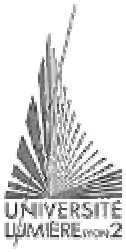

Rapport d'Activité

2004-2007

Laboratoire ERIC
Université Lumière Lyon 2
5, avenue Pierre Mendès-France
Bât L.
69600 Bron France

Tel. +33 4 78 77 23 76
Fax. +33 4 78 77 23 75
Web. http://eric.univ-lyon2.fr

LISTE DES MEMBRES DU LABORATOIRE

Direction

Djamel Abdelkader ZIGHED, Professeur, Directeur
Sabine LOUDCHER, Maître de conférences, Directeur adjoint

Administratifs

Valérie GABRIELE
Julien CREVEL

Professeurs

Stéphane LALLICH
Jean-Hugues CHAUCHAT

Maîtres de conférences

Fadila BENTAYEB
Omar BOUSSAID
Jérôme DARMONT
Nouria HARBI
Ricco RAKOTOMALALA
Julien VELCIN
Jacques VIALLANEIX

ATER

Anne Muriel ARIGON
Virginie LEFORT

Doctorants

Emna BAHRI
Anouck BODIN-NIEMCZUK
Sonia BOUATTOUR
Ahmad El SAYED
Cécile FAVRE
Rémi GAUDIN
Marouane HACHICHA
Hakim HACID
Hadj MAHBOUBI
Nora MAIZ
Simon MARCELLIN
Efthimios MAVRIKAS
Elie PRUDHOMME
Taimur QURESHI
Ony RAKOTOARIVELO
Jean-Christian RALAIVAO
Rashed Khalil SALEM
Anna STAVRIANOU
Julien THOMAS
Zhihoua WEI

Table des matières

A notre bien aimé et regretté Nicolas.

1 NOTE DE SYNTHESE

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)[1]. ERIC est actuellement composée de trois professeurs (CNU 27)[2], huit maîtres de conférences et d'une secrétaire à mi-temps. Le laboratoire accueille une vingtaine de doctorants, trois Attachés Temporaires d'Enseignement et de Recherche (ATER) et en moyenne trois professeurs invités pour un mois chacun par an. 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 Économiques et de Gestion.

Les travaux menés au sein du laboratoire s'articulent autour du processus d'Extraction des Connaissances à partir des Données (ECD) et s'attaquent à des verrous tant scientifiques que technologiques comme :

- La prise en compte des données complexes : hétérogènes (tableaux, multimédia, graphiques…), peu structurées, volumineuses, pouvant présenter ou non un caractère temporel ou spatial … ;
- La prise en compte du caractère empirique de la fouille de données et son impact sur les mesures de qualité et leur optimisation en apprentissage automatique, l'identification d'espaces de représentation efficients, la combinaison et l'agrégation de classifieurs…
- La prise en compte des connaissances du domaine soit pour l'enrichissement sémantique des corpus de données soit pour le déploiement des systèmes de prise de décision.

La mise en perspective de ces problématiques dans un processus d'ECD permet d'une part de les unifier et d'autre part de faire émerger des questions et des approches nouvelles.

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

- 9 chercheurs ont pu effectuer ou achever leur doctorat au sein d'ERIC ;
- 4 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 20 thèses en cours ;
- 3 projets de création d'entreprise ont été ou sont incubés au sein du laboratoire ;

[1] « Jeune Equipe » et « Equipe d'Accueil » sont des statuts attribués par le ministère aux équipes qu'il labélise.

[2] « CNU 27 » est la section du Conseil National des Universités qui regroupe les enseignants-chercheurs en informatique.

- 9 collègues étrangers ont été accueillis sur des postes de professeurs invités pour une durée minimale d'un mois ;
- 5 accords pédagogiques et scientifiques internationaux ;
- Une participation à des projets de recherche nationaux qui a généré 200 K€ ;
- Un partenariat en termes de contrats avec les entreprises privées qui a généré (146 €) ;
- 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 (19) ou nationales (9), communications dans des conférences internationales (90) ou nationales (72), chapitres (15) ou direction d'ouvrages (5), diffusion de logiciels, organisation 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 50% de ses ressources.
- ERIC maintient un lien fort et explicite avec l'enseignement, notamment, à travers l'animation du Master d'Informatique de Lyon 2 et la préparation d'un Master Erasmus Mundus en partenariat avec 6 universités de 4 pays européens (Italie, Espagne, Roumanie et France).

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 théorique, développement de logiciels, recherche de terrains d'application en particulier dans le domaine des Sciences Humaines, Sociales et Économiques ;
- renforcer la politique éditoriale et d'animation scientifique de l'équipe par plus de publications dans les journaux internationaux spécialisés ;
- développer nos relations avec la communauté scientifique locale, nationale et internationale autour à la fois d'activités 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 TRAVAUX SCIENTIFIQUES

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 de l'Informatique et des Mathématiques appliquées avec les Sciences Humaines et Sociales (Licence MISASHS)[1], ou avec les Sciences Economiques et de Gestion (Licence IDEA[2]), ainsi que celle d'un Master d'Informatique[3] avec trois spécialités professionnelles[4] et une spécialité Professionnelle et Recherche[5], sont une parfaite illustration de cette forte implication des collègues informaticiens et mathématiciens dans l'environnement spécifique de notre université.

2.2 Positionnement scientifique de ERIC

Le positionnement scientifique des recherches d'ERIC peut se faire selon deux points de vue que nous présentons de manière succincte.

[1] Mathématiques Informatique et Statistiques Appliquées aux Sciences Humaines et Sociales

[2] Informatique Décisionnelle et Économétrie Appliquées

[3] 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).

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

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

2.2.1 Point de vue historique

Une étude[1] de la société américaine *Disk/Trend* basée en Californie et spécialisée dans la veille industrielle a montré que le coût moyen de stockage d'un mégaoctet sur disque dur est passé de $11.54 en 1988 à $0.04 en 1998 et à $0.003 en en 2007. Cette décroissance exponentielle du coût de stockage, que chacun de nous a pu observer au quotidien, a eu pour effet d'accroître de manière exponentielle le volume des données stockées par les entreprises. En moyenne, selon les mêmes études, ce volume doublerait tous les 9 mois. Cette tendance à l'accumulation des données ne semble pas avoir atteint son niveau asymptotique et se trouve même accélérée sous l'effet du développement des réseaux de transmission comme Internet, qui sont de plus en plus puissants. En effet, ces derniers peuvent, de nos jours, atteindre les 10 gigaoctet par seconde et leur coût ne cesse de décroître. Ainsi, depuis 1975, le coût, au Mbit/s.km, a été divisé par 1000. Aujourd'hui, on peut affirmer que l'accès aux données et leur stockage reposent sur des technologies fiables et peu coûteuses. Autrement dit, le défi des années 60-70 visant la maîtrise des systèmes d'information a été pleinement relevé et gagné. Où se situent alors les nouveaux enjeux ? Ils se sont déplacés vers l'accès à l'information et à la connaissance cachée dans les immenses bases de données. Pour s'en convaincre, il suffit de regarder vers la plus grande réussite industrielle de l'informatique de la dernière décennie, Google et son moteur de recherche d'information sur le web. Ainsi, on peut affirmer que ce qui pose problème de nos jours, ce n'est ni l'accès aux données, ni leur stockage mais l'accès au contenu sémantique de celles-ci.

Les activités du laboratoire ERIC se situent dans ce vaste champ de recherche en forte croissance et dont les objectifs sont de définir des méthodologies et de proposer des outils informatiques permettant l'accès au contenu sémantique des grandes bases de données. C'est ce que nous appelons Extraction des Connaissances à partir des Données (ECD).

2.2.2 Point de vue pragmatique

La prise de décision est au cœur de toutes les activités qu'elles soient humaines, sociales, biologiques, économiques… Elle repose sur l'identification d'une situation et, en fonction d'un état désiré que l'on peut considérer comme l'objectif, le fait d'entreprendre les actions adéquates.

Considérons trois exemples simples :

- Un praticien qui examine un patient et qui observe une anomalie va prescrire une action thérapeutique dont l'objectif est l'élimination de la pathologie en vue d'améliorer l'état de santé du malade.

[1] Gamze Zeytinci, CSIS-550 History of Computing Spring-2001 ; Evolution of the Major Computer Storage Devices From Early Mechanical Systems to Optical Storage Technology.

- Un garde côte maritime qui, au moyen de ses radars, observe un mouvement suspect en mer, lance un processus de vérification pour savoir s'il s'agit d'un bateau qu'il faut intercepter.
- Un juge, au vu du récit des faits reconnus par un prévenu, identifie le type d'infraction et décide de la sanction.

Pour de multiples raisons de coût, d'efficacité, de rapidité, de complexité, de volume,… une large partie du processus de décision, notamment, le processus d'identification, est confiée à l'ordinateur. En effet, considérons des services de sécurité en charge d'Internet et dont l'objectif est par exemple d'intercepter les communications jugées « sensibles ». Comment envisager cette tâche quand on sait qu'il y a près de 700 millions d'internautes et qu'en moyenne, un internaute émet une dizaine de messages (e-mail, blog, requête…) par jour ? Dans ce contexte, il est impossible d'imaginer une organisation humaine capable d'assurer une surveillance sur les contenus de 7,5 milliards de messages journaliers sans parler des sites Web dont le contenu peut être ciblé.

On peut également observer cette situation dans d'autres domaines comme le marketing ou la santé. Par exemple, la généralisation du dépistage des cancers du sein à toutes les femmes dans la tranche d'âge 50-74 ans exigerait une infrastructure radiologique capable de traiter correctement près de 11 millions d'examens par an pour seulement 2000 radiologues en France qui sont déjà au bord de la saturation alors qu'ils ne traitent qu'environ 60%.

Dans de telles situations, le recours à la puissance des ordinateurs semble naturel pour assurer un passage à l'échelle. Mais pour pouvoir recourir à l'aide des ordinateurs, il faut être en mesure « d'expliquer » à celui-ci comment reconnaître un échange de nature suspecte parmi des millions de messages et comment identifier et localiser des cas suspects à partir d'un dossier médical comportant des images (mammographie), des examens cliniques et/ou biologiques, des comptes-rendus, etc.

Supposons que nous souhaitons mettre au point un système informatique capable d'aider les responsables de la sécurité publique dans l'identification des messages « sensibles ».

Le concepteur du système d'aide à l'identification collectera, auprès des experts en analyse des contenus, l'ensemble des règles conduisant au diagnostic. Ces règles décrivent le procédé d'analyse du contenu d'un message et les règles de déduction pour décider enfin de la catégorie sensible ou non d'un message. Pour y parvenir, il faut deux conditions qui reposent sur des hypothèses fortes :

- La première postule que le processus d'identification peut être décrit comme une suite d'opérations sur des structures symboliques qu'on appelle des règles d'inférence et que l'on assimile aux connaissances qu'utilisent les experts pour identifier les catégories de message.
- La seconde postule que l'expert est en mesure d'expliciter ses connaissances sous forme de règles dans un formalisme précis qui pourrait être codé en machine.

Si ces conditions sont réunies, alors les connaissances sont introduites dans la machine sous forme de règles pour en constituer la base de connaissances. On dote ensuite la machine d'un programme capable d'interpréter les règles à la manière d'un compilateur. Ce programme est appelé « moteur d'inférence » et il appliquera les règles dont il dispose en vue d'inférer sur des faits qui lui sont soumis. Si le système Expert (Base de connaissance + moteurs d'inférence) est jugé pertinent, on peut alors le dupliquer en une population d'agents dits « intelligents » qui seraient ensuite déployés sur Internet pour identifier et signaler la présence de tout message suspect. Dans la mesure où cette approche vise à mimer le raisonnement de l'expert face à des cas concrets, on pourrait la qualifier de démarche psycho-mimétique.

Malheureusement cette approche se heurte à deux difficultés majeures qui remettent en cause, au moins en partie, les hypothèses que l'on vient d'énoncer. En effet, il arrive que, sur des champs nouveaux, il n'y ait pas d'experts, donc aucune connaissance ne peut être transférée sur l'ordinateur. Que faut-il faire alors ? Faut-il attendre que des personnes confrontées à des situations de surveillance finissent par acquérir l'expertise nécessaire par un processus d'essai-erreur pour ensuite la communiquer à la machine ? Il arrive également qu'un expert sache parfaitement identifier les situations requises, mais qu'il soit en revanche incapable d'expliciter le procédé cognitif qu'il met en œuvre pour y parvenir. On sait tous reconnaître une personne dans une foule, mais sommes nous pour autant capables d'expliciter la manière dont nous y parvenons ?

C'est pour combler cette déficience que l'on fait appel aux méthodes d'Extraction des Connaissances à partir des Données. Les connaissances ne sont pas fournies par l'expert mais engendrées par la machine suite à un apprentissage automatique sur des situations passées. Par exemple, le médecin fournit un ensemble de données relatives à des patients cancéreux et non cancéreux déjà traités et, grâce aux méthodes d'ECD, on cherchera à déterminer les règles de diagnostic qui pourraient être appliquées sur les nouveaux cas à diagnostiquer. Une fois validées, ces connaissances pourraient, à leur tour, être insérées dans le Système Expert. Ce système Expert pourra également incorporer des fragments de connaissances venant d'experts humains. Cette démarche peut s'appliquer dans tous les domaines de la décision.

Le laboratoire ERIC travaille sur le processus d'ECD que nous allons décrire de manière un peu plus détaillée.

2.3 Présentation de l'ECD

Sans trop s'étendre sur les détails techniques, on peut dire que l'ECD fait appel à des méthodes et à des outils issus de différents domaines de l'informatique : bases de données, intelligence artificielle,

statistique, optimisation etc. en vue d'explorer des données volumineuses et hétérogènes à la recherche d'éléments structurants, d'invariants qui, une fois extraits et validés, pourraient être considérés comme des connaissances.

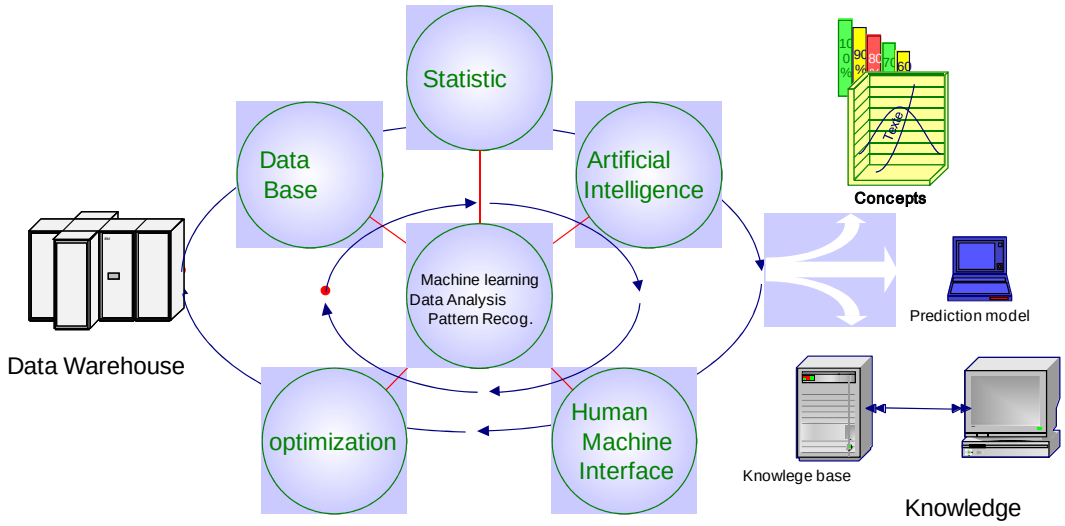

Figure 1 : KDD : From Data to Knowledge, Technologies involved

L'agencement de ces méthodes obéit à un processus logique en quatre étapes que nous décrivons très brièvement :

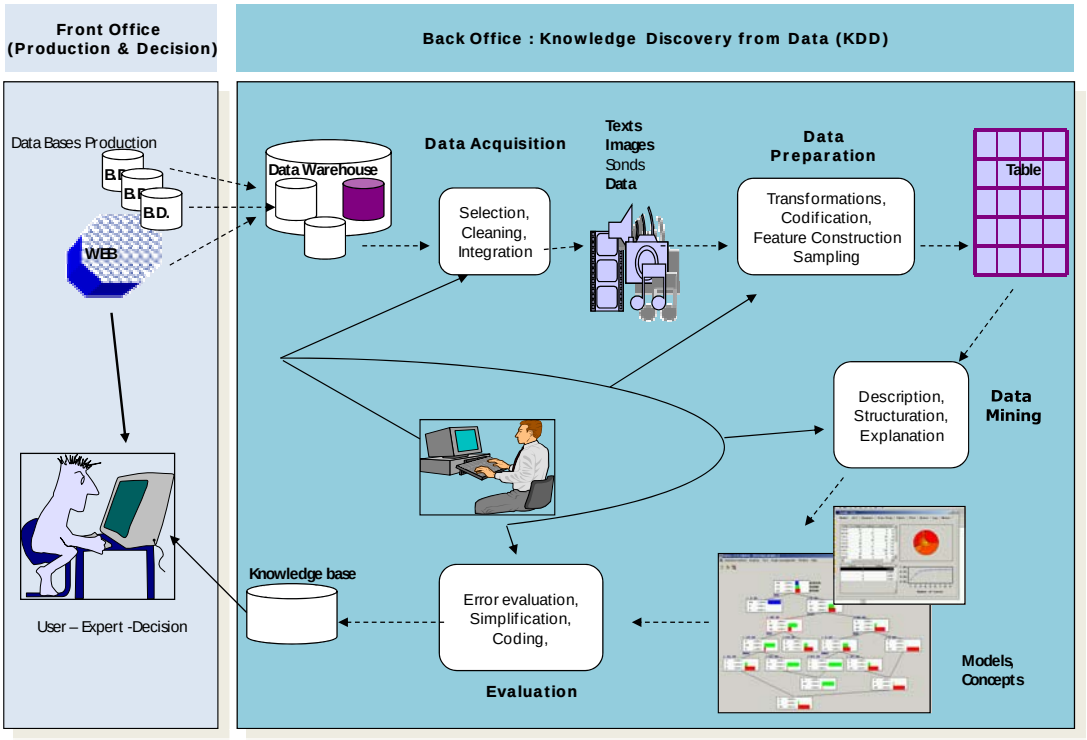

Figure 2: KDD workflow

Acquisition : elle a pour but de récupérer dans les entrepôts de données sources celles que l'on estime susceptibles d'aider dans la réalisation d'une tâche d'ECD. L'acquisition va s'effectuer sur des sources très volumineuses, distribuées géographiquement, enregistrées dans des environnements informatiques différents (bases de données relationnelles, bases de données XML, fichiers plats ou formats spécialisées comme le DICOM ou MPEG7). Outre l'accès et la sélection, il s'agit également d'organiser et d'intégrer les données dans un environnement local adapté à la fouille de données : bases de données XML, Entrepôts OLAP…

Préparation des données : les données acquises peuvent être de différents types : tableaux de chiffres, données textuelles, images etc. La phase de préparation a pour but de les structurer pour rendre possible la mise en œuvre des méthodes de fouille de données. La forme généralement la plus appropriée est un tableau de données à double entrée; cela peut être un tableau d'observations individus-variables, un tableau de contingence ; un tableau de similarité, etc. Il convient de préciser que cette opération est loin d'être simple et constitue généralement un véritable goulot d'étranglement. Par exemple si les données sources sont des textes, il convient de définir une série de prétraitements linguistiques comme la lemmatisation, la suppression de la ponctuation ou de la casse, le recours à une ontologie pour unifier le vocabulaire, l'extraction de concepts, etc. Il convient également d'identifier les individus statistiques : s'agit-il des textes, des paragraphes ou des concepts etc. Ce travail exige, généralement, une expertise et il est lié au domaine d'application. Les mêmes difficultés apparaissent quand les données sont des images. Quels attributs faut-il extraire pour décrire ces données sans structure mathématique évidente ? Quelles unités faut-il prendre en compte : images entières, imagettes issue d'un découpage ou d'une segmentation automatique etc. Quand la base d'exemples comporte des données de différents types : images, texte, courbes, tableaux de chiffres…, on parle de données complexes. Il faut non seulement définir un codage *ad hoc* pertinent pour ces données mais également les unifier. Le dossier médical d'un patient est une parfaite illustration de ce cas de figure car il contient souvent des radiographies, des courbes d'électrocardiogramme, des données quantitatives sur des mesures biologiques, des données textuelles qui décrivent un bilan clinique par exemple etc. Il faut par conséquent être en mesure de construire des mesures de similarité entre individus en considérant la totalité des données. C'est à ce niveau également que se pose le problème du traitement des données incomplètes et/ou imprécises, comment les prendre en compte en fouille de données. Cette phase est cruciale. Avec la phase de sélection, elle représente 80% du temps passé dans un cycle d'ECD.

Fouille de données : cette étape est le cœur de la démarche d'ECD. On y fait appel à une variété d'algorithmes soit pour la description des lignes et/ou colonnes des tableaux, soit pour la structuration des lignes et/ou colonnes du tableau, soit, enfin, pour établir un modèle de prédiction

au moyen des méthodes d'explication. On peut citer pêle-mêle : les méthodes d'analyse factorielles des données, les méthodes de classification dites d'apprentissage non supervisé, ou les méthodes de prédiction comme celles issues de l'apprentissage supervisé ou des règles d'association etc. La nature des données et du codage introduisent ensuite des variantes d'algorithmes qui enrichissent la gamme des méthodes de fouille. Par exemple si les données sont floues ou symboliques, il conviendrait de proposer des algorithmes spécifiques dans chacune des trois catégories d'algorithmes cités.

Validation : A ce stade, il convient d'apprécier les résultats selon leur fiabilité, leur intérêt pour l'utilisateur et le cas échéant de se poser le problème de leur intégration dans une base de connaissances auquel cas, il faut procéder à leur codage dans un formalisme approprié pour ensuite les déployer en situation concrète.

2.4 Avancées scientifiques 2004-2007

A la lecture des fiches d'activité des chercheurs (cf. Annexe I) et en parcourant la liste des publications sur la période considérée, on observe que les travaux d'ERIC s'étalent sur un spectre de sujets relativement large qui couvre tout le cycle de l'ECD. Mais, leur mise en perspective permet néanmoins d'observer qu'une attention particulière a été accordée à la prise en compte des données complexes dans le processus d'ECD. Ce cadre spécifique a permis de faire émerger de nouveaux problèmes et des défis tant théoriques que technologiques pour répondre aux besoins des applications réelles. Nous allons donc, surtout, mettre en avant cette originalité pour présenter les contributions des chercheurs d'ERIC dans le domaine de l'Extraction des Connaissances à partir des Données Complexes (ECDC). Nous verrons ensuite, à travers le cycle complet d'ECD, quels sont les problèmes abordés, où se situent les apports théoriques, les contributions méthodologiques et les applications traitées. On trouvera ensuite, dans l'annexe I, pour chaque chercheur, y compris les doctorants, une fiche qui résume de manière plus spécifique ces travaux.

2.4.1 Contributions à l'entreposage de données complexes

Les entrepôts de données classiques ont été développés autour du modèle des bases de données relationnelles. Ils ont donné naissance à des technologies telles que l'OLAP (*On Line Analysis Process*) permettant de naviguer dans la grande masse données qu'ils contiennent et d'en extraire des synthèses. Quels seraient alors les défis scientifiques et techniques à relever dans le contexte des données complexes ? Outre le problème de la représentation des données peu ou pas structurées, peut-on imaginer des modèles d'interrogation et d'exploration équivalents aux données tabulaires ? Quelles seraient ensuite les performances de tels systèmes une fois déployés ? Dans ce qui suit nous allons décrire les travaux menés pour tenter de répondre à ces questions.

2.4.1.1 Représentation et navigation dans les entrepôts de données complexes

Pour s'attaquer à ces verrous scientifiques, nous proposons un processus complet d'entreposage et d'analyse en ligne des données complexes. L'intégration et la modélisation physique consistent en l'intégration de données complexes dans une base de données agissant comme un ODS (Operational Data Storage). Dans la phase d'intégration, nous définissons des modèles conceptuels, logiques et physiques. Nous utilisons XML comme formalisme pour décrire les modèles logiques et physiques. Le modèle conceptuel est traduit au niveau logique sous la forme d'une DTD ou d'un schéma XML. Du modèle logique obtenu, nous générons une collection de documents XML comme modèle physique. Les documents XML générés sont valides et peuvent être stockés dans une base de données XML native ou une base de données relationnelle par mapping. De plus, nous avons proposé une approche pour construire un cube OLAP décrit par un schéma XML. Ce cube XML est généré automatiquement à partir des besoins de l'utilisateur exprimés par le modèle conceptuel multidimensionnel (MCM) et à partir d'un corpus de données complexes représentées par des documents XML. Le MCM et les documents XML sont exprimés en utilisant des schémas XML (XSD), puis ils sont transformés en arbres d'attributs afin d'être comparés. Certains algorithmes rendent cette comparaison possible grâce à des opérateurs de fusion par élagage ou greffe afin de traiter les arbres d'attributs et de générer un schéma XML du cube et des documents XML. Ce cube XML fournit un contexte d'analyse qui peut être analysé par des opérateurs OLAP ou par des méthodes de fouille de données.

Nous avons également développé une autre approche d'intégration de données basée sur un système de médiation utilisant des ontologies pour décrire chaque source de données. A partir des ontologies locales, l'objectif est de construire une ontologie globale qui permet au médiateur de proposer les données pertinentes pour la construction d'un cube OLAP. Cette ontologie globale est construite en utilisant une classification de l'ensemble des termes des ontologies locales.

Comme résultats de cette recherche, deux logiciels ont été développés. (1) SMAIDoC est un système multi-agents, articulé autour de cinq agents, afin d'intégrer les données complexes dans une base de données relationnelle ou une base de données XML native. (2) X-Warehousing est une plate-forme Java dédiée à la génération automatique de cubes XML. Les cubes XML sont obtenus à partir des besoins de l'utilisateur exprimés par le biais du modèle conceptuel multidimensionnel et à partir des documents XML contenant les données complexes.

2.4.1.2 Optimisation et évaluation des performances des entrepôts de données complexes

Dans ce contexte d'utilisation du langage XML comme support pour l'entreposage des données complexes, la performance des entrepôts de données reste plus que jamais une question cruciale. Les

principales structures physiques des données utilisées pour optimiser les temps d'accès aux données lors de l'exécution de requêtes complexes d'analyse sont les index, les vues matérialisées et les partitions. Sélectionner un ensemble optimal de ces objets est un problème NP-complet qui a été très largement traité. Toutefois, le passage à l'échelle demeure un problème car les approches existantes nécessitent soit une expertise humaine soit des structures de données coûteuses. En outre, les relations entre les index et les vues matérialisées ne sont jamais prises en compte, alors que ces structures de données vont mutuellement s'enrichir.

Pour répondre à ces questions, nous avons conçu une approche, générique, automatique, qui utilise des techniques de fouille de données, pour, à partir d'une charge (ensemble de requêtes) représentative de l'utilisation de l'entrepôt de données, en déduire une configuration quasi-optimale des index et / ou des vues matérialisées. Cette approche réduit considérablement l'espace de recherche des index des vues et donc améliore le passage à l'échelle. Ensuite, les modèles de coût aident à choisir les index et les vues matérialisées les plus efficaces en termes de gain de performance. Ces modèles prennent en compte les relations interdépendantes entre les index et les vues afin d'obtenir le meilleur compromis. Notre recherche d'optimisation des performances a été supportée et appliquée à deux projets pour optimiser l'accès aux données complexes : MAP (entrepôt de données biomédicales) et CLAPI (entrepôt XML de corpus de langue parlée).

De plus, afin d'évaluer et de comparer l'efficacité des techniques d'optimisation des performances, il est nécessaire de les tester avec différents jeux de données. Cette tâche est généralement réalisée à l'aide de bancs d'essai. Le « Transaction Processing Performance Council » (TPC) fournit des bancs d'essai standard, mais le schéma de bases de données et la charge sont fixes (seule la taille de l'entrepôt varie). Ces bancs d'essai présentent donc peu d'intérêt dans un contexte d'ingénierie et de conception. Par conséquent, pour valider expérimentalement notre approche d'optimisation des performances, nous avons conçu plusieurs bancs d'essai génériques. Leur principal principe de conception est la capacité d'adaptation : nos bancs d'essai permettent la production d'entrepôts de données avec des configurations différentes, ainsi que les charges associées. DWEB (Data Warehouse Engineering Benchmark) est le plus mature des outils développés. Il est actuellement le seul banc d'essai opérationnel, pour évaluer les performances des entrepôts de données, disponible en ligne.

2.4.2 Contributions à la recherche d'information dans les entrepôts de données complexes

La Recherche d'Information dans les Entrepôts de données Complexes (RIEC) pose des problèmes de nature spécifique. En effet, dans tous les processus de recherche d'information il est nécessaire de se doter soit d'une mesure de similarité entre les objets soit d'une structure topologique sur ces

mêmes objets sans avoir à expliciter la mesure de similarité sous jacente. Dans le cas où les données sont tabulaires, il existe une multitude d'indices de similarité. Mais, dans le cas où les données sont non structurées, comme par exemple pour les molécules chimiques, les textes, les images, les séries temporelles, les vidéos ou les documents multimédia, définir une proximité entre objets n'est pas aisé. Les nombreuses solutions proposées jusque là ne prennent qu'un seul type de données à la fois : soit textuel, soit image, soit la structure de la molécule en chimie,… mais peu de travaux ont été consacrés aux mesures qui prennent en compte cette hétérogénéité dans les données.

Nous avons exploré différentes stratégies pour construire une mesure de similarité entre objets complexes. Par exemple, par agrégation des mesures de similarité classiques issues de chaque type de données. L'agrégation est alors réalisée par une combinaison linéaire des similarités partielles qui résulte de chaque type de données. Nous avons également adopté d'autres approches qui combinent à la fois un point de vue topologique et un point de vue probabiliste. Pour cela, nous construisons par exemple une structure topologique dans chaque sous-ensemble homogène de données en utilisant des mesures de similarité classiques. La similarité globale entre deux objets sera d'autant plus forte que ces objets sont voisins dans les sous-espaces spécifiques. Par exemple deux patients seront d'autant plus voisins (semblables) qu'ils sont proches dans l'espace des données cliniques, proches dans l'espace des données image, proches dans l'espace des données biologiques etc. Ainsi, deux patients seront déclarés voisins s'ils sont voisins dans chacun des sous espaces topologiques induits. De là, on peut construire un indicateur qui une fois normalisé s'apparenterait à la probabilité pour que deux individus soient voisins. Si cette probabilité était égale à 1 alors les deux individus pourraient être considérés comme identiques ou quasi identiques. Ces approches, encore en cours, ont été mises en œuvre et testées avec efficacité dans des applications réelles, parmi lesquelles on peut citer :

- Le corpus langage parlé qui contient des données textuelles, audio et vidéo. Ce travail de recherche a été réalisé dans le cadre d'un projet conjoint avec le laboratoire ICAR[1] (UMR Lyon 2 ENS lettres), et a bénéficié d'un soutien financier du ministère de l'enseignement supérieur.
- Le corpus de textes juridiques liés au droit international du travail. Cette recherche a été conduite en collaboration avec l'Université de Genève et le Bureau International du travail.
- Un corpus de bases d'images indexées par des textes disponibles comme benchmark dans la communauté de Recherche d'Information.

