeur de thèse : Jérôme Darmont
Financement : Personnel
Date de début de thèse : Octobre 2005
Date de soutenance prévue : Octobre 2008
Titre de la thèse : *Optimisation des performances des entrepôts XML de données complexes*
Mots clés : Entrepôts de données XML, bases de données natives XML, optimisation des performances des requêtes XML, entreposage de données complexes.

Résumé du projet de thèse :

Les technologies entrant en compte dans les processus décisionnels, comme les entrepôts de données (*data warehouses*), l'analyse multidimensionnelle en ligne (*On-Line Analysis Process*, OLAP) et la fouille de données (*data mining*), sont désormais très efficaces pour traiter des données simples numériques ou symboliques. Cependant, les données exploitées dans le cadre des processus décisionnels sont de plus en plus complexes. L'avènement du Web et la profusion de données multimédia ont en grande partie contribué à l'émergence de cette nouvelle sorte de données.

Bien que l'entreposage de données permette essentiellement le stockage et l'analyse de données numériques et symboliques, les concepts qu'il introduit demeurent valides pour des données complexes. Dans ce contexte, des mesures, bien que non nécessairement numériques, demeurent les indicateurs d'analyse et cette analyse est toujours menée suivant différentes perspectives représentées par des dimensions. Les grandes masses de données à stocker ainsi que leur historisation sont également des arguments en faveur de cette approche. Un entrepôt de données peut également être le socle de différents types d'analyses : statistiques, multidimensionnelles ou de fouille de données.

Une réponse possible à la problématique d'entreposage de données complexes issues du Web s'appuie sur le langage XML, qui permet de représenter aisément des données complexes diverses et peu structurées. Cependant, deux problèmes se posent avant que ce type de solutions soit viable :

les rares entrepôts de données XML existants traitent des données classiques et non des données complexes ;

les systèmes actuels permettant de stocker des documents XML et plus précisément les bases de données natives XML, ne présentent pas des performances suffisantes pour permettre le stockage et l'analyse de grands volumes de données.

L'objectif de cette thèse est de proposer des solutions à ces deux problèmes. Il s'agit dans un premier temps de proposer des modèles multidimensionnels permettant d'entreposer des données complexes dans une base de données native XML ou d'étendre les modèles existants qui traitent des données classiques.

Il sera alors nécessaire de concevoir des techniques d'optimisation de performance adaptées (indexation, matérialisation de vues, partitionnement) pour rendre efficace le traitement de données complexes modélisées de façon multidimensionnelle et stockées sous forme de documents XML. Par ailleurs, il est désormais devenu crucial de réduire la fonction d'administration des systèmes de gestion de bases de données en général et des entrepôts de données en particulier. Les solutions proposées devront donc être autant que possible autoadaptatives.

Nora MAIZ

Date de naissance : 08/04/1979

Directeur de thèse : Professeur Nicolas NICOLOYANNIS

Co-directeur de thèse : Fadila BENTAYEB et Omar BOUSSAID

Financement : Non

Date de début de thèse : Octobre 2005

Date de soutenance prévue : Juin 2008

Titre de la thèse : Médiation basée sur les ontologies pour l'entreposage dynamique de données complexes

Mots clés : Médiation, ontologies, l'entreposage dynamique, données complexes.

Résumé de thèse :

Les grandes entreprises modernes sont dotées d'organisations qui utilisent différents systèmes pour stocker et rechercher les données. La concurrence, l'évolution des technologies, la distribution géographique et la croissance de l'inévitable décentralisation contribuent à cette diversité. Ces systèmes sont conçus indépendamment les uns des autres, avec des modèles et des langages qui sont différents, propriétaires et indépendants. La plupart d'entre eux n'ont pas été créés pour être interopérables. Mais les besoins de l'entreprise incitent ces systèmes hétérogènes à l'être.

Le besoin d'outils de médiation, entre les utilisateurs et les sources de données, dans les entreprises amplifie de plus en plus. Ces outils doivent dépasser les limites des moteurs de recherche actuels en permettant aux utilisateurs de poser des requêtes plus complexes et plus sophistiquées que de simples mots-clés. Ces médiateurs doivent aussi être capables d'agréger des éléments de réponses provenant de différentes sources pour construire une réponse globale à la requête de l'utilisateur. De plus, dans un contexte décisionnel, ces systèmes doivent prendre en charge une série de tâches destinées à la sélection, l'extraction et la mise à disposition des données, pour la construction d'un contexte d'analyse servant à la prise de décision et à la capitalisation des connaissances.

L'interopérabilité entre ces différents systèmes doit être possible à un niveau technique et informationnel. Pour les entreprises, le partage d'informations a non seulement créé les besoins d'accessibilité aux données à partir des différentes sources, mais il exige également le traitement et l'interprétation à distance de ces données par le médiateur. Les problèmes qui pourraient surgir

en raison de l'hétérogénéité des données sont déjà bien connus au sein de la communauté des systèmes de bases de données réparties, par exemple l'hétérogénéité structurelle (schématique) et l'hétérogénéité sémantique (de données).

Il existe plusieurs architectures pour mettre en place les ontologies dans un système d'intégration. Une architecture intéressante est de faire correspondre une ontologie locale à chaque source, et de construire une ontologie globale à partir de ces ontologies. Cependant, ces méthodes sont très générales et ne prennent pas en compte le contexte d'intégration. Il existe quelques approches pour la construction des ontologies dans le cadre de réalisation d'un médiateur. Mais ces approches proposent une démarche de construction descendante, c'est à dire construire l'ontologie globale puis les ontologies locales. Cette démarche ne signifie pas la résolution des problèmes d'hétérogénéité sémantique. L'approche qui nous intéresse consiste à créer l'ontologie globale en aval, et ce, à partir des ontologies locales, ce qui facilite et améliore la réconciliation sémantique entre les ontologies de source.

Plusieurs modèles structuraux peuvent être appliqués à cette architecture. Le modèle GAV (Global As View) suppose l'utilisation des vues au niveau global et simplifie le traitement de la requête. Le traitement de la requête se fait par une simple reformulation ou dépliement. Le modèle LAV (Local As View) suppose l'utilisation des vues au niveau local. Dans ce cas, le traitement de la requête est plus complexe et requière une réécriture. Plus récemment, le modèle GLAV (Generalized Local As View) est apparu. Ce modèle suppose l'utilisation des vues au niveau local et global. Le traitement de requête dans GLAV nécessite une réécriture et un dépliement et n'est pas toujours faisable. Cependant, le traitement de requête dans le cadre de l'architecture avec plusieurs ontologies modélisées selon GLAV est possible, si la requête est exprimée dans un langage qui prend en charge le niveau global et local. Dans ce contexte, nous pouvons proposer un langage de requête basé sur l'ontologie globale et les ontologies locales. Le problème de médiateur, utilisant plusieurs ontologies selon GLAV, est la façon de combiner les résultats obtenus.

En général, les systèmes d'intégration doivent (1) permettre à un utilisateur de poser des requêtes plus complexes que de simples mots clés et (2) être capable d'agréger des éléments de réponse provenant de différentes sources, pour construire une réponse globale à la requête de l'utilisateur.

Un système d'intégration se compose en général d'une ou plusieurs sources de données et d'un médiateur qui facilite l'accès aux données locales et réconcilie les conflits sémantiques entre ces systèmes locaux. Pour réaliser l'interopérabilité sémantique dans un système d'informations hétérogène, il faut que la sémantique des informations échangées soit comprise à travers tout le

système. Les conflits sémantiques se produisent lorsque deux contextes n'emploient pas la même interprétation d'informations. GOH identifie trois causes principales pour l'hétérogénéité sémantique:

- les conflits de confusion (Confounding conflicts) qui se produisent quand les concepts semblent avoir la même signification, mais différent en réalité. Ils sont dus par exemple aux différents contextes temporels;
- les conflits de graduation (Scaling conflicts) qui se produisent lorsque différents systèmes de référence sont employés pour mesurer une valeur. Un exemple est la monnaie.
- les conflits de nom (Naming conflicts) se produisent lors de l'attribution des noms dans des schémas qui diffèrent de manière significative. Un phénomène fréquent est la présence des homonymes et des synonymes.

L'utilisation des ontologies pour l'interprétation de la connaissance implicite et cachée est une approche possible pour surmonter le problème de l'hétérogénéité sémantique. Dans l'un des travaux, les auteurs mentionnent qu'une des principales applications des ontologies est l'interopérabilité. Beaucoup d'approches d'intégration basées sur les ontologies ont été développées afin de réaliser l'interopérabilité.

Il y a beaucoup d'avantages dans l'utilisation des ontologies pour l'intégration de données. L'ontologie fournit un vocabulaire riche et prédéfini qui sert d'interface conceptuelle stable pour l'accès aux bases de données, et qui est indépendante des schémas de base de données. La connaissance représentée par l'ontologie est suffisamment complète pour soutenir la traduction appropriée de toutes les sources d'informations. L'ontologie soutient une gestion conforme et une identification des données contradictoires.

Les objectifs de cette thèse est d'étudier et de proposer des méthodes et des outils pour concevoir des systèmes décisionnels basés sur un entreposage virtuel. Il s'agit de construire des contextes d'analyse (cubes de données) à la demande et de déployer des techniques d'analyse en recourant à la fouille de données. Pour cela, nous dégageons plusieurs pistes de recherche :

- Proposer des modèles d'intégration des données basés sur la médiation et utilisant des ontologies sur lesquels s'appuieront ces systèmes décisionnels.
- Opter pour un langage de représentation des ontologies.
- Proposer enfin un langage de requête basé sur les ontologies et des algorithmes de réécriture de requêtes exprimées dans ce langage.
- Développer de nouveaux opérateurs d'analyse pouvant opérer exploitant les ontologies de ces systèmes décisionnels.

Simon MARCELLIN

Date de naissance : 08/07/1981

Directeur de thèse : D.A. Zighed

Financement : Bourse CIFRE

Date de début de thèse : Oct. 2004

Date de soutenance prévue : Oct. 2007

Titre : Aide à la lecture des mammographies : localisation automatique des régions susceptibles d'être cancéreuses

Mots clés : Images, segmentation, apprentissage, mammographie, détection

Résumé de thèse :

Ce projet de thèse est le fruit d'une collaboration entre le laboratoire ERIC et la société Fenics, le centre anti-cancéreux Léon Bérard à Lyon et un cabinet de radiologie à Clermont-Ferrand : le centre d'imagerie médicale République. L'objectif de la thèse est de détecter automatiquement sur les mammographies les zones considérées par les experts comme étant susceptibles d'être cancéreuses. Ce projet met en jeu plusieurs domaines de compétences, en regroupant des informaticiens, des mathématiciens, des radiologues et des cancérologues. Il s'agit d'une part de mettre en place une méthodologie spécifique pour segmenter ce type d'images, et d'autre part de mettre en œuvre ces méthodes d'apprentissage en les intégrant à la plate forme logicielle développée par l'entreprise.

Cette thèse concerne principalement deux thèmes de recherche :

- L'imagerie, et en particulier la segmentation d'images. Il existe un grand nombre de méthodes permettant de détecter les zones homogènes d'une image. Il faudra présenter une vue d'ensemble complète des méthodes de segmentation, les appliquer, les comparer, et mettre au point une méthode spécifique permettant d'extraire d'une mammographie les zones dangereuses.
- La fouille de données : une fois les objets détectés, il sera nécessaire de les caractériser, c'est-à-dire de les résumer sous la forme d'un vecteur de caractéristiques. On peut calculer plusieurs variables sur les régions d'une image, en terme de forme, de couleur et de texture. L'objectif est d'appliquer les méthodes de Data Mining afin de proposer un modèle

permettant de discriminer les zones susceptibles d'être cancéreuses. Nous étudierons les méthodes de classifications, comme les méthodes de discrimination.

Publications :

Marcelin, S. Extraction d'objets à partir d'objets complexes, Mémoire DEA ECD, Université Lyon 2.

Efthimios MAVRIKAS

Date de naissance :

Directeurs de thèse : N. Nicoloyannis et S.Dascalopoulos (Co-tutelle, Université de l'Egée)

Financement : Bourse de la Grèce

Date de début de thèse : 2002

Date de soutenance prévue : 2005

Résumé de thèse :

Ce projet de thèse recherche une approche sémantique pour la publication et la présence électronique de documents culturels, praticable dans le milieu universitaire et les institutions gérant le patrimoine. Cette approche repose sur l'utilisation d'une ontologie de domaine et de méthodes de traitement semi-automatique du langage naturel et d'analyse du discours pour la classification et l'exploration narrative de documents culturels distribués dans une infrastructure de réseau peer-to-peer.

Le choix d'un modèle informatique participatif comme support à la création d'un tel repositoire sémantique vise à accommoder tout document annoté près de son point d'origine et ses points de consultation sur le réseau, suivant un calcul ad hoc. L'analyse de contenu et de structure de chaque nouveau document introduit en réseau se base sur une ontologie partagée, qui définit les pôles thématiques d'intérêt prenant la forme de clusters suivant une structure d'hypergraphe. Les annotations de contenu de l'ensemble des documents introduits par un certain peer servent à classifier ce peer dans un ou plusieurs clusters et (re)former leurs tables de routage intra-cluster et inter-cluster.