[1] Interaction Corpus Apprentissage Représentation

2.4.3 Contributions à la préparation des données

Dans l'ECDC, plus qu'ailleurs, la préparation des données pour la fouille s'avère ardue. Le principal problème qui se pose est celui du choix de l'espace de représentation. En effet, à l'origine, les instances d'objets enregistrés dans la base de données sont exprimées dans un formalisme qui ne se prête pas ou peu aux traitements mathématiques qui fondent la plus part des méthodes de fouille de données. Le passage par un codage unifié, généralement sous forme vectorielle, s'avère souvent incontournable. Comment alors passer de données textuelles, images, vidéo, temporelles vers des vecteurs ? Faut-il « vectoriser » et tout aligner dans un tableau unique ? Ces choix ne sont pas neutres car ils pourraient engendrer, à leur tour, d'autres difficultés. Par exemple, le choix de coder par un vecteur numérique les textes, exige des procédés linguistiques qui sont souvent très sophistiqués et où l'intuition, les connaissances *a priori* du domaine et même les choix arbitraires sont couramment mis à contribution. Et cela, sans, pour autant, être certain du bon choix. Parmi les problèmes qui peuvent surgir, par exemple, le vecteur résultant peut s'avérer d'une grande dimension ce qui impacterait fortement les temps de calcul mais également agirait sur la cohérence de l'interprétation des proximités entre points observations. Outre le cas des données textuelles et/ou image, on peut citer également les données génomiques où l'espace des variables est généralement nettement plus grand que celui des individus. Comment alors réduire la dimensionnalité avec une perte minimale d'information ? Par sélection ? Par élimination ? Par projection ? Et, surtout, comment évaluer la pertinence du nouvel espace de représentation ? De plus, les données qui arrivent sont parfois entachées de bruits et incomplètes. Par exemple, le contenu des courriers électroniques réunit bon nombre de ces problèmes. A défaut de redresser les anomalies, peut-on au moins prendre en compte l'incomplétude, l'incertitude et l'imprécision dans nos analyses ?

Nous avons réalisé de nombreux travaux autour d'applications réelles dans les domaines du marketing, de l'exploitation de textes juridiques, de l'identification de plancton à partir d'images, d'identification de structures de courbes dans les flots de séries chronologiques… où ces questions ont été centrales.

L'expertise acquise par le laboratoire dans ce domaine est conséquente et a produit des résultats théoriques et méthodologiques particulièrement intéressants. Par exemple, un test statistique non paramétrique pour mesurer la séparabilité des classes en vue d'un apprentissage supervisé, des stratégies de détection d'objets atypiques dans des espaces multidimensionnels, des mesures de similarités sur des textes permettant de s'affranchir d'une vectorisation explicite ou encore le recours à des approches de taxinomie comme les cartes de Kohonen pour réduire la dimensionnalité.

2.4.4 Contributions à la fouille de données

La fouille de données opère, le plus souvent, sur des structures tabulaires, préparées à la phase précédente. C'est, généralement, la partie la plus visible du processus de fouille car, c'est à ce stade que l'on produit les connaissances sous la forme de modèles : règles logiques, algébriques, probabilistes, topologiques etc. Et, pour cela, on fait appel aux méthodes d'apprentissage, qu'elles soient supervisées ou non, aux méthodes exploratoires comme les algorithmes de recherche des règles d'association, aux analyses factorielles, ou aux méthodes de modélisation comme les réseaux bayésiens etc. Nous allons décrire quelques uns des travaux réalisés.

- La mise en œuvre de méthodes de fouille dans le cadre de l'ECDC a fait apparaître, à la fois, des problèmes théoriques et des problèmes de mise en œuvre pratique sur machine. En effet, même si, au stade de la fouille, les données sont structurées sous forme tabulaire, leur volume peut être très grand et par conséquent peut poser des problèmes de temps de calcul. Dans ce cadre, nous avons été amenés à réfléchir dans deux directions. La première part du constat qu'un fichier, même de taille très grande, reste un échantillon issu d'une population plus large. Par conséquent, si le même traitement était effectué de façon itérative sur ce même fichier enrichi, itérativement, par de nouveaux cas, le modèle qui en résulterait serait très probablement différent par sa structure et par son taux d'erreur. D'où l'idée d'exploiter à fond cette piste en travaillant sur de petits échantillons dont on fait croître la taille par ajout aléatoire de cas jusqu'à ce que la variance du taux d'erreur par exemple devienne quasi nulle. On utilisera ainsi, de manière efficiente l'information disponible sans subir son poids massif. La seconde direction visait à mieux exploiter les technologies informatiques disponibles. Il s'agit de voir comment faire migrer les méthodes de fouille de données vers des structures systèmes et logiciels qui supportent l'aspect massif ? Dans ce contexte, nous avons réalisé des couplages forts entre les systèmes de gestion de bases de données capables de travailler sur des vues (tableaux) de taille quasi illimitée et les algorithmes de fouille comme les arbres de décision. On peut ainsi bénéficier des structures de données, notamment les indexes *bitmap*, pour implémenter de manière plus efficace et scalable de nombreux algorithmes de fouille. On s'oriente ainsi vers l'introduction de nouveaux opérateurs de fouille de données aux côtés des opérateurs SQL classiques au sein des systèmes de gestion des entrepôts de données ce qui peut, par là-même, conduire à des plates formes logicielles intégrées.
- Nous travaillons sur des données réelles et nous avons été souvent confrontés à la sous-représentation de certaines classes d'intérêt. Dans ce cas, la mise en œuvre des méthodes d'apprentissage supervisé nécessite une prise en compte et un contrôle de l'asymétrie des classes. A ce propos, nous avons effectué un travail quasi exhaustif sur les mesures

généralement utilisées dans les arbres de décision, dans l'extraction de règles d'association etc. L'étude à la fois des propriétés théoriques de ces mesures ainsi que leurs performances sur des Benchmarks a débouché sur de nouvelles mesures d'entropie généralisées. Nous avons également proposé de nouvelles axiomatiques pour ces mesures qui donnent par ailleurs des résultats plus probants sur des cas pratiques.

- A l'issue du processus de fouille de données, les utilisateurs préfèrent disposer de modèles de prédiction intelligibles comme ceux qui sont issus des arbres ou des graphes de décision et qui s'expriment sous la forme de règles logiques. Mais, l'utilisation de ces algorithmes se heurte à des difficultés quand par exemple les variables possèdent une large distribution qui nécessite des groupements de modalités ou quand la variable à prédire est d'un genre particulier par exemple une courbe de survie ou un vecteur. Nous avons proposé plusieurs extensions aux méthodes basées sur les arbres de décision.

2.4.5 Validation-intégration et déploiement

Les modèles issus de l'apprentissage doivent être validés avant d'être utilisés comme connaissance par l'utilisateur ou par un système de décision. La plupart des méthodes d'apprentissage proposent des procédures d'évaluation de la qualité des modèles qui sont généralement basées sur les taux d'erreur en resubstitution ou sur échantillon test. La réduction de ces taux d'erreur a conduit à de nouvelles stratégies d'apprentissage comme le *bagging*, le *boosting* ou encore l'apprentissage semi-supervisé. Les évaluations qui ont été conduites au sein du laboratoire ont montré, de façon évidente, l'intérêt d'exploiter ces techniques de ré-échantillonnage autour des algorithmes d'apprentissage et des extensions que nous avons ajoutées.

Les connaissances produites de façon automatique, ne constituent souvent qu'un fragment de connaissances pour bâtir de vrais systèmes d'aide à la décision. Dans ce cadre, et afin d'accroître les performances et l'intérêt de ces systèmes, nous avons développé et testé des méthodologies pour intégrer dans une même base de connaissances, celles qui proviennent du domaine et/ou de l'expert. Cette intégration se fait au moyen d'ontologies. Ces ontologies servent à la fois d'outil d'expansion et d'unification de fragments de connaissances provenant de sources différentes et de réceptacle de connaissances.

3 PERSPECTIVES DE RECHERCHE

3.1 Fouille de données complexes (FDC)

Depuis toujours, le rôle des données complexes (image, vidéo, texte non structuré ou combinaison de ces médias) n'a cessé de croître, pour être aujourd'hui le principal véhicule d'information. La problématique d'une diffusion massive et de qualité est quasi-réglée grâce aux technologies de l'entreposage massif des données et aux réseaux à haut débit. Nous disposons de volumineuses bases de données complexes dont la croissance est exponentielle mais dont la valorisation reste encore très faible.

Le défi qu'il faut relever est de tirer profit, dans tous les sens du terme, de ces données : recherche d'information, extraction de connaissances, création de valeur économique etc. La Fouille de données complexe tente de répondre à ce besoin. Elle propose de définir un cadre méthodologique et des outils pour structurer les données complexes, les analyser en vue d'extraire des connaissances ou des informations non accessibles par des moyens classiques. Le projet scientifique d'ERIC s'inscrit dans cette perspective.

3.2 Caractéristiques des données complexes

La plupart des données réelles disponibles, issues de la vie de tous les jours, sont complexes. Elles sont généralement :

- Volumineuses : plusieurs téraoctets. Par exemple le Dossier Médical Personnalisé (DMP) peut atteindre plusieurs dizaines de giga-octets par patient ;
- Distribuées : le DMP par exemple, peut être stocké dans différentes bases de données distribuées selon les services médicaux où le patient a séjourné ;
- Hétérogènes : les données peuvent être de différente nature. Dans le cas du DMP par exemple, on aura des images radiologiques, des comptes-rendus textuels, des tableaux de chiffres de mesures biologiques, des courbes d'électrocardiogramme, des enregistrements vidéo d'échographie, etc.
- Evolutives : différents enregistrements avec des contenus différents. Par exemple, le DMP contient divers examens réalisés à des instants différents et qui ne portent pas nécessairement sur les mêmes tests médicaux ;

- Non structurées : elles ne sont généralement pas modélisées dans le cadre d'un schéma de base de données mais stockées quasiment en vrac et dans le meilleur des cas dans des formats *ad hoc*.

3.3 Défis scientifiques dans la FDC

Les défis scientifiques que soulèvent ces particularités des données complexes sont multiples. Parmi ceux que l'on peut assez facilement identifier, on peut lister :

- La grande dimensionnalité. Les attributs issus des images, des textes et des autres modalités peuvent atteindre plusieurs centaines, voire des milliers de variables. Comment évaluer la pertinence de cet espace de représentation par rapport aux taches que l'on souhaite effectuer comme l'apprentissage supervisé ou la classification ? comment réduire l'espace si tant est que cela soit possible ?
- Absence de structure mathématique. Généralement les codages effectués sur les données complexes sont faits de sorte que les tableaux qui en résultent sont assimilés à des points plongés dans des espaces multidimensionnels et, dans le meilleur des cas, dans des espaces vectoriels. Dans ce cadre, l'outillage mathématique, issu notamment de l'algèbre linéaire et de la programmation mathématique, permet de traiter ces données. Or ce codage n'est pas toujours possible notamment en présence de données hétérogènes qualitatives (état civil, localisation géographique etc.) et quantitatives ou de données non structurées comme des graphes orientés ou de données symboliques (courbes, intervalles, distribution etc.). Comment alors analyser ces ensembles de données ? Quel codage faut-il adopter ? comment construire des indices de proximités qui sont des outils indispensables pour les tâches d'apprentissage notamment non supervisé ? Quelles propriétés mathématiques résultent de ces choix pour savoir si oui ou non des algorithmes classiques de fouille peuvent être utilisés?
- Différence de niveau sémantique. Les données qui se rapportent à un objet complexe ne se situent pas toujours toutes sur le même niveau d'abstraction. Ce phénomène bien connu dans le domaine de la représentation des connaissances et notamment dans les ontologies prend une nouvelle dimension encore plus difficile à maîtriser. Par exemple, un compte-rendu médical écrit par un médecin sur un patient peut être le résultat d'une interprétation de clichés et d'examens biologiques. Par conséquent, le niveau sémantique du texte est différent de celui des images. Dans ce cas, les attributs issus des comptes-rendus textuels auront du mal à être alignés sur ceux issus des images radiologiques ou des examens biologiques. Comment alors intégrer ces niveaux sémantiques pour ensuite pouvoir décrire des patients ou

les comparer ? Le texte devrait-il être vu comme un subsumant des images et des données biologiques ? difficile d'y répondre promptement.

- Fusion des données et intégration des connaissances du domaine. Souvent, dans nos processus d'interprétation des situations qui nous entourent, comme être humains, nous pouvons mieux inférer grâce à une contextualisation des données que nous recevons par rapport à d'autres qui leurs sont liées indirectement. Par exemple, un grand opérateur de téléphonie dont les entrepôts de données sont extrêmement volumineux cherchera à contextualiser ses clients par rapport aux caractéristiques de leur quartier d'habitation, par rapport aux spécificités des moments d'appel : fin de semaine, jour ou nuit, période de vacances etc. La base clients peut ainsi être enrichie par des informations indirectes qui fournissent un contexte susceptible d'améliorer l'interprétation. On peut également y introduire des connaissances formelles, par exemple, pour le diagnostic médical, certaines hypothèses peuvent être renforcées grâce aux connaissances médicales disponibles. Ainsi, compte tenu d'un certain profil, on doit pouvoir adjoindre d'autres informations dans le dossier patient. Ce procédé est particulièrement développé dans la fouille de données textuelles et est généralement destiné à améliorer la désambigüisation.

3.4 Défis technologiques de la FDC

Ils sont étroitement liés aux défis scientifiques et ils sont parfois des conséquences des difficultés scientifiques. Parmi les défis technologiques identifiés, nous allons poursuivre nos travaux sur :

- Le passage à l'échelle (scalabilité). Nous pouvons en effet disposer de solution formelle et même opérationnelle sans pour autant être en mesure de l'utiliser sur des corpus réels de fouille, soit pour des raisons de temps de calcul soit pour des raisons d'espace mémoire. Par exemple, et pour rester dans le cas DMP, une classification d'une population de patients s'avère quasi impossible de façon directe en prenant en compte la totalité des informations. Outre le problème de l'alignement des attributs, le mélange du type de données, la dimension élevée de l'espace de représentation, le grand nombre d'observations etc. rendent cette opération quasi-impossible de façon directe. On peut alors s'interroger sur : comment revenir sur les aspects méthodologiques pour développer des algorithmes appropriés incrémentaux par exemple ? Ou comment mieux exploiter les ressources physiques des machines ? Le *GRID computing* est l'une des réponses possibles, mais est-ce la REPONSE sur des applications réelles ? une autre approche serait : comment travailler sur des vues partielles des objets ? La fouille de données multi-tables tente d'y répondre.

- Un processus de fouille de données destiné à produire des connaissances à partir de données en perpétuelle évolution. Le processus de fouille devrait alors être continu pour identifier à temps les éventuelles modifications majeures au niveau des connaissances qui pourraient survenir sur un phénomène modélisé. Comment alors assurer le couplage fort entre inférence sur des cas et amélioration incrémentale des connaissances qui sous-tendent cette inférence ?

3.5 Projet scientifique

D'une façon générale, notre projet porte sur une méthodologie pour une approche dynamique et intégrée de la fouille des données et le déploiement des connaissances dans des applications réelles. Le plus souvent, les défis technologiques et scientifiques sont étroitement liés. Au sein du laboratoire ERIC nous avons toujours été soucieux de ce va-et-vient entre le scientifique et le technologique. Nous avons souhaité fédérer nos activités de recherche, quelles qu'elles soient, théoriques et/ou appliquées, autour d'une méthodologie intégrée dont l'objectif est de déboucher sur des méthodes et des outils informatiques pour assurer :

- L'acquisition et la gestion des données complexes avec l'ensemble des problèmes sous-jacents de stockage, d'accès, de mise à jour, de sécurité, d'anonymat le cas échéant etc.
- L'organisation et la représentation de ces données pour assurer une fouille efficace avec toutes les questions liées au codage, à l'indexation, à la mise en forme etc.
- Le traitement des données complexes par des algorithmes de fouille capable de couvrir les besoins en description, en structuration ou en explication-prédiction avec un intérêt majeur pour les questions liées à la prise en compte des phénomènes non linéaires ;
- L'évaluation et la validation des connaissances (modèles) produites que cela soit en termes de reproductibilité des modèles (statistique) ou de cohérence des modèles (logique).
- L'intégration automatique des connaissances dans un système à base de connaissances, lequel pouvant également recevoir des connaissances en provenance de l'expert. Cela conduit à imaginer des systèmes plus ouverts pour la gestion des connaissances, dotés de formalismes de représentation unifiés.
- La mise au point de moteur d'inférence capable de fournir une réponse à une requête d'utilisateur. La requête peut se faire dans un cadre de recherche d'information ou d'une prise de décision. Il convient en outre d'imaginer des stratégies d'inférence combinant le symbolique, le numérique et le multiexpert.

Le spectre de compétences, relativement large, des chercheurs d'ERIC (bases de données, apprentissage, statistique) permet en effet d'affecter nos ressources pour couvrir de façon transversale les différents besoins évoqués pour notre projet scientifique.

Les thèses en cours au laboratoire illustrent de manière concrète les projets de recherche à court et moyen terme et, qui tous, d'une façon ou d'une autre, trouvent leur place dans les perspectives de recherche du laboratoire telles qu'elles sont esquissées. Par conséquent, nous allons surtout définir les grands axes d'orientation majeurs pour le futur : moyen et long terme.

3.5.1 XML, cadre de référence pour la fouille de données complexes

Nous allons utiliser XML comme cadre de référence pour développer les technologies pour la gestion des données complexes, l'intégration des connaissances et le développement d'outils informatiques.

3.5.1.1 Intégration de données complexes (DC)

La problématique d'intégration de DC est abordée selon deux voies :

- **L'intégration classique des DC**. Il s'agit d'intégrer physiquement des données complexes et hétérogènes, éventuellement issues de sources variées, dans une base de données cible jouant le rôle d'un sas à un entrepôt de données. Pour cela, nous travaillerons sur un processus de modélisation en trois phases conceptuelle, logique puis physique. Il est important de noter que l'objectif n'est pas seulement de stocker les données, mais aussi de les préparer à l'analyse, conformément aux tâches d'ETL (Extract, Transform, Load) classiques. Pour cela, XML a été sélectionné en tant que formalisme pivot des volets logique et physique de notre processus de modélisation. En effet, XML encapsule à la fois les données et leur schéma, soit implicitement, soit dans une définition de ce schéma. Cette représentation se retrouve dans les entrepôts, qui stockent à la fois des données et des méta-données.
- **L'intégration de données complexes par médiation basée sur des ontologies**. Contrairement à la démarche précédente, celle-ci ne centralise pas les données dans une base cible, mais les conserve dans leurs sources originelles. Pour répondre à une requête décisionnelle, un dispositif de médiation se charge alors de recueillir les données et d'alimenter un cube de données représentant le contexte d'analyse de la requête décisionnelle. Il s'agit de considérer un ensemble de sources de données hétérogènes, composé à la fois de bases de données, mais également d'applicatifs dédiés. Pour généraliser la description de ces sources, nous les représentons chacune par une ontologie locale qui décrit l'ensemble des

concepts et leurs liens. Nous construisons ensuite une ontologie globale par des techniques de fusion d'ontologies.

3.5.1.2 Modélisation multidimensionnelle en XML des DC

Les DC sont décrites par des documents XML. Le choix de construire des entrepôts XML s'est imposé naturellement. La base de données de documents XML, obtenue à l'issue de la phase d'intégration, permet d'abord une exploitation transactionnelle pour la gestion de ces documents XML. Dans notre démarche, nous recommandons une couche supplémentaire de modélisation des DC pour mieux les préparer à l'analyse. La modélisation multidimensionnelle répond à ce besoin. Les schémas en étoile illustrent bien des contextes d'analyse. Cependant, les documents XML décrivant les DC ne sont pas évidents à modéliser sous forme multidimensionnelle. Comment rapprocher alors les grammaires (DTD ou schémas XML) de documents XML du modèle en étoile ?

L'autre voie de recherche qui nous intéresse tout particulièrement concerne les entrepôts de données dynamiques. Il s'agit de définir des règles d'analyse, de les structurer et de les stocker dans l'entrepôt. L'utilisation des services web sera utile pour la programmation de scénarios d'analyses dans les entrepôts de données actifs. Nous explorons les concepts de règles d'analyse et de graphes d'analyse.

3.5.1.3 Fouille de documents XML

Afin d'exploiter autrement une base de documents XML obtenue à l'issue de l'intégration des données complexes, nous nous intéressons à l'extraction des connaissances à partir de la structure de documents XML. En particulier, nous souhaitons dégager les liens existants entre les balises d'un document XML. Nous exploiterons à cette fin les méthodes de fouille de données et plus particulièrement les règles d'association.

3.5.2 Variétés non linéaires et prétopologie pour la fouille de données

La plupart des outils mathématiques utilisés, notamment en fouille de données, sont issus, en grande partie, de l'algèbre linéaire. On pense par exemple aux méthodes d'analyse factorielles, à certaines approches en classification automatique ou en apprentissage où l'on fait appel à des distances euclidiennes qui reposent sur des hypothèses qui ne sont presque jamais discutées par les utilisateurs au moment de leur mise en œuvre. Certes, quand les données se situent dans des variétés linéaires, ces méthodes donnent des résultats souvent probants. Mais que se passe-t-il quand les données sont issues de variétés non linéaires. Pensons par exemple à des données qui se situeraient le long de deux spirales torsadées semblables aux molécules d'ADN comme sur la figure a, ou organisées sur une spirale plane comme en Figure b, ou même simplement en spirale unidimensionnelle comme en figure c.

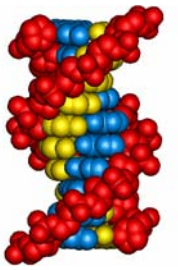

Figure a

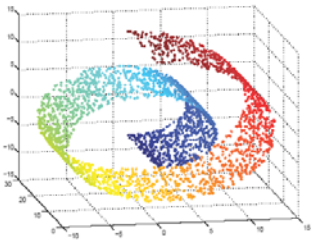

Figure b

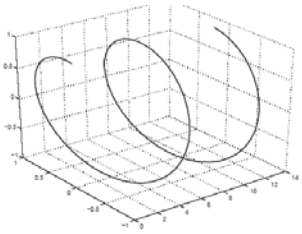

Figure c

Malheureusement dans ce contexte, la plupart des outils de fouille de données échouent dans la restitution du modèle caché. Parmi les principales causes de cet échec on peut en identifier au moins trois.

- L'usage des métriques euclidiennes qui ignorent la topologie des données. Par exemple, la distance entre deux points sur un cercle sera mesurée par la corde qui relie les deux points (distance euclidienne) alors que, naturellement, cela devrait être la longueur de l'arc de cercle (distance géodésique) qui les relie comme indiqué sur la figure d.

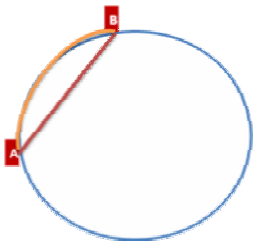

Figure d

- La non prise en compte de la dimension dans laquelle les objets se trouvent réellement. En effet, regardons les Figure b et c où chaque point i sur les variétés est caractérisé par des coordonnées dans un espace à trois dimensions (x_i, y_i, z_i). Or, dans la figure b tous les points sont dans une variété géométrique à deux dimensions (sans épaisseur) on pourrait alors imaginer de les caractériser par des coordonnées dans un plan (a_i, b_i) sans pour autant dire que les points sont dans un plan. Il en est de même de la figure c où, cette fois ci, les données sont dans une variété à 1 dimension (la courbe) et il suffirait alors d'une seule coordonnée (t_i) pour repérer un point dans son vrai voisinage. Il faut rester conscient du fait que si on ramenait les objets à leur vraie dimension on perdrait alors la morphologie du nuage de points, en revanche on conserverait la topologie exprimée par le voisinage local. Autrement dit, ce que l'on conserverait de manière assez fidèle serait la topologie au détriment de la morphologie. Ainsi, deux points voisins dans l'espace d'origine le resteraient-ils dans l'espace

de projection. Comme si, d'une certaine manière, on plongeait les objets dans des variétés linéaires.

- Enfin, en troisième, vient l'énorme privilège accordé à l'aspect métrique au détriment de l'aspect topologique. En effet la notion de voisinage dont on s'attend à ce qu'elle soit symétrique ne l'est plus. En effet, dès lors que nous sommes en présence de dispersions locales différentes la symétrie du voisinage se perd. Par exemple sur la figure e, le point y est voisin de x, au sens du plus proche voisin, alors que le voisin de y est t, toujours selon cette même notion. Pourtant, une propriété de voisinage symétrique nous semble naturellement requise même en présence de dispersions locales différentes.

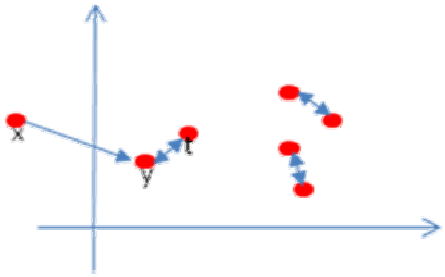

Figure e

Les trois causes cumulées rendent, du moins dans certains cas, les problèmes de classification ou d'apprentissage plus difficiles qu'ils ne le sont. Pour casser cette complexité nous faisons alors appel aux méthodes non linéaires où par exemple, les distances euclidiennes sont remplacées par les distances géodésiques. Par la suite le passage vers un cadre linéaire s'effectue via des méthodes de décomposition spectrales comme le MultiDimensional Scaling (MDS).

Jusque là nous avons imaginé que les objets analysés étaient plongés dans un espace de forte dimension et que cette dimension initiale est donnée par le nombre de composantes du vecteur de description. Or, dans de nombreuses applications nous ne connaissons même pas l'espace de description original. Par exemple, sur une ontologie, nous pouvons calculer des proximités sémantiques entre concepts, avoir besoin et pouvoir faire des classifications sur des concepts, sans avoir de vecteur associé à chaque concept. Le calcul des valeurs de proximité repose essentiellement sur la topologie exprimée par l'arbre, ou de manière plus générale le treillis, de concepts.

En conclusion, on peut dire qu'une reconsidération des méthodes de fouille de données dans un cadre mathématique qui repose d'une part sur les variétés géométriques et, d'autre part, sur la topologie nous paraît constituer une voie et un cadre mathématique de recherche riches et prometteurs.

Pour terminer, il convient de préciser que, d'ores et déjà, de nombreux résultats existent tant sur les volets des variétés mathématiques que sur les aspects de la topologie. Par exemple, il existe des méthodes d'analyse factorielle non linéaire, comme il existe de nombreux travaux sur l'apprentissage

sur variétés géométriques (manifold learning). On assiste à l'émergence d'un courant, au sein de la communauté d'apprentissage, autour de l'apprentissage topologique. Nos futurs travaux s'inséreront, en grande partie, dans ce cadre où nous pensons pouvoir apporter des contributions originales.

3.6 Plate-forme logicielle

Pour appuyer les recherches théoriques, afin de donner un caractère concret à nos travaux et assurer nos expérimentations, nous allons développer une plate-forme complète, libre et ouverte permettant la fouille de données complexes, l'intégration et le déploiement des connaissances et qui fonctionne de manière incrémentale. Ainsi les données pourraient être intégrées en continu, les modèles produits et évalués et s'ils présentent un intérêt ils seront intégrés à la base de connaissances. Une interface sera également mise en place afin de permettre à l'expert d'introduire des connaissances de façon « manuelle ». Cette plate-forme s'appuiera sur les logiciels libres comme MySql, XML, Protégé, Tanagra, Weka… de sorte à faciliter aussi bien l'intégration des outils que la comparaison des résultats issus d'autres méthodes et la diffusion de nos solutions.

3.7 Recherche de projets applicatifs en univers SHS

Nous allons privilégier des applications dans le domaine des Sciences Humaines et Sociales pour profiter des compétences qui existent au sein de Lyon 2 et ainsi favoriser des synergies de recherche locales. L'une des applications visées est la synthèse vocale à partir de corpus qui intègre la prosodie. Des linguistes ainsi que des littéraires de Lyon 2 s'intéressent à ce sujet et nous pensons construire avec eux, et des équipes spécialisées en traitement de la parole et la synthèse vocale, un projet ANR.

4 VALORISATION SCIENTIFIQUE

4.1 Publications

Le tableau ci-dessous résume le bilan scientifique quantitatif sur la période 2004-2007. Une liste complète des publications est donnée dans la section 5.

Publications	**2004**	**2005**	**2006**	**2007**	**Total**
Revues internationales	6	4	4	5	**19**
Revues nationales	4	2	2	1	**9**
Conférences internationales	14	21	33	22	**90**
Conférences nationales	17	16	20	19	**72**
Ouvrages	0	2	1	2	**5**
Chapitres d'ouvrages	1	2	3	9	**15**
Total	**42**	**47**	**63**	**58**	**210**
Chercheurs statutaires	11	11	11	11	**/**
Thèses soutenues	3	1	4	1	**9**
HDR soutenues	1	0	2	1	**4**
Etudiants en DEA/Master ECD	31	30	25	35	**121**

Tableau 1 : Bilan scientifique 2004-2007

4.2 Activités Editoriales

- D. A. Zighed (ERIC, Lyon2) et G. Venturini (LI, Tours) sont co-directeurs 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. La liste complète des numéros parus depuis 2004 se trouve dans l'annexe II.
- J. Darmont est membre des comités éditoriaux suivants :

  - International Journal of Biomedical Engineering and Technology (IJBET, see http://www.inderscience.com/browse/index.php?journalCODE=ijbet)

  - Idea Groupe Inc. Editorial Advisory Review Board

- Editorial Review Board of the Advances in Data Warehousing and Mining (ADWM, see http://users.monash.edu.au/%7Edtaniar/book_series_warehousing.html)

- F. Bentayeb, O. Boussaid, J. Darmont et S. Loudcher font partie du comité de pilotage de la conférence "Entrepôts de Données et Analyse en ligne (EDA)".

4.3 Animations scientifiques

Les membres du laboratoire ERIC sont régulièrement impliqués dans de nombreuses manifestations scientifiques, conférences, séminaires, groupes de travail, …

4.3.1 Conférences et ateliers

- Conférence "Extraction et Gestion des Connaissances" (EGC 2000-2008).
- Conférence "Entrepôts de Données et Analyse en Ligne" (EDA 2005-2008).
- Atelier "Mining Complex Data" en association avec ICDM IEEE International Conference (2005-2006) et PKDD International conference (2007).
- Atelier "Qualité des Données et des Connaissances", en association avec la conférence "Extraction et Gestion des Connaissances" (2007-2008).
- Atelier "Fouille de Données Complexes", en association avec la conférence "Extraction et Gestion des Connaissances" (2007-2008).
- Atelier "Mesure de similarité sémantique" en association avec la conférence "Extraction et Gestion des Connaissances" (2007-2008).
- Atelier "Systèmes Décisionnels" (ASD 2006-2008).

Le détail de ces manifestations est dans l'annexe III.

4.3.2 Groupes de travail

Le laboratoire ERIC a créé et organise régulièrement le groupe de travail sur la "Fouille de Données Complexes" (http://eric.univ-lyon2.fr/~gt-fdc/).

4.3.3 Séminaires

Le laboratoire ERIC organise, deux fois par mois, des séminaires avec des intervenants d'horizons différents (http://eric.univ-lyon2.fr/index.php?section=6&soussection=14 et http://dea-ecd.univ-lyon2.fr/?page=seminaire&section=0). Les séminaires 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.

La liste complète des séminaires figure en annexe III.

4.4 Projets de recherche appliquée

Les collaborations des membres d'ERIC (voir la liste complète en annexe IV et V) 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 ;
- des équipes nationales ou dans le cadre de projets nationaux ACI : TELECOM, LSIIT et LIV, IRC;
- l'industrie : Crédit Lyonnais, Védior Bis, l'hôpital Léon Berard, … ;
- 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.

Le laboratoire ERIC est également impliqué dans l'incubation et la création d'entreprises innovantes. Il leur offre l'expertise scientifique par le biais de différents types de collaborations.

Avec l'incubateur CREALYS, ERIC est actuellement impliqué dans la création de trois sociétés dans différents secteurs :

• MAP (2003-2004) : archivage, structuration et interrogation de données médicales pour l'aide au diagnostic et à la prescription ;

• TradingBots (2006-2007) : outils pour l'analyse des données financières et l'aide à la décision boursière ;

• Tapeo (2007-2008) : gestion de portefeuilles virtuels d'actions détenues par des communautés d'utilisateurs sur le Web.

4.5 Développement de logiciels

TANAGRA est un logiciel (*open source*) de fouille de données destiné à l'enseignement et à la recherche. Le projet a commencé en Janvier 2004. Le logiciel est disponible gratuitement sur le Web[1]. TANAGRA met en œuvre diverses méthodes statistiques et d'apprentissage automatique. A ce jour, il y a environ 130 méthodes implémentées.

Le principal objectif du projet est de proposer aux chercheurs et aux enseignants un outil qui respecte les normes du domaine de la fouille de données. Les utilisateurs utilisent le logiciel pour des études universitaires, pour leurs activités de recherche, mais aussi pour leurs publications. TANAGRA est maintenant reconnu par la communauté. Il est référencé dans les études comparatives et dans les projets avec des données réelles (par exemple, X. Chen, Y. Ye, G. Williams, X. Xu, "A Survey of Open Source Data Mining Systems", Industrial Track Workshop, PAKDD-2007, 3-14.).

Le logiciel est également sur un site avec de nombreux tutoriels (en français et en anglais) sur l'extraction de données et l'analyse exploratoire des données. Il y a environ 70 tutoriels en ligne. Les pages Web concernant les didacticiels sont les pages les plus visitées du site. Au cours de l'année 2007, il y a eu en moyenne près de 4000 visiteurs par mois (# 130 visiteurs par jour).

4.6 Synergie entre enseignement et recherche

ERIC a toujours connecté ses activités d'enseignement avec ses travaux de recherche, plus particulièrement avec des masters. Nous avons mis en place une palette complète de formations afin de répondre aux besoins des entreprises et de fournir au laboratoire des chercheurs et des doctorants.

Au niveau de la licence, les deux voies principales sont :

- Informatique décisionnelle et économétrie appliquée (IDEA) à la faculté de Sciences Economiques et de Gestion ;
- Mathématiques, Informatique et Statistiques Appliquées aux Sciences Humaines et Sociales (MISASHS).

Au niveau du master, nous avons mis en place des formations qui s'articulent autour des systèmes d'aide à la décision et de la statistique. Quatre spécialisations sont possibles pour la deuxième année de master :

- Statistique, Informatique et Socio-Economique (SISE). Son contenu forme les étudiants au traitement statistique des données dans les domaines du marketing ou de l'industrie

[1] http://eric.univ-lyon2.fr/~ricco/Tanagra/

(pharmacie, etc.) Cette spécialité est également réalisée dans le contexte d'un master commun avec l'Université de Kharkov (Ukraine).

- Ingénierie Informatique pour la Décision et l'Evaluation Economiques (IIDEE) dont le contenu est plus ciblé sur le développement d'outils pour les systèmes d'aide à la décision. Cette spécialité forme chaque année une deuxième promotion en cours du soir pour des professionnels.
- Organisation et Protection des Systèmes d'Information dans les Entreprises (OPSIE) dont le programme a trois facettes : technologies de l'information, la gestion et le droit. Cette spécialité forme également chaque année une deuxième promotion en cours du soir pour des professionnels.
- Extraction des Connaissances à partir des Données (ECD) qui met plus l'accent sur nos activités de recherche et qui forme la plus grande partie de nos futurs doctorants. Cette spécialité est co-habilitée depuis sa création par l'École Polytechnique de l'Université de Nantes et jusqu'en 2006 par l'université d'Orsay Paris 11. Les cours sont assurés en visioconférence ce qui permet à certains de nos étudiants de suivre les cours depuis leur lieu de résidence en France ou à l'étranger. C'est le cas, par exemple, pour les étudiants roumains de Bucarest et les étudiants vietnamiens de Cantho.

Le détail des cours est disponible sur le site de l'université et en particulier sur celui du département d'informatique et de statistique de la faculté de sciences économiques et de gestion (http://dis.univ-lyon2.fr/).

Nous poursuivons nos efforts d'ouverture pour établir des partenariats avec d'autres universités, y compris européennes, et attirer des bons étudiants. Dans cette perspective, et pour donner un impact plus fort, nous allons soumettre à la commission européenne, un projet de création d'un master européen dans le cadre du programme Erasmus Mundus. Il sera positionné dans le domaine de l'extraction de données et de la gestion des connaissances. Il sera mis en place avec 6 universités dont 3 étrangères (Italie, Espagne et Roumanie). D'autres actions sont également prévues dans le domaine professionnel avec des écoles d'ingénieurs ou de commerce, ou avec des universités étrangères pour des formations délocalisées.

5 RESSOURCES

5.1 Bilan financier

Les deux tableaux suivants présentent le bilan financier pour la période 2004-2007.

Ressources annuelles (HT)	2004	2005	2006	2007	**Total**	**Moyenne**
Crédits du Ministère Direction de la Recherche (BQR déduit)	41 650 €	41 650 €	41 650 €	37 400 €	**162 350 €**	**40 588 €**
Ressources supplémentaires provenant de Lyon 2		738 €	1 000 €	1 000 €	**2 738 €**	**913 €**
Ressources propres (contrats de recherche, prestations, …)	73 738 €	13 593 €	33 784 €	25 510 €	**146 625 €**	**36 656 €**
Collectivités territoriales	16 600 €	4 250 €	4 500 €	33 850 €	**59 200 €**	**14 800 €**
Communauté européenne		20 898 €	30 602 €		**51 500 €**	**25 750 €**
Fonds National pour la Science (ACI,)	34 078 €	20 700 €	27 700 €		**82 478 €**	**27 493 €**
Fonds pour la Recherche et la Technologie					**0 €**	
Total	**166 066 €**	**101 829 €**	**139 236 €**	**97 760 €**	**504 891 €**	**126 223 €**

Tableau 2 : Recettes financières pour 2004-2007

Dépenses annuelles (TTC)	2004	2005	2006	2007	**Total**	**Moyenne**
Fonctionnement	108 087 €	81 275 €	83 944 €	62 558 €	**335 864 €**	**83 966 €**
Equipement	16 844 €	18 958 €	39 620 €	12 636 €	**88 058 €**	**22 015 €**
Personnel	27 495 €	25 204 €	13 993 €	6 512 €	**73 203 €**	**18 301 €**
Total	**152 426 €**	**125 437 €**	**137 557 €**	**81 705 €**	**497 126 €**	**124 281 €**

Tableau 3 : Dépenses pour 2004-2007

5.2 Ressources humaines au 31 décembre 2007

5.2.1 Enseignants-chercheurs statutaires

Nom, Prénom, Date de Naissance	Corps grade	Section CNU	Date d'arrivée
Bentayeb Fadila, 15 mai 1966	MCF	27	oct-01
Bousaïd Omar, 2 juin 1954	MCF	27	janv-95
Chauchat Jean-Hugues, 6 juillet 1946	PR2	27	juin-97
Darmont Jérôme, 15 janvier 1972	MCF	27	oct-99
Harbi Nouria, 27 août 1961	MCF	27	oct-05
Lallich Stéphane, 20 septembre 1947	PR2	27	juin-97
Loudcher Rabaséda Sabine, 27 octobre 1969	MCF	27	oct-98
Rakotomalala Ricco, 19 juillet 1967	MCF	27	oct-98
Velcin Julien, 09 mars 1978	MCF	27	oct-07
Viallaneix Jacques, 6 juillet 1963	MCF	27	janv-95
Zighed Abdelkader, 12 mars 1955	PR1	27	janv-95

5.2.2 ATER

Nom, Prénom	Année universitaire
Arigon Anne-Muriel	2006-2007 2007-2008
Favre Cécile	2007-2008
Lefort Virginie	2006-2007 2007-2008
Mahboubi Hadj	2007-2008
Maïz Nora	2007-2008

5.2.3 Thèses en cours

Nom, Prénom	Début	Directeur	Co directeur	Financement
Bahri Emma	2006	S. Lallich		Bourse MENRT
Bodin-Niemczuk Anouck	2007	O. Boussaid	S. Loudcher	Bourse MENRT Moniteur
Bouatour Sonia	2007	O. Boussaid (Co-tutelle avec la Tunisie)		Ressources propres
Charbel Julien	2004	D. Zighed L. Saitta (Co-tutelle avec l'Italie)		Ressources propres
El Sayed Ahmad	2004	D. Zighed	F. Bentayeb	Ressources propres
Gaudin Rémi	2004	D. Zighed		Bourse MENRT Moniteur
Hacid Hakim	2004	D. Zighed		Bourse Région
Hachicha Marouane	2007	J. Darmont		Bourse MENRT Moniteur
Mahboubi Hadj	2005	J. Darmont		Ressources propres
Maïz Nora	2005	O. Boussaid	F. Bentayeb	Ressources propres
Marcellin Simon	2004	D. Zighed		Bourse CIFRE
Mavrikas Efthimios	2002	D. Zighed S. Dascalopoulos (Co-tutelle avec l'Université de l'Egée)		Bourse de la Grèce
Prudhomme Elie	2005	S. Lallich		Bourse MENRT
Qureshi Taimur	2006	D. Zighed		Bourse du gouvernement pakistanais
Ralaivao Jean-Christian	2003	S. Lallich V. Manantsoa (Co-tutelle avec l'Université de Fianarantsao)	J. Darmont	Bourse de l'Ambassade de France
Rakotoarivelo Ony	2006	J. Darmont	F. Bentayeb	Bourse MENRT
Salem Rashed Kh.	2007	O. Boussaid et J. Darmont		Bourse du gouvernement égyptien
Stavrianou Anna	2005	JH. Chauchat		Bourse MENRT
Thomas Julien	2005	D. Zighed		Bourse CIFRE
Wei Zhihua	2006	JH. Chauchat		Bourse du gouvernement chinois

5.2.4 Thèses soutenues

Nom, Prénom	Année	Directeur	Co directeur	Devenir
Aouiche Kamel	2005	D. Zighed	J. Darmont	Post Doc au Canada
Baume Laurent	2004	N. Nicoloyannis	C. Mirodatos	Post Doc en Espagne
Ben Messaoud Riadh	2006	N. Nicoloyannis	O. Boussaid S. Loudcher	MCF en Tunisie
Clech Jérémy	2004	D. Zighed		Privé
Clerc Frédéric	2006	N. Nicoloyannis	R. Rakotomalala	Privé
Erray Walid	2006	D. Zighed		Privé
Fangseu Badjio Edwige	2006	D. Zighed	F. Poulet	
Favre Cécile	2007	O. Boussaid	F. Bentayeb	ATER
Legrand Gaëlle	2004	N. Nicoloyannis		Privé

5.2.5 Habilitations à diriger des recherches

Nom, Prénom	Année	Directeur	Devenir
Lenca Philippe	2007	D. Zighed	
Boussaid Omar	2006	D. Zighed	Lyon 2
Darmont Jérôme	2006	D. Zighed	Lyon 2
Poulet François	2004	D. Zighed	

5.2.6 Personnel admninistratif

Nom, Prénom	Corps grade	Quotité recherche	Date d'arrivée
Gabrièle Valérie	IATOS	0,5	Sept-00
Crevel Julien	Technicien	0,5	Sept-07

5.2.7 Récapitulatif au 31 décembre 2007

Catégorie	Effectif
Enseigants-chercheurs statutaires	11
ATER	5
Thèses en cours	20
Thèses soutenues	9
HDR soutenues	4
Personnel administratif	2

5.2.8 Personnes ayant terminé leur contrat ou quitté le laboratoire

Enseignants-chercheurs statutaires

Nom, Prénom	Corps grade	Section CNU	Date d'arrivée	Date de départ
Viallefont Anne	MCF	26	oct-00	sept-06

ATER

Nom, Prénom	Année universitaire
Ben Messaoud Riadh	2006-2007
Clech Jérémy	2003-2004
Effantin Dit Toussaint Brice	2004-2005
Kouomou Choupo Anicet	2005-2006
Legrand Gaëlle	2004-2005
Muhlenbach Fabrice	2002-2003
Scuturici Marian	2003-2004
Scuturici Michaela	2002-2003 et 2003-2004
Suchier Maxime	2006-2007
Tweed Tiffany	2002-2003 et 2003-2004
Walid Erray	2003-2004 et 2004-2005

Post Doc rattaché au laboratoire

Nom, Prénom	Année universitaire
Jouve Pierre	2004-2005

Personnels admnistratifs

Nom, Prénom	Financement	Quotité recherche	Date d'arrivée	Date de départ
Delhomme Lydie	Fonds propres	1	oct-02	août-04

6 PUBLICATIONS 2004-2007

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

6.1 Revues internationales

[ADBB07] J. Darmont, F. Bentayeb, O. Boussaïd, "Benchmarking Data Warehouses", *International Journal of Business Intelligence and Data Mining*, Vol. 2, No. 1, 2007, 79-104.

[ABDFU07] F. Bentayeb, J. Darmont, C. Favre, C. Udréa, "Efficient On-Line Mining of Large Databases", International Journal of Business Information Systems, Vol. 2, No. 3, 2007, 328-350.

[ABTBD07] O. Boussaïd, A. Tanasescu, F. Bentayeb, J. Darmont, "Integration and Dimensional Modelling Approaches for Complex Data Warehousing", Journal of Global Optimization, Vol. 37, No. 4, April 2007, 571-591.

[ASAN07] A. Stavrianou, P. Andritsos, N. Nicoloyannis, "Overview and Semantic Issues of Text Mining", SIGMOD Record, Vol. 36, No. 3, September 2007, 23-34.

[ALLV07] S. Lallich, P. Lenca, B. Vaillant, "Probabilistic framework towards the parametrisation of association rule interestingness measures", Methodology and Computing in Applied Probability, Vol. 9, No. 3, 2007, 447–463.

[ACFRNM06] F. Clerc, D. Farrusseng, R. Rakotomalala, N. Nicoloyannis, C. Mirodatos, "Meta Modeling for Combinatorial Catalyst Optimization", International Journal of Computer Science and Network Security, Vol. 6, No. 10, 2006, 256-262.

[ARM06] R. Rakotomalala, F. Mhamdi, "Supervised and Unsupervised Feature Reduction for Protein Classification", WSEAS Transactions on Information Science and Applications, Vol. 3, No. 12, 2006, 2448-2455.

[AMRE06] F. Mhamdi, R. Rakotomalala, M. Elloumi, "A Compromise Between N-gram Length and Classifier Characteristics for Protein Classification", International Journal of Computer Science and Network Security, Vol. 6, No. 4, 2006, 82-87.

[ABBL06] R. BenMessaoud, O. Boussaïd, S. Loudcher-Rabaseda, "A Data Mining-Based OLAP Aggregation of Complex Data: Application on XML Documents", International Journal of Data Warehousing and Mining, Vol. 2, No. 4, Oct.-Dec. 2006, 1-26.

[AHD05] Z. He, J. Darmont, "Evaluating the Dynamic Behavior of Database Applications", Journal of Database Management, Vol. 16, No. 2, April-June 2005, 21-45.

[AZRES05] D. Zighed, G. Ritschard, W. Erray, V. Scuturici, "Decision tree with optimal join partitioning", Journal of Intelligent Information Systems, Vol. 20, 2005, 1-26.

[AZLM05] 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.

[ASCSZ05] 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.

[AGVCCM04] 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.

[ABFLM04] L. Baumes, D. Farrusseng, M. Lengliz, C. Mirodatos, "Using Artificial Neural Networks for boosting discovery in High", QSAR & Combinatorial Science, 2004.

[AKFBMS04] 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.

[APCFWMM04] 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.

[AGVLF04] 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.

[AMLZ04] F. Muhlenbach, S. Lallich, D. Zighed, "Identifying and Handling Mislabelled Instances", Journal of Intelligent Information Systems, Vol. 22, No. 1, January 2004, 89-109.

6.2 Revues nationales

[BR07] R. Rakotomalala, "Data Mining : Spécificités et outils", *Actes de Chimiométrie*, Novembre 2007, 108 - 110 (Lyon).

[BRRMJ06] M. Raimbault, R. Rakotomalala, X. Morandi, P. Jannin, "Mise en évidence d'invariants dans une population de cas chirurgicaux", *Revue des Nouvelles Technologies de l'Information*, Vol. E-5, 2006, 339-348.

[BLBFCRR06] J. Labarère, J. Bosson, D. Farrusseng, B. Crémilleux, R. Rakotomalala, C. Robert, "Arbres d'Induction : méthodes et exemple d'application", *Journal d'Economie Médicale*, Vol. 24, No. 2, 2006, 115-129.

[BADBB05] K. Aouiche, J. Darmont, O. Boussaïd, F. Bentayeb, "Auto-administration des entrepôts de données complexes", *Revue des Nouvelles Technologies de l'Information*, Vol. E-4, Septembre 2005, 47-70.

[BR05] R. Rakotomalala, "TANAGRA, une plate-forme d’expérimentation pour la fouille de données", *MODULAD*, No. 32, 2005, 70-85.

[BHD04] 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.

[BLMVPL04] 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.

[BLN04] G. Legrand, N. Nicoloyannis, "Sélection de variables et agrégation d'opinions", *Revue des Nouvelles Technologies de l'Information*, Vol. C1, 2004, 89-101 (ISBN 2.85428.667.7.).

[BLT04] S. Lallich, O. Teytaud, "Evaluation et validation de l'intérêt des règles d'association", *Revue des Nouvelles Technologies de l'Information*, No. 2, 2004, 193-217.

6.3 Conférences internationales

[CCMR07] J. Chauchat, A. Morin, R. Rakotomalala, "Correcting the error rate estimation bias in Data Mining when the dataset comes from a two-stage sampling", *Statistics for Data Mining, Learning and Knowledge Extraction (IAST 07), Aveiro, Portugal*, August 2007.

[CEHZ07] A. ElSayed, H. Hacid, D. Zighed, "Mining semantic distance between corpus terms", *1st Ph.D. Workshop, 16th ACM Conference on Information and Knowledge Management (PIKM-CIKM 07), Lisbon, Portugal*, November 2007, 49-54; ACM.

[CRD07] J. Ralaivao, J. Darmont, "Knowledge and Metadata Integration for Warehousing Complex Data", *6th International Conference on Information Systems Technology and its Applications (ISTA 07), Kharkiv, Ukraine*, May 2007; *Lecture Notes in Informatics (LNI)*, Vol. P-107, GI-Edition, Bonn, Germany, 164-175.

[CTJN07] J. Thomas, P. Jouve, N. Nicoloyannis, "Asymmetric measure for supervised learning models assessment, application to breast cancer detection", *International Conference on Industrial Engineering and Systems Management (IESM 07), Beijing, China*, May 2007.

[CEHZ07b] A. ElSayed, H. Hacid, D. Zighed, "A Multisource Context-Dependent Semantic Distance Between Concepts", *18th International Conference on Database and Expert Systems Applications (DEXA 07), Regensburg, Germany*, September 2007; *LNCS*, Vol. 4653, Springer, Heidelberg, Germany, 54-63.

[CEHZ07c] A. ElSayed, H. Hacid, D. Zighed, "Using Semantic Distance in a Content-based Heterogeneous Information Retrieval System", *3rd International Workshop on mining complex data (MCD 07), Warsaw, Poland*, 2007; *LNAI*, Springer, Heidelberg, Germany.

[CALLV07] J. Azé, P. Lenca, S. Lallich, B. Vaillant, "A Study of the Robustness of Association Rules", *2007 International Conference on Data Mining (DMIN 07), Las Vegas, USA*, 2007, 163–169; CSREA Press.

[CEHZ07d] A. ElSayed, H. Hacid, D. Zighed, "A New Approach Towards Content-based Heterogeneous Information Retrieval", *ECML/PKDD Workshop on Mining Complex Data*, 2007.

[CHMD07] M. Hachicha, H. Mahboubi, J. Darmont, "Vers une algèbre XML-OLAP : État de l'art", *2ème Atelier Systèmes Décisionnels (ASD 07), Sousse, Tunisie*, Octobre 2007.

[CSCBM07] A. Silic, J. Chauchat, B. Basic, A. Morin, "N-Grams and Morphological Normalization in Text Classification: A Comparison on a Croatian-English Parallel Corpus", *13th Portuguese Conference on Artificial Intelligence (EPIA 2007), Guimaraes, Portugal*, December

2007; *LNCS*, Vol. 4874, Springer, Heidelberg, Germany, 671-682.

[CEHZ07e] A. ElSayed, H. Hacid, D. Zighed, "A New Context-Aware Measure for Semantic Distance Using a Taxonomyand a Text Corpus", *IEEE International Conference on Information Reuse and Integration (IRI 07), Las Vegas, USA*, August 2007, 279-284; IEEE Systems, Man, and Cybernetics Society.

[CBMGNC07] L. Baumes, M. Moliner, R. Gaudin, N. Nicoloyannis, A. Corma, "Genetic Algorithms in Materials Science and Engineering", *2007 E-MRS Fall Meeting, Warsaw, Poland*, September 2007.

[CLLV07] S. Lallich, P. Lenca, B. Vaillant, "Construction of an off-centered entropy for supervised learning", *XIIth International Symposium on Applied Stochastic Models and Data Analysis (AMSDA 07), Chania, Crete, Greece*, 2007.

[CEHZ07f] A. ElSayed, H. Hacid, D. Zighed, "Combining Text and Image for Content-Based Information Retrieval", *2007 International Conference on Information andKnowledge Engineering (IKE 07)*, 2007; CSREA Press.

[CARGPSG07] E. Antajan, R. Rakotomalala, S. Gasparini, M. Picheral, L. Stemmann, G. Gorsky, "Automatic quantification and recognition of major zooplankton groups in a North Sea time series using the Zooscan imaging system", *4th International Zooplankton Production Symposium*, 2007, 189 - 190 (Hiroshima, Japan).

[CDDFWGBP07] N. Durand, S. Derivaux, G. Forestier, C. Wemmert, P. Gançarski, O. Boussaïd, A. Puissant, "'Ontology-based Object Recognition for Remote Sensing Image Interpretation", *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 07), Patras, Greece*, October 2007.

[CBNM07] E. Bahri, N. Nicoloyannis, M. Maddouri, "Improving boosting by exploiting former assumptions", *3rd International Workshop on mining complex data (MCD 07), Warsaw, Poland*, 2007.

[CAAD07] S. Azefack, K. Aouiche, J. Darmont, "Dynamic index selection in data warehouses", *4th International Conference on Innovations in Information Technology (Innovations 07), Dubai, United Arab Emirates*, November 2007; IEEE.

[CFBB07] C. Favre, F. Bentayeb, O. Boussaïd, "Dimension Hierarchy Updates in Data Warehouses: a User-driven Approach", *9th International Conference on Enterprise Information ystems (ICEIS 07), Funchal, Madeira, Portugal*, June 2007, 206 - 211.

[CEHZ07g] A. ElSayed, H. Hacid, D. Zighed, "A Context-Dependent Semantic Distance Measure", *19th International Conference on Software Engineering and Knowledge Engineering (SEKE 07), Boston, USA*, July 2007, 432-437; Knowledge Systems Institute Graduate School.

[CMBB07] N. Maiz, F. Bentayeb, O. Boussaïd, "Ontology based mediation system", *18th Information Ressource Management Association International Conference (IRMA 07), Vancouver, Canada*, May 2007; IRMA, Hershey, USA.

[CFBB07b] C. Favre, F. Bentayeb, O. Boussaïd, "Evolution of data warehouses' optimization: a workload perspective", *9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007)*, 2007; *LNCS*, Vol. 4654, Springer, Heidelberg, Germany, 13 - 22.

[CRCP06] R. Rakotomalala, J. Chauchat, F. Pellegrino, "Accuracy Estimation with Clustered Dataset", *The Australasian Data Mining Conference (AusDM 06), Sidney, Australia*, November 2006; *Conferences in Research and Practice in Information Technology*, Vol. 61.

[CBBL06] R. BenMessaoud, O. Boussaïd, S. Loudcher-Rabaseda, "Efficient Multidimensional Data Representation Based on Multiple Correspondence Analysis", *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 06), Philadelphia, USA*, August 2006.

[CLVL06] P. Lenca, B. Vaillant, S. Lallich, "On the Robustness of Association Rules", *IEEE International Conferences on Cybernetics and Intelligent Systems and Robotics, Automation and Mechatronics (CIS-RAM 06), Bangkok, Thailand*, June 2006, 596-601.

[CMRE06] F. Mhamdi, R. Rakotomalala, M. Elloumi, "A Hierarchical N-Grams Extraction Approach for Classification Problem", *IEEE International Conference on Signal-Image Technology and Internet-Based Systems (SITIS 06), Tunisia*, 2006, 310-321.

[CFBB06] C. Favre, F. Bentayeb, O. Boussaïd, "A Knowledge-driven Data Warehouse Model for Analysis Evolution", *13th ISPE International Conference on Concurrent Engineering: Research and Applications (CE 06), Antibes, France*, September 2006; *Frontiers in Artificial Intelligence and Applications*, Vol. 143, IOS Press, 271-278.

[CDO06] J. Darmont, E. Olivier, "A complex data warehouse for personalized, anticipative medicine", *17th Information Resources Management Association International Conference (IRMA 06), Washington, USA*, May 2006, 685-687; Idea Group Publishing.

[CRZ06] G. Ritschard, D. Zighed, "Implication Strength of Classification Rules", *Foundations of Intelligent Systems (ISMIS 06), Bari, Italy*, September 2006; *LNAI*, Vol. 4203, Springer, Heidelberg, Germany, 463-472.

[CGN06] R. Gaudin, N. Nicoloyannis, "An Adaptable Time Warping Distance for Time Series Learning", *5th International Conference on Machine Learning and Applications (ICMLA 06), Orlando, USA*, December 2006.

[CTGLP06] O. Teytaud, S. Gelly, S. Lallich, E. Prudhomme, "Quasi-random bootstrap, with applications to rule extraction and (sub)bagging", *International Workshop on Intelligent Information Access (IIIA 06), Helsinki, Finland*, July 2006.

[CMBB06] N. Maiz, O. Boussaïd, F. Bentayeb, "Ontology-Based Mediation System", *13th ISPE International Conference on Concurrent Engineering: Research and Applications (CE 06), Antibes, France*, September 2006; *Frontiers in Artificial Intelligence and Applications*, Vol. 143, IOS Press, 181-189.

[CMR06] A. Morineau, R. Rakotomalala, "The TVpercent Criteria to Eliminate Uninformative Models among Association Rules", *Knowledge Extraction and Modeling IASC-INTERFACE-IFCS Workshop (KNEMO 06), Anacapri, Italy*, 2006.

[CBBL06b] R. BenMessaoud, O. Boussaïd, S. Loudcher-Rabaseda, "Using a Factorial Approach for Efficient Representation of Relevant OLAP Facts", *Seventh International Baltic Conference on Databases and Information Systems (DB&IS 06), Vilnius, Lithuania*, July 2006.

[CMD06] H. Mahboubi, J. Darmont, "Benchmarking XML data warehouses", *Atelier Systèmes*

Décisionnels (ASD 06), 9th Maghrebian Conference on Information Technologies (MCSEAI 06), Agadir, Maroc, December 2006.

[CC06] J. Chauchat, "Microeconomics Forecast: Learning by Doing, A Ten Years Graduate Level Experience", *7th International Conference On Teaching Statistics (ICOTS7), Salvador, Bahia, Brazil*, July 2006.

[CAJD06] K. Aouiche, P. Jouve, J. Darmont, "Clustering-Based Materialized View Selection in Data Warehouses", *10th East-European Conference on Advances in Databases and Information Systems (ADBIS 06), Thessaloniki, Greece*, September 2006; *LNCS*, Vol. 4152, Springer, Heidelberg, Germany, 81-95.

[CMAD06] H. Mahboubi, K. Aouiche, J. Darmont, "Materialized View Selection by Query Clustering in XML Data Warehouses", *4th International Multiconference on Computer Science and Information Technology (CSIT 06), Amman, Jordan*, April 2006, 68-77.

[CGBNB06] R. Gaudin, S. Barbier, N. Nicoloyannis, M. Banens, "Clustering of Bi-Dimensional and Heterogeneous Time Series: Application to Social Sciences Data", *2006 International Conference on Data Mining (DMIN 06), Las Vegas, USA*, June 2006, 10-16.

[CRM06] R. Rakotomalala, F. Mhamdi, "Combining feature selection and feature reduction for protein classification", *2nd WSEAS International Symposium on Data Mining, Lisbon, Portugal*, 2006.