Au fur et à mesure que les clusters s'élargissent et évoluent, les passages représentatifs de leur contenu sont identifiés par un processus MDS (multi-document summarisation) continu, faisant usage des annotations de structure pour générer une narrative intra-cluster. Les narratives intra-cluster de l'ensemble des clusters du réseau sont ensuite combinées en une narrative générale inter-cluster, proposée aux utilisateurs du repositoire comme un outil de progression structurée de la collection dans sa totalité.

Publications :

E.C. Mavrikas, E. Kavakli and N. Nicoloyannis (2004) Ontology-based Narrations from Cultural Heritage Texts, *5th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST 2004)*, Ename, Belgium, December 2004, submitted for review.

E.C. Mavrikas, N. Nicoloyannis and E. Kavakli (2004) Cultural Heritage Information on the Semantic Web, *14th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2004)*, Northamptonshire, UK, October 2004, Springer LNAI, vol. 3257, pp. 477-478.

N. Vernicos, G. Pavlogeorgatos, D.C. Papadopoulos, E. Kavakli, E.C. Mavrikas and S. Bakogianni (2004) FCS_WORD Project: Wiki-based Ongoing Research Data Management, *32nd International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA 2004)*, Prato, Italy, April 2004.

D.C. Papadopoulos and E.C. Mavrikas (2003) Peer-to-Peer Ways to Cultural Heritage, *31st International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA 2003)*, Vienna, Austria, April 2004.

E.C. Mavrikas and N. Vernicos (2002) Cultural Knowledge Management in a Human-centered, Multi-ethnic Environment as a Tool for Improving the Coherence of the NATO Alliance, *11th International Conference and Exhibition*: *NATO Regional Conference on Military Communications and Information Systems (RCMCIS 2002)*, Zegrze, Poland, October 2002.

E.C. Mavrikas (2001) Object Detection Under Heavy Noise Environments Using Statistical Techniques, *MEng Thesis*, Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, University of London, June 2001.

Elie PRUDHOMME

Date de naissance : 23.01.1979

Directeur de thèse : Stéphane Lallich

Co-directeur de thèse :

Financement : Allocation de Recherche

Date de début de thèse : 01.10.2005

Date de soutenance prévue : fin 2008

Titre de la thèse : Fouille de données volumineuses en grandes dimensions

Mots clés : apprentissage, grandes dimensions, données volumineuses, cartes de Kohonen, distances, risque multiple.

Résumé de thèse :

La fouille des données volumineuses en grandes dimensions pose deux types de problèmes :

- des problèmes liés à la complexité algorithmique des méthodes utilisées, en raison du caractère volumineux des données ;
- des problèmes liés au fléau de la dimension, notamment le contrôle du risque multiple, la perte de sélectivité de la distance euclidienne et le phénomène de l'espace vide.

Cette double difficulté affecte les différentes étapes de la procédure de fouille des données :

- le recours à la construction préalable d'un graphe de voisinage issu des prédicteurs, comme outil tout à la fois de navigation et de préparation des données. Dans ce cas, nous avons déjà montré que l'une des solutions possibles consiste à substituer au graphe de voisinage une carte de Kohonen issue des prédicteurs (Prudhomme, Lallich 2005a, 2005b) ; l'utilisation de distances fractionnaires est en cours d'étude.
- la préparation des données : sélection des variables et détection des exemples atypiques, en utilisant une carte de Kohonen ;
- l'apprentissage, pour lequel on privilégiera des algorithmes de complexité linéaire suivant le nombre d'exemples ;
- la validation, en raison de la multiplicité des tests effectués, ce qui pose le problème du contrôle du risque multiple pour lequel nous avons proposé une solution par bootstrap (Lallich, Prudhomme, Teytaud 2004) pour la recherche de règles d'association significatives.

Dans le cadre de cette thèse, on recherchera différentes solutions à ces problèmes, en accordant une place toute particulière à la représentation des données par les cartes de Kohonen.

Publications :

Lallich S., Prudhomme E. et Teytaud O. (2004), Contrôle du risque multiple en sélection de règles d'association significatives, Revue des Nouvelles Technologies de l'Information, RNTI-E-2, vol. 2, pp. 305-316, *actes 4e Conférence Extraction et Gestion des Connaissances 2004*, Clermont-Ferrand.

Prudhomme E. et Lallich S. (2005), Validation statistique des cartes de Kohonen en apprentissage supervisé, Revue des Nouvelles Technologies de l'Information, RNTI-E-3, vol. 1, pp. 79-90, *actes 5e Conférence Extraction et Gestion des Connaissances 2005*, Paris.

Prudhomme E. et Lallich S. (2005), Quality measure based on Kohonen maps for supervised learning of large high dimensional data, *actes Conference International Symposium on Applied Stochastic Models and Data Analysis, ASMDA 2005*, Brest.

Jean-Christian RALAIVAO

Date de naissance : 03/07/1966

Directeurs de thèse : Stéphane LALLICH (ERIC), Victor MANANTSOA (Madagascar)

Co-directeur de thèse : Jérôme DARMONT

Financement : SCAC - Ambassade de France à Madagascar

Date de début de thèse : Octobre 2003

Date de soutenance prévue : Fin 2006

Titre de la thèse : Performance des entrepôts de données complexes

Mots clés : Données complexes, entrepôts de données, XML, performance, métadonnées, connaissances

Résumé de thèse :

Au sein des processus décisionnels, la technologie des entrepôts de données (*data warehouses*) a désormais fait ses preuves pour traiter des données simples numériques ou symboliques. Cependant, diverses sources, dont le Web, présentent des données sous des formes variées et hétérogènes : textes, images, sons, vidéos ou bases de données, données temporelles ou géographiques, exprimées dans différentes langues, stockées dans différents formats en différents endroits et sur différentes plateformes, etc. Ces données, qualifiées de complexes, sont largement porteuses d'information et donc intéressantes à traiter au sein d'un processus décisionnel. Cependant, cela pose de nombreux problèmes de structuration, de stockage et d'interrogation.

L'objectif de cette thèse est d'aborder l'entreposage de données complexes sous l'angle de la performance. Diverses techniques existent pour optimiser le stockage et l'accès à des données simples au sein d'un entrepôt. Cependant, elles ne s'appliquent plus ou mal à des données complexes. Il s'agit donc, d'une part, de définir des modèles d'entrepôts de données complexes adaptés à la nature des données stockées et, d'autre part, de concevoir des outils d'optimisation des performances de ces entrepôts de données complexes : stratégies d'indexation, de matérialisation de vues, de partitionnement, de regroupement, de gestion de cache, etc.

Par ailleurs, l'utilisation du langage XML pour la gestion des entrepôts de données procure plusieurs avantages notamment dans l'intégration des données hétérogènes. XML permet également la représentation à la fois du contenu et de la structure. Il est donc intéressant d'étudier la mise en œuvre d'un entrepôt de données complexes en XML. Cependant, il est primordial d'assurer la performance d'un tel entrepôt. Dans cette optique, l'intégration et l'utilisation

conjointe de metadonnées et de connaissances du domaine dans l'entrepôt doivent permettre de maîtriser la gestion de la complexité des données et surtout des processus d'optimisation de performance de l'entrepôt de données complexes.

Publications :

J.-C. Ralaivao, "Améliorer la performance d'un entrepôt de données complexes par l'utilisation de métadonnées et de connaissances du domaine", 2ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances, EGC 05, Paris, Janvier 2005, 81-84.

J. Darmont, O. Boussaid, J.-C. Ralaivao, K. Aouiche, "An Architecture Framework for Complex Data Warehouses", 7th International Conference on Enterprise Information Systems (ICEIS 05), Miami, USA, May 2005.

Anna STAVRIANOU

Date de naissance : 23/02/1977
Directeur de thèse : Nicolas Nicoloyannis
Co-directeur de thèse : -
Financement : Bourse MRT
Date de début de thèse : Septembre 2005
Date de soutenance prévue : 2008
Titre de la thèse : Passage à l'échelle en fouille de données complexes
Mots clés : Fouille de données complexes, apprentissage, Dimensionnalité

Résumé de thèse :

L'argument sur lequel s'est construite la problématique de l'Extraction de Connaissances à partir des Données (ECD) repose sur un postulat d'exhaustivité des données ou du moins de représentativité des données. Cette vision s'avère cependant impraticable pour des raisons à la fois économiques mais aussi algorithmiques. Prenons le cas de la base de données qui recense les comptes rendus de malades hospitalisés en France (PMSI). Elle contient actuellement 18 millions d'enregistrements alors que la généralisation à toutes les structures de soins n'interviendra qu'en 2006. Exploiter cette base de données de façon exhaustive nécessite une autre approche, plus statistique, qui exploiterait la redondance disponible dans de tels volumes. De plus, cette base de données n'est qu'un échantillon puisqu'elle continue de s'enrichir.

Sur le plan algorithmique, dès lors que l'on considère n objets complexes décrits par p variables ou p concepts selon les modes d'intégration privilégiés, avec $n \geq 10^6$ et $p \geq 10^3$, les algorithmes polynomiaux en $O(n^k)$ avec $k \geq 2$ ne sont plus directement applicables. Seuls résistent ceux qui sont « presque » linéaires en n ou au plus quadratiques en p. Ainsi, si la maîtrise des performances reste incontournable, les limites théoriques nécessitent d'autres approches.

Publications :

User Modelling in a GUI, M. Virvou & A. Stavrianou, 1999, Proceedings of HCI International Conference 99, Vol. 1, Eds. H.-J. Bullinger and J. Ziegler, Lawrence Erlbaum Associate Publishers, Munich (Germany), pp. 262-265.

Julien THOMAS

Date de naissance : 23 février 1982

Directeur de thèse : Nicolas Nicoloyannis

Co-directeur de thèse : ---

Financement : CIFRE (société FENICS-SAS)

Date de début de thèse : 19 septembre 2005

Date de soutenance prévue : fin 2008

Titre de la thèse : Apprentissage automatique de données dynamiques complexes

Mots clés : Apprentissage, données complexes, ECD

Résumé de thèse :

La fouille de données peut être désormais considérée comme un champ scientifique stable possédant de nombreuses méthodes efficientes. Néanmoins le contexte a aujourd'hui fortement évolué, et les connaissances doivent désormais être extraites à partir de données de plus en plus complexes. Des extensions méthodologiques au traitement de données tabulaires classiques ont été proposées pour s'adapter aux nouvelles sources de données, comme par exemple dans le domaine du langage naturel (text mining) et de l'image (image mining). Ces nouveaux ensembles de données fortement hétérogènes et non structurées s'appelle de manière consensuelle données complexes, regroupant à la fois des données usuelles (numériques, qualitatives à valeur discrète), des données moins élémentaires (intervalle, distributions floues, imprécises), des données temporelles, ou encore des données à contenu sémantique riche pour l'humain comme les données à support média. Cette qualification de données complexes est davantage qu'une simple généralisation des familles de données, et l'extraction de connaissances à partir de données complexes nécessite une modélisation spécifique et des méthodes d'accès très avancées. Dès lors la question est de savoir comment combiner des informations de différentes natures au sein d'une même identité sémantique, en tirant parti des spécificités des objets complexes.

Notre objectif est de travailler plus précisément sur la spécificité de dimensionnalité des données et d'utilisation simultanée de différents espaces de représentation. Nous nous attacherons certainement dans un premier temps à des problèmes présentant différents niveaux conceptuels hiérarchiques. La population est définie par des individus simples décrits dans un espace de représentation, puis ceux-ci peuvent être regroupé selon une loi de composition ou un concept, de niveau hiérarchique supérieur, lui-même projeté dans un nouvel espace de représentation et ainsi de suite selon la nature du problème. Un apprentissage classique se fait

alors souvent niveau par niveau, comme des problèmes distincts chacun lié à un unique espace de représentation, risquant une perte de cohérence entre les résultats finaux et les données de départs et des résultats éloignés en termes de performance des souhaits initiaux.

Pour pallier ces difficultés notre axe de recherche principal portera sur l'instauration d'une communication normalisée entre les procédés d'apprentissage travaillant sur chaque niveau, et d'une interaction dynamique entre eux, ceci devant permettre d'exploiter au mieux la complexité des données due à leurs représentations dans différents espaces relatifs à chaque concept. La notion de dynamique employée ne repose donc pas sur une évolution temporelle des données mais bien sur des échanges entre espaces de représentation. Ces recherches devraient permettre d'aboutir au développement d'un méta modèle prenant en compte toutes ces considérations spécifiques, pouvant travailler à l'aide de n'importe quel algorithme classique de fouille de données afin d'optimiser considérablement les résultats aussi bien en terme de cohérence, que de performance, de part la communication dynamique mise en place.