[CMZR06] S. Marcellin, D. Zighed, G. Ritschard, "Detection of breast cancer using an asymmetric entropy measure", *Computational Statistics (COMPSTAT 06)*, 2006; *Computational Statistics*, Vol. XXV, Springer, Heidelberg, Germany, 975-982 (On CD).

[CBBL06c] R. BenMessaoud, O. Boussaïd, S. Loudcher-Rabaseda, "Mining Association Rules in OLAP Cubes", *International Conference on Innovations in Information Technology (ITT 06), Dubai*, November 2006.

[CLTP06] S. Lallich, O. Teytaud, E. Prudhomme, "Statistical inference and data mining: false discoveries control", *17th COMPSTAT Symposium of the IASC, Rome, Italy*, August 2006, 325-336.

[CMBB06b] N. Maiz, F. Bentayeb, O. Boussaïd, "Un système de médiation basé sur les ontologies pour l'entreposage des données", *Atelier Systèmes Décisionnels (ASD 06), 9th Maghrebian Conference on Information Technologies (MCSEAI 06), Agadir, Maroc*, Décembre 2006.

[CZMR06] D. Zighed, S. Marcellin, G. Ritschard, "An asymmetric entropy measure for decision trees", *Knowledge Extraction and Modeling Workshop (KNEMO 06), Capri, Italy*, September 2006.

[CBBCA06] O. Boussaïd, R. BenMessaoud, R. Choquet, S. Anthoard, "X-Warehousing: an XML-Based Approach for Warehousing Complex Data", *10th East-European Conference on Advances in Databases and Information Systems (ADBIS 06), Thessaloniki, Greece*, September 2006; *LNCS*, Vol. 4152, Springer, Heidelberg, Germany, 39-54.

[CMZR06b] S. Marcellin, D. Zighed, G. Ritschard, "An asymmetric entropy measure for decision trees", *11th Information Processing and Management of Uncertainty in knowledge-based systems (IPMU 06), Paris, France*, July 2006, 1292-1299.

[CRM06b] R. Rakotomalala, F. Mhamdi, "Improved Singular Value Decomposition for Supervised Learning in a High Dimensional Dataset", *6th International Workshop on Pattern Recognition in Information Systems (PRIS 06), Paphos, Cyprus*, May 2006, 38-47.

[CZH06] D. Zighed, H. Hacid, "Proximity graphs and separability of classes", *11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 06), Paris, France*, July 2006, 1488-1495; IPMU.

[CFBB06b] C. Favre, F. Bentayeb, O. Boussaïd, "WEDriK : une plateforme pour des analyses personnalisées dans les entrepôts de données évolutifs", *Atelier Systèmes Décisionnels (ASD 06), 9th Maghrebian Conference on Information Technologies (MCSEAI 06), Agadir, Maroc*, Décembre 2006.

[CHZ06] H. Hacid, D. Zighed, "Content-Based Image Retrieval in Large Image Databases", *IEEE International Conference on Granular Computing (GrC 2006), Atlanta, USA*, May 2006.

[CVLL06] B. Vaillant, S. Lallich, P. Lenca, "Modelling of the counter-examples and association rules interestingness measures behavior", *2nd International Conference on Data Mining (DMIN 06), Las Vegas, USA*, June 2006, 132-137.

[CTJN06] J. Thomas, P. Jouve, N. Nicoloyannis, "Optimisation and evaluation of random forests for imbalanced datasets", *16th International Symposium on Methodologies for Intelligent Systems (ISMIS 06), Bari, Italy*, September 2006; *LNAI*, Vol. 4203, Springer, Heidelberg, Germany, 642-651.

[CHZ06b] H. Hacid, D. Zighed, "Content-Based Image Retrieval Using Topological Models", *12th International MultiMedia Modelling Conference (MMM 06), Beijing, China*, 2006.

[CBLBM06] R. BenMessaoud, S. Loudcher-Rabaseda, O. Boussaïd, R. Missaoui, "Enhanced Mining of Association Rules from Data Cubes", *9th ACM International Workshop on Data Warehousing and OLAP (DOLAP 06), Arlington, USA*, November 2006.

[CPHZ05] V. Pisetta, H. Hacid, D. Zighed, "Automatic Juridical Texts Classification and Relevance Feedback", *First IEEE International Workshop on Mining Complex Data (IEE MCD05), Texas, USA*, 2005.

[CHZ05] H. Hacid, D. Zighed, "An Effective Method for Locally Neighborhood Graphs Updating", *16th International Conference on Database and expert Systems Applications (DEXA 05)*, 2005; *LNCS*, Vol. 3588, Springer, Heidelberg, Germany, 930-939.

[CLN05] G. Legrand, N. Nicoloyannis, "A new feature selection method", *8th International Conference on Pattern Recognition and Information Processing (PRIP05), Minsk Belarus*, 2005.

[CPL05] 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.

[CRME05] 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, Springer, Heidelberg, Germany, 610-617.

[CBBL05] R. BenMessaoud, O. Boussaïd, S. Loudcher-Rabaseda, "Evaluation of a MCA-Based

Approach to Organize Data Cubes", *ACM Fourteenth Conference on Information and Knowledge Management (CIKM 05), Bremen, Germany*, 2005.

[CFB05] C. Favre, F. Bentayeb, "Bitmap index-based decision trees", *15th International Symposium on Methodologies for Intelligent Systems (ISMIS 05), New York, USA*, May 2005; *LNAI*, Vol. 3488, Springer, Heidelberg, Germany, 65-73.

[CADBB05] 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, Springer, Heidelberg, Germany, 64-73.

[CMKC05] A. Morin, A. Kouomou-Choupo, J. Chauchat, "Dimension reduction and clustering for query-by-example in huge image databases", *3rd IASC World Conference on Computational Statistics and Data Analysis, Limassol, Cyprus*, October 2005.

[CLN05b] 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.

[CMRE05] 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*, Springer, Heidelberg, Germany, 611-617.

[CTBB05] 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.

[CDBRA05] 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.

[CDBB05] 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, Springer, Heidelberg, Germany, 85-94.

[CLN05c] G. Legrand, N. Nicoloyannis, "Feature selection method using preferences aggregation", *International Conference on Machine Learning and Data Mining (MLDM 05), Leipzig Germany*, 2005; *LNCS*, Vol. 3587, Springer, Heidelberg, Germany, 9-11.

[CLVL05] 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.

[CCRF05] 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.

[CHZ05b] H. Hacid, D. Zighed, "An Incremental Algorithm for Neighborhood Graphs Construction", *3rd World Conference on Computational Statistics & Data Analysis, Cyprus*, October 2005.

[CMKC05b] A. Morin, A. Kouomou-Choupo, J. Chauchat, "Dimension reduction and clustering for query-by-example in huge image databases", *3rd World Conference on Computational Statistics and Data Analysis (CSDA 05), Cyprus*, November 2005.

[CHZ05c] H. Hacid, D. Zighed, "Neighborhood Graphs for Image Databases Indexing and Content-Based Retrieval", *First IEEE International Workshop on Mining Complex Data (IEE MCD05), Texas, USA*, 2005.

[CCPC05] 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.

[CMNK04] 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, Springer, Heidelberg, Germany, 477-478.

[CTBB04] 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.

[CBBR04] 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.

[CJCR04] 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 (A paraître).

[CBDU04] 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.

[CMKN04] 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 (Submitted for review).

[CHC04] 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.

[CMER04] 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.

[CBRBB04] 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.

[CVPPKMB04] 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.

[CMLZ04] 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, Springer, Heidelberg, Germany, 314-321.

[CMER04b] 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.

[CJCD04] 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.

[CVLL04] B. Vaillant, P. Lenca, S. Lallich, "A clustering of interestingness measures", *7th International Conference Discovery Science, Padova, Italy*, October 2004; *LNAI*, Vol. 3245, Springer, Heidelberg, Germany, 290-297.

6.4 Conférences nationales

[DFBB07] C. Favre, F. Bentayeb, O. Boussaïd, "Intégration des connaissances utilisateurs pour des analyses personnalisées dans les entrepôts de données évolutifs", *7èmes Journées Francophones Extraction et Gestion des Connaissances (EGC 07), Namur, Belgique*, Janvier 2007; *Revue des Nouvelles Technologies de l'Information*, Cépaduès, Toulouse, 217 - 222.

[DPRZ07] V. Pisetta, G. Ritschard, D. Zighed, "Choix des conclusions et validation des règles issues d'arbresde classification", *7ème Conférence Extraction et Gestion des Connaissances (EGC 07), Namur, Belgique*, 2007; *Revue des Nouvelles Technologies de l'Information*, Vol. E-9, Cépaduès, Toulouse, 485-496.

[DSRBGM07] M. Studer, G. Ritschard, L. Baccaro, I. Georgiou, N. Muller, "Relations entre types de violation des libertés syndicales garantiespar les conventions de l'OIT : Une analyse statistique implicativedes résultats d'une fouille de texte", *Nouveaux apports théoriques à l'analyse statistique implicative et applications*, 2007, 111-122; Departament de Matematiques, Universitat Jaume I.

[DVMMLL07] B. Vaillant, S. Menou, S. Moga, P. Lenca, S. Lallich, "Qualité des règles d'association : étude de données d'entreprise", *3ème Atelier Qualité des Connaissances à partir des Données (QDC-EGC 07), Namur, Belgique*, Janvier 2007, 55-64.

[DZPR07] D. Zighed, V. Pisetta, D. Ratsimba, "Separability of Classes in a multidimensional Space", *Classification and Data Analysis*, September 2007; *Meeting of the classification and Data Analysis Group of the Italian Statistical Society*, Eum edizioni università di macerata, 147-150.

[DTJN07] J. Thomas, P. Jouve, N. Nicoloyannis, "Mesure non symétrique pour l'évaluation de modèles, utilisation pour les jeux de modèles, utilisation pour les jeux de données déséquilibrés", *7ème Conférence Extraction et Gestion des Connaissances (EGC 07), Namur, Belgique*, Janvier 2007; *Revue des Nouvelles Technologies de l'Information*, Vol. E-9, Cépaduès, Toulouse, 509-519.

[DRB07] O. Rakotoarivelo, F. Bentayeb, "Evolution de schéma par classification automatique pour les entrepôts de données", *4ème atelier Fouille de Données Complexes dans un Processus*

d'Extraction des Connaissances (FDC-EGC 07), Namur, Belgique, Janvier 2007.

[DMBB07] N. Maiz, O. Boussaïd, F. Bentayeb, "Clustering method for semi-automatically ontologies fusion", *4ème atelier Fouille de Données Complexes dans un Processus d'Extraction des Connaissances (FDC-EGC 07), Namur, Belgique*, Janvier 2007.

[DMD07] H. Mahboubi, J. Darmont, "Fragmentation des entrepôts de données XML", *3èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 07), Poitiers*, Juin 2007; *Revue des Nouvelles Technologies de l'Information*, Vol. B-3, Cépaduès, Toulouse, 177-190.

[DRZM07] G. Ritschard, D. Zighed, S. Marcellin, "Données déséquilibrées, entropie décentrée et indice d'implication", *Nouveaux apports théoriques à l'analyse statistique implicative et applications*, 2007, 315-327; Departament de Matematiques, Universitat Jaume I; ASI4.

[DBNM07] E. Bahri, N. Nicoloyannis, M. Maddouri, "Amélioration du Boosting par combinaison des hypothèses antérieures", *14èmes Rencontres de la Société Francophone de Classification (SFC 07), Paris*, Septembre 2007.

[DFBB07b] C. Favre, F. Bentayeb, O. Boussaïd, "Evolution de modèle dans les entrepôts de données : existant et perspectives", *3èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 07), Poitiers*, Juin 2007; *Revue des Nouvelles Technologies de l'Information*, Vol. B-3, Cépaduès, Toulouse, 21-36.

[DZMR07] D. Zighed, S. Marcellin, G. Ritschard, "Mesure d'entropie asymétrique et consistante", *7ème Conférence Extraction et Gestion des Connaissances (EGC 07), Namur, Belgique*, 2007; *Revue des Nouvelles Technologies de l'Information*, Vol. E-9, Cépaduès, Toulouse, 81-86.

[DPL07] E. Prudhomme, S. Lallich, "Ensemble prédicteur fondé sur les cartes auto-organisatrices adapté aux données volumineuses", *7ème Conférence Extraction et Gestion des Connaissances (EGC 07), Namur, Belgique*, Janvier 2007; *Revue des Nouvelles Technologies de l'Information*, 473-484.

[DRB07b] O. Rakotoarivelo, F. Bentayeb, "Evolution de schéma par classification automatique pour les entrepôts de données", *3èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 07), Poitiers*, Juin 2007; *Revue des Nouvelles Technologies de l'Information*, Vol. B-3, Cépaduès, Toulouse, 99-112.

[DEHZ07] A. ElSayed, H. Hacid, D. Zighed, "Recherche d'Information par le Contenu des Données Hétérogenes", *RIAS*, 2007; IRIT, Université de Toulouse.

[DFBB07c] C. Favre, F. Bentayeb, O. Boussaïd, "Evolution et personnalisation des analyses dans les entrepôts de données : une approche orientée utilisateur", *XXVème congrès Informatique des organisations et systèmes d'information et de décision (INFORSID 07), Perros-Guirec*, Mai 2007, 308 - 323.

[DLLV07] S. Lallich, P. Lenca, B. Vaillant, "Construction d'une entropie décentrée pour l'apprentissage supervisé", *3ème Atelier Qualité des Connaissances à partir des Données (QDC-EGC 07), Namur, Belgique*, Janvier 2007, 45-54.

[DBBGPD07] R. Brisson, O. Boussaïd, P. Gançarski, A. Puissant, N. Durand, "Navigation et appariement d'objets géographiques dans une ontologie", *7ème Conférence Extraction et Gestion des Connaissances (EGC 07), Namur, Belgique*, Janvier 2007; *Revue des Nouvelles Technologies*

de l'Information, Cépaduès, Toulouse.

[DFBB06] C. Favre, F. Bentayeb, O. Boussaïd, "Evolution de schémas dans les entrepôts de données : modèle à base de règles", *2ème journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA 06), Versailles*, Juin 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. B-2, Cépaduès, Toulouse, 175-176.

[DH06] H. Hacid, "Annotation semi-automatique de grandes BD images : Approche par graphes de voisinage", *COnférence en Recherche d'Informations et Applications (CORIA 06), Lyon*, Mars 2006.

[DMBB06] N. Maiz, O. Boussaïd, F. Bentayeb, "Un système de médiation basé sur les ontologies", *3ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances, EGC 06, Lille*, Janvier 2006, 27-38.

[DBJTC06] A. Brémond, P. Jouve, J. Thomas, J. Clech, "Résultats Préliminaires d'une étude comparative de deux CAD", *Innovations Technologiques et Bonnes Pratiques en Sénologie :Dépistage - Diagnostic - Traitement*, Juin 2006, 92-94; Fusium; Sofmis.

[DMER06] F. Mhamdi, M. Elloumi, R. Rakotomalala, "Extraction et Sélection des n-grammes pour le Classement de Protéines", *Atelier Extraction et gestion de connaissances appliquées aux données biologiques (Bio-EGC 06), Lille*, Janvier 2006, 25-37.

[DBBCA06] O. Boussaïd, R. BenMessaoud, R. Choquet, S. Anthoard, "Conception et construction d'entrepôts XML", *2ème journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA 06), Versailles*, Juin 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. B-2, Cépaduès, Toulouse, 3-22.

[DMAD06] H. Mahboubi, K. Aouiche, J. Darmont, "Un index de jointure pour les entrepôts de données XML", *6èmes Journées Francophones Extraction et Gestion des Connaissances (EGC 06), Lille*, Janvier 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. E-6, Cépaduès, Toulouse, 89-94.

[DMR06] A. Morineau, R. Rakotomalala, "Critère VT-100 de sélection des règles d'association", *6èmes Journées Francophones Extraction et Gestion des Connaissances (EGC 06), Lille*, Janvier 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. E-6, Cépaduès, Toulouse, 581-592.

[DPHZ06] V. Pisetta, H. Hacid, D. Zighed, "Multi-catégorisation de textes juridiques et retour de pertinence", *6èmes Journées Francophones Extraction et Gestion des Connaissances (EGC 06), Lille*, Janvier 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. E-6, Cépaduès, Toulouse, 235-246.

[DZ06] D. Zighed, "Aspects conceptuels : Différentes Méthodologies des Systèmes Experts pour la Détection ou la Caractérisation", *Innovations Technologiques et Bonnes pratiques en Sénologie : Dépistage - Diagnostic - Traitement*, 2006; *Sofmis*, Fusium, 76-81.

[DBL06] O. Boussaïd, S. Loudcher-Rabaseda, "Intégration des méta-données dans la fouille de données", *XXIVème Congrès Informatique des organisations et systèmes d'information et de décision (INFORSID 06), Hammamet, Tunisie*, 2006.

[DFBB06b] C. Favre, F. Bentayeb, O. Boussaïd, "Modèle d'entrepôt de données à base de règles", *3ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances,*

EGC 06, Lille, 2006, 39-50.

[DFBB06c] C. Favre, F. Bentayeb, O. Boussaïd, "A Rule-based Data Warehouse Model", *23rd British National Conference on Databases (BNCOD 2006), Belfast, Northern Ireland*, July 2006; *LNCS*, Vol. 4042, Springer, Heidelberg, Germany, 274-277.

[DGN06] R. Gaudin, N. Nicoloyannis, "Séries temporelles : Vers une mesure de distance optimale", *Fouille de données temporelles, 6èmes Journées d'Extraction et de Gestion des Connaissances (EGC 06), Lille*, Janvier 2006, 67-75.

[DHZ06] H. Hacid, D. Zighed, "Graphes de Proximité pour l'Indexation et l'Interrogation d'Images par le Contenu", *6èmes Journées Francophones Extraction et Gestion des Connaissances (EGC 06), Lille*, Janvier 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. E-6, Cépaduès, Toulouse, 11-22.

[DMAD06b] N. Maiz, K. Aouiche, J. Darmont, "Sélection automatique d'index et de vues matérialisées dans les entrepôts de données", *2ème journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA 06), Versailles*, Juin 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. B-2, Cépaduès, Toulouse, 89-104.

[DPHBR06] V. Pisetta, H. Hacid, F. Bellal, G. Ritschard, "Traitement automatique de textes juridiques", *Semaine de la Connaissance (SdC 06), Nantes*, Juin 2006 (CDrom).

[DBINZ06] A. Brémond, A. Isnard, N. Nicoloyannis, D. Zighed, "Numérisation secondaire et Lecture sur Ecran : Evaluation desPerformances", *Innovations Technologiques et Bonnes Pratiques en Sénologie : Dépistage - Diagnostic - Traitement*, 2006, 22-28; Fusium.

[DRM06] R. Rakotomalala, F. Mhamdi, "Evaluation des Méthodes Supervisées pour le Classement de Protéines", *13èmes Rencontres de la Société Française de Classification (SFC 06), Metz*, Septembre 2006, 181-184.

[DZI06] D. Zighed, A. Isnard, "Projet NORDOM (Numérisation, Optimisation, Rationalisation du Dépistage Organisé en Mammographie", *Innovations Technologiques et Bonnes Pratiques en Sénologie : Dépistage - Diagnostic - Traitement*, 2006, 29-34; Fusium; Sofmis.

[DFBBN05] 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.

[DBLB05] 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.

[DPL05] 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*, Cépaduès, Toulouse, 79-90.

[DGN05] 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*, Cépaduès, Toulouse, 201-212.

[DFB05] 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*, Cépaduès, Toulouse, 319-330.

[DLLV05] S. Lallich, P. Lenca, B. Vaillant, "Variations autour de l'intensité d'implication", *Colloque Analyse Statistique Implicative (ASI 2005), Palerme, Sicile*, Octobre 2005, 237-246.

[DR05] 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*, Cépaduès, Toulouse, 697-702.

[DUB05] 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*, Cépaduès, Toulouse, 356.

[DBRB05] 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.

[DLN05] 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.

[DRRMJ05] 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.

[DR05b] 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.

[DJN05] 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*, Cépaduès, Toulouse, 31-42.

[DLN05b] G. Legrand, N. Nicoloyannis, "Gestion de la phase de prétraitement des données et coefficient Kappa", *XXXVIIèmes Journées de Statistique, Pau*, 2005, 6-10; SFdS.

[DBAF05] 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, Cépaduès, Toulouse, 34-50.

[DVMPLLB05] 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.

[DJLN04] 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, Cépaduès,

Toulouse, 185-196.

[DLN04] G. Legrand, N. Nicoloyannis, "Sélection de variables et agrégation d'opinions", *11èmes Rencontres de la Société Francophone de Classification (SFC 04), Bordeaux*, 2004.

[DVLL04] 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, Cépaduès, Toulouse, 341-352.

[DUBDB04] 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, Cépaduès, Toulouse, 83-94.

[DLN04b] 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, Cépaduès, Toulouse, 204 (Poster).

[DLN04c] 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*, Cépaduès, Toulouse.

[DLM04] 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, 23-28.

[DSSCZ04] 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.

[DSCSZ04] 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, Cépaduès, Toulouse, 409-414.

[DE04] 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.

[DLPT04] 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, Cépaduès, Toulouse, 305-316.

[DJC04] R. Jalam, J. Chauchat, "Catégorisation de textes multilingues: quelques solutions", *Atelier Fouille de Textes, EGC 04, Clermont-Ferrand*, Janvier 2004, 27-36.

[DDBLB04] 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, Cépaduès, Toulouse, 519-524.

[DDBB04] 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.

[DADB04] 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.

[DLN04d] G. Legrand, N. Nicoloyannis, "Nouvelle méthode de construction de variables", *11èmes Rencontres de la Société Francophone de Classification (SFC 04), Bordeaux*, 2004.

[DBRBB04] 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, Cépaduès, Toulouse, 35-46.

6.5 Chapitres d'ouvrage

[EAD07] K. Aouiche, J. Darmont, "Index and Materialized View Selection in Data Warehouses", *Encyclopedia of Database Technologies and Applications, Second Edition*, Idea Group Publishing, 2007.

[ELVML07] P. Lenca, B. Vaillant, P. Meyer, S. Lallich, "Association rule interestingness measures: experimental and theoretical studies", *Quality Measures in Data Mining*, Springer, Heidelberg, Germany, 2007, 51-76.

[EBL07] R. BenMessaoud, S. Loudcher-Rabaseda, "OLEMAR: an On-Line Environment for Mining Association Rules in Multidimensional Data", *Advances in Data Warehousing and Mining*, Vol. 2, Idea Group Publishing, 2007.

[EFBB07] C. Favre, F. Bentayeb, O. Boussaïd, "A Survey of Data Warehouse Model", *Encyclopedia of Database Technologies and Applications, Second Edition*, Idea Group Publishing, 2007.

[EZ07] D. Zighed, "Induction Graphs for Data Mining", *Studies in Classification, Data Analysis and Knowledge Organisation*, Springer, Heidelberg, Germany, 2007, 419-430 (In Selected Contributions in Data Analysis and Classification).

[EBBL07] R. BenMessaoud, O. Boussaïd, S. Loudcher-Rabaseda, "A multiple correspondence analysis to organize data cubes", *Databases and InformationSystems IV - Frontiers in Artificial Intelligence and Applications*, Vol. 155(1), IOS Press, 2007, 133-146.

[EMD07] H. Mahboubi, J. Darmont, "Indices in XML databases", *Encyclopedia of Database Technologies and Applications, Second Edition*, Idea Group Publishing, 2007.

[EBAGB07] M. Bouet, M. Aufaure, P. Gançarski, O. Boussaïd, "Pattern Mining and Clustering on Image Databases", *Successes and New Directions in Data Mining*, Idea Group Publishing, 2007, 187-212.

[ERL07] R. Rakotomalala, T. LeNouvel, "Interactive Clustering Tree : Une méthode de classification descendante adaptée aux grands ensembles de données", *Revue des Nouvelles*

Technologies de l'Information, Vol. A1, Cépaduès, Toulouse, 2007, 75-94 (In Data Mining et apprentissage statistique : application en assurance, banque et marketing).

[EHD06] Z. He, J. Darmont, "Evaluating the Performance of Dynamic Database Applications", *Advanced Topics in Database Research*, Vol. 5, Idea Group Publishing, 2006, 294-319.

[EHZ06] H. Hacid, D. Zighed, "A Machine Learning Based Model For Content Based Image Retrieval", , 2006.

[ELTP06] S. Lallich, O. Teytaud, E. Prudhomme, "Association rules interestingness: measure and validation", *Quality Measures in Data Mining*, Springer, Heidelberg, Germany, 2006.

[ED05] J. Darmont, "Object Database Benchmarks", *Encyclopedia of Information Science and Technology*, Vol. 1, Idea Group Publishing, January 2005, 2146-2149.

[EMR05] F. Muhlenbach, R. Rakotomalala, "Discretization of Continuous Attributes", *Encyclopedia of Data Warehousing and Mining, Second Edition*, Idea Group Publishing, 2005, 397-402.

[EBA04] O. Boussaïd, M. Aufaure, "Spatial Data Warehouses: a methodological framework", *Advances in Spatial Analysis and Decision Making*, A.A. Balkema, 2004, 275-282.

[EAD07] K. Aouiche, J. Darmont, "Index and Materialized View Selection in Data Warehouses", *Encyclopedia of Database Technologies and Applications, Second Edition*, Idea Group Publishing, 2007.

6.6 Ouvrages et actes (Eds.)

[FLP07] S. Lallich, D. Pastor, *Special Issue on the ASMDA International Symposium on Applied Stochastic Models and Data Analysis*, *Communications in Statistics - Theory and Methods*, Vol. 36(14), Taylor & Francis, January 2007 (Edited special issue).

[FLLG07] S. Lallich, P. Lenca, F. Guillet, *Actes du 3ème Atelier Qualité des Données et des Connaissances (QDC-EGC 07), Namur, Belgique*, EGC, Janvier 2007.

[FDB06] J. Darmont, O. Boussaïd, *Managing and Processing Complex Data for Decision Support*, Idea Group Publishing, April 2006.

[FBBDL05] F. Bentayeb, O. Boussaïd, J. Darmont, S. Loudcher-Rabaseda, *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, Toulouse, Juin 2005.

[FBGMT05] O. Boussaïd, P. Gançarski, F. Masseglia, B. Trousse, *Fouille de Données Complexes*, *Revue des Nouvelles Technologies de l'Information*, Vol. 3, Cépaduès, Toulouse, 2005.

6.7 Thèses et HDR

[GL07]P. Lenca, "Des données à la décision : apprentissage, validation et exploitation de règles", Université Lumière Lyon 2, Novembre 2007 (Mémoire scientifique d'Habilitation à Diriger des Recherches).

[GF07] C. Favre, "Évolution de schémas dans les entrepôts de données : mise à jour de hiérarchies de dimension pour la personnalisation des analyses", Université Lumière Lyon 2, Décembre 2007(Thèse de doctorat).

[GD06] J. Darmont, "Optimisation et évaluation de performance pour l'aide à la conception et à l'administration des entrepôts de données complexes", Université Lumière Lyon 2, Novembre 2006 (Mémoire scientifique d'Habilitation à Diriger des Recherches).

[GB06b] O. Boussaïd, "Evolution de l'entreposage des données complexes", Université Lumière Lyon 2, Décembre 2006 (Mémoire scientifique d'Habilitation à Diriger des Recherches).

[GB06] R. BenMessaoud, "Couplage de l'analyse en ligne et de la fouille de données pour l'exploration, la classification et l'explication des données complexes", Université Lumière Lyon 2, Novembre 2006 (Thèse de doctorat).

[GE06] W. Erray, "Extensions et nouvelles approches en Extraction des Connaissances à partir des données", Université Lumière Lyon 2, Décembre 2006 (Thèse de doctorat).

[GC06] F. Clerc, "Optimization and datamining for catalysts design", Université Lumière Lyon 2, septembre 2006 (Thèse de doctorat).

[GA05] K. Aouiche, "Techniques de fouille de données pour l'optimisation automatique des performances des entrepôts de données", Université Lumière Lyon 2, Décembre 2005 (Thèse de doctorat).

[GF05] E. P. Fangseu Badjio, "Evaluation qualitative et guidage des utilisateurs en Fouille visuelle de données", Université Lumière Lyon 2, 2005 (Thèse de doctorat).

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

[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", Université Lumière Lyon 2, 2004 (Thèse de doctorat).

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

[GP04]F.Poulet, "Visualisation et extraction de connaissances", Université Lumière Lyon 2, Novembre 2004 (Mémoire scientifique d'Habilitation à Diriger des Recherches).

ANNEXES

I. FICHES INDIVIDUELLES D'ACTIVITE

Anne-Muriel ARIGON

Current Position : Assistant professor
E-mail : anne-muriel.arigon@univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr/~amarigon
Birth Date : 26/11/1980
Arrival Date : 01/10/2007
Administrative Charges :

Research topics

The first theme of my research topics is in bioinformatics area. The number of available biological sequences is growing very fast, due to the development of massive sequencing techniques. Sequence classification is needed and contributes to the assessment of gene and species evolutionary relationships. Classification methods are thus necessary to carry out these identification operations in an accurate and fast way. I develop a classification method dedicated to homologous sequence family databases, allowing to attribute a new sequence to a cluster using similarity measures. I used this classification method to implement two applications, HoSeqI (Homologous Sequence Identification) and MultiHoSeqI. They allow to automatically identify biological sequences and to rapidly add several sequences to a database. HoSeqI is accessible through a Web interface (http://pbil.univ-lyon1.fr/software/HoSeqI/) allowing to identify one or several sequences and to visualize resulting alignments and phylogenetic trees. MultiHoSeqI makes it possible to quickly add a large set of sequences to a family database in order to identify them, to update the database, or to help automatic genome annotation. Lately, I developed a chimera detection method and implement an application, ChiSeqI (Chimeric Sequence Identification), to automate the processes of classification of specific biological data, the bacterial 16S ribosomal RNA sequences, and of detection of chimeric sequences.

The second theme of my research topics is in information system area and, more precisely, the multimedia data warehouse. Data warehouses are dedicated to collecting heterogeneous and distributed data in order to perform decision analysis. In numerous fields, like in medical or bioinformatics, multimedia data are used as valuable information in the decisional process. One of the problems when integrating multimedia data as facts in a multidimensional model is to deal with dimensions built on descriptors that can be obtained by various computation modes on raw multimedia data. I propose a new multidimensional model that integrates functional dimension versions allowing the descriptors of the multidimensional data to be computed by different functions. With this approach, the user is able to obtain and choose multiple points of view on the data he analyses. This model is used to develop an OLAP application for navigation into a hypercube integrating various functional dimension versions for the calculus of descriptors in a medical use case.

Publications

Arigon A.-M., Perrière G. and Gouy M., Automatic identification of large collections of protein-coding or rRNA sequences, A paraître dans Biochimie (2007), doi:10.1016/j.biochi.2007.08.006

Arigon A.M., Miquel M. and Tchounikine A. Multimedia data warehouses: a multiversion model and a medical application. Multimedia Tools Appl. 2007 October; 35(1): 91-108

Arigon A.M., Tchounikine A. and Miquel M. Handling multiple points of views in a multimedia data warehouse. ACM Transactions on Multimedia Computing, Communications and Applications. 2006 August; 2(3):199-218

Arigon A.M., Perriere G., Gouy M. (2006) HoSeqI: automated homologous sequence identification in gene family databases. Bioinformatics. 2006 Jul 15; 22(14):1786-7

Emna BAHRI

Current Position : PhD student
E-mail : Emna.bahri@univ-lyon2.fr
Web site :

Birth Date : 15/04/1981
Arrival Date : 17/10/2006

Research supervisor : Stéphane Lallich

Research topics

The recent advances in Communication and Information Technologies led to huge amounts of data, which exceeds the human processing and understanding capabilities. These data repositories contain an enormous amount of information, but require the development of intelligent tools in order to transform this information to knowledge. Those needs gave rise to data mining, which is an active area of research today.

In spite of great theoretical and practical achievements, data mining still lacks from low-scalability to large and real-world datasets. Two major problems are thus, the treatment of large data volumes, and the intolerance to the presence of noisy data. Even if these two problems seem classic, they still constitute major challenges in the area of machine learning.

My PhD goal is to design more powerful prediction systems, able to reach better success rates (seldom but not perfect), while being insensitive to noisy data. We can divide our prospects for this thesis into two parts. The first, which will be investigated this year, consists of providing a general and an exact definition of noise in order to handle it. The second part, which will be carried out later during my PhD, consists in finding new approaches and new algorithms to detect and manage the noise already modeled.

Publications

E.bahri, N.Nicoloyannis, M..Maddouri, « Amélioration du Boosting par combinaison des hypothèses antérieures », 14èmes Rencontres de la Société Francophone de Classification (SFC07), Paris, Septembre 2007.

E.bahri, N.Nicoloyannis, M..Maddouri « improving boosting by exploiting former assumptions », Third International Workshop on mining complex data (MCD07),warsaw,Poland.

FADILA BENTAYEB

Current Position : Associate professor since 2001
E-mail : bentayeb@eric.univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr/~bentayeb

Birth Date : 15/05/1966
Arrival Date : 01/09/1999

Administrative Charges : Head of the bachelor of science in computer science and statistics (Informatique Décisionnelle et Statistique –IDS-)
Member of the recruitment commission for mathematic-informatics and automatic at the university Lyon 2

Research topics

My current research interests regard complex data warehousing, integration of data mining techniques into data warehouses that we call on-line data mining and schema evolution in data warehouses. The special nature of complex data poses different and new requirements to data warehousing technologies, over those posed by conventional data warehouse applications. Indeed, current multidimensional data models fail to model the complex data found in some real-world application domains. Our main contribution is, then, the definition of a general framework to warehouse complex data. We used XML as the canonical standard to transform and store complex data from original data sources and we used the XML Schema to define the global ODS (Operating Data Storage) schema. Our approach presents several advantages. We can mention the unified format of complex data with XML and the use of data mining techniques for extracting relevant information necessary for building dimensional models.

On-line data mining: Data mining research has made many efforts to apply various mining algorithms efficiently on large databases. However, a serious problem in their practical application is the long processing time of such algorithms since they operate in main memory. We propose then a complete integrated solution for mining large databases into DBMSs without size limit in acceptable processing times. We think that data mining and databases should not remain separate components of the decision support. Indeed, data mining tools need integrated, consistent, and clean data. A database is constructed exactly by such preprocessing steps. Our first contribution consists in reducing the size of the learning database by building its contingency table, and our second contribution consists in reducing the number of database accesses by using bitmap indices. As a perspective of this work, we intend to extend our integrated approach to deal with multi-relational tables.

Data warehouse evolution : Due to the role of data warehouses in the daily business work of an enterprise, the requirements for the design and the implementation are dynamic and subjective. Therefore, data warehouse design is a continuous process which has to reflect the changing environment of a data warehouse, in other words, the data warehouse schema must evolve in reaction to the enterprise's evolution. My research focuses in integrating user's new analysis needs in the data warehouse process. We propose, then, a global approach composed by (1) the user's knowledge acquisition, (2) the user's needs integration, (3) a data warehouse schema update, and (4) an on-line analysis. Our main contribution consists in defining a user-driven approach that enables a data warehouse schema update. We integrate the specific user's knowledge representing new aggregated data under the form of If-Then rules into the data warehouse model. These rules are used to dynamically and automatically create new granularity levels in dimension hierarchies.

Publications

1. F. Bentayeb, J. Darmont, C. Favre, C. Udréa, "Efficient On-Line Mining of Large Databases", *International Journal of Business Information Systems*, Vol. 2, No. 3, 2007, 328-350.
2. J. Darmont, F. Bentayeb, O. Boussaïd, "Benchmarking Data Warehouses", *International Journal of Business Intelligence and Data Mining*, Vol. 2, No. 1, 2007, 79-104.
3. O. Boussaïd, J. Darmont, F. Bentayeb, S. Loudcher-Rabaseda, "Warehousing complex data from the Web", *International Journal of Web Engineering and Technology*, 2007.
4. C. Favre, F. Bentayeb, O. Boussaïd, "A Survey of Data Warehouse Model", *Encyclopedia of Database Technologies and Applications, Second Edition*, Idea Group Publishing, 2007.
5. C. Favre, F. Bentayeb, O. Boussaïd, "Evolution of data warehouses' optimization: a workload perspective", *9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007), Regensburg, Germany*, September 2007; *LNCS*.

Scientific activities and valorisation

Scientific programs and/or industrial collaborations	Phd Program Cécile Favre, « Data Warehouse evolutions », 2004-2007; Nora Maiz, « Integration by Mediation for data warehousing », 2005-2008; Ony Rakoarivélo, «On-line data mining for schema evolution in data warehouses», 2006-2009 Industrial collaborations LCL-Le Crédit Lyonnais (Rhône-Alpes Auvergne), 2004-2007 (Cécile Favre's thesis) Scientific programs ACI FodoMust (Fouille de données Multi-stratégie) 2005-2007
Editorial boards and program committees	- International Journal of Information Technology and Web Engineering Idea Group Publishing, 2007 - International Workshop « Ateliers sur les Systèmes Décisionnels », 2006-2007 Processing and Managing Complex data for Decision Support, Idea Group Publishing, 2005 - Journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 2006, EDA 2007) Editorial board and Committee steering member - International Journal of Biomedical Engineering and Technology (IJBET). SPECIAL EDITION "Warehousing and Mining Complex Data: Applications to Biology, Medicine, Behavior, Health and Environment", 2007 - French conference « Journées francophones sur les Entrepôts de Données et l'Analyse en ligne » (EDA), since 2005 International Multiconference on Computer Science and Information Technology (CSIT 06), Amman, Jordan , 2006 (Chair of session) Organizing Committee member French Conference EDA, Lyon, 2005 International Conference on Flexible Query Answering Systems (FQAS), Lyon 2004

Anouck BODIN-NIEMCZUK

Current Position : PhD student
E-mail : anouck.bodin-niemczuk@eric.univ-lyon2.fr
Web site :
Birth Date : 28/07/1984
Arrival Date : 01/09/2007

Research supervisor : Omar Boussaid and Sabine Loudcher Rabaséda

Research topics

The on-line analysis OLAP (On-line Analytical Processing) is a technology which comes to supplement data warehouses by proposing tools for visualization, exploration and navigation in data cubes in order to discover interesting information.
The user finds manually potential knowledge contained in data cubes. Indeed, OLAP technology makes it possible to visualize facts, to structure them according to analysis axes and to explore them but does not allow classification, explanation and prediction.
On the other hand, data mining employs machine learning techniques for visualization and description, for the structuring and classification, and for explanation and prediction.
During the last years, several works showed that it was possible to enrich the decision-making process using the coupling of on-line analysis and data mining [Imieliński 1996], [Han 1997], [Messaoud 2006].
Our approach consists in defining a new concept of on-line analysis by integrating data mining methods into OLAP data cubes.

Han J., OLAP Mining: An Integration of OLAP with Data Mining, Proceedings of the 7th IFIP Conference on Data Semantics, 1997, Leysin, Switzerland

Imieliński T. and Mannila H., A Database Perspective on Knowledge Discovery, Communications of the ACM, vol. 39, n°11, 1996, pages 58-64.

Messaoud R.B., Couplage de l'analyse en ligne et de la fouille de données pour l'exploration, l'agrégation et l'explication des données complexes, Thèse de doctorat informatique, Université Lumière Lyon 2, novembre 2006.

Sonia BOUATTOUR

Current Position : PhD student
E-mail : bouattoursonya@yahoo.fr
Web site :
Birth Date : 03/07/1983
Arrival Date : 09/10/2006

Research supervisor : Omar Boussaid

Research topics

In the space domain, the construction of an operandi information and its availability to different types of users including mobile clients (embedded systems, PDAs, mobile phones, etc.) requires a change in traditional architectures of data warehouses. It is necessary to take charge, through computerization may leave some with interactivity, analysis Scenario by integrating them into the same process of storage. This results in the form of shares that may be triggered under given conditions including on the same sources OLTP, allowing access and act on detailed data. These treatments will be expressed in the form of analysis rules. They can make an effective contribution to improving the performance of these new architectures as did the pre-aggregated data in a classic OLAP cube. By integrating analysis rules in warehouses, they become active. The active data warehouses are new architectures, which create a dynamic where the OLAP cube is no longer an end in itself but on the contrary an intermediate step to design and produce information decision at the request with a return on decision-making sources, ETL, the sources OLTP ...
There are several approaches for designing such an architecture data warehouse of spatial data:
In the first approach, it is a traditional configuration centered on a warehouse or a datamart with a device of ETL from OLTP sources. The multidimensional model (star diagrams or snowflakes or flakes facts (Fact flake)) may have one or more spatial dimensions and / or measures space. A number of cubes can be constructed from complaints decision already identified. To give a dynamic to this setup, it must be complemented by a set of analysis rules corresponding to decision-making queries well established.
In the second approach, in contrast to the first one, there is no centralized multidimensional source (warehouse or datamart). The ETL device consists of a system of mediation to build at the request of cubic spatial data from space or non-space OLTP sources. The goals of analysis are supported by a mediator who identifies relevant sources, selects and extracts the data and propose a cube (or a set of cubes). The analysis scenarios have been identified and defined as analysis rules.
The third approach presents a solution that combines the first and the second approaches to take into account a set of multidimensional or OLTP sources, which are assumed to exist. The demand of a spatial information regarded as making a request may be met by a multidimensional structure (warehouse datamart, or cube) already existed. Otherwise, we have to build this cube from existing multidimensional sources or even from OLTP sources. This approach requires, of course, a mediation device which must support the request of spatial information. The analysis rules will complete of this configuration to provide an active character to this solution.

Publications

S. Bouattour, R.Ben messaoud, O. Boussaïd, "Modélisation de règles d'analyse dédiées aux entrepôts de données actifs", 2ème édition de l'atelier des systèmes décisionnels (*ASD 07*), Sousse, 2007.

Omar BOUSSAID

Current Position : Assistant professor
E-mail : Omar.boussaid@univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr.~boussaid/

Birth Date : 02/06/1954
Arrival Date : 01/09/1990

Administrative Charges : In charge of the IIDEE (Informatic Enginnering and Economic Evaluation of Decision Support Systems) in Master IDS (Business Intelligence and Statistic)

Research topics

My research tasks relate to the complex data warehousinge and on-line analyzing. The decision-making processes are based on the technology of the data warehouses and the OLAP. This technology is considered as mature in particular when the data are simple data. The challenge of today is to make evolve this technology by applying it to the complex data. To achieve this objective, I organized my work according to three resarch orientations:
1°) Integration of the complex dat. After proposing an approach for describing complex data with the aid of UML and XML languages in order to store them into a target database, nowadays, we are working, as part of a thesis, on an approach of data integration based on a mediation system using ontologies for each of the data sources. The aim is to provide analysis contexts (datacubes built on the fly) and achieve on-line analysis.
2°) Modeling of complex data. We have chosen to use XML as a language of complex data modeling. We are currently working on methods of dimensional modeling to build XML-based complex data warehouse. As part of a thesis, we develop some works on the conceptual and dimensional modeling of complex data. We proposed a dimensional conceptual model of complex objects -representing complex data- which we describe at logic level with XML schemas. This model is being validated. The optimization of the physical models in XML warehouses is another objective of this thesis. Furthermore, we have developed an approach, which starting from a conceptual and multidimensional mode, to generate an XML complex data cube automatically. On the other hand, we are focused on further work to address the problem of performances in XML warehouses. To do that, we currently experienced a new method of fragmentation of the complex data warehouses. This work is being developed as part of another thesis.
3°) On-line analysis of complex data. In order to reinforce the capabilities of OLAP and expand its capacities to the explanation and the prediction, we work on the coupling of OLAP with data mining. As part of a thesis, we tried out different methods of coupling allowing to aggregate data, to improve the data representation in an OLAP cube and to apply the association rules as an analytical tool in an OLAP cube. We have proposed the theoretical foundations of these OLAP and data mining coupling. We are generalizing this formal framework to define any approach of coupling. We continue this work to extend this coupling in order to achieve the prediction analysis in OLAP cubes.

Publications

O. Boussaïd, Adrian Tanasescu , Fadila Bentayeb, Jerome Darmont, "Integration and Dimensional Modelling Approaches for Complex Data Warehousing", in Journal of Global Optimization, Vol. 37, No. 4, pp 571-591, Springer Netherlands, 2007
O. Boussaïd, R. Ben Messaoud, R. Choquet, S. Anthoard, "X-Warehousing : an XMLBased

Approach for Warehousing Complex Data", 10th East-European Conference on Advances in Databases and Information Systems (ADBIS 06), in LNCS Vol. 4152, 39-54, Thessaloniki, Greece, September 2006

R. Ben Messaoud, S. Loudcher Rabaséda, O. Boussaïd, R. Missaoui, "Enhanced Mining of Association Rules from Data Cubes", Proceedings of the 9th ACM International Workshop on Data Warehousing and OLAP (DOLAP'06), Arlington, VA, USA, ACM Press, November 2006, pp 11-18

O. Boussaïd, J. Darmont, F. Bentayeb, S. Loudcher-Rabaseda, "Warehousing complex data from the Web", International Journal of Web Engineering and Technology, 2007.

J. Darmont, O. Boussaïd, Eds., "Processing and Managing Complex Data for Decision Support", Idea Group Publishing, April 2006

Scientific activities and valorization

Scientific programs and/or industrial collaborations	2004-2007: FoDoMuSt (multistrategy data mining). Project with the LSIIT computer science and LIV geography labs (Strasbourg) for automatically identifying vegetation from satellite images. Funding from the Ministry of Research (ACI project) *2002-2005: CLAPI (spoken language corpus).* Project with the ICAR linguistics lab for building, managing and exploiting a complex database of spoken language corpora. Funding from the Ministry of Research (ACI project).
Editorial boards and program committees	*Editorial boards*: International Journal of Biomedical Engineering and Technology, Advances in Data Warehousing and Mining book series; EDA conferences steering committee *Journal and book paper reviewing*: Journal of Intelligent Information Systems, International Journal of Foundations of Computers Science, International Journal of Software and Systems Modeling, The International Journal of Computers and Applications, "Multimedia Systems and Applications" book, Kluwer Academic Publishers, Ingénierie des Systèmes d'Information, numéro spécial : "Elaboration des entrepôts de données", Encyclopedia of Data Warehousing and Mining 2nd Edition. *Conference program committees*: CE 06, EDA 05-07, ASD 06-07, MDDE, 01-02, SFdS, 03, FDC 04-08, SimSem 08, EGC 08, CSIT 06 *Conference organizing committees*: EDA 05, SFdS 03, ISMIS 02, ReTIS 01
International activities	Scientific stay as invited professor to the university Laval (Québec-Canada) at Laboratory CRG (Research center in Geomatic) of Pr. Yvan Bédard March 2006 ; Scientific Stay as invited professor to the university of Quebec in Outaouais (Canada) at Laboratory LARIM (Research Laboratory on Multi-media Information) of Pr. Rokia Missaoui in June-July 2006

Jean-Hugues CHAUCHAT

Current Position : Full professor
E-mail : jean-hugues.chauchat@univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr/%7Echauchat/
Birth Date : 06 July 1946
Arrival Date :

Administrative Charges : In charge of the strand SISE (Statistics & Informatics) in Master IDS (Business Intelligence and Statistic)
In charge of the double diploma Master/Magister (Statistics & Informatics) of University Lyon2 and the National University of Economics in Kharkov, Ukraine

Research topics

Statistics and Data Mining: models and validation
validation methods when the dataset is not collected using a two-stage, or a clustered, or a strata sampling design,
sampling in the whole dataset,
visualization.

Text mining.
Complex surveys analysis.
Applied statistics for managers.
Teaching statistics.

Publications

CHAUCHAT J.H., A. MORIN & R. RAKOTOMALALA, 2007. "Correcting the error rate estimation bias in Data Mining when the dataset comes from a two-stage sampling", *Statistics for Data Mining, Learning and Knowledge Extraction (IAST'07), Aveiro, Portugal*.

RAKOTOMALALA, R., JH CHAUCHAT & F. PELLEGRINO, 2006. Accuracy Estimation With Clustered Dataset. In Proc. *Fifth Australasian Data Mining Conference (AusDM2006), Sydney, Australia. CRPIT, 61*. Peter, C., Kennedy, P. J., Li, J., Simoff, S. J. and Williams, G. J., Eds., ACS. 17-22.

MORIN A, A. KOUOMOU-CHOUPO, JH CHAUCHAT, 2005, Dimension reduction and clustering for query-by-example in huge image databases. Proc. *3rd world conference on Computational Statistics and Data Analysis, Limassol, Cyprus*, October 2005.

RADWAN J., CHAUCHAT J.-H. and DUMAIS J. 2004 "Automatic Recognitiion of Keywords using N-grams". In Jaromir A., editor, *Compstat'04 - Proceedings in Computational Statistics,* 1245-1254. Physica Verlag, Heidelberg, Germany.
PELLEGRINO F., CHAUCHAT J.H. & R. RAKOTOMALALA, 2002, "Can Automatically Extracted Rhythmic Units Discriminate among Languages?", *Proceedings of Speech Prosody 2002*, pp.562-565.

Scientific activities and valorisation	
Scientific programs and/or industrial collaborations	Scientific programs **2004-2005** Program EGIDE Econet (France – Croatia – Slovenia) « Fouille de données intelligente pour l'aide à la décision avec applications en médecine » - « Intelligent Data Mining in order to help decision making with applications in the medical field». **2007-2008** Program EGIDE COGITO (France – Croatia) and PROTEUS (France – Slovenia) "Knowledge discovery and visualization for textual data" Industrial collaborations **2006 Institut Fournier** statistical and computer techniques for the analysis of large files that contain financial data received by insurance companies. **2004 Commissariat Général au Plan** Analysis and implementation of a national survey regarding the changes in the public sector. **2004 Laboratoire SERVIER**. Data Mining Advisor for the research of undesirable effects of new drugs.. **2002-2003 Région Rhône-Alpes.** Computer-based modelization for the estimation of the total number of commuters between the towns of the Rhone-Alpes region. **2000-2001 Crédit Agricole Centre-Est.** Data mining for marketing : update of an online banking tool. Design and analysis of the clients' satisfaction in different market segments.
Editorial boards and program committees	Referee for the conference IASC 07 Referee for the conference IASE'06
International activities	**1997-98**. Visiting Professor, University of Delaware, USA, College of Economics and Business Administration, Course taught : Data Analysis (Master in Economics) **2005-2006-2007** Scientific Expert pour the research funds of Quebec Elected member of the International Statistical Institute Member of the International Association of Computing Statistics (IASC), Member of the International Association of Surveys Statisticians (IASS), Member of the International Association For Statistical Education (IASE). Member of the French Statistical Society (SFdS), Member of the French Classification Society (SFC),

Jérôme DARMONT

Current Position	Associate professor (HDR)
E-mail	jerome.darmont@univ-lyon2.fr
Web site	http://eric.univ-lyon2.fr/~jdarmont/
Birth date	15/01/1972
Arrival date	01/09/1999
Administrative charges	Since 2003: Director, Computer Science and Statistics Department (DIS), School of Economics and Business Administration; U. Lyon 2 Since 2000: Head, Decision Support Databases group, ERIC lab

Research topics

Since my arrival at ERIC, I have been working on the border of databases and data mining. More precisely, I have lead my research following two complementary axes: data warehouse performance optimization and evaluation. The mix between databases and data mining is particularly obvious in the performance optimization part, since the automatic indexing and view materialization approach we proposed in K. Aouiche's PhD thesis (defended in 2005) is based on data mining techniques. Moreover, this research has lead to the design of generic benchmarks for data warehouse performance evaluation. Both these research topics allowed me to pass my "HDR" (qualification for supervising research) in 2006. Three PhD theses follow up this work. The first one (J.C. Ralaivao, started in 2003) aims at identifying performance factors in complex data warehouses. We have also proposed an XML-based reference architecture for complex data warehouses.

The second thesis' subject (H. Mahboubi, started in 2005) is dedicated to XML-native database management systems' performance optimization, and especially addresses two critical issues: response time and data volume. To help solve them, we have proposed XML data warehouse indexing, view materialization, fragmentation and distribution (over a grid) techniques.

The third thesis' objective (M. Hachicha, started in 2007) is to allow On-Line Analytical Processing over complex data stored in an XML warehouse. In this context, we have already proposed to formulate OLAP operators in an XML algebra, which helps execute classical OLAP queries over XML-native data (XML-OLAP or XOLAP).

On a longer term, my research project lies on the idea that XML must definitely become a pivot language for complex data warehousing, and I envisage three research axes: new, Web-based data warehouses architectures (Web 2.0, Web services); analytical extensions of the XQuery language for decision support; and exploiting semantic information about complex data for analysis. To restrict the scope of these research axes, I shall keep on addressing them from a performance point of view.

Publications

J. Darmont, F. Bentayeb, O. Boussaïd, "Benchmarking Data Warehouses", *International Journal of Business Intelligence and Data Mining*, Vol. 2, No. 1, 2007, 79-104

F. Bentayeb, J. Darmont, C. Favre, C. Udréa, "Efficient On-Line Mining of Large Databases", *International Journal of Business Information Systems*, Vol. 2, No. 3, 2007, 328-350

J. Darmont, O. Boussaïd, Eds., *Processing and Managing Complex Data for Decision Support*, Idea Group Publishing, April 2006

K. Aouiche, P. Jouve, J. Darmont, "Clustering-Based Materialized View Selection in Data Warehouses", *10th East-European Conference on Advances in Databases and Information Systems (ADBIS 06), Thessaloniki, Greece*, September 2006; *LNCS*, Vol. 4152, 81-95

Z. He, J. Darmont, "Evaluating the Dynamic Behavior of Database Applications", *Journal of Database*

Scientific activities and valorization

Scientific programs and/or industrial collaborations	*2007-2008: TAPEO*. Project with a young company for expressing a collaborative, collective intelligence from a Web site managing virtual stock exchange portfolios. Funding from the Rhône-Alpes Region: 29,500 €. *2004-2007: FoDoMuSt (multistrategy data mining)*. Project with the LSIIT computer science and LIV geography labs (Strasbourg) for automatically identifying vegetation from satellite images. Funding from the Ministry of Research (ACI project): 69,000 €. *2002-2005: CLAPI (spoken language corpus)*. Project with the ICAR linguistics lab (Lyon 2) for building, managing and exploiting a complex database of spoken language corpora. Funding from the Ministry of Research (ACI project): 36,000 €. *2003-2004: MAP (anticipative, personalized medicine)*. Project with Dr Ferret, former physician of the French national soccer team, for storing, managing and analyzing complex medical data to optimize the health capital of high-level athletes. Funding from the Rhône-Alpes Region: 29,000 €.
Editorial boards and program committees	*Editorial boards:* International Journal of Biomedical Engineering and Technology, Advances in Data Warehousing and Mining book series, IGI Editorial Advisory Review Board; EDA conferences steering committee *Journal and book paper reviewing:* Data & Knowledge Engineering, Journal of Intelligent Information Systems, Ingénierie des Systèmes d'Information – Special Issue: Information retrieval and information mining; Encyclopedia of Database Technologies and Applications 2nd Edition, Encyclopedia of Information Science and Technology 1st and 2nd Editions *Conference program committees:* IRMA 2005-2007, ASD 2006-2007, SAC-WT 2007, PICCIT 2007, CSIT 2006, ISWC 2004, FQAS 2004; EGC 2001-2008, EDA 2006-2007, INFORSID 2007, FDC-EGC 2006-2007, BDA 2003, SFdS 2003 Conference organizing committees: EDA 2005, SFdS 2003
International activities	*Since 2006:* Pedagogical director, French-Ukrainian double diploma (MSc in Computer Science and Statistics), in collaboration with the Kharkiv National University of Economics, Ukraine. PhD co-supervisions are planned. *2003-2005:* Collaboration with Dr. Zhen He, La Trobe University, Australia: joint research project and publications about database dynamic performance evaluation. *2003:* Collaboration with Pr. Le Gruenwald, University of Oklahoma, USA: joint research project and publications about database auto-indexing, student exchanges.

Ahmad El SAYED

Current Position : PhD student
E-mail : asayed@eric.univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr/~asayed

Birth Date : 11/03/1982
Arrival Date : 1/11/2004
Research supervisor : Djamel Abdelkader Zighed

Research topics

My PhD's goal essentially consists on developing intelligent methods and tools for allowing a semantic content-based information retrieval on heterogeneous documents (including texts and images). At a first stage, "semantics" were acquired using hand-crafted resources like Wordnet or domain ontologies in order to allow a query/document matching on the highest semantic level. We explored an approach where image and text contents in a document are analyzed automatically to represent each part by a set of terms. The deduced terms will be redirected into a fuzzy ontology enabling a conceptual representation of the whole document content. At a second stage, "semantics" were acquired automatically by means of a developed technique for knowledge acquisition from text. Furthermore, we designed a framework for learning taxonomy from a target text corpus. To achieve this, semantic relations between terms are first mined from text using a hybrid approach combining pattern-based and text mining techniques. Then, the entire set of relations is used for clustering terms into sense-bearing units that will be regarded to some extent as concepts. Following this, taxonomic relations will be deduced between the obtained concepts in order to build the hierarchy. To improve accuracy, the learned taxonomy is finally involved in our information retrieval environment where users interactions with the system are taken into account in order to launch a relevance feedback mechanism able to adapt the taxonomy to the user vision over text. As for future works, we're intending to use the acquired knowledge to perform a semantic parsing of text in order to represent it in predicate-argument structure rather than a bag of words. This can lead to the development of more sophisticated text-based applications for many areas like Question-Answering and Text Summarization.

Publications

A. El Sayed, H. Hacid, A.D. Zighed. "Mining Semantic Distance Between Corpus Terms", In Proceedings of the ACM CIKM 1st Ph.D. Workshop in Information and Knowledge Management, PIKM 07, November 2007, *Lisboa, Portugal.*

A. El Sayed, H. Hacid, A. D. Zighed "A Multisource Context-Dependent Semantic Distance Between Concepts", *In Proceedings of the 18th International Conference on Database and Expert Systems Applications DEXA'07,* September 2007- Regensburg, Germany

A. El Sayed, H. Hacid, A. D. Zighed "Combining Text and Image for Content-Based Information Retrieval"*, In Proceedings of the 2007 International Conference on Information and Knowledge Engineering, IKE 2007*, June 2007, Las Vegas, Nevada, USA.

A El Sayed, H. Hacid, A. D. Zighed "A New Context-Aware Measure for Semantic Distance Using a Taxonomy and a Text Corpus", *In Proceedings of the 2007 IEEE International Conference on Information Reuse and Integration, IEEE IRI'07,* August 2007, Las Vegas, USA.

A. El-Sayed, H. Hacid, D.A. Zighed, "Recherche d'Information par le Contenu des Données Hétérogènes", *in Actes des 3émes Rencontres Inter-Associations RIA's 07,* March 2007, Toulouse.

Cécile FAVRE

Current Position : PhD student
E-mail : cecile.favre@univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr/~cfavre
Birth Date : 19/08/1980
Arrival Date : 15/01/2004
Research supervisor : Fadila Bentayeb, Omar Boussaid

Research topics

After working on data mining integration into DBMSs, my current research works focus on data warehouse evolution.
Data warehouses store aggregated data issued from different sources to meet users' analysis needs in terms of decision support. As a matter of fact, user's requirements change over time and never reach a final state. Therefore, a data warehouse model cannot be designed in one step, it usually has to evolve progressively. We are thus interested in data warehouse model evolution. More specifically, we aim at involving users in the evolution process in order to supply them with personalized answers to their analysis needs.
Data warehouse evolution usually means evolution of its model. Meanwhile, a decision support system is composed of the data warehouse along with several other components, such as optimization structures, e.g. indices or materialized views. Thus, dealing with the data warehouse evolution also implies dealing with the maintenance of these structures. However, propagating evolution to these structures thereby maintaining the coherence with the evolutions on the data warehouse is not always enough. In some cases, redeployment of optimization strategies is required. Thus, we are interested in finding efficient solutions to ensure good performances, taking into account the model evolution. Since selection of optimization strategies is mainly based on workload according to user queries, one perspective is to lead the workload to evolve in order to test performances without waiting for a new workload for taking decisions on the optimization strategy.

Publications

C. Favre, F. Bentayeb, O. Boussaid, Evolution of Data Warehouses' Optimization: a Workload Perspective, 9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 07), Regensburg, Germany, September 2007 ; LNCS, Vol. 4654, 13-22.
C. Favre, F. Bentayeb, O. Boussaid, Dimension Hierarchies Updates in Data Warehouses: a User-driven Approach, 9th International Conference on Enterprise Information Systems (ICEIS 07): Databases and Information Systems Integration, Funchal, Madeira, Portugal, June 2007 ; 206-211.
F. Bentayeb, J. Darmont, C. Favre, C. Udréa, Efficient On-Line Mining of Large Databases, International Journal of Business Information Systems, Vol.2, N°3, 2007, 328-350.
C. Favre, F. Bentayeb, O. Boussaid, A Knowledge-driven Data Warehouse Model for Analysis Evolution, 13th ISPE International Conference on Concurrent Engineering: Research and Applications (CE 06), Antibes, France, September 2006 ; Frontiers in Artificial Intelligence and Applications, Vol. 143, 271-278.
C. Favre, F. Bentayeb, Bitmap Index-based Decision Trees, 15th International Symposium on Methodologies for Intelligent Systems (ISMIS 05), New York, USA, May 2005 ; LNAI, Vol. 3488, 65-73.