Publications :

Julien Thomas, Simon Marcellin, *«Fouille de bases d'images mammographiques»,* Groupe de Travail sur la Fouille de Données Complexes, Lyon, Septembre 2005.

5. VALORISATION SCIENTIFIQUE

Dans les références des publications, la 1ère lettre désigne le type de publication (ex A pour les ouvrages, …), les lettres suivantes correspondent aux intiales des auteurs et les chiffres à l'année de publication.

5.1. PUBLICATIONS SUR LES CINQ DERNIÈRES ANNÉES

Ouvrages

[ADB05] J. Darmont, O. Boussaïd, *Managing and Processing Complex Data for Decision Support*, Idea Group Publishing, 2005.

[ABGMT05] O. Boussaïd, P. Gançarski, F. Masseglia, B. Trousse, *Fouille de Données Complexes*, *Revue des Nouvelles Technologies de l'Information*, Vol. 3, Cépaduès Editions, 2005.

[ABBDL05] F. Bentayeb, O. Boussaïd, J. Darmont, S. Loudcher, *Actes de la 1ère journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA 05)*, *Revue des Nouvelles Technologies de l'Information*, Vol. B-1, Cépaduès Editions, Juin 2005.

[ABL03] O. Boussaïd, S. Lallich, *Entreposage et Fouille de Données. Numéro spécial*, *Revue des Nouvelles Technologies de l'Information*, Vol. 1, Cépaduès Editions, 2003.

[AHZ02] D. Herin, D. Zighed, *Actes des 2èmes Journées Francophones d'Extraction et de Gestion des Connaissances*, *Extraction de Connaissance et Apprentissage*, Hermès, 2002.

[AHRZK02] M. Hacid, Z. Ras, D. Zighed, Y. Kodratoff, *Foundations of Intelligent Systems*, *LNAI*, Vol. 2366, Springer Verlag, 2002.

Chapitres d'ouvrages

[BD05] J. Darmont, "Object Database Benchmarks", *Encyclopedia of Information Science and Technology*, Vol. 1, Idea Group Publishing, January 2005, 2146-2149.

[BMR05] F. Muhlenbach, R. Rakotomalala, "Discretization of Continuous Attributes", *Encyclopedia of Data Warehousing and Mining*, Idea Group Publishing, 2005, 397-402.

[BHD05b] Z. He, J. Darmont, "Evaluating the Performance of Dynamic Database Applications", *Advanced Topics in Database Research*, Vol. 5, Idea Group Publishing, 2005.

[BBA04] O. Boussaïd, M. Aufaure, "Spatial Data Warehouses: a methodological framework", *Advances in Spatial Analysis and Decision Making*, A.A. Balkema Publishers, 2004, 275-282.

[BDBBRZ03] J. Darmont, O. Boussaïd, F. Bentayeb, S. Rabaseda, Y. Zellouf, "Web multiform data structuring for warehousing", *Multimedia Systems and Applications*, Vol. 22, Kluwer Academic Publishers, 2003, 179-194.

[BDS02] J. Darmont, M. Schneider, "Object-Oriented Database Benchmarks", *Advanced Topics in Database Research*, Vol. 1, Idea Group Publishing, 2002, 34-57.

[BZR02] D. Zighed, R. Rakotomalala, "Data Mining", *Techniques de l'ingénieur*, Vol. H3 744, Editions Techniques de l'Ingénieur, 2002, 1-26.

[BZR02b] D. Zighed, R. Rakotomalala, "Graphes d'induction : apprentissage et data mining", *Bases de Données et Statistique*, Dunod, 2002, 98-124.

[BRZN01b] G. Ritschard, D. Zighed, N. Nicoloyannis, "Maximiser l'association par agrégation dans un tableau croisé", *La fouille dans les données par la méthode d'analyse statistique implicative*, Ecole Polytechnique de l'Université de Nantes, IRI, 2001, 219-233.

[BCR01] J. Chauchat, R. Rakotomalala, "Sampling Strategy for Building Decision Trees from Very Large Databases Comprising Many Continuous Attributes", *Instance Selection and Construction-A Data Mining Perspective*, Kluwer Academic Publishers, 2001, 179-188.

Revues internationales

[CZRES05] D. Zighed, G. Ritschard, W. Erray, V. Scuturici, "Decision tree with optimal join partitioning", *Journal of Intelligent Information Systems*, Vol. 20, 2005, 1-26.

[CZLM05] D. Zighed, S. Lallich, F. Muhlenbach, "A statistical approach of class separability", *Applied Stochastic Models in Business and Industry*, Vol. 21, No. 2, 2005, 187-197.

[CHD05] Z. He, J. Darmont, "Evaluating the Dynamic Behavior of Database Applications", *Journal of Database Management*, Vol. 16, No. 2, April-June 2005, 21-45.

[CSCSZ05] M. Scuturici, J. Clech, V. Scuturici, D. Zighed, "Topological representation model for image databases query", *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 17, No. 1-2, 2005, 145-160.

[CLN05] G. Legrand, N. Nicoloyannis, "Feature Selection Method Using Preferences Aggregation", *LNCS*, Vol. 3587, 2005, 203-217.

[CMLZ04] F. Muhlenbach, S. Lallich, D. Zighed, "Identifying and Handling Mislabelled Instances", *Journal of Intelligent Information Systems*, Vol. 22, No. 1, January 2004, 89-109.

[CGVLF04] J. Gaillard, A. Viallefont, A. Loison, M. Festa-Bianchet, "Assessing senescence patterns in populations of large mammals", *Animal Biodiversity and Conservation*, Vol. 27, No. 1, 2004, 47-58.

[CGVCCM04] O. Gimenez, A. Viallefont, E. Catchpole, R. Choquet, B. Morgan, "Methods for investigating parameter redundancy", *Animal Biodiversity and Conservation*, Vol. 27, No. 1, 2004, 561-572.

[CBFLM04] L. Baumes, D. Farrusseng, M. Lengliz, C. Mirodatos, "Using Artificial Neural Networks for boosting discovery in High", *QSAR & Combinatorial Science*, 2004.

[CKFBMS04] C. Klanner, D. Farrusseng, L. Baumes, C. Mirodatos, F. Schüth, "The Development of Descriptors for Solids: Teaching "Catalytic", *Angewandte Chemie International Edition*, Vol. 43, No. 40, 2004, 5347-5349.

[CPCFWMM04] S. Pereira, F. Clerc, D. Farrusseng, J. Waal, T. Maschmeyer, C. Mirodatos, "Effect of the Genetic Algorithm parameters on the optimisation of heterogeneous catalysts", *QSAR & Combinatorial Science*, September 2004.

[CSCZ04] M. Scuturici, J. Clech, D. Zighed, "Topological Query in Image Databases", *LNCS*, Vol. 2905, 2004, 145-152.

[CLMJ03] S. Lallich, F. Muhlenbach, J. Jolion, "A test to control a region growing process within a hierarchical graph", *Pattern Recognition*, Vol. 36, No. 10, 2003, 2201-2211.

[CKFBMS03] C. Klanner, D. Farrusseng, L. Baumes, C. Mirodatos, F. Schüth, "How to Design Diverse Libraries of Solid Catalysts?", *QSAR & Combinatorial Science*, Vol. 22, 2003, 729-736.

[CSNL02] M. Sebban, R. Nock, S. Lallich, "Stopping criterion for boosting-based data reduction technics : from binary to multiclass problems", *Journal of Machine Learning Research*, Vol. 3, 2002, 863-885.

[CCMV02] E. Catchpole, B. Morgan, A. Viallefont, "Solving problems in parameter redundancy using computer algebra", *Journal of Applied Statistics*, Vol. 29, No. 1-4, 2002, 625-636.

[CGCSND02] P. Godard, P. Chanez, L. Siraudin, N. Nicoloyannis, G. Duru, "Costs of asthma are correlated with severity : a 1-yr prospective study", *European Respiratory Journal*, Vol. 19, No. 1, 2002.

[CDABLPN02] G. Duru, J. Auray, A. Beresniak, M. Lamure, A. Paine, N. Nicoloyannis, "Limitations of the methods used for calculating Quality-Adjusted Life-Year values", *Pharmacoeconomics*, Vol. 20, No. 7, 2002, 463-473.

[CSNCR01] M. Sebban, R. Nock, J. Chauchat, R. Rakotomalala, "Impact of Learning Set Quality and Size on Detection Tree Performances: a Comparative Study", *International Journal of Computers, Systems and Signals (IJCSS)*, Vol. 1, No. 1, 2001, 85-105.

[CFT01] F. Feschet, L. Tougne, "Discrete Waves on Cellular Automata", *International Journal on Pattern Recognition and Artificial Intelligence*, Vol. 15, No. 7, 2001, 1007-1021.

Conférences internationales avec comité de lecture et actes

[DLVL05] S. Lallich, B. Vaillant, P. Lenca, "Parametrised measures for the evaluation of association rule interestingness", *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005), Brest, France*, 2005, 220-229.

[DTBB05] A. Tanasescu, O. Boussaïd, F. Bentayeb, "Preparing Complex Data for Warehousing", *3rd ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 05), Cairo, Egypt*, January 2005.

[DHZ05] H. Hacid, D.A. Zighed, "An Effective Method For Locally Neighborhood Graphs Updating", *16th International Conference on Database and Expert Systems Applications (DEXA 05),* 2005; *LNCS*, Vol. 3588, 930-939.

[DPL05] E. Prudhomme, S. Lallich, "Quality measure based on Kohonen maps for supervised learning of large high dimensional data", *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005), Brest, France*, 2005, 246-255.

[DFB05] C. Favre, F. Bentayeb, "Bitmap index-based decision trees", *15th International Symposium on Methodologies for Intelligent Systems (ISMIS 05), New York, USA*, May 2005.

[DDBRA05] J. Darmont, O. Boussaïd, J. Ralaivao, K. Aouiche, "An Architecture Framework for Complex Data Warehouses", *7th International Conference on Enterprise Information Systems (ICEIS 05), Miami, USA*, May 2005, 370-373.

[DADBB05] K. Aouiche, J. Darmont, O. Boussaïd, F. Bentayeb, "Automatic Selection of Bitmap Join Indexes in Data Warehouses", *7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 05), Copenhagen, Denmark*, August 2005; *LNCS*, Vol. 3589, 64-73.

[DDBB05] J. Darmont, F. Bentayeb, O. Boussaïd, "DWEB: A Data Warehouse Engineering Benchmark", *7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 05), Copenhagen, Denmark*, August 2005; *LNCS*, Vol. 3589, 85-94.

[DRME05] R. Rakotomalala, F. Mhamdi, M. Elloumi, "Hybrid Feature Ranking for Protein Classification", *1st International Conference on Advanced Data Mining and Applications (ADMA'05)*, 2005; *LNAI*, Vol. 3584, 610-617.

[DMRE05] F. Mhamdi, R. Rakotomalala, M. Elloumi, "Feature Ranking for Protein Classification", *4th International Conference on Computer Recognition Systems (CORES'05)*, 2005; *Advances in Soft Computing*, 611-617.

[DCRF05] F. Clerc, R. Rakotomalala, D. Farrusseng, "Learning Fitness Function in a Combinatorial Optimization Process", *International Symposium on Applied Stochastic Models and Data Analysis*, 2005, 535-543.

[DLN05b] G. Legrand, N. Nicoloyannis, "A new feature selection method", *8th International Conference on Pattern Recognition and Information Processing (PRIP05), Minsk Belarus*, 2005.

[DLN05c] G. Legrand, N. Nicoloyannis, "Feature selection and preferences aggregation", *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005), Brest, France*, 2005, 305-312.

[DLN05d] G. Legrand, N. Nicoloyannis, "Feature selection method using preferences aggregation", *International Conference on Machine Learning and Data Mining (MLDM05), Leipzig Germany*, 2005, 9-11.

[DCPC05] J. Chauchat, M. Pacaut-Troncin, A. Cuerq, "Model Assessment and Selection : a Case Study on Risk Factors for Acute Suicidality in Psychiatric Patients", *Applied Statistics, Ribno (Bled), Slovenia*, 2005.

[DJCR04] R. Jalam, J. Clech, R. Rakotomalala, "Un cadre pour la catégorisation de textes multilingues", *7èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT 04), Louvain-la-Neuve, Belgique*, 2004, 650-660.

[DBDU04] F. Bentayeb, J. Darmont, C. Udréa, "Efficient Integration of Data Mining Techniques in Database Management Systems", *8th International Database Engineering and Applications Symposium (IDEAS 04), Coimbra, Portugal*, July 2004, 59-67.

[DBRBB04] R. BenMessaoud, S. Rabaseda, O. Boussaïd, F. Bentayeb, "OpAC: A New OLAP Operator Based on a Data Mining Method", *Sixth International Baltic Conference on Databases and Information Systems (DB&IS 04), Riga, Latvia*, June 2004.

[DJCD04] R. Jalam, J. Chauchat, J. Dumais, "Automatic Recognitiion of Keywords using N-grams", *16th Symposium of IASC (COMPSTAT 04), Prague, Czech Republic*, August 2004, 1245-1254.

[DTBB04] A. Tanasescu, O. Boussaïd, F. Bentayeb, "Towards Complex Data Warehousing: A new approach for integrating and modeling Complex data", *5th International Conference on Modelling,*

Computation and Optimization in Information Systems and Management Sciences (MCO 04), Metz, France, July 2004, 619-626.

[DBBR04] R. BenMessaoud, O. Boussaïd, S. Rabaseda, "A New OLAP Aggregation Based on the AHC Technique", *ACM 7th International Workshop on Data Warehousing and OLAP (DOLAP 04), Washington DC, USA*, November 2004, 65-72.

[DMLZ04b] F. Muhlenbach, S. Lallich, D. Zighed, "Outlier Handling in the Neighbourhood-Based Learning of a Continuous Class", *7th International Conference Discovery Science, Padova, Italy*, October 2004; *LNAI*, Vol. 3245, 314-321.

[DVLL04] B. Vaillant, P. Lenca, S. Lallich, "A clustering of interestingness measures", *7th International Conference Discovery Science, Padova, Italy*, October 2004; *LNAI*, Vol. 3245, 290-297.

[DMNK04] E. Mavrikas, N. Nicoloyannis, E. Kavakli, "Cultural Heritage Information on the Semantic Web", *14th International Conference on Knowledge Engineering and Knowledge Management (EKAW 04), Northamptonshire, UK*, October 2004; *LNAI*, Vol. 3257, 477-478.

[DHC04] V. Hopirtean, J. Chauchat, "Knowledge Discovery on Clinical Trials to Explore the Overall Safety of the Medical Products - A case study", *International Workshop on Intelligent Data Analysis and Data Mining, Application in Medecine (SRCE), Zagreb, Croatia*, June 2004.

[DMKN04] E. Mavrikas, E. Kavakli, N. Nicoloyannis, "Ontology-based Narrations from Cultural Heritage Texts", *5th International Symposium on Virtual Reality Archaeology and Cultural Heritage (VAST 2004), Ename, Belgium*, December 2004.

[DVPPKMB04] N. Vernicos, G. Pavlogeorgatos, D. Papadopoulos, E. Kavakli, E. Mavrikas, S. Bakogianni, "FCS_WORD Project : Wiki-based Ongoing Research Data Management", *32nd International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA 2004), Pratoy, Italy*, April 2004.

[DMER04] F. Mhamdi, M. Elloumi, R. Rakotomalala, "Textmining, feature selection and datamining for proteins classification", *2nd International Conference on Information and Communication Technologies (ICICT 04), Cairo, Egypt*, 2004, 457-458.

[DMER04b] F. Mhamdi, M. Elloumi, R. Rakotomalala, "Descriptors Extraction for proteins classification", *3rd Conference on Neuro-Computing and Evolving Intelligence (NCEI 04), Auckland, New Zealand*, December 2004.

[DBBD03] F. Bentayeb, O. Boussaïd, J. Darmont, "Multi-Link Lists as Data Cube Structures in the MOLAP Environment", *14th IRMA International Conference, Philadelphia, USA*, May 2003, 35-37.

[DADG03] K. Aouiche, J. Darmont, L. Gruenwald, "Frequent itemsets mining for database auto-administration", *7th International Database Engineering and Application Symposium (IDEAS 03), Hong Kong, China*, July 2003, 98-103.

[DBBD03b] O. Boussaïd, F. Bentayeb, J. Darmont, "A Multi-Agent System-Based ETL Approach for Complex Data", *10th ISPE International Conference on Concurrent Engineering: Research and Applications (CE 03), Madeira, Portugal*, July 2003, 49-52.

[DHD03] Z. He, J. Darmont, "DOEF: A Dynamic Object Evaluation Framework", *14th International Conference on Database and Expert Systems Applications (DEXA 03), Prague, Czech Republic*, September 2003; *LNCS*, Vol. 2736, 662-671.

[DBBDC03] O. Boussaïd, F. Bentayeb, A. Duffoux, F. Clerc, "Complex Data Integration based on a Multi-Agent System", *1st International Conference on Industrial Applications of Holonic and Multi-Agent Systems (HoloMAS 03), Prague, Czech Republic*, September 2003; *LNAI*, Vol. 2744, 201-212.

[DABBD03] K. Aouiche, F. Bentayeb, O. Boussaïd, J. Darmont, "Conception informatique d'une base de données multimédia de corpus linguistiques oraux : l'exemple de CLAPI 2", *36ème Colloque International de la Societas Linguistica Europaea, Lyon, France*, Septembre 2003, 11-12.

[DSCZ03] M. Scuturici, J. Clech, D. Zighed, "Topological Query in Image Databases", *8th Ibero-American Congres on Pattern Recognition (CIARP 03), Havana, Cuba*, November 2003, 144-151.

[DZRES03] D. Zighed, G. Ritschard, W. Erray, V. Scuturici, "Abogodaï, a New approach for Decision Trees", *7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 03), Dubrovnik, Croatia*, September 2003; *LNAI*, Vol. 2838, 495-506.

[DRZ03] G. Ritschard, D. Zighed, "Goodness-of-Fit Measures for Induction Trees", *14th International Symposium on Methodologies for Intelligent Systems (ISMIS 03), Maebashi, Japan*, October 2003; *LNAI*, Vol. 2871, 57-64.

[DRZ03b] G. Ritschard, D. Zighed, "Simultaneous Row and Column Partitionning : Evaluation of a heuristic", *14th International Symposium on Methodologies for Intelligent Systems (ISMIS 03), Maebashi, Japan*, October 2003; *LNAI*, Vol. 2871, 468-472.

[DBJFLNM03] L. Baumes, P. Jouve, D. Farrusseng, M. Lengliz, N. Nicoloyannis, C. Mirodatos, " Dynamic control of the browsing-exploitation ratio for iterative", *7th International Conference on Knoqledge-Based Intelligent Information & Engineering Systems (KES'03)*, September 2003; *LNCS*, Vol. 2773, 265-270.

[DBFMGCKHB03] L. Baumes, D. Farrusseng, C. Mirodatos, G. Grubert, L. Cholinska, S. Kolf, M. Holena, M. Baerns, "Neural Networks-based techniques for boosting high-throughput", *EuropCat-VI*, September 2003.

[DHRCH03] V. Hopirtean, M. Rouprêt, J. Chauchat, B. Hubert, "Concept Descritpion and Unsupervised Classification as Datamining Types to Analyse the Circadian Variation of Blood Pressur Mesuare by Ambulatory Blood Pressure Monitoring (ABPM)", *23rd Annual Conference of International Society for Clinical Biostatistics, Dijon, France*, 2003.

[DJN03] P. Jouve, N. Nicoloyannis, "KEROUAC : an Algorithm for Clustering Categorical Data Sets with Practical Advantages", *International Workshop on Data Mining for Actionable Knowledge (DMAK'2003, in conjunction with PAKDD03)*, 2003.

[DJN03b] P. Jouve, N. Nicoloyannis, "A Method for Aggregating Partitions, Applications in Knowledge Discovery in Databases", *In Advances in Knowledge Discovery and Data Mining, 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD03). Seoul, Korea*, 2003; *LNCS*, Vol. 2637, 411-422.

[DJN03c] P. Jouve, N. Nicoloyannis, "The 'Who is it ?' Problem, Application for customizable Web Sites", *In Web Intelligence, First International Atlantic Web Intelligence Conference (AWIC'03), Madrid, Spain*, 2003; *LNCS*, Vol. 2663, 83-93.

[DJN03d] P. Jouve, N. Nicoloyannis, "A New Method for Combining Partitions, Applications for Distributed Clustering", *International Workshop on Paralell and Distributed Machine Learning and Data Mining (ECML/PKDD03)*, 2003, 35-46.

[DJN03e] P. Jouve, N. Nicoloyannis, "Classification Non Supervisée pour Données Catégorielles", *XXXVèmes Journées de Statistique, Session spéciale Entreposage et Fouille de Données, Lyon*, 2003; *Revue des Nouvelles Technologies de l'Information*, 87-98.

[DC03] J. Chauchat, "Teachning Statistical Inference Using Many Samples from a Real Large Dataset (invited paper)", *54ème Congrès de l'Institut International de Statistique, Berlin, Allemagne*, 2003.

[DPM03] D. Papadopoulos, E. Mavrikas, "Peer-to-Peer Ways to Cultural Heritage", *31st International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA 2003), Vienna, Austria*, April 2003.

[DZES03] D. Zighed, W. Erray, V. Scuturici, "Arbogodaï : segmentation généralisée", *Sciences of Electronics, Technology of Information and Telecommunications (SETIT2003), Sousse, Tunisia*, 2003.

[DDBB02] J. Darmont, O. Boussaïd, F. Bentayeb, "Warehousing Web Data", *4th International Conference on Information Integration and Web-based Applications and Services (iiWAS 02), Bandung, Indonesia*, September 2002, 148-152.

[DBD02] F. Bentayeb, J. Darmont, "Decision tree modeling with relational views", *XIIIth International Symposium on Methodologies for Intelligent Systems (ISMIS 2002), Lyon, France*, June 2002; *LNAI*, Vol. 2366, 423-431.

[DJC02] R. Jalam, J. Chauchat, "Pourquoi les n-grammes permettent de classer des textes ? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques", *6èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT 02), St. Malo, France*, March 2002; *Lexicometrica*, Vol. 1, 381-390.

[DHRCH02] V. Hopirtean, M. Rouprêt, J. Chauchat, B. Hubert, "Databases Structures for Data Mining in Therapeutic Research", *Conference for Computational Statistics (COMPSTAT'02), Germany*, 2002.

[DBFNM02] L. Baumes, D. Farrusseng, N. Nicoloyannis, C. Mirodatos, "Advanced Data Management for Combinatorial Heterogeneous Catalysis", *European Workshop on Combinatorial Catalysis (EUROCOMBICAT02). Ishia, Italy*, 2002.

[DKBFS02] C. Klanner, L. Baumes, D. Farrusseng, F. Schüth, "Evaluation of Descriptors for Solid Catalysts", *European Workshop on Combinatorial Catalysis (EUROCOMBICAT02). Ishia, Italy*, 2002.

[DBJNF02] L. Baumes, P. Jouve, N. Nicoloyannis, D. Farrusseng, "Hybrid Algorithm for Iterative Optimization", *International Symposium on Combinatorial Optimization (CO02), Paris-France*, April 2002.

[DZRE02] D. Zighed, G. Ritschard, W. Erray, "Bi dimensional partitioning with limited loss of information", *Societa Italiana di Statistica, Milan*, 2002, 319-328.

[DLMZ02] S. Lallich, F. Muhlenbach, D. Zighed, "Improving classification by removing or relabeling mislabeled instances", *Foundations of Intelligent Systems, 13th International Symposium on Methodologies for Intelligent Systems (ISMIS 2002), Lyon, France*, June 2002; *LNAI*, Vol. 2366, 5-15.

[DZLM02] D. Zighed, S. Lallich, F. Muhlenbach, "Separability Index in Supervised Learning", *Principles of Data mining and Knowledge Discovery, 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02), Helsinki, Finland*, August 2002; *LNAI*, Vol. 2431, 475-487.

[DC02] D. Coeurjolly, "Visibility in Discrete Geometry: an application to discrete geodesic paths", *Discrete Geometry for Computer Imagery, 10th International Conference (DGCI 2002)*, 2002; *LNCS*, Vol. 2301, 326-327.

[DCFTT02] D. Coeurjolly, F. Flin, O. Teytaud, L. Tougne, "Multigrid convergence and Surface Area Estimation", *Theoretical Foundations of Computer Vision, Geometry, Morphology, and Computational Imaging*, 2002; *LNCS*.

[DSC02] I. Sivignon, D. Coeurjolly, "From digital plane segmentation to polyhedral representation", *Theoretical Foundations of Computer Vision, Geometry, Morphology, and Computational Imaging*, 2002; *LNCS*.

[DMR02] F. Muhlenbach, R. Rakotomalala, "Multivariate supervized discretization, a neighborhood graph approach", *3rd International Conference on Data mining (ICDM'02)*, 2002, 314-321.

[DBFNM02b] L. Baumes, D. Farrusseng, N. Nicoloyannis, C. Mirodatos, "Advanced data management for combinatorial heterogeneous catalysis", *Eurocombicat 2002, European workshop on Combinatorial Analysis, Ischia Italy*, 2002.

[DJC02b] R. Jalam, J. Chauchat, "Pourquoi les n-grammes permettent de classer des textes ? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques", *6th International Conference on Textual Data Statistical Analysis, France*, 2002, 381-390.

[DTS02] T. Tweed, A. Saadane, "On the contribution of different spatial pooling models to the performance of a perceptual image quality metric", *Advanced Concepts for Intelligent Vision Systems (ACIVS 2002), Ghent, Belgique*, 2002.

[DTM02] T. Tweed, S. Miguet, "Automatic detection of regions of interest in mammographies based on a combined analysis of texture and histogram", *International Conference on Pattern Recognition (ICPR 2002), Québec City, Canada*, 2002.