Rémi GAUDIN

Current Position : PhD student
E-mail : remi.gaudin@univ-lyon2.fr
Web site :
Birth Date : 01/06/1981
Arrival Date : 01/09/2004
Research supervisor : Djamel A. Zighed

Research topics

Most machine learning algorithms for classification problems are similarity/dissimilarity-based approaches. The similarity between instances is often explicitly expressed by a distance. In addition to the classical p-norm distance, other measures have been studied for the special case of time series, and among them the Dynamic Time Warping. Due to the observed performances variability of the previously proposed solutions considering various applications and benchmarks, a new distance called Adaptable Time Warping (ATW) is investigated. ATW is a form of generalization of both the classical Euclidian distance and DTW. A learning process which uses a genetic algorithm allows ATW to reach optimal solutions. We can prove that ATW efficiency is always at least equivalent to other distances use whatever the classification problem to be handled. We also demonstrate empirically the efficiency of ATW through different applications. Some are classical benchmarks for allowing comparative tests with previous studies, and two others are dealing with material science and more precisely zeolite crystalline structure. Effectiveness and stability are the two key advantages of ATW, which made it a promising methodology within our young research area.

Publications

R. Gaudin, L. A. Baumes, S. Jimenez, N. Nicoloyannis and A. Corma. "Improving Time Series Classification Using an Adaptable Distance and a Genetic Algorithm: Application to Automatic Classification of Zeolite Structures from X-Ray Diffraction". *The Third International Conference on Advanced Data Mining and Applications (ADMA'07)*, Harbin, China, August 6-8, 2007, LNCS series, Springer Press.

L. A. Baumes, M. Moliner, R. Gaudin, N. Nicoloyannis, A. Corma. "A robust methodology for high throughput identification of mixture of crystallographic phases from powder diffraction data". *Invited at E-MRS Fall Meeting 2007, Symposium on Genetic algorithms in materials science and engineering*, Warsaw, Poland, September 17-21, 2007.

R. Gaudin and N. Nicoloyannis. "An Adaptable Time Warping Distance for Time Series Learning". *Fifth International Conference on Machine Learning and Applications (ICMLA'06)*, Orlando, USA, December 14-16, 2006. IEEE Press, Pages 213–-218.

R. Gaudin, S. Barbier, N. Nicoloyannis and M. Banens. "Clustering of Bi-Dimensional and Heterogeneous Time Series: Application to Social Sciences Data". *2006 International Conference on Data Mining (DMIN'06)*, Las Vegas, USA, June 26-29, 2006. CSREA Press, pages 10–-16.

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". *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC'05),* Paris, Janvier 2005. Presse RNTI, pages 201–212.

Marouane HACHICHA

Current Position : PhD student
E-mail : Marouane.Hachicha@univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr/~mhachicha/

Birth Date : 7th June, 1983
Arrival Date : 3st September, 2007

Research supervisor : Jérôme Darmont

Research topics

The objective of this thesis is to allow OLAP analysis of complex data structured in XML data warehouses. This work consists on the design of a XML-OLAP (or XOLAP) algebra in the order to carry out traditional OLAP queries on native XML data.

This new formal framework represents the first step of our work. Then, it will be necessary to enrich this XOLAP algebra with new specific operators to the XML context. Then, it appears necessary to be able to carry out some operations like roll up and drill down on complex hierarchies of dimensions such as the ragged hierarchies [1]. Operators coupling the principles of OLAP and Data Mining could also allow the treatment (aggregation) of multimedia data resulting from the Web [2]. This work also aims at supporting the efforts of the extension of the XQuery language for the decisional applications.

In addition, to have a XOLAP algebra for the decisional complex data processing must allow the optimization of OLAP queries expressed in XQuery. Native XML Data Bases Management Systems (DBMS), although in a constant progress, present some limitations in term of performance and would profit largely from an automatic queries' optimization, particularly the decisional queries which are, generally, very expensive.

Finally, an implementation of this work on XOLAP is envisaged within the framework of a platform of complex XML data storage, under development at the ERIC laboratory [3]. The objective is to allow, through a simple and accessible interface since the Web, the construction and the handling of XML cubes of complex data.

Hakim HACID

Current Position : PhD student
E-mail : hhacid@eric.univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr/~hhacid/
Birth Date : 05/03/1979
Arrival Date : 01/10/2004

Research supervisor : Pr. Abdelkader Djamel Zighed

Research topics

These last five years, particularity, due to the emerging new data acquisition technologies: scanner, satellite, video, web, etc. the available data related to the same problematic become in the same time larger, richer, and more heterogeneous, in one word, more complex. This situation concerns all the human activities such as medicine, astronomy, marketing, etc. The challenge of the next decade is the valorisation of these collected data. Access to the hidden knowledge in these large, heterogeneous, and unstructured databases is the most important task from a scientific and a technological point of view. Combining techniques issued from different domains, databases and data mining in our context, is a crucial question to face the new challenges in analysis, interrogation, and efficient access. Proceeding like that (combination of different techniques), databases can benefit from data mining to improve the quality of the answers. The data mining as for it can benefit from the optimisation strategies offered by the domain of databases to access larger datasets. This is very interesting since data mining models become more efficient when they learn on lager datasets.

Thus, we propose in this thesis to adapt a data mining, an instance-based learning particularly, structure to index large databases. This is done in order to incorporate certain intelligence in the indexing process and extending the functionalities of the indexing structures. Indeed, by doing that, the index can be useful not only to answer queries quickly but to offer also other possibilities like classification. From a data mining point of view, since neighbourhood graphs are hardly scalable to large datasets, we propose optimisations, issued from the databases domain, in order to make them operational on large datasets.

Another major problem in data mining and databases is that data collection doesn't hold in the memory. The databases techniques offer an efficient data access, sorting techniques, grouping techniques, and query optimisation techniques which are the basis of system's scalability. The majority of the methods issued from statistics, automatic learning, etc. consider that data hold completely in memory and do not consider the case where they do not satisfy this condition. We address also this problem in this thesis and we propose some solutions to manage it in the context of our study.

Publications

Hakim Hacid, Tetsuya Yoshida: Incremental Neighborhood Graphs Construction for Multidimensional Databases Indexing. Canadian Conference on AI 2007: 405-416

Ahmad El Sayed, Hakim Hacid, Djamel A. Zighed: A Multisource Context-Dependent Semantic Distance Between Concepts. DEXA 2007: 54-63

Ahmad El Sayed, Hakim Hacid, Djamel A. Zighed: A New Context-Aware Measure for Semantic Distance Using a Taxonomy and a Text Corpus. IRI 2007: 279-284

Hakim Hacid: Neighborhood Graphs for Semi-automatic Annotation of Large Image Databases. MMM (1) 2007: 586-595

Ahmad El Sayed, Hakim Hacid, Djamel A. Zighed: A Context-Dependent Semantic Distance Measure. SEKE 2007: 432-437

Nouria HARBI

Current Position : Assistant professor
E-mail : Nouria.harbi@univ-lyon2.fr
Birth Date : 27/08/1961
Arrival Date : 01/06/2006

Administrative Charges : In charge of the OPSI in Master IDS (Business Intelligence and Statistic)

Research topics

My research interests are about the decisional Information systems which integrate different categories of information processing: data warehouses, data marts, multidimensional databases. The main dimensions are:

Methodologies of the decisional Information Systems design and Conception: the objective is to build methods of conception and development for decisional information systems. These methods should allow conceiving, establishing, feeding and updating the different areas of storing data for decisional Information Systems. As a result, these methods will offer concepts, formalisms and steps adapted to decision systems, oriented towards the decision-makers. The proposed methodology will be validated by the respective tools.

Decisional data systems modelling: definition of data models based on a multidimensional data representation. The models should allow for a dependable, uniform and secured presentation of decisional data derived from various sources (Databases, Files, HTML, XML). These models should also permit the representation and archiving of all or part of the data warehouse.

Publications

Nouria Harbi,Omar Boussaid, Fadila Bentayeb, Propriétés d'un modèle conceptuel multidimensionnel pour les données complexes, Communication, EGC 2008, Nice Sophia Antipolis, Janvier 2008, 12 pages

Nouria HARBI , Henri SAVALL, Véronique ZARDET, , Spectral analysis of socio-economic diagnoses: qualimetrical treatment of qualitative data, Communication, AOM HONOLULU, Août 2005, 20 p.

HARBI Nouria, SAVALL Henri, ZARDET Véronique, Analyse spectrale de diagnostics socio-économiques : traitement qualimétrique de données qualitatives, Communication, Colloque International AOM-RMD, Mars 2004, 26 p.

Nouria HARBI, Henri SAVALL, Véronique ZARDET, , Spectral analysis of socioeconomic diagnoses : qualimetrical treatment of qualitative data, Communication, Colloque international AOM-RMD, Mars 2004, 17 p.

Editorial boards and program committees	ASD 2007 (IJBET) International Journal of Biomedical Engineering and Technology

Charbel JULIEN

Current Position : PhD student

E-mail : charbeljulien@hotmail.com

Web site : eric.univ-lyon2.fr/~jcharbel

Birth Date : 31/08/1978

Arrival Date : 01/10/2004

Research supervisor : Prof. Djamel Zighed university Lyon2 and Prof. Lorenza Saitta university of Turin

Research topics

I work on image modeling, Unsupervised and Semi-supervised learning. My research activity is usually related to statistical learning. My current interests are unsupervised classification of digital images.

Images may include different kinds of content descriptors from different levels. Until now no direct way has been found to extract high level semantic descriptors from images. Many low-level visual descriptor schemes have been proposed in the literature to extract visual content from images. Using these low-level visual descriptors, we can get high semantic level by inference.

Mixture distributions such as signatures or Gaussian Mixture Model (GMM) of color and texture are very interesting to describe the global composition of image. Mixture distributions, unlike histograms, try to abstract the content of image, color and texture, by a number of classes depending on the complexity of a particular image and this by using clustering techniques. To compute the distance between signatures linear optimization techniques are needed such as Mallows distance or Earth Mover's distance.

Unlike fixed size vector feature, we are interested of using a set of signatures to represent the image low-level visual content. The clusters in signatures representation mode are defined for each image individually. Simple images have a short signatures while complex images have long ones. Two clustering algorithms were tested to extract signatures from image: k-means algorithm and Expectation Maximization (EM) using Gaussian Mixture Model (GMM) with minimum description length to find the optimal number of clusters

This set of signatures that abstract the color and the texture of images is used afterward to compute the distance between pairs of images. The Earth Mover's Distance (EMD) is applied to each pairs of signatures of color and texture independently. The distance between two centers of clusters is computed using the Euclidean distance, this distance is used internally by the EMD algorithm. Afterward, the distance between two images is worked out using a linear combination of individual distances.

While signatures are very interesting to abstract the color and texture of images, continuous distribution like GMM offer a powerful way to abstract the color with correlation to spatial coordinates (x, y). We appended the (x, y) to the color features and we compute a GMM of color plus the spatial correlation.

Image modeling can be extended to image-set modeling using mixture distributions. By image-set, we mean a collection of images that exhibit visual similarity in color content and/or in spatial relationships between colored regions. Image-sets are generated either by supervised categorization or by unsupervised clustering of image collection into groups. Modeling an image-set can be done by computing a mixture distribution that minimizes the distance to all mixture distributions of images within the image-set, as can be modeled by a mixture of mixture distributions; in this case the image-

set is partitioned into homogenuous subsets, and for every subset a prototype is computed.
Unlike fixed-size feature vector, where the centroid that minimizes the distance to a set of vectors can be computed by averaging the values in the feature vectors, mixture distribution's centroid needs a more complex technique to be computed. We use linear optimization algorithm, to compute a mixture distribution that minimizes the distance to all distributions in the image-set.

Publications

C. Julien, L. Saitta, "Image Database Browsing by Unsupervised Learning", 17th International Symposium on Methodologies for Intelligent Systems (ISMIS 08), Toronto, Canada, 2008; LNAI, Springer, Heidelberg, Germany.

C. Julien, L. Saitta, "Automatic Handling of Digital Image Repositories: A Brief Survey", 17th International Symposium on Methodologies for Intelligent Systems (ISMIS 08), Toronto, Canada, May 2008; LNAI, Springer, Heidelberg, Germany.

Stéphane LALLICH

Current Position : Full professor
E-mail : Stephane.lallich@univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr/~lallich/

Birth Date : 20/09/1947
Arrival Date : 01/09/1997

Administrative Charges : Head of the master IDS (Business Intelligence and Statistic)
Head of the PhD program in Computer Sciences of University Lyon 2 (since 2007)

Research topics

In the past four years, I have developed my research around two main topics, the measures in data mining and the ensemble methods.
Concerning the measures, I was first interested by the measures which allow evaluating the quality of association rules, mainly in collaboration with Philippe Lenca, ENST Bretagne. We have identified various criteria for classifying usual objective measures, which allowed us to propose an automated procedure for assistance in choosing the most appropriate to the needs of a user. On the basis of these same criteria, we have also built a formal typology of the usual measures resulting from their properties according to the different criteria. This typology was compared to an experimental typology associated with the experimental behavior of these measures on different sets of test. We have also developed a presentation of usual measures parameterized according to the reference value of the confidence (independence value in case of targeting or 0.5 in case of prediction). This presentation allows to at the same time to emphasize the links between the measures which differ only by the value of the parameter and generate new control measures corresponding to a desired reference value of the confidence, which is particularly useful in case of targeting.
As a consequence of this latter work, we proposed a method to off-center the various entropies used in supervised learning to select at each step the best predictive attribute, for example Shannon entropy in C4.5 or Gini quadratic entropy in CART algorithm. In fact, at each node of a decision tree, we off-center the entropy in order that the off-centered entropy takes its maximum value for the distribution of the class in the node and not for the. This strategy improves systematically the precision on the class minority without degrading the results on the class majority.
Several of my works deals with ensemble methods : In the case of large high dimensional databases, with Elie Prudhomme (PhD ongoing), we proposed to replace neighborhood graphs by self organized maps to represent information from predictors. This substitution is moving from a complexity O (n3) to linear complexity depending on the number of individuals and variables, while retaining the capability of representation and navigation that is the interest of neighborhood graphs and putting in before a statistical cross-product basis of the map and taking into account the class of predictive performance generalization. To escape the dimensionality of the data, we propose to use a combination of Kohonen maps compiled from a limited number of predictors, thus viewing while improving the accuracy generalization. With Bissan Audeh (master thesis), we have developed a strategy that combines ensemble approach and sampling approach, which makes its complexity independent of the number of individuals and of the number of dimensions. With Romain Billot (master thesis), we have undertaken to adapt the boosting to clustering, and we propose UBLA, a method which leads to improve the value of the clustering quality coefficients.

Publications

Lenca P., Meyer P., Vaillant B., Lallich S. (2008), On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid, *EJOR, European Journal of Operational Research*, 184(2), 610–626.
Lallich S., Lenca P., Vaillant B. (2007), Probabilistic framework towards the parametrisation of association rule interestingness measures, *MCAP, Methodology and Computing in Applied Probability*, 9(3), 447–463.
Lallich S., Teytaud O. Prudhomme E. (2007), Association rules interestingness: measure and validation. In *Quality Measures in Data Mining*, pp. 251-275, Springer.
Zighed D.A., Lallich S., Muhlenbach F. (2005), A statistical approach of classes separability, *Applied Stochastic Models in Business and Industry*. Vol. 21, No. 2, 2005, pp. 187-197.
Lallich S., Muhlenbach F., Jolion J.-M.(2003), A test to control a region growing process within a hierarchical graph, *Pattern Recognition*, Pergamon, 36 (10), pp. 2001-2011.

Scientific activities and valorization

Scientific programs and/or industrial collaborations	Collaboration with *Hospices Civils de Lyon*, devoted to data mining and hospital acquired infections, 2007 Collaboration with *Banque Populaire Rhône et Loire* to initiate the staff of the bank to data mining methods, 2005
Editorial boards and program committees	**Editorial activity** Lallich S., Pastor D. (2007), Special Issue on the ASMDA International Symposium on Applied Stochastic Models and Data Analysis, *Communications in Statistics - Theory and Methods*, Volume 36, Issue 14 January 2007 , pages 2475 – 2671 S. Lallich, P. Lenca et F. Guillet (2007, 2008), Proceedings of the workshop *Qualité des Données et des Connaissances*, QDC 07, in association with Conference *Extraction et Gestion des Connaissances*, EGC 2007 and 2008. **Program Committee** *International Conference on Data Mining*, DMIN, Las Vegas, USA : DMIN 2006, DMIN 2007 ; Conference *International Symposium on Applied Stochastic Models and Data Analysis* : ASMDA 2005 (Brest, France), ASMDA 2007 (Chania, Crète, Grèce) Conference *Extraction et Gestion des Connaissances*, EGC 2005 Paris, EGC 2006 Lille, EGC 2007 Namur, EGC 2008 Nice Workshop *Qualité des Données et des Connaissances* , associated with Conference EGC (2005 Paris, 2006 Lille, 2007 Paris, 2008 Nice)
International activities	Collaboration with Dragan Gamberger (Chercheur Rudger Boskovic Institute, Zagreb), as part of Program Egide, with Jean-Hugues Chauchat, ERIC Lyon 2 et Annie Morin, IRISA, Rennes ; one week in Zagreb, sept. 05 (overfitting in machine learning). Collaboration with *Faculté des Sciences Economiques et de Gestion de Jendouba*, to welcome during 4 months a master research student (Nejmeddine Ben Ouarred, 2007). Expert's valuation for the Fonds québécois de recherche sur la nature et les technologies technologies à Québec (2007)

Virginie LEFORT

Current Position : Assistant professor
E-mail : virginie.lefort@univ-lyon2.fr
Web site : web-lefort.net

Birth Date : 24/09/1980
Arrival Date : 01/09/2007
Administrative Charges :

Research topics

Second order evolution (or indirect selection) corresponds to a situation where the individuals are not only selected on their fitness to an environment, but also on their ability to evolve "better". Even if such a mechanism seems***, a priori,*** very interesting in artificial evolution, it is not permitted by the structure of evolutionary algorithms because the evolutionary processes are fixed. Therefore, we propose a new evolutionary algorithm, RBF-Gene. It includes an intermediate level, the proteom (made of "proteins"), between the phenotype of an individual and its genotype, that allows for changes in the structure of the genome without changing the phenotype. We show the existence of an indirect selection in our algorithm, acting on genomes by changing the size of non coding sequences or the order of the genes.

Publications

Simultaneous optimization of weights and structure of an RBF Neural Network, V. Lefort, C. Knibbe, G. Beslon, J. Favrel, Talbi et al., *Artificial Evolution, proceedings of the 7th International Conference, EA 2005, Revised and selected papers, Lille, October 2005*, LNCS 3871, Springer

A bio-inspired genetic algorithm with a self-organizing genome: The RBF-Gene model, V. Lefort, C. Knibbe, G. Beslon, J. Favrel, Kalyanmoy Deb et al., *Genetic and Evolutionary Computation – GECCO 2004, Part II*, LNCS 3103, Springer, p. 406-407

Introducing « proteins » into genetic algorithms, V. Lefort, C. Knibbe, G. Beslon, J. Favrel, dans les actes de la conférence *Complex Systems, Intelligence and Modern Technology Applications (CSIMTA'04, Cherbourg)*, p. 181-186

Scientific activities and valorisation

International activities	Targeted Thematic Action (TTA) week, organized by François Kepès (Génopole d'Evry), on « New ideas for Genetic and Evolutionary Computation inspired by Recent Developments in Biology ». We were 8 : François Kepès (Genopole d'Evry), Jeremy Ramsden (Cranfield University, UK), James Foster (University of Idaho, Moscow, USA), Julian Miller (University of York, Heslington, UK), Wolfang Banzhaf (University of Newfoundland, Canada), Steffen Christensen (Carleton University, Ottawa, Canada), Guillaume Beslon and me (INSA Lyon). After this week, we have written a paper published in *Nature Reviews Genetics*.

Sabine LOUDCHER RABASEDA

Current Position : Associate professor
E-mail : Sabine.Loudcher@univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr/~sabine/
Birth Date : 27/10/1969
Arrival Date : 01/10/1998

Administrative Charges : Since 2003 : Assistant director, ERIC laboratory
1998-2002 : Head of the Computer Science and Statistics department, Institute of Technology, University of Lyon 2

Research topics

My research area is based on combining online analytical processing and data mining in order to improve the decision-making process, especially in the case of complex data. OLAP and data mining could be two complementary fields that interact together within a unique analysis process. The aim of this research is to propose new approaches based on coupling online analytical processing and data mining for exploration, aggregation, explanation and prediction of complex data in OLAP cubes.

In order to do so, we have established four main proposals :
The visualization of sparse data. According to the multiple correspondence analysis, we have reduced the negative effect of sparsity by reorganizing the cells of a data cube.
A new aggregation of facts in a data cube by using agglomerative hierarchical clustering. The obtained aggregates are semantically richer than those provided by traditional multidimensional structures.
An explanation of the possible relationships within multidimensional data by using association rules. We have designed a new algorithm for a guided-mining of association rules in data cubes.
An extension to prediction capacities. Our approach is based on the regression trees and consists in predicting the value measure of new data aggregates. By exploiting the decision rules, the user can anticipate the realization of future events. Moreover, the model makes it possible to improve the knowledge of the relations existing in the data.

Publications

R. Ben Messaoud, S. Loudcher Rabaséda, R. Missaoui, O. Boussaid. OLEMAR: an On-Line Environment for Mining Association Rules in Multidimensional Data. *Advances in Data Warehousing and Mining*, vol. 2. Idea Group Inc., 2007.
O. Boussaïd, J. Darmont, F. Bentayeb, S. Loudcher-Rabaseda, "Warehousing complex data from the Web", *International Journal of Web Engineering and Technology*, 2007.
Riadh Ben Messaoud, Omar Boussaid, Sabine Loudcher Rabaséda, "A Data Mining-Based OLAP Aggregation of Complex Data: Application on XML Documents", *International Journal of Data Warehousing and Mining*, 2(4) : 1-26. Idea Group Inc., 2006.
Riadh Ben Messaoud, Sabine Loudcher Rabaséda, Omar Boussaid, Rokia Missaoui, "Enhanced Mining of Association Rules from Data Cubes", *In Proceedings of the 9th ACM International Workshop on Data Warehousing and OLAP (DOLAP'2006)*, pp. 11-18, Arlington, VA, USA : ACM Press. November, 2006.
Riadh Ben Messaoud, Omar Boussaid, Sabine Loudcher Rabaséda, "Efficient Multidimensional Data Representation Based on Multiple Correspondence Analysis", *In Proceedings of the 12th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining (KDD'2006), pp. 662-667, Philadelphia, PA, USA : ACM Press. August, 2006.

Scientific activities and valorisation	
Scientific programs and/or industrial collaborations	*Since 2003:* Person in charge for FORMASUP RA for the annual inquire of the becoming to training students in the Rhone-Alpes area: 3,800 € per year. *2004-2007: FoDoMuSt (multistrategy data mining).* Project with the LSIIT computer science and LIV geography labs (Strasbourg) for automatically identifying vegetation from satellite images. Funding from the Ministry of Research (ACI project): 69,000 €.
Editorial boards and program committees	*Editorial boards:* International Journal of Biomedical Engineering and Technology ; EDA conferences steering committee *Journal and book paper reviewing:* Revue des Nouvelles Technologies de l'Information (RNTI), Processing and Managing Complex Data for Decision Support, 2005 Conference program committees : ASD 2006-2007, EDA 2006-2007, PKDD 2004, ISWC 2004, JDS-SFdS 2003 Conference organizing committees: EDA 2005, SFdS 2003

Hadj MAHBOUBI

Current Position : PhD student
E-mail : hadj.mahboubi@eric.univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr/~hmahboubi
Birth Date : 17/03/1981
Arrival Date : 01/09/2005

Research supervisor : Jérôme Darmont

Research topics

Decision-support applications currently exploit more and more heterogeneous data from various sources. In this context, XML can greatly help in their integration and warehousing. However, decision-support queries, exploiting XML data warehouses, are generally complex because they involve several join and aggregation operations. In addition, XML-native database management systems present poor performances when data volume is very large and queries are complex.

Several studies address the issue of designing and building XML data warehouses. These works propose different architectures and they differ on the way they represent facts and dimensions. Hence, we define a unified XML data warehouse model. Entirely based on XML formalism, this model is actually the translation of a classical snowflake schema. It presents better performance compared to the existing models.

In order to guarantee the best performance when accessing warehouse data, we propose a new index that is specifically adapted to the multidimensional architecture of XML warehouses [5]. It eliminates join operations. We also design and implement an automatic strategy for the selection of XML materialized views that exploit a data mining technique (clustering of the query workload) [2].

We actually focus on designing a distributed XML data warehouse system to reduce warehouse storage cost and to perform parallel execution of queries. Traditionally, this process involves data fragmentation and fragments repartition. So, we propose to adapt existing fragmentation techniques (as defined in the relational context) to partition XML data warehouses [4]. After that, a repartition architecture (distributed system, peer to peer network or data grid) must be defined. The choice of the architecture is based on the query performance evaluation over these architectures. Hence, a distributed decision-support query processing mechanism is also defined.

Publications

[1] Hadj Mahboubi and Jérôme Darmont. *Indices in XML databases* . Encyclopedia of Database Technologies and Applications, Second Edition. Idea Group Publishing. 2007.

[2] Hadj Mahboubi, Kamel Aouiche, Jérôme Darmont, *Materialized View Selection by Query Clustering in XML Data Warehouses* , 4th International Multiconference on Computer Science and Information Technology (CSIT 06), Amman, Jordan , 2006, pages 68-77

[3] Hadj Mahboubi, Jérôme Darmont, *Benchmarking XML data warehouses* , Atelier Systèmes Décisionnels (ASD 06), 9th Maghrebian Conference on Information Technologies (MCSEAI 06), Agadir, Maroc , December 2006

[4] Hadj Mahboubi, Jérôme Darmont, *Fragmentation des entrepôts de données XML*, 3èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 07), Poitiers , 2007, pages 177-190

[5]Hadj Mahboubi, Kamel Aouiche, Jérôme Darmont, *Un Index de Jointure pour les Entrepôts de données XML*, 6èmes Journées Francophones Extraction et Gestion des Connaissances (EGC 06), Lille , 2006, pages 89-94

Nora MAIZ

Current Position : PhD student
E-mail : nmaiz@eric.univ-lyon2.fr
Web site :
Birth Date : 08/04/1979
Arrival Date : 01/10/2005

Research supervisor : Omar Boussaid and Fafila Bentayeb

Research topics

In a data warehousing process, data integration is one of the most important phases. Centralized data warehouse is a solution for companies that handle static data. However, when data change, this solution is not practical because of the refreshment cost. We believe that data integration by mediation can solve this problem by allowing the construction of a mediation system for building an analysis context on-the-fly using data from their real sources.

The use of ontologies in the mediation process allows semantic and structural integration. In our work, we try to propose a new mediation system based on a hybrid architecture of ontologies modelled according to GLAV (Generalized Local As View) model. The hybrid architecture defines a local ontology for each data source and a global ontology viewed as a mediator. The integration model defines how sources, local and global ontologies are mapped. So we propose an ascending method for building ontologies, which starts by building local ontologies. After that, we use data mining technics to merge local ontologies. This method facilitates the semantic reconciliation between data sources. We use OWL (Ontology Web Language) for defining ontologies and mappings between data sources and ontologies. Moreover, user queries are expressed in the specific language that we propose which handles global ontology concepts and local ontology properties because we assume that the user is expert in his domain. User queries are decomposed by the rewriting algorithm in order to obtain a set of equivalent subqueries that are sent to the corresponding data sources to be executed. After that, the subqueries are recomposed to obtain the final result.

Publications

N. Maiz, O. Boussaid and F. Bentayeb, "Ontology-based mediation system", 13th ISPE International Conference on Concurrent Engineering: Research and Applications (CE06), Antibes, France, September, 2006.

N. Maiz, O. Boussaid and F. Bentayeb, "Ontology-based data integration in datawarehouses", 18th Information Ressources Management Association (IRMA) International Conference, Vancouver, Canada. 2007.

N. Maiz, O. Boussaid and F. Bentayeb, " Un système de médiation basé sur les ontologies pour l'entreposage de données". Atelier Systèmes Décisionnels (ASD 06), 9th Maghrebian Conference on Information Technologies (MCSEAI 06), Agadir, Maroc, December, 2006.

N. Maiz, O. Boussaid and F. Bentayeb, " Fusion automatique des ontologies-OWL par classification hiérarchique pour la conception d'un entrepôt de données", 4ème atelier Fouille de Données Complexes dans un Processus d'Extraction des connaissances (FDC-EGC 07), Namur, Belgique, January 2007.

N. Maiz, K. Aouiche, J. Darmont, " Sélection automatique d'index et de vues matérialisées dans les entrepôts de données". 2ème journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA 06), Versailles, Juin 2006; Revue des Nouvelles Technologies de l'Information, Vol. B-2, 89-104.

Simon MARCELLIN

Current Position : PhD student
E-mail : smarcellin@eric.univ-lyon2.fr
Web site : -

Birth Date : 08/07/1981
Arrival Date : 15/09/2004
Research supervisor : Djamel A. Zighed

Research topics

Our main research topic is machine learning on imbalanced datasets (when an important class is weakly represented). We propose some methods to deal with this kind of data, especially using decision trees-based algorithms :

An asymmetric entropy measure for decision trees: a new splitting criterion taking into account the class imbalance. We also propose a framework for asymmetric entropy measures.

An adaptation of Laplace estimation of probabilities, adapted to imbalanced datasets.

Decision rules adapted to imbalanced data: to obtain a prediction model from a decision tree, a decision rule must be applied on each leaf. We propose different decision rules.

Performance measures of prediction models adapted to imbalanced data, and empirical comparison of asymmetric splitting criteria using ROC curves.