[DPCR02] F. Pellegrino, J. Chauchat, R. Rakotomalala, "Can Automatically Extracted Rhythmic Units Discriminate among Languages?", *Speech Prosody*, 2002, 562-565.

[DCRP02] J. Chauchat, R. Rakotomalala, F. Pellegrino, "Modelling Survey Data for Social and Economic Research : Accuracy Estimation with Clustered Dataset", *XIX International Methodology Symposium of Statisctics Canada ; Modelling Survey Data for Social and Economic Research, Ottawa*, 2002.

[DC02b] J. Chauchat, "Une application de la fouille de données, au marketing : construction et mise à jour d'une segmentation de clients", *Premier colloque franco-libanais de Statistique et Analyse des Données, Beyrouth*, 2002.

[DC02c] J. Chauchat, "Survey Sampling : Learning by Doing. A twenty years graduate level teaching experience", *Invited Paper ICOTS'2002, International Conference on Teaching Statistics, Durban, South Africa*, 2002.

[DMV02] E. Mavrikas, N. Vernicos, "Cultural Knowledge Management in a Human-centered, Multi-ethnic Environment as a Tool for Improving the Coherence of the NATO Alliance", *11th International Conference on Exhibition : NATO Regional Conference on Military Communications and Information Systems (RCMCIS 2002), Zegrze, Poland*, October 2002.

[DMDB01] S. Miniaoui, J. Darmont, O. Boussaïd, "Web data modeling for integration in data warehouses", *First International Workshop on Multimedia Data and Document Engineering (MDDE 01), Lyon, France*, July 2001, 88-97.

[DJT01] R. Jalam, O. Teytaud, "Kernel based text categorization", *12h International Joint Conference on Neural Networks (IJCNN 01), Washington, USA*, 2001, 1891-1896.

[DCS01] S. Clippe, D. Sarrut, "Patient positioning in radiotherapy", *92nd Annual Meeting - American Association for Cancer Research*, 2001.

[DCMT01] D. Coeurjolly, S. Miguet, L. Tougne, "Discrete Curvature based on Osculating Circle Estimation", *4th international workshop on visual form (IWVF4'2001)*, 2001; *LNCS*, Vol. 2059, 303-312.

[DGT01] G. Gavin, O. Teytaud, "Equivalence in the worst case between the training error estimate and the Leave-One-Out estimate", *International Joint Conference on Neural Networks (IJCNN)*, 2001, 1238-1243.

[DTJ01] O. Teytaud, R. Jalam, "Kernel based text categorization", *12th International Joint Conference on Neural Networks (IJCNN), Washington, US*, 2001, 1892-1897.

[DTS01] O. Teytaud, D. Sarrut, "Convergence speed of deformable models", *12th International Joint Conference on Neural Networks (IJCNN), Washington, US*, 2001, 2850-2855.

[DSNL01] M. Sebban, R. Nock, S. Lallich, "Boosting Neighborhood-Based Classifiers", *The Eighteenth International Conference on Machine Learning – ICML'2001, Massachusetts*, 2001, 505-512.

[DTS01b] O. Teytaud, D. Sarrut, "Kernel Based Image Classification", *International Conference on Artificial Neural Networks (ICANN 2001), Vienna, Austria*, 2001, 269-375.

[DCGRT01] D. Coeurjolly, Y. Gérard, J. Reveilles, L. Tougne, "An elementary algorithm for digital arc segmentation", *International Workshop on Combinatorial Image Analysis (IWCIA'2001)*, 2001; *Electronic Notes in Theoritical Computer Science*, Vol. 46.

[DSSMP01] V. Scuturici, M. Scuturici, S. Miguet, J. Pinon, "Measuring Web Servers Performance in VoD", *International Conference on Telecommunications (ICT2001), Romania, Bucharest*, 2001.

[DCDT01] D. Coeurjolly, I. Debled-Rennesson, O. Teytaud, "Segmentation and Length Estimation of 3D Discrete Curves", *Digital and Image Geometry*, 2001; *LNCS*, Vol. 2243, 295-313.

[DC01] D. Sarrut, S. Clippe, "Geometrical transformation approximation for 2D/3D intensity-based registration of portal images and CT scan", *Medical Image Computing and Computer- Assisted Intervention (MICCAI'2001), Utrecht (Netherlands)*, 2001; *LNCS*, Vol. 2208, 532-540.

[DCRPC01] J. Chauchat, R. Rakotomalala, C. Pelletier, M. Carloz, "Targeting Groups Using Gain and Cost Matrix; a Marketing Application", *"Datamining and Marketing" Whorkshop, 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01), Fribourg, Allemagne*, 2001, 1-14.

Revues nationales

[EADBB05b] K. Aouiche, J. Darmont, O. Boussaïd, F. Bentayeb, "Auto-administration des entrepôts de données complexes", *Revue des Nouvelles Technologies de l'Information*, Vol. 3, 2005.

[ER05] R. Rakotomalala, "TANAGRA, une plate-forme d'expérimentation pour la fouille de données", *MODULAD*, No. 32, 2005, 70-85.

[ER05b] R. Rakotomalala, "TANAGRA, une plateforme d'expérimentation pour la fouille de données", *MODULAD*, Vol. 32, 2005, 70-85.

[ELMVPL04] P. Lenca, P. Meyer, B. Vaillant, P. Picouet, S. Lallich, "Evaluation et analyse multicritère des mesures de qualité des règles d'association", *Revue des Nouvelles Technologies de l'Information*, No. 2, 2004, 219-246.

[ELT04] S. Lallich, O. Teytaud, "Evaluation et validation de l'intérêt des règles d'association", *Revue des Nouvelles Technologies de l'Information*, No. 2, 2004.

[EHD04] Z. He, J. Darmont, "Une plate-forme dynamique pour l'évaluation des performances des bases de données à objets", *Ingénierie des Systèmes d'Information (RSTI série ISI)*, Vol. 9, No. 1, 2004, 109-127.

[ELN04] G. Legrand, N. Nicoloyannis, "Sélection de variables et agrégation d'opinions", *Revue des Nouvelles Technologies de l'Information*, Vol. C1, 2004, 89-101.

[EADG03b] K. Aouiche, J. Darmont, L. Gruenwald, "Vers l'auto-administration des entrepôts de données", *Revue des Nouvelles Technologies de l'Information*, No. 1, 2003, 1-12.

[ECDRBB03] F. Clerc, A. Duffoux, C. Rose, F. Bentayeb, O. Boussaïd, "SMAIDoC : Un Système Multi-Agents pour l'Intégration des Données Complexes", *Revue des Nouvelles Technologies de l'Information*, No. 1, 2003, 13-24.

[EBBDR03] O. Boussaïd, F. Bentayeb, J. Darmont, S. Rabaseda, "Vers l'entreposage des données complexes : structuration, intégration et analyse", *Ingénierie des Systèmes d'Information (RSTI série ISI)*, Vol. 8, No. 5-6, 2003, 79-107.

[ELMPVL03] P. Lenca, P. Meyer, P. Picouet, B. Vaillant, S. Lallich, "Critères d'évaluation des mesures de qualité des règles d'association", *Revue des Nouvelles Technologies de l'Information*, No. 1, 2003, 123-134.

[ETM02b] T. Tweed, S. Miguet, "Vers une aide au diagnostic du cancer du sein par une analyse multicritère de bases de données mammographiques", *Santé et Systémique*, 2002, 305-320.

[ERL02] R. Rakotomalala, S. Lallich, "Construction d'arbres de décision par optimisation", *Revue des Sciences et Technologies de l'Information*, Vol. 16, No. 6, 2002, 685-703.

[ERZN01] G. Ritschard, D. Zighed, N. Nicoloyannis, "Maximisation de l'association par regroupement de lignes ou de colonnes d'un tableau croisé", *Revue Mathématique Sciences Humaines*, No. 154-155, 2001, 81-97.

[ERZ01] G. Ritschard, D. Zighed, "Réseaux neuronaux: applications potentielles à l'économétrie", *Monde en Développement*, Vol. 18, No. 721990, 2001, 71-79.

Conférences nationales avec comité de lecture et actes

[FLLV05] S. Lallich, P. Lenca, B. Vaillant, "Variations autour de l'intensité d'implication", *Colloque Analyse Statistique Implicative (ASI 2005), Palerme, Sicile*, Octobre 2005.

[FFB05b] C. Favre, F. Bentayeb, "Intégration efficace des arbres de décision dans les SGBD : utilisation des index bitmap", *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC 05), Paris*, Janvier 2005; *Revue des Nouvelles Technologies de l'Information*, 319-330.

[FUB05] C. Udréa, F. Bentayeb, "Fouille de données relationnelles dans les SGBD", *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC 05), Paris*, Janvier 2005; *Revue des Nouvelles Technologies de l'Information*, 356.

[FR05c] J. Ralaivao, "Améliorer la performance d'un entrepôt de données complexes par l'utilisation de métadonnées et de connaissances du domaine", *2ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances, EGC 05, Paris*, Janvier 2005, 81-84.

[FFBBN05] C. Favre, F. Bentayeb, O. Boussaïd, N. Nicoloyannis, "Entreposage Virtuel de demandes marketing : de l'acquisition des objets complexes à la capitalisation des connaissances", *2ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances, EGC 05, Paris*, Janvier 2005, 65-68.

[FBRB05] R. BenMessaoud, S. Rabaseda, O. Boussaïd, "L'analyse factorielle pour la construction de cubes de données complexes", *2ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances, EGC 05, Paris*, Janvier 2005, 53-56.

[FJN05] P. Jouve, N. Nicoloyannis, "Forage distribué des données : une comparaison entre l'agrégation d'échantillons et l'agrégation de règles", *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC 05), Paris*, Janvier 2005; *Revue des Nouvelles Technologies de l'Information*, 31-42.

[FPL05b] E. Prudhomme, S. Lallich, "Validation statistique des cartes de Kohonen en apprentissage supervisé", *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC 05), Paris*, Janvier 2005; *Revue des Nouvelles Technologies de l'Information*, 79-90.

[FGN05] R. Gaudin, N. Nicoloyannis, "Apprentissage non supervisé de séries temporelles à l'aide des k-Means et d'une nouvelle méthode d'agrégation de séries", *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC 05), Paris*, Janvier 2005; *Revue des Nouvelles Technologies de l'Information*, 201-212.

[FR05d] R. Rakotomalala, "TANAGRA : un logiciel gratuit pour l'enseignement et la recherche", *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC 05), Paris*, Janvier 2005; *Revue des Nouvelles Technologies de l'Information*, 697-702.

[FRRMJ05] M. Raimbault, R. Rakotomalala, X. Morandi, P. Jannin, "Mise en évidence d'invariants dans une population de cas chirurgicaux", *2ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances, EGC 05, Paris*, Janvier 2005, 149-158.

[FVMPLLB05] B. Vaillant, P. Meyer, E. Prudhomme, S. Lallich, P. Lenca, S. Bigaret, "Mesurer l'intérêt des règles d'association", *Atelier Qualité des Données et des Connaissances (DQK 05), EGC 05, Paris*, Janvier 2005, 69-78.

[FBLB05] G. Brunet, S. Lallich, A. Bideau, "Analyse quantitative des réseaux généalogiques ascendants, l'exemple des lignées familiales de la vallée de la Valserine (Jura français)", *XXVe Congrès international de la Population (UIESP), Tours*, Juillet 2005.

[FBAF05] R. BenMessaoud, K. Aouiche, C. Favre, "Une approche de construction d'espaces de représentation multidimensionnels dédiés à la visualisation", *1ère journée sur les Entrepôts de Données et l'Analyse en ligne (EDA 05), Lyon*, Juin 2005; *Revue des Nouvelles Technologies de l'Information*, Vol. B-1, 34-50.

[FLN05e] G. Legrand, N. Nicoloyannis, "Etat de l'art des méthodes de construction de variables", *12èmes Rencontres de la Société Francophone de Classification (SFC 05), Montréal*, 2005, 182-185.

[FLN05f] G. Legrand, N. Nicoloyannis, "Gestion de la phase de prétraitement des données et coefficient Kappa", *XXXVIIèmes Journées de Statistique, Lyon*, 2005, 182-185.

[FUBDB04] C. Udréa, F. Bentayeb, J. Darmont, O. Boussaïd, "Intégration efficace de méthodes de fouille de données dans les SGBD", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04), Clermont-Ferrand*, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, 83-94.

[FSCSZ04] M. Scuturici, J. Clech, V. Scuturici, D. Zighed, "Modèle topologique pour l'interrogation des bases d'images", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04), Clermont-Ferrand*, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, 409-414.

[FDBLB04] A. Duffoux, O. Boussaïd, S. Lallich, F. Bentayeb, "Fouille dans la structure de documents XML", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04), Clermont-Ferrand*, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, 519-524.