This thesis is financed by the French Ministry of Research and Industry (*CIFRE* financing)

Publications

D.A. Zighed, S. Marcellin, G. Ritschard « Mesure d'entropie asymétrique et consistante », *Revue des Nouvelles Technologies de l'Information*, E-9 (Vol. I), EGC'2007, 81-86.
S. Marcellin, D. Zighed, G. Ritschard, "Une mesure d'entropie asymétrique pour les arbres de décision", *38ème journées des statistique (JDS 06), Clamart, France*, Mai - Juin 2006
S. Marcellin, D. Zighed, G. Ritschard, "An asymmetric entropy measure for decision trees", *11th Information Processing and Management of Uncertainty in knowledge-based systems (IPMU 06), Paris, France*, July 2006, 1292-1299.
S. Marcellin, D. Zighed, G. Ritschard, "Detection of breast cancer using an asymmetric entropy measure", *17th Computational Statistics (COMPSTAT 2006), Rome, Italy*, August - September 2006, 975 - 984.
D. Zighed, S. Marcellin, G. Ritschard, "An asymmetric entropy measure for decision trees", *Knowledge Extraction and Modeling, Island of Capri, Italy*, September 2006.

Efthimios MAVRIKAS

Current Position : PhD student
E-mail : efthimios.mavrikas@eric.univ-lyon2.fr
Web site : -
Birth Date : 04/03/1979
Arrival Date : 06/01/2003
Research supervisor : Djamel. ZIGHED

Research topics

Cultural Heritage documents deal with objects/artifacts and the people that created, owned, used, or (re)discovered them. Their fates are intertwined in unique and complex stories forming a cumulative body of knowledge, often fragmented across large online document collections. While our collective memory has explicitly documented these stories, the heterogeneity of the available sources creates islands of information that can only be implicitly connected by a limited, expert audience.

My current research work aims to define a semantically consistent framework for the online presence of Cultural Heritage document collections, set upon a participatory centre stage and supported by a shared knowledge model. In this framework, Cultural Heritage document contributors benefit from knowledge-rich document processing modules which analyse and classify each contribution, capture the notion of time and the unfolding of events spanning single or multiple documents, and establish meaning connectivity over the entire collection. Overall, this framework assists a scholarly audience with the exploration of online Cultural Heritage document collections, and offers an informed tap into the collective memory scattered therein.

Keywords: Discourse Analysis, Ideology, CIDOC CRM, WorldNet, PLSA, Scripts.

Publications

Efthimios C. Mavrikas, Evangelia Kavakli and Nicolas Nicoloyannis (2005) The Story between the Lines: Exploring Online Cultural Heritage Document Collections using Ontology-based Methods, *Annual Conference of the International Committee for Documentation of the International Council of Museums (CIDOC 2005)*, Zagreb, Croatia, May 2005.

Efthimios C. Mavrikas, Vagelis Stournaras and Christis Konnaris (2005) Historical Memory Preservation on the Semantic Web: the Case of the Historical Archive of the Aegean - Ergani, *33rd International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA 2005)*, Tomar, Portugal, March 2005.

Efthimios C. Mavrikas, Evangelia Kavakli and Nicolas 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.

Efthimios C. Mavrikas, Nicolas Nicoloyannis and Evangelia 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.

Dimitris C. Papadopoulos and Efthimios 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.

Elie PRUDHOMME

Current Position : PhD student

E-mail : eprudhomme@eric.univ-lyon2.fr
Web site : eric.univ-lyon2.fr/~eprudhomme/
Birth Date : 23/01/1979
Arrival Date : 01/10/2005
Research supervisor : Stéphane Lallich

Research topics

The learning process encounters many difficulties to analyze large amount of data. Indeed, algorithms must be of linear complexity to handle these datasets and some theoretical problems, related to high dimensional spaces, appear and degrade their predictive capacity. Furthermore, end-user needs to understand and interact with the prediction.

The selection of data "features" - variables or association rules that can be derived from them - is a simple response to this problem, applied at the pre-processing stage. In high-dimensional space, this selection requires a large number of tests from which arise a number of false discoveries. We have proposed an original non-parametric control method. A new criterion, UAFWER, defined as the risk of exceeding a pre-set number of false discoveries, is controlled by BS-FD, a bootstrap based algorithm that can be used on one- or two-sided problems. We have illustrated the usefulness of that procedure by the selection of differentially interesting association rules on genetic data.

High-dimensional space prevents algorithms from doing a data representation. Nevertheless, in some applications, this representation can help the user to make good use of the learning model. For that purpose, we propose an ensemble approach to overcome problems related to high dimensional spaces. Self-organized map, which allows both a fast learning and a navigation through the data is used like base classifiers to learn several features subspaces. Further, a genetic algorithm is used to optimize diversity of the ensemble by relying on an adapted error measure. This approach offers similar representation capabilities and competitive prediction performance with boosting and random forests.

Publications

Lallich S., Teytaud O., Prudhomme E. (2006), Statistical inference and data mining: false discoveries control. *Proc. of 17th COMPSTAT Symposium of the IASC*, La Sapienza, Rome, août 2006, pp. 325-336.
Prudhomme E. and Lallich S. (2005), Quality measure based on Kohonen maps for supervised learning of large high dimensional data, *Proc. of ASMDA 2005* , pp. 246-255, Brest.
Prudhomme E. et Lallich S. (2007), Ensemble prédicteur fondé sur les cartes auto-organisatrices adaptées aux données volumineuses, *Actes EGC'07, RNTI-E-9*, vol. 2, pp. 473-484, Namur, Belgique.
Prudhomme E. et Lallich S. (2005), Validation statistique des cartes de Kohonen en apprentissage supervisé, *Actes EGC 2005, RNTI-E-3*, vol. 1, pp. 79-90, Paris.
Lallich S.,Prudhomme E. et Teytaud O. (2004), Contrôle du risque multiple en sélection de règles d'association significatives, Actes EGC'04, RNTI-E-2, vol. 2, pp. 305-316, Clermont-Ferrand.

Taimur QURESHI

Current Position : PhD student
E-mail : taimur.q80@gmail.com
Web site : NA
Birth Date : 10/09/1980
Arrival Date : 01/10/2006

Research supervisor : D.A.Zighed

Research topics

The goal of this PhD can be divided into two parts:
Practical Part:
Design and implementation of a generalized algorithm for decision trees and induction graphs.
Development of a software that incorporates the above algorithm and thus, creation of a test bed for experimentation using vast data and understanding of various algorithms and techniques.
Development of an internet based collaborative tool that implements a huge data store for decision trees and induction graph based resources e.g. articles with their summaries, tools, books etc.
Theoretical Part: The theoretical part concerns with the development of new techniques and methods in the area of decision trees and experimentation with huge data on the created test bed and obtaining applicable results.

1) Practical Part:
In this portion a generalized decision tree and induction graph algorithm has been conceived and designed using flowcharts and object oriented designing techniques. The concept of the generalized algorithm is to create such an algorithm which is generic and can be used to implement anyone of the existing decision tree or induction graph techniques. We have implemented various discretization algorithms such as Chi merge, FUSBIN, FUSINTER, MDLPC, CONSTRAST and also various decision trees as ID3, C4.5, CART and Arbogodai. We have implemented our algorithm in R and once the object oriented implementation is completed, we will transfer it into our software which is implemented in Java. The software has a user friendly interface that converts many types of data e.g. text, xml, db etc into a table structure. After that the user can select the type of technique to use on that data and the results shall be given as a graphical output. The third phase of the implementation is development of an internet based collaborative platform for decision tree and induction graph based resource sharing. We have developed a "Wikipedia" like tool for information sharing and editing. It shall contain resumes and sources of many articles, tools and platforms and a test bed; thus forming a complete resource for decision tree and induction graphs.

2) Theoretical Part:
From various studies done earlier, we know that the learning sample is an approximation of the whole population, so the optimal discretization built on a single sample set is not necessarily the global optimal one. Whereas, we proved that resampling gives a better estimate of the distretization point distribution in terms of acheiving a well-defined distribution. We have created a discretization point selection protocol which selects cut points from a certain frequency distribution achieved by resampling. This protocol selects the discretization points from a given frequency point distribution

having higher probability of occurrences and splits on those points if a certain criterion (e.g. entropy) is met. When we apply that protocol, it significantly improves the quality of discretization and prediction rate and thus, nearing to a global optimal solution. Moreover, the same protocol when applied to the frequency point distribution of random samples, achieved much lesser improvements in the prediction rate as compared to bootstrap. We applied the discretization point selection protocol (after resampling) to various methods on the breiman waveform dataset. Except for Chi-Merge, all the other methods provide small variations in terms of prediction rates. MDLPC performs the best and FUSBIN achieves the best time complexity, which is a key point when dealing with a lot of examples.

We applied the above mentioned resampling methodology in the context of fuzzy or soft discretization in decision trees. Our soft discretization technique gives better prediction rates than the hard discretization based methods. As ongoing work, we are applying this soft discretization in building soft decision trees and thus, will try to show that this method will also improve the classification accuracy of decision trees.

Publications

IEEE ICSEA-2004: Integration of Mobile IP and Adhoc Networks with Multi-homing and Smooth Handoff capabilities.
IEEE 16th IST Mobile Summit, Budapest, Hungry: A Network Layer based Hard Real Time Protocol for Wireless Sensor Networks.

Ony RAKOTOARIVELO

Current Position : PhD student
E-mail : orakoto@eric.univ-lyon2.fr
Web site :
Birth Date : 27/08/1981
Arrival Date : 07/11/2006

Research supervisor : Fadila Bentayeb

Research topics

Data warehouses provide an integrated and consistent view on all enterprise data which are relevant for the OLAP analysis. This analysis requires time-variant and non-volatile data. Thus, dimension updates and schema evolutions on the data warehouse are prohibited because they can induce data loss or erroneous results. However, needs and data change constantly. As a result, requirements are not, then, satisfied and some trends are not explored. This is the reason why data warehouse schema evolution becomes an important research topic. In our research, we are interested in the following issue: how can this problem be treated by using data mining techniques? We have proposed a schema evolution operator based on the k-means clustering algorithm. This leads us to the very interesting research topic of online data mining which is how to integrate effectively data mining methods in a RDBMS (Relational DataBase Management System).

Publications

O. Rakotoarivelo, F. Bentayeb, "Evolution de schéma dans les entrepôts de données: utilisation de la méthode des k-means", 4ème atelier Fouille de Données Complexes dans un Processus d'Extraction des Connaissances (FDC-EGC 07), Namur, Belgique, Janvier 2007.

O. Rakotoarivelo, F. Bentayeb, "Evolution de schéma par classification automatique pour les entrepôts de données", 3èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 07), Poitiers, Juin 2007; Revue des Nouvelles Technologies de l'Information, Vol. B-3, 99-112

Ricco RAKOTOMALALA

Current Position : Assistant professor

E-mail : Ricco.rakotomalala@univ-lyon2.fr

Web site : http://eric.univ-lyon2.fr/~ricco/

Birth Date : 19 July 1967

Arrival Date : 01 Sept 1995

Administrative charges : Joint manager of the strand SISE (Statistics & Informatics) in Master IDS (Business Intelligence and Statistic)

Research topics

My research's activities are mainly the applications of data mining. We try to characterize the outline of a successful data mining process, in the area, needed to be precisely defined. One of our goal, but not restricted to, is the determination of the most effective strategies in this context.
One of my highlighted project, with many publications, is the automatic classification of protein from their primary structure. Carried out in collaboration with Mr. Elloumi of the Faculty of Science of Tunis (Tunisia), this work is an important step in the defense of PhD thesis of Mr. Mhamdi at the beginning of 2008. The principal task is the comprehension of data comprising a few observations but a very large number of descriptors, that are being automatically generated from very rough techniques such as the n-grams. We developed rapid approaches for dimensionality reduction by carrying out a very aggressive feature selection without reducing the accuracy of the classifiers. The experiments show, without surprise, that the margin maximization methods such as SVM (Support Vector Machines) are powerful. Surprisingly, other strongly regularized approaches such as PLS regression, not well known in the machine learning community, are also very accurate.

Another project is the automatic classification of planktons from scanned images. Carried out in collaboration with the team of Mr. Gorsky of the laboratory of Oceanography of Villefranche-sur-Mer (France), this project aims to industrialize the automatic categorization of planktons (Plankton Identifier Project -- http://www.obs-vlfr.fr/). We also handle unstructured datasets here, with the original data description being the image. Beyond the research of the most effective strategy, the question of validation of performances arise. Indeed, the composition of the marine environment is very dynamic, according to the location, and the period. We must take a new view of the validation problem, as the traditional indicators (accuracy rate in particular) are not really adequate. We must produce results which are transposable in various contexts.
This project is accompanied by important software development activity software which is placed freely at the disposal of the scientific community (http://www.obs-vlfr.fr/~gaspari/Plankton_Identifier/index.php).

Then, the last highlighted project, more personal, is the development of the TANAGRA data mining software, freely available with source code on the web (http://chirouble.univ-lyon2.fr/~ricco/tanagra/). The project, started in 2003, comprises of more than 200.000 lines of source code today. With about 100 implemented methods, it covers a very broad field of the data mining techniques, starting from the statistical approaches (parametric and not-parametric tests), to

the machine learning algorithms (supervised and unsupervised, association rule mining), while passing by the traditional techniques of the exploratory data analysis (factorial methods, etc.).

The diffusion of the software is accompanied by about sixty tutorials in English and in French. Our Web site has a rather good frequentation. On average, we count 130 visitors daily since the beginning of the year 2007.
It is a very important project for me. To give a larger base to the dissemination of the knowledge, I started to put on-line detailed course notes. The described techniques can be applied directly via the free software, via TANAGRA of course, but also using tools such as the R-project software or a spreadsheet. We count nearly 50 visitors per day since the starting of the website (January, 2007). This value is all the more interesting since all the documents are in French (http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html). In addition, we developed a website which directs on the most interesting courses notes that one can find on the Web about the various subjects which are related to the data mining process (http://eric.univ-lyon2.fr/~ricco/data-mining/).

Publications

E. Antajan, R. Rakotomalala, S. Gasparini, M. Picheral, L. Stemmann, G. Gorsky, « Automatic quantification and recognition of major zooplankton groups in a North Sea time series using the Zooscan imaging system", in the Proceedings of the 4th International Zooplankton Production Symposium, pp. 189-190, Hiroshima, Japan, 2007.
Chauchat J.H., A. Morin & R. Rakotomalala, 2007. "Correcting the error rate estimation bias in Data Mining when the dataset comes from a two-stage sampling", Statistics for Data Mining, Learning and Knowledge Extraction (IAST'07), Aveiro, Portugal.
F. Clerc, D. Farrusseng, R. Rakotomalala, N. Nicoloyannis, C. Mirodatos, "Meta Modeling for Combinatorial Catalyst Optimization", International Journal of Computer Science and Network Security, vol. 6, n°10, pp.256-262, 2006.
R. Rakotomalala, F. Mhamdi, "Supervised and Unsupervised Feature Reduction for Protein Classification", WSEAS International Journal -- WSEAS Transactions on Information Science and Applications, vol. 3, n°12, pp. 2448-2455, 2006.
A. Morineau, Rakotomalala R. "The TVpercent Criteria to Eliminate Uninformative Models among Association Rules", in Electronic Proceedings of Knowledge Extraction and Modeling, IASC-INTERFACE-IFCS Workshop, KNEMO'06, Anacapri, Italy, 2006.

Jean-Christian RALAIVAO

Current Position : PhD student
E-mail : jean-christian.ralaivao@eric.univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr/~jcralaivao/

Birth Date : 03/07/1966
Arrival Date : 01/11/2003
Research supervisor : Jérôme Darmont

Research topics

Within decision processes, data warehousing technologies are now mature to handle simple, numerical or symbolic data. However, various sources including the Web store, contain very heterogeneous data: texts, images, sounds, videos, databases, temporal or geographical data; which may be expressed in several languages, stored in various formats, located in different places and frameworks, etc. These so-called complex data carry a lot of information and are thus interesting to include within a decision process. However, numerous issues relate to structuring, storing and querying complex data.

The aim of my PhD thesis is to address the issue of complex data warehouse performance. Several techniques do exist to optimize simple data access and storage in a warehouse. However, they cannot be applied very efficiently onto complex data. Thus, we have to define complex data warehouse models that are adapted to the nature of stored data, and to design custom performance optimization tools for these warehouses: indexing, view materialization, partitioning, clustering, buffering, etc.

Aside, using the XML language for managing data warehouses has several advantages, especially when integrating heterogeneous data. XML indeed helps represent both structure and contents. Hence, we have proposed an XML-based complex data warehouse architecture [2] that helps benefit from optimization techniques developed in the database and XML communities.

Performance optimization always needs well-defined measures and metrics. In order to identify them, we listed performance indicators for complex data warehouses. This list helped identify performance factors that are used to determine metrics.

Finally, integrating metadata and domain-related knowledge in the complex data warehouse has a positive impact on managing data complexity, especially in the process of performance optimization [1, 3].

Publications

1. J.C. Ralaivao, J. Darmont, "Knowledge and Metadata Integration for Warehousing Complex Data", *6th International Conference on Information Systems Technology and its Applications (ISTA 07), Kharkiv, Ukraine*, May 2007; *Lecture Notes in Informatics (LNI)*, Vol. P-107, 164-175.
2. J. Darmont, O. Boussaïd, 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, 370-373.
3. 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.

Rashed Khalil SALEM

Current Position : PhD student
E-mail : rashed.salem@eric.univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr/~rsalem/
Birth Date : 20/10/1981
Arrival Date : 01/11/2007

Research supervisor : Jérôme Darmont & Omar Boussaïd

Research topics

Decision-support technologies, including data warehouses and OLAP (On-Line Analytical Processing), are nowadays technologically mature. However, their complexity makes them unattractive to many companies; hence, some vendors develop simple Web-based interfaces (Lawton, 2006). Furthermore, many decision-support applications necessitate external data sources. For instance, performing competitive monitoring for a given company requires the analysis of data available only from its competitors. In this context, the Web is a tremendous source of data, and may be considered as a farming system (Hackathorn, 2000).

There is indeed a clear trend toward on-line data warehousing, which gives way to new approaches such as virtual warehousing (Belanger et al., 1999) or XML warehousing (Pokorny, 2002; Hümmer, 2003; Park et al., 2005; Boussaïd, Darmont et al., 2007). This research is backed up by new technologies such as Web services, a set of protocols and norms that help exchange data between applications over the Web (Eckert, 2005), or Active XML, a declarative framework that harnesses Web services for data integration in a peer-to-peer architecture (Abiteboul et al., 2002).

The ERIC lab is currently designing and developing a whole XML warehousing platform. In this context, the objective of this thesis is to design and include into this platform active features (Thalhammer et al., 2001) to turn it into an active XML warehouse. This work includes integrating analysis scenarios into the warehouse, automatically. Such scenarios may be based on ECA (Event, Condition, Action) rules similar to that used active databases (Dayal et al., 1995). To devise ECA rules within the OLAP framework, they may be coupled with analysis graphs. This technique shall help break up an OLAP cube with classical OLAP operators to express the targeted on-line analysis scenario. Eventually, an active XML warehouse may be viewed as a set of distributed data sources over a peer-to-peer architecture. Deploying on-line operations for this kind of warehouse shall be based on Web services (Bonifati et al., 2000).

The objective of my PhD thesis are to:
propose an approach to integrate ECA rules into the warehousing process,
design an algebra for analysis graphs,
define a framework to handle problems linked to different analysis scenarios,
propose a Web service-based architecture for automatic, Web-based, on-line analyses.

Publications

- Rashed Khalil, Wail Elkilani, Nabil Ismail, Mohie Hadhoud, "A Cost Efficient Location Management Technique for Mobile Users with Frequently Visited Locations ", Proceedings of the 4th Annual Communication Networks and Services Research Conference (CNSR'06) - Volume 00, p.p. 259 - 266
- Rashed Khalil, Wail Elkilani, Nabil Ismail, Mohie Hadhoud, "A Cost Efficient Location Management Technique using Replicated Database ", INFOS2006, March 2006, Cairo, Egypt.

Anna STAVRIANOU

Current Position : PhD student
E-mail : Anna.Stavrianou@univ-lyon2.fr
Web site : -
Birth Date : 23/02/1977
Arrival Date : 01/10/2005

Research supervisor : Jean-Hugues CHAUCHAT

Research topics

Text mining is an interdisciplinary field that combines techniques and methodologies from various areas such as information extraction, information retrieval, computational linguistics and categorization. In our work, we concentrate on the semantic rather than the statistical techniques since it seems that the statistics alone are not sufficient for the mining of a text. More specifically, our objective is to make an initial step in combining text mining and database methodologies for the purpose of categorizing and retrieving knowledge from text.
We base our work on the LIMBO [1] algorithm which is a hierarchical clustering algorithm for categorical data, based on the Information Bottleneck framework. It has been used to cluster both tuples and categorical attribute values, discover duplication in a set of tuples and identify structures in data that may contain erroneous information or duplicates. Within LIMBO, the similarity between the values of the same attribute is measured on the basis of the distribution they induce on the remaining attributes. The semantics of the attribute values are not taken into account. In this work, our goal is to identify the similarities between the values of each tuple attribute and feed this semantic information into LIMBO in order to perform clustering of the tuples. A comparison between the clustering results while using or not the semantic information provided, will allow us to identify whether semantics can benefit the clustering task or not.
For the purpose of incorporating semantic knowledge into the tuple representation, our objective becomes two-fold: a)find the semantic relations among the values of a particular attribute and b)use these relations in order to re-distribute the weights in each tuple. A Java application has been implemented called "SemanticLIMBO" in order to allow for the application of various semantic measures on the values of an attribute. The available measures include those proposed by Seco et al. [2] and the measures that appear in the WordNet-Similarity package [3].

References

Andritsos, P., Tsaparas, P., Miller R.J., Sevcik K.C. 2004. LIMBO: Scalable Clustering of Categorical Data. In *9th International Conference on Extending Database Technology (EDBT),* March 2004.
Seco, N., Veale, T., and Hayes, J. 2004. An intrinsic information content metric for semantic similarity in WordNet. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, Valencia, Spain, 1089-1090.
WordNet-Similarity. http://www.d.umn.edu/~tpederse/similarity.html

Publications

Stavrianou, A., Andritsos, P., and Nicoloyannis, N. Overview and Semantic Issues of Text Mining. In SIGMOD Record, 36(3), September 2007, 23-34.

Julien THOMAS

Current Position : PhD student
E-mail : jthomas@fenics-sas.com
Web site :
Birth Date : 02/23/1982
Arrival Date : 09/19/2005
Research supervisor : D. Zighed (N. Nicoloyannis)

Research topics

Measure for supervised learning models assessment, taking into account user needs specificities and working well on imbalanced datasets. (PRAGMA : Precision and RecAll Guided Model Assessment)

Adaptive sampling strategy for imbalanced datasets and random forest. (FUNSS : Fitting User Needs Sampling Strategy)

Association rules base fuzzification.

Features construction and reduction of high dimensional space using association rules fuzzyfication.

Evolutionary features space for random forest. (G2S : Gradual Shaping Space)

Supervised visual and interactive clustering.

Search of similarity between objects using random forest.

Publications

J. Thomas, S. Marcellin "Fouille de bases d'images mammographiques", Groupe de Travail sur la Fouille de Données Complexes, Lyon, France, Septembre 2005.
A. Brémond, P.E. Jouve, J. Thomas, J. Clech, D.A. Zighed "Résultats préliminaires d'une étude comparative de deux CAD", Innovations Technologiques et bonnes pratiques en sénologie, Congrès de la SOciété Française de Mastologie et d'Imagerie du Sein (SOFMIS 06), Clermont-Ferrand, France, Mai 2006; pp 92-94.
J. Thomas, P.E. Jouve, N. Nicoloyannis "Optimisation and evaluation of random forests for imbalanced datasets", 16th International Symposium on Methodologies for Intelligent Systems (ISMIS 06), Bari, Italy, September 2006; Springer LNAI, Vol. 4203, pp 642-651.
J. Thomas, P.E. Jouve, N. Nicoloyannis "Mesure non symétrique pour l'évaluation de modèles, utilisation pour les jeux de données déséquilibrés", Extraction et Gestion des Connaissances (EGC 07), Namur, Belgique, Janvier 2007; Cepadues RNTI, Vol E-9.
J. Thomas, P.E. Jouve, N. Nicoloyannis "Asymmetric measure for supervised learning models assessment, application to breast cancer detection", International Conference on Industrial Engineering and Systems Management (IESM 07), Beijing, China, May 2007.

Julien VELCIN

Current Position : Assistant professor
E-mail : Julien.Velcin@univ-lyon2.fr
Web site : http://eric.univ-lyon2.fr/~jvelcin/
Birth Date : 09/03/1978
Arrival Date : 01/11/2007

Administrative Charges : Joint manager of the strand ECD in Master IDS (Business Intelligence and Statistic)

Research topics

As a whole, my researches deal with artificial intelligence, machine learning and data mining. More precisely, I'm working on concept extraction from symbolic and sparse datasets. This task is done in a non-supervised way, task which is known as *conceptual clustering*, as defined by Michalski, Diday et al. A lot of applications can be addressed like online news analysis, technological survey and database summary. I also take into account the relationship of my work with other areas, such as psychology and sociology.

My current work is on topic extraction from binary, sparse and high-dimensional datasets. I use an optimization approach and especially the meta-heuristic of tabu search defined by Glover and Laguna in order to go through this very combinatorial search space. The preliminary results I obtained, both on artificial datasets and on real news sources, are really promising and lead to publications at the MLDM and ADMA international conferences. This work is done in collaboration with sociologists from the EHESS in Paris who are studying press content and controversies through the media. I'm also working on non-supervised learning evaluation, both with a theoretical point of view and considering a pragmatic approach. Hence, a clustering evaluation software was implemented and presented at EGC in 2007.

Publications

VELCIN, J. and GANASCIA, J.-G.. Default Clustering with Conceptual Structures. In *Journal on Data Semantics VIII*, LNCS 4380, Springer-verlag (2007), p. 1-25.
VELCIN,J. and GANASCIA, J.-G.. Topic Extraction with AGAPE. In: *Proceedings of the International Conference on Advanced Data Mining and Applications* (ADMA 2007). GANASCIA, J.-G. and VELCIN,J.. Modeling Stereotype Construction with Artificial Intelligence. *Annual scientific meeting of the International Society of Political Psychology (ISPP)*. Portland, Oregon, USA (2007).
VELCIN, J. and GANASCIA, J.-G.. Stereotype Extraction with Default Clustering. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. Edinburgh, Scotland (2005).
VELCIN, J. and GANASCIA, J.-G.. Modeling default induction with conceptual structures. In *ER 2004 Conference Proceedings*. Lu, Atzeni, Chu, Zhou, and Ling editors. Springer-Verlag. Shangai, China (2004).

Scientific activities and valorisation	
Scientific programs and/or industrial collaborations	Project "metadata extraction from textual data using machine learning techniques", in collaboration with the LIP6 (Paris 6 University) and Alcatel-Lucent.

Jacques VIALLANEIX

Current Position :	Associate Professor	
E-mail :	jacques.viallaneix@univ-lyon2.fr	
Web site :	http://eric.univ-lyon2.fr	
Birth Date :	06/07/1963	
Arrival Date :	01/09/1993	
Administrative and Pedagogic Charges :	Within the Faculty of Sociology and Anthropology of the University Lyon 2 : In charge of teaching in data processing for the 1st years of Bachelor of Sociology and of Bachelor of Anthropology until in 2004 (representing on average 450 hours TD per year) ; In charge of teaching in Data processing for the 2nd year of Bachelor of Sociology and of Bachelor of Anthropology until in 2005 (representing on average 400 hours TD per year) ; Co-responsible since 2005 ; In charge of the 2nd year of the Course Bachelor of MISASHS (Mathematics, Computer Science and Applied Statistics for Humanities and Social Sciences) (representing on average 470 hours per year) ; In charge of the 3rd year of the Course Bachelor of MISASHS (representing on average 280 hours per year) ; Being added to on average 300 hours per year of personal teaching, these responsibilities are heavy : ranging from the development (and update) of teaching contents until recruitment of professors or lecturers, adding the management of computer rooms, organizing schedules, and so on ; Member of the recruitment committee in mathematics and computer science of the university Lyon 2 (CSE 26-27-61) ; Member of the recruitment committee in computer science of INSA de Lyon (CSE 27-61).	
Research topics	Because of my teaching and administrative charges, I unfortunately cannot currently lead a substantial research activity.	

Zhihua WEI

Current Position : PhD student
E-mail : zhihua.wei@univ-lyon2.fr
Web site :
Birth Date : 22/01/1979
Arrival Date : 01/12/2006

Research supervisor : Jean-Hugues CHAUCHAT

Research topics

My research area is Chinese text mining, including Chinese text categorization and extracting knowledge from texts based on statistical learning and natural language understanding.
My research objectives include:
1. Text representation methods which are based on bag-of-words, n-grams, keywords or noun phrases and verb phrases. Besides n-grams, the other methods are all based on natural language analysis. Different from most Latin languages, there is no delimiter between two characters in Chinese texts. As a result, most of researches based on the text content need the process of word segmentation as prerequisite. Disambiguation and recognition of unknown words are the most difficult in this process.
2. Text feature selection methods which include improving the traditional selection algorithm and exploring better methods for measuring the dissimilarities among texts in different classes.
3. Multi-class and multi-label classification methods which mainly aim to solve the classification problem in some large corpora. My work is exploring the methods to improve the performance of classifier in the complex multi-class and multi-label conditions and decrease the effects from unbalanced distribution among real corpora.

Publications

1. Book chapter: *"Chinese Language Understanding Algorithms and Applications"* Duoqian Miao, Zhihua WEI, 2007 by Tsinghua University Press (in China).
2. *A New Structure-based Bill Location Technique*, Zhihua WEI, Duoqian MIAO, Fuchun XIA, Hongyun ZHANG, Computer Application.No.10.2006.