[FBRBB04b] R. BenMessaoud, S. Rabaseda, O. Boussaïd, F. Bentayeb, "OpAC : Opérateur d'analyse en ligne basé sur une technique de fouille de données", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04), Clermont-Ferrand*, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, 35-46.

[FVLL04b] B. Vaillant, P. Lenca, S. Lallich, "Etude expérimentale de mesures de qualités de règles d'association", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04), Clermont-Ferrand*, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, 341-352.

[FLM04] S. Lallich, F. Muhlenbach, "Apprentissage à partir de voisinages et fouilles d'images", *Workshop Analyse de données, Statistique et Apprentissage pour la Fouille d'Images, 14e Conference Francophone AFRIF AFIA*, Janvier 2004.

[FLPT04] S. Lallich, E. Prudhomme, O. Teytaud, "Contrôle du risque multiple en sélection de règles d'association significatives", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04), Clermont-Ferrand*, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, 305-316.

[FADB04] K. Aouiche, J. Darmont, O. Boussaïd, "Sélection automatique d'index dans les entrepôts de données", *1er atelier Fouille de Données Complexes dans un processus d'extraction des connaissances, EGC 04, Clermont-Ferrand*, Janvier 2004, 91-102.

[FJLN04] P. Jouve, G. Legrand, N. Nicoloyannis, "Sélection rapide en apprentissage supervisé", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04), Clermont-Ferrand*, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, 185-196.

[FLN04b] G. Legrand, N. Nicoloyannis, "Construction de variables et arbres de décision", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04), Clermont-Ferrand*, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, 204.

[FSSCZ04] M. Scuturici, V. Scuturici, J. Clech, D. Zighed, "Navigation dans une base d'images à l'aide de graphes topologiques", *XXIIème Congrès Informatique des organisations et systèmes d'information et de décision (INFORSID 04), Biarritz*, Mai 2004.

[FJC04] R. Jalam, J. Chauchat, "Catégorisation de textes multilingues: quelques solutions", *Atelier Fouille de Textes, EGC 04, Clermont-Ferrand*, Janvier 2004, 27-36.

[FE04] W. Erray, "WF : Une méthode de sélection de variables combinant une méthode filtre rapide et une approche enveloppe", *11èmes Rencontres de la Société Francophone de Classification (SFC 04), Bordeaux*, Septembre 2004.

[FDBB04] J. Darmont, F. Bentayeb, O. Boussaïd, "Conception d'un banc d'essais décisionnel", *20èmes Journées Bases de Données Avancées (BDA 04), Montpellier*, Octobre 2004, 493-511.

[FLN04c] G. Legrand, N. Nicoloyannis, "Nouvelle méthode de construction de variables", *11èmes Rencontres de la Société Francophone de Classification (SFC 04), Bordeaux*, 2004.

[FLN04d] G. Legrand, N. Nicoloyannis, "Sélection de variables et agrégation d'opinions", *11èmes Rencontres de la Société Francophone de Classification (SFC 04), Bordeaux*, 2004.

[FLN04e] G. Legrand, N. Nicoloyannis, "Sélection de variables et agrégation d'opinions", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04), Clermont-Ferrand*, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*.

[FCDRBB03b] F. Clerc, A. Duffoux, C. Rose, F. Bentayeb, O. Boussaïd, "SMAIDoC : Un Système Multi-Agents pour l'Intégration des Données Complexes", *XXXVèmes Journées de Statistique, Session spéciale Entreposage et Fouille de Données, Lyon*, Juin 2003, 337-340.

[FADG03c] K. Aouiche, J. Darmont, L. Gruenwald, "Vers l'auto-administration des entrepôts de données", *XXXVèmes Journées de Statistique, Session spéciale Entreposage et Fouille de Données, Lyon*, Juin 2003, 105-108.

[FHD03b] Z. He, J. Darmont, "Une plate-forme dynamique pour l'évaluation des performances des bases de données à objets", *19èmes Journées de Bases de Données Avancées (BDA 03), Lyon*, Octobre 2003, 423-442.

[FHRCRHDB03] V. Hopirtean, V. Ravery, J. Chauchat, M. Rouprêt, J. Hermieu, V. Delmas, L. Boccon-Gibod, "Réseaux de neurones, apprentissage automatique ou statistique dans les nomogrammes ?", *93ème Congrès Français d'Urologie*, 2003.

[FLBZ03] B. Léger, N. Belkhiter, D. Zighed, "Conception d'un outil graphique et visuel des données", *Journées francophones d'Extraction et de Gestion des Connaissances (EGC 03), Lyon*, Janvier 2003; *Revue des Sciences et Technologies de l'Information*, Vol. 17, 189-200.

[FLMZ03] S. Lallich, F. Muhlenbach, D. Zighed, "Traitement des individus atypiques en apprentissage par la régression", *Journées francophones d'Extraction et de Gestion des Connaissances (EGC 03), Lyon*, Janvier 2003; *Revue des Sciences et Technologies de l'Information*, Vol. 17, 399-410.

[FRZ03c] G. Ritschard, D. Zighed, "Modélisation de tables de contingences par arbres d'induction", *Journées francophones d'Extraction et de Gestion des Connaissances (EGC 03), Lyon*, Janvier 2003; *Revue des Sciences et Technologies de l'Information*, Vol. 17, 381-392.

[FHCCZ03] C. Hammami, Y. Chahir, L. Chen, D. Zighed, "Détection de couleurs de peau dans l'image", *Journées francophones d'Extraction et de Gestion des Connaissances (EGC 03), Lyon*, Janvier 2003; *Revue des Sciences et Technologies de l'Information*, Vol. 17, 219-231.

[FADG03d] K. Aouiche, J. Darmont, L. Gruenwald, "Extraction de motifs fréquents pour l'auto-administration des bases de données", *Journées francophones d'Extraction et de Gestion des Connaissances (EGC 03), Lyon*, Janvier 2003; *Revue des Sciences et Technologies de l'Information*, Vol. 17, 547.

[FCZ03] J. Clech, D. Zighed, "Data Mining et analyse des CV, une expérience et des perspectives", *Journées francophones d'Extraction et de Gestion des Connaissances (EGC 03), Lyon*, Janvier 2003; *Revue des Sciences et Technologies de l'Information*, Vol. 17, 83-92.

[FLMPVL03b] P. Lenca, P. Meyer, P. Picouet, B. Vaillant, S. Lallich, "Critères d'évaluation des mesures de qualité en ECD", *XXXVèmes Journées de Statistique, Lyon*, Juin 2003, 647-650.

[FCRJ03] J. Clech, R. Rakotomalala, R. Jalam, "Séléction multivariée de termes", *XXXVèmes Journées de Statistique, Lyon*, 2003, 933-939.

[FJN03f] P. Jouve, N. Nicoloyannis, "Classification Non Supervisée pour Données Catégorielles", *XXXVèmes Journées de Statistique, Lyon*, 2003, 579-582.

[FLJN03] G. Legrand, P. Jouve, N. Nicoloyannis, "Chaos Game Representation et Traitement des Séries Temporelles", *Xèmes Rencontres de la Société Francophone de Classification (SFC03), Neuchatel, Suisse*, 2003.

[FHRCBJCTDM03] V. Hopirtean, M. Rouprêt, J. Chauchat, J. Bazin, A. Jaquemard, Y. Chretien, N. Thioun, B. Dufour, A. Mejean, "Intérêt des techniques de simulation par rééchantillonage dans l'analyse de survie des cancers à faible incidence : application dans le carcinome à cellules rénales bilatéraux", *XXIIIème Forum de Cancérologie*, Juillet 2003.

[FMLZ02] F. Muhlenbach, S. Lallich, D. Zighed, "Amélioration d'une classification par filtrage des exemples mal étiquetés", *Journées Francophones d'Extraction et Gestion des Connaissances (EGC02), Montpellier*, 2002; *Extraction de Connaissance et Apprentissage*, Vol. 1(4), 155-166.

[FCRP02b] J. Chauchat, R. Rakotomalala, F. Pellegrino, "Estimation du taux d'erreur sur données en grappes - Application à la reconnaissance de la parole", *Journées Francophones d'Extraction et Gestion des Connaissances (EGC02), Montpellier*, 2002; *Extraction de Connaissance et Apprentissage*, Vol. 1(4), 269-280.

[FHCZS02] M. Hammami, L. Chen, D. Zighed, Q. Song, "Définition d'un modèle de peau et son utilisation pour la classification des images", *MediaNet2002*, 2002, 186-197.

[FZLM02b] D. Zighed, S. Lallich, F. Muhlenbach, "A statistical approach for separability of classes", *Conference : Statistical Learning, Theory and Applications, CNAM - Paris*, 2002, 58-65.

[FTL02] O. Teytaud, S. Lallich, "Contribution de l'apprentissage statistique à l'apprentissage non supervisé", *Conférence Apprentissage, CAp2002, Orléans France*, 2002, 87-98.

[FLMZ02b] S. Lallich, F. Muhlenbach, D. Zighed, "Test de structure pour la prédiction de variable numérique", *IXème Congrès de la Société Francophone de Classification (SFC'02), Toulouse*, 2002, 235-238.

[FCK02] D. Coeurjolly, R. Klette, "Estimateurs de longueur discrets", *Denis Richard 60th Birthday Conference*, 2002.

[FMR02b] F. Muhlenbach, R. Rakotomalala, "Utilisation d'amas pour la discrétisation de variables", *IXème Congrès de la Société Francophone de Classification (SFC'02), Toulouse*, 2002, 283-286.

[FLEB02] G. Legrand, W. Erray, M. Boulle, "Un survey des méthodes de sélection d'attributs dans le Data mining", *IXème Congrès de la Société Francophone de Classification (SFC'02), Toulouse*, 2002, 263-266.

[FCSMC02] S. Clippe, D. Sarrut, S. Miguet, C. Carrie, "Aide au positionnement du patient en radiothérapie conformationnelle par l'usage de techniques de recalage d'image 2D/3D basées sur l'intensité des pixels", *13ème Congrès de la SFRO, Paris*, 2002.

[FJN02] P. Jouve, N. Nicoloyannis, "Nouvelle Approche pour l'Extraction des Rectangles Maximaux d'une Relation Binaire, Applications en Extraction de Connaissances à partir de Données", *Journées Francophones d'Extraction et Gestion des Connaissances (EGC02), Montpellier*, 2002; *Extraction de Connaissance et Apprentissage*, Vol. 1(4), 47-58.

[FBBN01] F. Bentayeb, O. Boussaïd, N. Nicoloyannis, "Un cadre topologique pour OLAP", *Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 01), Nantes*, Janvier 2001; *Extraction des Connaissances et Apprentissage*, Vol. 1, 355.

[FMZD01] F. Muhlenbach, D. Zighed, S. D'Hondt, "Génération de règles par compression", *Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 01), Nantes*, 2001; *Extraction de Connaissance et Apprentissage*, Vol. 1(1-2), 93-104.

[FJT01b] R. Jalam, O. Teytaud, "Identification de la langue et catégorisation de textes basées sur les N-grammes", *Journées Francophones d'Extraction et Gestion des Connaissances (EGC01), Nantes*, 2001; *Extraction de Connaissance et Apprentissage*, Vol. 1(1-2), 227-238.

[FCRPC01b] J. Chauchat, R. Rakotomalala, C. Pelletier, M. Carloz, "TQC: un indice d'évaluation du ciblage d'événements rares; une application en marketing", *Journées Francophones d'Extraction et Gestion des Connaissances (EGC01), Nantes*, 2001; *Extraction de Connaissance et Apprentissage*, Vol. 1(1-2), 155-159.

[FTL01] O. Teytaud, S. Lallich, "Bornes uniformes en extraction de règles d'association", *Conférence Apprentissage CAp2001, Grenoble, France*, 2001, 133-148.

[FSC01b] D. Sarrut, S. Clippe, "Positionnement de patient en radiothérapie par recalage 2d/3d sans segmentation", *Radiothérapie Assistée par l'Image (RAI 2001), Centre Antoine-Lacassagne, Nice*, Mai 2001.

[FTM01] T. Tweed, S. Miguet, "Sélection automatique de régions d'intérêt pour la détection de zones cancéreuses dans les mammographies", *Coopération Analyse d'Image et Modélisation, Université Claude Bernard, Lyon*, Juin 2001, 2-5.

[FTS01c] T. Tweed, A. Saadane, "L'apport d'un bloc de segmentation d'erreur dans l'évaluation de la qualité d'images codées", *GRETSI 2001, Toulouse*, Septembre 2001.

[FADLN01] J. Auray, G. Duru, M. Lamure, N. Nicoloyannis, "Extension du concept de métrique et structures topologiques associées", *VIIIème Congrès de la Société Francophone de Classification (SFC'01), Pointe-à-Pitre, Guadeloupe, France*, 2001, 14-18.

[FTM01b] T. Tweed, S. Miguet, "Analyse conjointe de l'histogramme et de la texture pour la sélection de régions d'intérêt dans les mammographies", *VIIIème Congrès de la Société Francophone de Classification (SFC'01), Pointe-à-Pitre, Guadeloupe, France*, 2001, 339-347.

[FZLM01] D. Zighed, S. Lallich, F. Muhlenbach, "Séparabilité des classes dans IRp", *VIIIème Congrès de la Société Francophone de Classification (SFC'01), Pointe-à-Pitre, Guadeloupe, France*, 2001, 356-363.

Thèses et HDR

[GC04] J. Clech, "Contribution Méthodologique à la Fouille de Données Complexes", Thèse de doctorat, Université Lumière Lyon 2, 2004.

[GL04] G. Legrand, "Approche méthodologique de sélection et construction de variables pour l'amélioration du processus d'extraction de connaissances à partir de grandes bases de données", Thèse de doctorat, Université Lumière Lyon 2, 2004.

[GB04] L. Baumes, "Combinatorial Stockastic Iterative Algorithms and High Throughput : from discovery to optimisation of heterogeneous catalysts", Thèse de doctorat, Université Lumière Lyon 2, 2004.

[GJ03] R. Jalam, "Apprentissage automatique et catégorisation de textes multilingues", Université Lumière Lyon 2, 2003.

[GJ03b] P. Jouve, "Apprentissage Non Supervisé et Extraction de Connaissances à partir de Données", Thèse de doctorat, Université Lumière Lyon 2, 2003.

[GL02] S. Lallich, "Mesure et validation en extraction des connaissances à partir des données", Mémoire d'Habilitation à Diriger les Recherches, Université Lumière Lyon 2, 2002.

[GL02b] M. Lazar, "Contribution aux techniques orientées objet de gestion des séquences vidéo pour les serveurs Web", Thèse de doctorat, INSA, 2002.

[GM02] F. Muhlenbach, "Evaluation de la qualité de la représentation en fouille de données", Thèse de doctorat, Université Lumière Lyon 2, 2002.

[GC02d] D. Coeurjolly, "Algorithmique et géométrie discrète pour la caractérisation des courbes et des surfaces", Thèse de doctorat, Université Lumière Lyon 2, 2002.

[GC01] J. Chauchat, "Echantillonnage, validation et généralisation en extraction des connaissances à partir des données", Mémoire d'Habilitation à Diriger les Recherches, Université Lumière Lyon 2, 2001.

[GG01] G. Gavin, "Etude du modèle d'apprentissage probablement approximativement correct : Application aux méthodes d'agrégation", Thèse de doctorat, Université Lumière Lyon 2, 2001.

[GS01] V. Scuturici, "Utilisation efficace des serveurs Weben tant que serveurs vidéo pour des applications vidéo à la demande", Thèse de doctorat, Université Lumière Lyon 2, 2001.

[GB01] N. Belkhiter, "Communication homme-machine et décomposition des relations binaires avec application à divers domaines en informatique", Habilitation à Diriger les Recherches, Université Lumière Lyon 2, 2001.

5.2. ANIMATIONS SCIENTIFIQUES

5.2.1. Participation à des manifestions nationales et internationales

Nom de la manifestation	Membre du Comité de Programme	Période	Observations
Joint Conference on information Science, Durham (**CIS**)	Zighed D. A.	Depuis 1998	Organisateur de deux sessions
International Conference on Reverse Engineering for Information Systems (**ReTIS**)	Boussaid O. Darmont J.	2001 2001	 Comité d'organisation 2003
International Conference on Machine Learning, **(ICML)**	Zighed D. A.	2001	
European Conference on Machine Learning, **(ECML)**	Zighed D. A.	Depuis 2000	
European Conference on Principles of *Data mining* and Knowledge Discovery **(PKDD)**	Zighed D. A. Rakotomalala R. Chauchat J.-H. BentayebF. Rabaseda S.	Depuis 1998 2000 2000 2004 2004	Co-Président de la conférence et Président du comité d'Organisation en 2000 Evaluation d'articles Evaluation d'articles
International Symposium on Modelling Intelligent Systems **(ISMIS)**	Zighed D. A. Nicoloyannis N. Boussaid O.	Depuis 2001 Depuis 2001 	Co-Président Comité d'organisation 2002 Comité d'organisation 2002
Pacific Asia Knowledge Discovery in Data Bases **(PAKDD)**	Zighed D. A.	Depuis 2001	
9th **WSEAS** International Conference on CIRCUIT	Zighed D. A.	Depuis 2002	
Extraction et Gestion des connaissances **(EGC)**	Zighed D. A. Nicoloyannis N. Darmont J. Chauchat J.-H. Miguet S. Lallich S.	Depuis 2001 Depuis 2001 Depuis 2001 Depuis 2001 2001 2004	Co-fondateur de la conférence Membre du comité de pilotage Co-président de la conférence 2002 Vice président de l'association
Journées Francophones de la Société Française de Classification **(SFC)**	Zighed D. A.	Depuis 1997	Vice président depuis juin 2004
Multimedia Data and Document Engineering (**MDDE**)	Boussaid O.	2001, 2002	
Technologies for Information System	Boussaid O.	2001	
Multimedia Systems and Applications, Vol. 22, Kluwer Academic Publishers	Boussaid O.	2002	Evaluation d'articles

The International Journal of Computers and Applications	Boussaid O.	2003	Evaluation d'articles
Journal of Intelligent Information Systems (**JIIS**)	Darmont J.	2003	Comité de lecture
Encyclopedia of Information Science and Technology (IDEA Group Publishing)	Darmont J.	2003	Comité de lecture
Ingénierie des Systèmes d'Information (**ISI**), Special Issue: Information retrieval and information mining	Darmont J.	2003	Comité de lecture
Information Resources Management Association International Conference (**IRMA**)	Darmont J.	Depuis 2003	
Journées de la Société Française de Statistique (**SFdS**)	Zighed D. A. Boussaid O. Lallich S. Labo. ERIC	2003 2003 2003 2003	Président du comité d'organisation Comité d'organisation 2003
Session spéciale Entreposage et Fouille de Données (**EFD**), Journées de la Société Française de Statistique	Bentayeb F. Darmont J. Rabaseda S. Boussaid O. Lallich S.	2003 2003 2003 2003 2003	 Coéditeur Coéditeur
Journées Bases de Données Avancées (**BDA**)	Darmont J.	2003	
International Conference On Flexible Query Answering Systems, (**FQAS**)	Bentayeb F. Darmont J.	2004 2004	Membre du comité d'organisation
Fifth International Workshop on Multimedia Data Mining (**MDM/KDD**)	Boussaid O.	2004	
Workshop sur la Fouille de Données Complexes, (**FDC 2004**)	Boussaid O.	2004	
3rd International Semantic Web Conference (**ISWC**)	Darmont J. Rabaseda S. Zighed D. A.	2004 2004 2004	 Evaluation d'articles
International Journal of Information technology and Web Engineering	Bentayeb F.	2005	Comité de lecture
International Workshop on Intelligent Data Analysis and Data Mining, Application in Medicine Workshop Zagreb	Chauchat JH.	2004	

International Symposium on Applied Stochastic Models and Data Analysis	Lallich S. Zighed D. A.	2004	Session invitée
Intelligent Information Systems 2005 : New Trends in Intelligent Information Processing and Web Mining	Zighed D. A.	Depuis 2004	

5.2.2. Valorisations scientifiques organisées par ERIC

Journées de la Société Française de Statistique

Du 2 au 6 juin 2003, se sont tenues à Lyon, les 35ème journées de Statistique (JDS03) de la Société Française de Statistique (SFdS). Cette conférence constitue la principale manifestation francophone en Statistique. Le comité d'organisation était composé des membres du laboratoire de recherche ERIC de l'université Lyon 2. La conférence a remporté un important succès avec 433 participants (de 15 pays différents) et 243 présentations (183 communications orales et 20 posters, 9 conférences invitées, 3 sessions spéciales avec 25 papiers présentés, 6 présentations de logiciels statistiques).

Dans ce cadre, le laboratoire ERIC a également organisé une session spéciale « Entreposage et Fouille de données » (EFD). Cette session a été l'occasion pour les chercheurs et les praticiens d'échanger des idées et de faire le point des récents développements relatifs à l'entreposage des données et à l'extraction des connaissances. Les papiers longs acceptés par le comité de lecture de la session EFD ont été publiés dans un numéro spécial de la Revue des Nouvelles Technologies de l'Information, Cépaduès Editions.

Groupe de Travail sur la Fouille de Données Complexes

Le laboratoire ERIC est à l'origine de la création du groupe de travail Fouille de Données complexes (http://morgon.univ-lyon2.fr/GT-FDC/) dont l'objectif est de réunir la communauté des chercheurs en data mining qui s'attaquent aux données très volumineuses, hétérogènes. Ce groupe se réunit tous les trimestres. Il est adossé à la conférence EGC.

Les séminaires d'ERIC

Depuis octobre 1994, des séminaires sont organisés au sein du laboratoire ERIC (http://eric.univ-lyon2.fr/bderic/seminaires/), avec une fréquence bimensuelle entre le mois

d'octobre et le mois de juin. Ces séminaires, ouverts au public, mobilisent des intervenants d'horizons différents et ont pour objectifs de :

- mettre en relation les membres du laboratoire avec d'autres chercheurs dont certains viennent de l'étranger,
- obtenir un point de vue différent sur des problèmes qui entrent dans le cadre de l'activité de recherche du laboratoire,
- permettre aux chercheurs, notamment les doctorants, de présenter leurs travaux récents,
- mieux connaître des sujets connexes aux préoccupations du laboratoire.

5.2.3. Activité éditoriales

Zighed A. (ERIC, Lyon2), G. Venturini (LI, Tours) étaient co-directeurs de la revue **Extraction des Connaissances et Apprentissage** (ECA) publiée par Hermès. Cette revue dont la thématique s'est élargie aux domaines de l'extraction des connaissances est passée au format papier. Elle a pour ambition de traiter et de débattre des thèmes liés à l'extraction des connaissances, aussi bien d'un point de vue théorique que d'un point de vue pratique.

Zighed A. (ERIC, Lyon2) et G. Venturini (LI, Tours) sont co-directeur de la **Revue des Nouvelles Technologies de l'Information** (RNTI) publiée par Cépaduès. Les publications RNTI (http://www.antsearch.univ-tours.fr/rnti) sont des numéros spéciaux autour des problématiques de la fouille de données et de l'Extraction des Connaissances à partir des Données. 5 numéros sont déjà parus :

- Extraction et Gestion des Connaissances 2005. Rédacteurs invités : Suzanne Pinson (Lamsade, Université Dauphine Paris IX), Nicole Vincent (Crip5, Université René Descartes Paris 5). 2005.
- Classification et Fouille de données. Rédacteurs invités : M. Chavent, M. Langlais. 2004.
- Extraction et Gestion des Connaissances 2004. Rédacteurs invités : G. Hébrail, L. Lebart, J.-M. Petit. 2004.
- Mesures de Qualité pour la Fouille de Données. Rédacteurs invités : Henri Briand, IRIN Nantes, Michèle Sebag, LRI Orsay, Régis Gras, IRIN Nantes, Fabrice Guillet, Ecole Polytechnique de Nantes. 2004.
- Entreposage et Fouille de Données. Rédacteurs invités : Omar Boussaid, Stéphane Lallich, ERIC Université de Lyon 2. 2003

6. Déclaration de politique scientifique pour la période 2007-2010

6.1. L'ECD à partir de données complexes : un défi

Dès son émergence à la fin des années 80, l'Extraction de Connaissances dans les Données (ECD) s'est centrée sur l'exploitation des nouvelles sources d'informations pour mieux comprendre les mécanismes sous-jacents régissant les phénomènes observés et fournir des connaissances explicites nouvelles pour l'aide à la décision. « *Knowledge Discovery is the nontrivial extraction of implicit, previously unknown and potentially useful information of data* », ainsi que l'énonçait un des articles fondateurs de ce champ récent. Sous l'impact de l'explosion des capacités de stockage des informations (doublement tous les neuf mois) et de la structuration des organisations visant à une réactivité croissante dans un univers de plus en plus concurrentiel, l'exploitation des données est devenue cette dernière décennie un enjeu stratégique à la fois scientifique et industriel. L'ECD est maintenant un champ reconnu doté de conférences spécialisées, de revues dédiées et d'une communauté internationale active et structurée.

Aujourd'hui, deux défis majeurs sont associés aux nouvelles évolutions conjointes des modèles de représentation de données, et des technologies de stockage, d'exploitation et de restitution : (i) la fouille dans les données complexes qui ne se modélisent plus simplement sous la forme de tableaux IndividusxVariables (ii) l'entreposage de données complexes.

Dans la lignée de l'Analyse des Données et de l'Apprentissage Automatique, des logiciels performants ont été développés pour extraire des connaissances à partir de données représentées sous forme tabulaire aisément déductible des enregistrements de bases de données relationnelles. Des extensions méthodologiques ont été proposées pour appréhender des données initialement récoltées sous d'autres supports, notamment en langage naturel (fouille de textes), et en images (fouille d'images). L'ECD s'est ainsi développée selon un schéma uni-modal se déclinant suivant

le type de données traitées : données tabulaires, données textuelles, données images, … et se ramenant bien souvent *in fine* à un format classique à double entrée.