Djamel Abdelkader ZIGHED

Current Position : Full professor,
E-mail : Abdelkader.zighed@univ-lyon2.fr
Web site : http://morgon.univ-lyon2.fr

Birth Date : 12/March/1955
Arrival Date : 1/Oct./1987

Administrative Charges :
Head of the ERIC's Lab (1995-2002; since 2007)
President of the recruitment commission for mathematic-informatics and automatic at the university Lyon 2
Member of the scientific council of the university Lyon 2 (2007-...)
Head of the Master "Extraction des Connaissances à partir des Données (ECD)" on Data Mining (Univ. Of Lyon 2 and Univ. Of Nantes) (since 1999)
Head of the PhD program in Computer Sciences of University Lyon 2 (1995-2007)
Member of the steering committee at the faculty of economics

Research topics :

My research interests focus mainly on data mining problems, involving particularly complex data (heterogeneous, semi or unstructured, large data sets). The objective is to study the different representation spaces of the data and how the domain knowledge can be incorporated to better manage data mining tasks. This is done by proposing new machine learning algorithms that better take into account the real world applications constraints. More particularly, the current research work focuses on the following problems:
In the area of machine learning, two directions are being explored. The first one is part of the PhD thesis of Simon Marcellin. It aims to identify approaches to directly tackle the problems of mining datasets with imbalanced classes. This led us to review the properties of entropy measures and to define a new more appropriate one. The second direction is related to continuous attributes discretization. Here, we introduce re-sampling based approaches, such as the Boostrap. These works, under investigation in Taimur Qureshi's PhD thesis, led to new algorithms for fuzzy trees. In the area of the mining complex data, two directions are also followed. The first exploits topological approaches. It uses the neighborhood graphs based models for navigating, in a more appropriate and natural way, in multimedia databases. This work, done during Hakim Hacid's PhD thesis, will be presented for defense in early 2008. The second direction, followed in Ahmad El Sayed's PhD thesis, aims to capture domain knowledge (taxonomies, ontologies) automatically from text corpora. That is, new techniques were proposed for more effective clustering on textual data, that will be used in a general framework for taxonomy learning from text. As a continuity of the work that I have conducted on the use of neighborhood graphs in machine learning, and, more generally, in data mining, a new project is being launched. The finality is to apply our results to semi supervised learning. This will help us to connect, at the same time, to other emerging works in the field of topological learning.

Publications

Berka, P., Rauch, J. and Zighed, D. A., (eds.) *Case studies in medical data mining*, Idea Group, 2008 .
Ciampi, O., Zighed, D. A. and Ritschard, G. *Prediction Trees*, Wiley, 2008 (To appear).
Zighed, D. A. "Induction Graphs for Data Mining", *in* Brito, P., Bertrand, P., Cucumel, G. and de Carvalho, F., ed.,'Selected Contributions in Data Analysis and Classification', Springer, 2007, pp. 419-430.
Sayed, A. E., Hacid, H. and Zighed, D. A. "A New Context-Aware Measure for Semantic Distance Using a Taxonomy and a Text Corpus"'Proceedings of the IEEE International Conference on Information Reuse and

Integration, IRI 2007, 13-15 August 2007, Las Vegas, Nevada, USA', IEEE Systems, Man, and Cybernetics Society, 2007, pp. 279-284.
Zighed, D. A. and Hacid, H. "Proximity graphs and separability of classes"'Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2006, Paris', IPMU, Paris, 2006, pp. 1488-1495.

Scientific activities and valorisation

Scientific programs and/or industrial collaborations

International Labour Organisation (BIT) and University of Geneva: Mining Expert Comments on the Application of ILO Conventions on Freedom of Association and Collective Bargaining.
Breast Cancer Center Leon Berard Lyon : Computer Aided Diagnosis on mammograms
Sanofi-pasteur : Improving the production of vaccines
INTERREG IIIA France-Suisse INTERREG IIIA France-Suisse
INTERREG IIIA France-Suisse : Study of the interdependence of markets residential property on Lake Geneva

Editorial boards and program committees

	since (Year)								
	00	01	02	03	04	05	06	07	08
International Conference on Advanced Data Mining and Applications (ADMA)						■			
International Society devoted to the advancement of the theory and practice of stochastic models and data analysis techniques (ASMDA)						■	■	■	
joint meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Society of Statistics (CLADG-SFC)									■
International Conference on Computational Statistics (COMPSTAT)									■
Data Warehousing and Knowledge Discovery (DaWak)							■	■	■
International Conference on Discovery Science (DS)					■	■	■		
European Conference on Machine Learning (ECML)		■	■	■	■	■	■	■	■
Journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA)							■	■	■
European Semantic Web Conference (ESWC)									■
Atelier "Fouille de Données Complexes" associé à EGC									■
Flexible Query Answering Systems					■				
International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)								■	
Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GfKL)							■	■	■
International Association for Statistical Computing (IASC); Statistics for Data Mining, Learning and Knowledge Extraction, Satellite meeting of International Statistic Institute (ISI)								■	
International Conference on Natural Computation (ICNC)								■	
International Conference on NonConvex Programming (ICN)								■	
International Conference Intelligent Information Systems (IIS)				■	■	■	■	■	■
International Symposium on Methodologies for Intelligent Systems (ISMIS)			■	■	■	■	■	■	■
International Semantic Web Conference					■	■	■	■	
Journées Francophones sur les Réseaux Bayésiens (JFRB)					■	■	■	■	■
Mining Complex Data Workshops (IEEE ICDM & PKDD/ECML)						■	■	■	■
International Workshop on Multimedia Data Mining "Mining Integrated Media and Complex Data"					■	■			
Atelier « Mesures de similarité sémantique » associé à EGC									■
Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)			■	■	■	■	■	■	■
European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)	■	■	■	■	■	■	■	■	■
Rencontre sur la Statistique Implicative et ses Applications (SIA)						■	■	■	
Workshop on Visual Data Mining VDM@ICDM				■					

Other activities

Co-founder and co-director of the journal RNTI (since 2001)

President of the Association EGC "Extraction et Gestion des Connaissances" (since 2006) ; co-founder and VP of EGC (since 2001).

Vice-President of SFC « Société Francophone de Classification »

Member elected of the International Statistical Institute

Member of the board of the European Section of (IASC) "International Association for Statistical Computational" in charge of the relationship with machine learning community.

II. ACTIVITE EDITORIALE

Depuis 2003, en co-direction, D.A. Zighed a publié chez Cépaduès les numéros suivants pour la revue RNTI :

Guillet, F. & Trousse, B. *(ed.)*
Extraction et gestion des connaissances (EGC'2007), Actes des huitièmes journées Extraction et Gestion des Connaissances, Nice, France, 29 janvier - 1er février 2008, 2 Volumes
Cépaduès-Éditions, **2008**, RNTI-E-10

Noirhomme-Fraiture, M. & Venturini, G. *(ed.)*
Extraction et gestion des connaissances (EGC'2007), Actes des septièmes journées Extraction et Gestion des Connaissances, Namur, Belgique, 23-26 janvier 2007, 2 Volumes
Cépaduès-Éditions, **2007**, RNTI-E-9

Bénani, F.; Béra, M.; Patrat, C. & Saporta, G. *(ed.)*
Data Mining et Apprentissage Statistique : Application en Assurance, Banque et Marketing
Cépaduès-Éditions, **2007**, A-1

Bénani, Y. & Viennet, E. *(ed.)*
Apprentissage Artificiel et Fouille de Données
Cépaduès-Éditions, **2007**, A-2

Bellatreche, L.; Giacometti, A. & Marcel, P. *(ed.)*
Entrepôt de Données et Analyse en Ligne (3)
Cépaduès-Éditions, **2007**, B-3

Prince, V.; Kodratoff, Y.; Azé, Jé. & Roche, M. *(ed.)*
Défi Fouille de Textes
Cépaduès-Éditions, **2007**, E-10

Aït-Ameur, Y.; Boniol, F. & Wiels, V. *(ed.)*
Isola 2007 Workshop on Leveraging Applications of Formal Methods, Verification and Validation
Cépaduès-Éditions, **2007**, SM-1

Reynaud, C. & Venturini, G. *(ed.)*
Fouille du Web
Cépaduès-Éditions, **2007**, W-1

Ritschard, G. & Djeraba, C. *(ed.)*
Extraction et gestion des connaissances (EGC'2006), Actes des sixièmes journées Extraction et Gestion des Connaissances, Lille, France, 17-20 janvier 2006, 2 Volumes
Cépaduès-Éditions, **2006**, RNTI-E-6

Grigori, D.; Lopes, S.; Nguyen, B. & Zeitouni, K. *(ed.)*
Entrepôt de Données et Analyse en Ligne (2)
Cépaduès-Éditions, **2006**, B-2

Poulet, F. & Kuntz, P. *(ed.)*
Visualisation en Extraction de Connaissances
Cépaduès-Éditions, **2006**, E-7

Khenchaf, A. *(ed.)*
Systèmes d'Information pour l'Aide à la Décision en Ingénierie des Systèmes
Cépaduès-Éditions, **2006**, E-8

Pinson, S. & Vincent, N. *(ed.)*
Extraction et gestion des connaissances (EGC'2005), Actes des cinquièmes journées Extraction et Gestion des Connaissances, Paris, France, 18-21 janvier 2005, 2 Volumes
Cépaduès-Éditions, **2005**, RNTI-E-3

Bentayeb, F.; Boussaid, O.; Darmont, Jé. & Rabaséda, S. *(ed.)*
Entrepôts de Données en Ligne
Cépaduès-Éditions, **2005**, B-1

Boussaid, O.; Gançarski, P.; Masseglia, F. & Trousse, B. *(ed.)*
Fouille de Données complexes
Cépaduès-Éditions, **2005**, E-4

Cloppet, F.; Pettit, J. & Vincent, N. *(ed.)*
Extraction des Connaissances : État et Perspectives
Cépaduès-Éditions, **2005**, E-5

Hébrail, G.; Lebart, L. & Petit, J. *(ed.)*
Extraction et gestion des connaissances (EGC'2004), Actes des quatrièmes journées Extraction et Gestion des Connaissances, Clermont Ferrand, France, 20-23 janvier 2004, 2 Volumes
Cépaduès-Éditions, **2004**, RNTI-E-2

Chavent, M. & Langlais, M. *(ed.)*
Classification et Fouille de Données
Cépaduès-Éditions, **2004**, C-1

Briand, H. & Sebag, M. *(ed.)*
Mesures de Qualité pour la Fouille de données
Cépaduès-Éditions, **2004**, E-1

Boussaid, O. & Lallich, S. *(ed.)*
Entreposage et Fouille de données
Cépaduès-Éditions, **2003**, 1

III. ORGANISATION DE MANIFESTATIONS SCIENTIFIQUES

a. Conférences, ateliers et groupes de travail

EGC : La conférence Extraction et Gestion des Connaissances (EGC) a pour objet de rassembler les chercheurs des disciplines de l'informatique décisionnelle, de l'Extraction de Connaissances dans les Données (ECD, KDD) et de la Gestion des connaissances (GC, KM). Plus précisément, elle se fixe pour objectif de promouvoir les échanges multidisciplinaires (apprentissage, statistiques et analyse et données, systèmes d'information et bases de données, ingénierie des connaissances), en connexion avec les spécialistes d'entreprises qui déploient les méthodes et les outils adaptés à leurs besoins, afin de contribuer à la formation d'une communauté scientifique dans le monde francophone autour de cette double thématique de l'extraction et de la gestion de connaissances. Les chercheurs du laboratoire ERIC ont été à l'origine de la création de cette conférence en 2000 et continuent aujourd'hui d'animer cette manifestation d'envergure nationale.

EDA : La conférence nationale sur les Entrepôts de Données et l'Analyse en-ligne a été créée par les chercheurs du laboratoire d'ERIC. L'objectif de la 1ère journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA 05) est de créer et de pérenniser un cadre exclusivement réservé à ces travaux, afin de favoriser la rencontre des chercheurs, des industriels et des utilisateurs français et francophones afin de discuter de l'avancement de la recherche ainsi que d'expériences de développement dans le domaine des entrepôts de données. Cette journée a pour vocation de devenir un rendez-vous national régulier sur le thème des entrepôts de données. 3 éditions ont déjà eu lieu. La prochaine est prévue à Toulouse le 5 et 6 juin 2008.

EDA 2008: http://www.irit.fr/EDA08/contact.html

EDA 2007: http://eda2007.sir.blois.univ-tours.fr/

EDA 2006: http://www.prism.uvsq.fr/~eda06/

EDA 2005: http://eric.univ-lyon2.fr/~eda05/

L'atelier Qualité des Données et des Connaissances, en assocation avec la conférence Extraction et Gestion des Connaissances (2007-2008) a été organisé en collaboration avec P. Lenca (Telecom Bretagne) et F. Guillet (LINA, Nantes).

http://conferences.enst-bretagne.fr/qdc2007/ et http://conferences.enst-bretagne.fr/qdc2008/

Atelier FDC, en assocation avec la conférence Extraction et Gestion des Connaissances, nous avons produit cinq éditions de cet atelier :

- 29 Janvier 2008 : Cinquième Aatelier sur la "Fouille de Données Complexes dans un processus d'extraction des connaissances", Nice, Sophia-Antipolis, à l'occasion d'EGC 2008
- 23 Janvier 2007 : Quatrième Aatelier sur la "Fouille de Données Complexes dans un processus d'extraction des connaissances", Namur, Belgique, à l'occasion d'EGC 2007
- 16 Janvier 2006 : Troisième Aatelier sur la "Fouille de données complexes dans un processus d'extraction des connaissances", Lille, à l'occasion d'EGC 2006
- 18 Janvier 2005, : Deuxième atelier sur la "Fouille de données complexes dans un processus d'extraction des connaissances, Paris, à l'occasion d'EGC 2005."
- 20 Janvier 2004 : Premier Atelier sur la "Fouille de données complexes dans un processus d'extraction des connaissances", Clermont-Ferrand, à l'occasion d'EGC 2004

Atelier Mesure de similarité sémantique (SimSem 2008)

http://www-rocq.inria.fr/axis/SimSem/AtelierEGC2008/AtelierEGC.html

Journée de travail thématique sur les mesures de similarité sémantiques : Cette première journée de travail sur les données complexes se donne comme objectif de réfléchir et de discuter autour des travaux présentés sur une première partie d'un bilan sur les approches existantes.

http://eric.univ-lyon2.fr/%7Enmaiz/cmss07/

Atelier sur les Systèmes Décisionnels (ASD) : L'objectif de ce premier atelier maghrébin sur les systèmes décisionnels est de contribuer, en collaboration avec le laboratoire ERIC, à dynamiser la recherche dans ce domaine et à créer une synergie entre les chercheurs maghrébins travaillant dans leur pays ou dans des laboratoires de recherche à l'étranger. Cet atelier s'adresse à tous les experts de la communauté internationale travaillant sur les entrepôts de données pour venir exposer, discuter leurs travaux de recherche (qu'elle soit fondamentale ou appliquée), s'échanger des points de vue et présenter leurs outils décisionnels. C'est également l'occasion d'encourager l'ensemble des doctorants maghrébins concernés par ces questions à participer à cette manifestation et se faire connaître afin de faire émerger une véritable communauté travaillant dans le domaine des systèmes décisionnels. La prochaine édition aura lieu en 2008 après les 2 précédentes :

- ASD 2008, Mohammadia, Maroc le 10et 11 octobre 2008 : http://eric.univ-lyon2.fr/~asd/asd2008/

- ASD 2007, Sousse, Tunisie le 19 et 20 octobre 2007 : http://eric.univ-lyon2.fr/~asd/asd2007/
- ASD 2006, Agadir, Maroc le 6et 8décembre 2006 : http://eric.univ-lyon2.fr/~asd/asd2006/

Co-organisation de la 6th International Conference on Flexible Query Answering Systems (FQAS 2004), 24-26 june, 2004 (Lyon).

b. Séminaires du master ECD

2003-2004

- Jean-Paul Rasson, LIGSAT, Facultés Universitaires N-D de la Paix, Namur, Belgique, De deux méthodes de stratification avant discrimination, Jeudi 11 décembre 2003
- Alain Dussauchoy, Laboratoire PRISMA, Université Lyon 1, Un siècle de modèles de processus stochastiques appliques aux phénomènes boursiers et autres, Jeudi 18 décembre 2003
- Amedeo Napoli, Équipe Orpailleur, LORIA Nancy, Extraction de/et connaissances, Jeudi 8 Janvier 2004
- Michel Verleysen, Université catholique de Louvain, Engineering Faculty, DICE - Microelectronics laboratory Apprentissage par réseaux de neurones: Le problème des données en grande dimension, Jeudi 29 Janvier 2004
- Jean Pierre Barthélémy, ENST Bretagne, Groupe des Ecoles de Télécommunications, Classifications binaires : une introduction, Jeudi 5 février 2004
- Christine Guinot, Unité de Biométrie et Epidémiologie, C.E.R.I.E.S., Neuilly sur Seine, France, Statistique exploratoire multidimensionnelle : Application à la recherche d'une typologie de la peau humaine saine du visage, Jeudi 26 février 2004

2004-2005

- Sylvie Philipp-Foliguet, ETIS (Equipes Traitement des Images et du Signal), CNRS UMR 8051, ENSEA, Cergy-Pontoise, France, Recherche d'images dans des bases à partir de signatures visuelles, Jeudi 15 octobre 2004
- Dragan Gamberger, Rudjer Boskovic Institute, Division of Electronics, Laboratory for Information Systems, Zagreb, Croatie, Avoiding data overfitting in scientific discovery : Experiments in functional genomics, Jeudi 25 novembre 2004

- Georges Hébrail, LTCI-UMR 5141 CNRS, Département Informatique et Réseaux, ENST Paris, Transformation de longues séries temporelles en descriptions symboliques, Jeudi 13 janvier 2005
- Gilles Venturini, Laboratoire d'Informatique, Université François Rabelais, Tours, Un survol des algorithmes biomimétiques pour la classification, Jeudi 27 janvier 05
- Christian Derquenne, R&D EDF, Clamart, France, Méthodes de fusion mises en oeuvre dans le cadre de l'enrichissement de base de données clientèle EDF , Jeudi 03 février 2005
- Lorenza Saitta, Università del Piemonte Orientale Amedeo Avogadro Dipartimento di Informatica, Complexity of Learning and Phase Transitions, Jeudi 10 février 2005

2005-2006

- Jean-Marc Petit, Laboratoire LIRIS, INSA Lyon, Recherche adaptative de bordures, Jeudi 9 février 06
- Marc Sebban, Laboratoire EURISE, Faculté des Sciences, Université de Saint-Etienne, Apprentissage non biaisé d'une distance d'édition stochastique sous la forme d'un transducteur déterministe, Jeudi 17 novembre 2005
- Jean-Michel POGGI, Laboratoire de Mathématiques, Equipe de Probabilités, Statistique et Modélisation, Université Paris-Sud Orsay, Détection de Données Aberrantes par Boosting, Jeudi 9 mars 2006

2006-2007

- Alain Morineau, Directeur de la Revue MODULAD, Préhistoire, histoire et perspectives du DM – le point-de-vue d'un statisticien, Jeudi 5 octobre 2006
- Grégoire de Lassence, Consultant Expert Académique SAS Institute, Exemples d'application de Data Mining et retour d'expérience, Jeudi 12 octobre 2006
- Michel Tenenhaus, HEC School of Management (GRECHEC), Approche PLS et analyse de tableaux multiples, Jeudi 6 octobre 2006
- Abdelaziz Faraj, Ingénieur de recherche, Institut Français du Pétrole, Sélection de modèle en régression PLS, Jeudi 9 novembre 2006
- Marc Boullé, France Telecom R&D, Spécificités du Data Mining dans les Télécoms, Jeudi 16 novembre 2006
- Abderrafih Lehmam, PERTINENCE MINING Sarl, Solution de text Mining et Linguistique, Jeudi 23 novembre 2006

- Gilbert Saporta, CEDRIC, CNAM, Paris, Classification supervisée et credit scoring, Jeudi 7 décembre 2006
- Malick Paye, Biomathématicien, bioMérieux, Grenoble, Using Data Mining for Biomarker Identification, Jeudi 14 décembre 2006
- Francois Wahl, Institut Francais du Petrole, Analyse d'incertitude et de sensibilité des modèles, Jeudi 11 janvier 2007
- Christophe Roche, ERT Condillac, LISTIC, Université de Savoie, Introduction aux problématiques des ontologies : état et perspectives en recherche et en applications, Jeudi 18 janvier 2007
- Serge Muller, Ingénieur Principal General Electric, Healthcare, Technologies Applications avancées en mammographie numérique, Jeudi 1er février 2007
- Roland Marion-Gallois, Expert Consultant, Statelis, La biostatistique dans les essais cliniques, Jeudi 8 février 2007
- Alexandre Aussem, Laboratoire PRISMA, Lyon 1, Apprentissage sous contraintes de la structure des réseaux bayésiens : Applications au cancer du Nasopharynx, Jeudi 15 février 2007
- Attilio Giordana, Università del Piemonte Orientale, Dipartimento di Informatica, Modeling Complex events by means of Structured Hidden Markov Models, Jeudi 8 mars 2007
- Jean-Gabriel Ganascia, LIP6 - Université Pierre et Marie Curie (Paris VI), Apprentissage non supervisé sur des données très partiellement décrites, Jeudi 15 mars 2007
- Bertrand Chabbat, CNAF-CNEDI Lyon, L'entreprise informationnelle - Exemple : la Branche Famille de la Sécurité Sociale et les documents réglementaires, Jeudi 5 avril 2007
- Yves Lechevallier, INRIA Paris – Rocquencourt, Autour des données d'intervalles, Jeudi 26 avril 2007

2007-2008

- Jean Riondet, Directeur de l'Institut International de Formation des Cadres de Santé, IFSCS, HCL, La statistique administrative et les questionnements sociaux, de Vauban à l'INSEE, Jeudi 13 décembre 2007
- Pablo Jensen, Laboratoire IXXI, ENS Lyon, Analyser la répartition des commerces en ville, 20 décembre 2007

- Gilles Bisson, Laboratoire TIMC-IMAG, Equipe Apprentissage Modèle et Algorithmes, Grenoble, Clustering d'objets structurés, application au traitement des molécules et à celui des données de criblage haut débit, Jeudi 10 janvier 2007
- Christian Derquenne, EDF R&D, Clamart, Méthodes de fusion mises en oeuvre dans le cadre de l'enrichissement de base de données clientèle EDF, Jeudi 17 janvier 2007
- Stefan Trausan-Matu, Equipe RACAI, "POLITEHNICA" University of Bucharest, Extraction de connaissances à partir de conversation chat, Jeudi 23 janvier 2007
-

c. Séminaires du laboratoire ERIC

2003-2004

- Pierre-Alain LAUR, Recherche de structures typiques au sein d'une collection de données semi-structurées ; 13/06/2004
- Djamel Zighed, Arbogodaï : Decision tree with optimal joint partitioning, 22/03/2004
- Dan J. Smith; Construction of domain-specific digital libraries, 26/01/2004
- Ricco Rakotomalala, TANAGRA : un logiciel libre pour l'enseignement et la recherche, 15/12/2003
- Florent Masseglia, Fouille de données : algorithmes et applications pour l'extraction de motifs séquentiels, 01/12/2003
- Kamel Aouiche, Utilisation des Index Bitmap pour la Fouille de Données, 17/11/2003
- Amandine Duffoux, Fouille de données à partir de la structure de documents XML, 17/11/2003
- Pierre Gançarski, L'approche multi-stratégies pour la classification non supervisée; la sélection automatique non-supervisée d'attributs pour la classification d'objets hétérogènes, 27/10/2003 à 10h00
- Didier PUZENAT, Visualisation et fouille de données, 13/10/2003

2004-2005

- Jerzy Korczak, Fouille interactive de séquences d'images IRMf, 30/05/2005
- Brice Effantin, Extraction de communautés dans le graphe du Web, 14/03/2005
- Chantal Reynaud, Comprendre le Web sémantique, 07/03/2005
- Nicole Vincent, La loi de Zipf en analyse d'images, 14/02/2005

- Sébatien Lefèvre, Introduction à la Morphologie Mathématique : principaux outils et applications 31/01/2005
- Karine Zeitouni, Entreposage et fouille de données spatiales et spatio-temporelles, 29/11/2004
- Zdenko Sonicki, Intelligent Data Analysis and Data Mining – Application in Medicine, 29/11/2004
- Edwige Fangseu Badjio, Qualité des IHM pour la fouille visuelle des données, 27/09/2004

2005-2006

- Michel Simonet, Ontologies, bases de connaissances et bases de données, 10/04/2006
- Ricco Rakotomalala, Les logiciels gratuits de DATA MINING pour l'enseignement, 12/12/2005
- Kamel Aouiche, Techniques de fouille de données pour l'optimisation automatique de performance des entrepôts de données, 28/11/2005
- Silvia Biffignandi, Shift-Share Analysis, 17/10/2005

2006-2007

- Miriam Alvariez, Plans d'expériences pour un modèle de simulation, 18/06/2007
- Rokia Missaoui, Opérateurs algébriques pour la manipulation des treillis de concepts, 23/04/2007
- Anne-Muriel Arigon, Développements d'applications pour l'identification de séquences génomiques, 12/03/2007
- Henri-Maxime Suchier, Nouvelles contributions du boosting en apprentissage automatique, 12/02/2007
- Frédéric Château, Inférence pour la Statistique Structurale, 15/01/2007
- Omar Boussaid, Evolution de l'entreposage des données complexes, 27/11/2006
- Yvan Bédard, Complexité des données géospatiales et peuplement de cubes de données : problématique, besoins et solutions, 27/11/2006
- Riadh Ben Messaoud, Couplage de l'analyse en ligne et de la fouille de données pour l'exploration, l'agrégation et l'explication des données complexes, 24/11/2006
- Jérôme Darmont, Optimisation et évaluation de performance pour l'aide à la conception et à l'administration des entrepôts de données complexes, 20/11/2006
- Djamel Zighed, Variation autour des mesures d'entropie, 16/10/2006

IV. PROJETS DE RECHERCHE APPLIQUEE

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

Identification des partenaires	Région Rhône-Alpes Formasup Rhône-Alpes (IPRA) Rectorats de l'académie 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 (depuis 2004)

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

Identification des partenaires	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 développer 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 des partenaires	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 des partenaires	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 des partenaires	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 : 43325 €

DataMining pour la recherche pharmaceutique

Identification des partenaires	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 : - 3700 € en fonctionnement - bourse de thèse de 30444 € pour 3 ans (Hakim HACID)

Interdépendance des marchés immobiliers résidentiels (INTERREG)

Identification des partenaires	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 cartographies 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 législation du travail.
Durée et financement	Durée : 2005-2007 Financement par le BIT : 12000 €

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

Identification des partenaires	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 €

Analyse automatique des cours de bourse (Tradingbots)

Identification des partenaires	Nicolas Macherey porteur d'un projet de création d'entreprise accueilli au sein de l'incubateur CREALYS

Objectifs recherchés	Conception et développement d'un système financier qui permet d'analyser les cours de bourse ou de change afin de prendre des décisions automatiques. Mise en place d'un progiciel.
Durée et financement	Durée : 2007-2008 Financement par la Région Rhône-Alpes : 29000 €

Génération de règles d'association

Identification des partenaires	SPAD
Objectifs recherchés	Implémentation d'un module de création de règles d'association dans la dernière version 7.0 du logiciel.
Durée et financement	Durée : 2005 Financement par SPAD : 4000 €

PMSI Privé

Identification des partenaires	UMR LIRIS, Université Claude Bernard Lyon 1 EA PRISMA, INSA de Lyon et Université Claude Bernard Lyon 1
Objectifs recherchés	Méthodologie d'analyse à visée décisionnelle des grandes bases de données médico-économiques : le PMSI privé
Durée et financement	Durée : 2005 Financement par la Région Rhône-Alpes : 25000 €

Gestion et de visualisation interactive de règles d'association

Identification des partenaires	DEENOV
Objectifs recherchés	Réalisation d'un module dans un logiciel de data mining : gestion et de visualisation interactive de règles d'association
Durée et financement	Durée : 2006-2007 Financement par DEENOV : 8000 €

Etudes marketing

Identification des partenaires	DATAXPRESSO
Objectifs recherchés	Conseil et expertise pour des études dans le domaine du marketing
Durée et financement	Durée : 2005-2007 Financement par DATAXPRESSO : 25000 €

Modélisation du processus de fabrication du vaccin de la coqueluche acellulaire

Identification des partenaires	SANOPHI-PASTEUR
Objectifs recherchés	Méthodologie pour formaliser la connaissance des procédés afin de définir les propriétés requises pour des productions optimales. Utilisation des méthodes de Data Mining et d'Ingénierie des Connaissances pour analyser le processus de fermentation de la production du vaccin « coqueluche acellulaire » ; pour construire un modèle de contrôle du processus ; pour expliquer les dérives observées sur les durées de culture industrielle et pour tenter de maîtriser ce facteur important dans la production des vaccins.
Durée et financement	Durée : 2007 Financement par SANOPHI-PASTEUR : 6000 €

V. 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 master 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 Laval à Québec, Canada

Identification du partenaire	Laboratoire CRG (Centre de Recherche en Géomatique) Professeur Yvan Badard
Collaboration en recherche	Entrepôts actifs de données spatiales.

Université du Québec en Outaouais, Canada

Identification du partenaire	Laboratoire LARIM Professeur Rokia Missaoui
Collaboration en recherche	Notre collaboration sur le couplage OLAP - Data Mining lors de séjours en France ou au Canada a déjà donné lieu à des publications communes.

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 master recherche 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. 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.

Université de Zagreb, Croatie

Identification du partenaire	Professeur Bojana DALBELO BASIC. Avec l'aide du Ministère des Affaires Etrangères (programme EGIDE depuis 2004)
Collaboration en recherche	Méthodes de fouille des données médicales Organisation conjointe de « International Workshop on Intelligent Data Analysis and Data Mining, Application in Medicine » depuis plusieurs années. Séjours en France pour fouiller de nouvelles données épidémiologiques et confronter les méthodes.

Université de Ljubljana, Slovénie

Identification du partenaire	Professeur Blaz ZUPAN Avec l'aide du Ministère des Affaires Etrangères (programme EGIDE depuis 2004)
Collaboration en recherche	Méthodes de fouille des données médicales Organisation conjointe de « International Workshop on Intelligent Data Analysis and Data Mining, Application in Medicine » depuis plusieurs années. Séjours en France pour fouiller de nouvelles données épidémiologiques et confronter les méthodes.

Université nationale d'Economie de Kharkov, Ukraine

Identification du partenaire	Professeurs Olexandr PUSKAR et Irina ZOLOTORIEVA
Collaboration en enseignement	Mise en place d'un master franco-ukrainien. Financement par le Ministère des Affaires Etrangères depuis 2005-2006.

Université d'Alessandria, Italie

Identification du partenaire	Professeur Lorenza Saitta
Collaboration en enseignement	Le Professeur Lorenza Saitta intervient depuis plusieurs années en tant que professeur invité dans un cours en master recherche ECD. Partenaire dans le projet de master européen Erasmus Mundus
Collaboration en recherche	Co-encadrement de la thèse de Julien Charbel

Université Polytechnique de Barcelone, Espagne

Identification du partenaire	Professeur Tomas Aluja
Collaboration en enseignement	Le Professeur Tomas Aluja intervient en 2007-2008 en tant que professeur invité dans un cours en master recherche ECD. Partenaire dans le projet de master européen Erasmus Mundus

Ecole Polytechnique de Bucarest, Roumanie

Identification du partenaire	Professeurs Eugenia Kalisz et Stefan Trausan
Collaboration en enseignement	Partenaire dans le projet de master européen Erasmus Mundus