Les nouvelles stratégies de fouille et d'entreposage devront intégrer la spécificité des objets complexes (unités auxquelles se rattachent les données complexes) qui peut se décliner en cinq points :

- Nature différente : les données observées sur un objet sont de nature différente. Aux cas désormais classiques de descripteurs numériques, catégoriels ou symboliques, s'ajoutent notamment le cas de données textuelles, image ou audio vidéo.
- Diversité des sources : les données proviennent de sources hétérogènes. Comme le montre l'exemple des dossiers médicaux, les données recueillies peuvent être issues de questionnaires remplis par le médecin, des comptes rendus textuels, de mesures acquises par des appareils médicaux couplés à des ordinateurs, d'images radiologiques ou échographiques, etc.
- Evolutives et distribuées : il arrive souvent que l'on dispose de plusieurs caractérisations du même objet à des époques et/ou en des localisations différentes. Un patient est généralement suivi périodiquement par plusieurs médecins dont chacun produit une information spécifique. Ces informations s'intègrent autour d'un même sujet.
- Liées à des connaissances externes : la fouille intelligente des données s'appuie sur la prise en compte des connaissances externes, dites du domaine, celle-ci pouvant se faire par le biais d'une ontologie. Dans le domaine de la cancérologie par exemple, les connaissances diagnostiques et thérapeutiques sont organisées sous forme d'arbres de décision et mis à disposition des praticiens sous la forme d'un « guide des bonnes pratiques » appelés S.O.R. Standard Option Recommandation.
- Dimensionnalité des données : l'association de différentes sources de données à différents moments multiplie les points de vue, et par là, le nombre de descripteurs potentiels. Les nouvelles dimensionnalités mises en jeu se heurtent alors à des difficultés algorithmiques et méthodologiques.

Nous envisageons une conduite de nos travaux sur trois niveaux.

Niveau méthodologique. Bien que notre problématique puisse s'instancier dans tout le continuum allant des sources de données aux modèles d'aide au diagnostic, notre projet portera essentiellement sur les phases suivantes :

a)Modélisation de données complexes. L'exploitation simultanée de bases de données mixtes (structurées et non structurées) et multimédia permet d'envisager des représentations plus larges

et plus performantes car susceptibles : de traiter des informations hétérogènes en grand volume, de faire émerger de nouveaux types de corrélations ou d'associations, de faciliter la recherche et l'accès dans les bases de données en exploitant des effets de redondance véhiculés par des éléments d'information imprécis et incomplets mais globalement cohérents, de tenir compte de modèles de connaissances ou d'expertise. L'objectif est de modéliser et de gérer les types de dépendances (e.g. temporelles, spatiales) susceptibles de caractériser les données complexes.

b)Proposition de modèles permettant de réduire la dimension des espaces de représentation des données, ceci afin de maîtriser la complexité algorithmique des traitements envisagés.

c)Proposition d'heuristiques permettant la fusion de critères de recherche hétérogènes (texte, images, vidéo/audio, relations attributs/valeurs) débouchant *in fine* sur la construction de nouveaux algorithmes d'indexation et de recherche approchée dédiés aux données complexes.

Notre démarche consiste à étendre des méthodes existantes à la prise en compte du nouveau facteur de complexité, ou à proposer en cas de caractérisation d'impossibilité de nouvelles méthodes.

Niveau logiciel. Nos algorithmes seront intégrés dans une plateforme logicielle basée sur XML.

Niveau applicatif. Nos collaborations avec le monde industriel nous permettront sans aucun doute de trouver des terrains d'application propices à nos recherches.

6.2. OBJECTIFS ET VERROUS SCIENTIFIQUES

6.2.1. La fouille dans les données complexes

Un objet complexe peut être considéré comme un agrégat hétérogène de documents qui, une fois réunis, forment une unité sémantique. Cette unité sémantique ne se dégage généralement pas de l'addition des contenus de chaque document ; elle se construit dans une compréhension systémique sur un ensemble d'éléments interconnectés.

Notre projet vise à développer une méthodologie associée à des outils novateurs qui permettent la fouille dans les données complexes. Il s'article autour de quatre points :

L'évaluation des modèles de représentation des données et notamment les modèles d'indexation et de codage d'objets complexes. Cette évaluation doit intégrer les différents formats de représentation de connaissances développés dans des communautés différentes :

d'une part évidemment les nouveaux standards d'interopérabilité (tels que le méta-langage XML), d'autre part, les formats spécifiques aux données audio-visuelles (tels que MPEG7) et au domaine médical (tel que DICOM).

L'intégration des données et des connaissances du domaine. La fouille dans les bases de données complexes doit pouvoir s'effectuer sur la totalité des données disponibles. Notre projet étant orienté vers l'aide à la décision, le point de vue retenu consiste à libérer l'utilisateur-décideur des contraintes liées à l'organisation, au codage, au format, et à la représentation des données. Les techniques de fouille de données opérant généralement sur des structures de données tabulaires, se pose alors la question de leur adaptation aux données complexes? Les ontologies peuvent servir de cadre de référence à cette intégration. Comment dans un domaine tel que celui des données de patients atteints de cancers, une ontologie issue des standards option recommandation pourrait servir cet objectif ?

Les modèles de recherche d'information et d'extraction des connaissances. L'ambition de l'ECD est de dépasser l'assistance technologique à la manipulation et à l'interprétation des données et des connaissances en considérant explicitement l'utilisateur comme une composante à part entière du processus de découverte. Cet aspect peu abordé dans la première phase de l'ECD est entrain de prendre une importance croissante. Parmi les approches actuellement en cours de développement, l'apprentissage par retour de pertinence constitue un angle d'approche prometteur car d'une part, il permet une adaptation incrémentale des modèles de prédiction, et d'autre part il intègre l'utilisateur dans la boucle. Il peut permettre ainsi de récupérer, sans verbalisation préalable, une information d'expertise complémentaire non disponible initialement dans les données.

La validation des connaissances. La volumétrie des données constitue sans aucun doute un obstacle. Les performances d'un système d'apprentissage qu'il soit à but descriptif ou prédictif dépendent de plusieurs facteurs qui se trouvent en amont : (i) L'espace de représentation dans lequel sont plongés les cas. Comment peut-on lui associer une structure d'espace vectoriel et d'espace métrique ? (ii) Le nombre d'individus est-il de taille suffisante au regard de la dimension ? Est-il peut être trop grand, si oui serait-il possible de le réduire ? (iii) Toutes les variables descriptives sont-elles pertinentes ? Est-il possible d'identifier celles qui apportent du bruit ou celles qui sont redondantes. Les instruments de validation prendront tout leur sens dès

lors qu'ils deviendront capables de guider l'utilisateur. Nous serons conduit à traiter de manière particulière ces questions d'ensemble d'apprentissage et d'espace de représentation.

6.2.2. L'entreposage de données complexes

La multiplication des données complexes engendre une volumétrie considérable des données. Les entrepôts apportent une solution aux problèmes d'organisation, de stockage et d'analyse de ces grandes masses de données. Une approche très intéressante est de représenter les données complexes dans des documents XML. Le problème de l'entreposage des documents XML, de la modélisation multidimensionnelle et de l'analyse de données organisées de façon semi-structurée se pose alors et constitue un véritable verrou scientifique.

Les travaux existants dans ce domaine sont peu nombreux, très récents et n'abordent que quelques aspects des problèmes à résoudre. Certains d'entre eux portent sur la description à l'aide de schémas XML de cubes de données classiques pour l'échange données sur le Web. D'autres concernent la description de modèles logiques et physiques des entrepôts de documents XML. Pour notre part, nous choisissons une approche de conception et de mise en place d'entrepôts XML de données complexes. Pour prendre en compte les besoins d'analyse de l'utilisateur, nous les représentons par un modèle conceptuel en étoile que nous décrivons à l'aide d'un schéma XML. Le schéma XML des données complexes est ensuite appareillé avec le schéma XML des besoins d'analyse pour générer le schéma XML du cube de données complexes. Enfin, le cube de données complexes est alimenté par les documents XML.

Par ailleurs, le choix de construire le cube de données complexes dans un seul ou plusieurs documents XML engendre des problèmes de performance en termes de temps de construction, d'accès, de maintenance… Ces problèmes d'optimisation des performances sont encore insuffisamment abordés dans les bases de données natives XML.

Développer des techniques d'optimisation des performances des entrepôts de données XML posera également le problème de leur évaluation et de leur validation. À l'heure actuelle, les bancs d'essais XML se focalisent sur la performance des bases de documents et des bases de données transactionnelles. Il apparaît donc utile d'étendre le banc d'essais DWEB développé au laboratoire pour, d'une part, permettre l'évaluation d'entrepôts de données XML et, d'autre part, d'introduire la dimension « données complexes » dans ces tests.

Le fait de construire des contextes d'analyse sous la forme de cubes de données complexes stockés dans des documents XML induit de repenser les techniques d'analyse en ligne et de fouille de données. Cependant, le formalisme XML permet de présenter un ensemble de données complexes comme une seule entité informationnelle (par exemple, le dossier médical d'un patient) et de l'analyser dans sa globalité et non pas comme un ensemble d'informations dissociées.

Par ailleurs, avec la complexité des données à analyser et l'évolution des besoins d'analyse, il est nécessaire de réfléchir à de nouveaux modèles d'entrepôts de données qui garantissent les performances des systèmes décisionnels. L'une des pistes à exploiter consiste à concevoir des entrepôts de données à la demande selon l'évolution des besoins d'analyse permettant ainsi l'introduction des connaissances du domaine dans le processus d'entreposage. L'utilisateur peut faire évoluer l'entrepôt en créant de nouveaux axes d'analyse et devient ainsi un réel acteur du processus décisionnel. Dans ce cadre, nous proposons une nouvelle méthodologie de conception des entrepôts de données basée sur les règles. Un entrepôt de données est composé par une table de faits définie en extension et des dimensions définies par des règles. Les agrégats sont alors calculés grâce à ces règles permettant de déterminer les hiérarchies. Les règles permettant d'inférer sur les dimensions et leur hiérarchie créant ainsi de nouveaux axes d'analyse ; ce qui rend le modèle de l'entrepôt évolutif.

6.3. Evolution de la synergie enseignement-recherche

La mise en place des Master constitue pour nous une nouvelle opportunité pour créer de nouvelles offres d'enseignement originales s'appuyant sur la recherche. Avec des collègues des universités de Liège (Bruxelles), du Piemont (Italie), de Nantes, de Paris et de Lyon, nous avons décidé de mettre en place un master commun sur l'ECD. Cette formation qui s'appuiera sur le savoir faire actuel du master ECD en matière d'enseignement à distance permettra de constituer une offre de très haut niveau car elle réunit des laboratoires européens de premier plan. Un dossier de demande de labellisation comme Master Européen est en train d'être constitué pour être soumis à la commission européenne en janvier prochain 2006 pour un démarrage effectif de ce master européen en 2006-2007.

Dans le même esprit, nous venons de mettre en place pour le Master un double diplôme franco-ukrainien entre l'Université Lyon 2 et l'Université Nationale d'Economie de Kharkov. Financée par le Ministère des Affaires Etrangères dès 2005-2006 cette collaboration vise à s'étendre à d'autres université européennes et notamment à l'Université de Bergame avec laquelle des contacts ont été noués.

6.4.ERIC, STRUCTURE D'INCUBATION INDUSTRIELLE

Nicolas Macherey et Gaël Plantier élèves ingénieurs CPE Lyon sont porteurs du projet de création d'entreprise (TradingBots) accueilli au sein de l'incubateur CREALYS. La société conçoit des solutions logicielles et matérielles d'assistance à la prise de décision à partir de traitements de séries temporelles hautes fréquences, destinées à de nombreuses activités économiques et plus particulièrement aux secteurs de la finance que sont les traders indépendants, les banques et les assurances.

L'appui du laboratoire ERIC permet au projet de disposer d'une expertise scientifique et d'infrastructures (locaux et matériels informatique). ERIC apporte au projet l'encadrement scientifique et matériel nécessaire. L'immersion au sein du laboratoire doit notamment permettre d'analyser théoriquement les modules de la technologie actuelle, déjà développée, afin de l'améliorer, de la compléter et d'en éprouver pleinement la puissance et la validité. Les domaines d'expertise concernent les points suivants :

- L'analyse des séries financières avec extraction de signaux faibles dans un signal bruité (utilisation de techniques liées au traitement du signal : analyse en ondelettes, etc…),
- La détection de régularités en vue de la prévision par utilisation des techniques de l'apprentissage automatique (*Machine Learning*) sur les séries temporelles.
- L'expression de ces régularités sous la forme de règles ou de diagrammes causaux qui en permettent l'analyse et la compréhension.

La durée du projet est d'un an (2005-2006) et le financement par la Région Rhône-Alpes est de l'ordre de 39600 €.

7. TABLE DES MATIÈRES