

# TATOO

Extraction de connaissances dans les bases de données : motifs séquentiels et ontologies

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier

Département Informatique

Responsable :

Pascal Poncelet  
poncelet@lirmm.fr  
tel: +33 4 67 41 86 53

EXTRACTION ET GESTION DE CONNAISSANCES, FOUILLE DE DONNÉES, MOTIFS SÉQUENTIELS, ENTREPÔTS DE DONNÉES, LOGIQUE FLOUE, ONTOLOGIES, ANNOTATION AUTOMATIQUE, BASES DE DONNÉES

## ● Présentation

Ce groupe de recherche vise à regrouper les compétences autour de l'extraction de connaissances et plus particulièrement sur la recherche de motifs fréquents, sur la découverte d'ontologies et enfin sur l'utilisation des méta-données associées aux ontologies comme support de qualité et de validité des motifs extraits. Les domaines d'application privilégiés sont le web sémantique, l'environnement et la santé.

## ● Composition de l'équipe (Mai 2009)

Permanents	• Bringay Sandra	01/09/2007	MCF UM3
	• Laurent Anne	01/09/2003	MCF UM2
	• Hérin Danièle	01/01/1995	PR UM2
	• Pompidor Pierre	01/01/1995	MCF UM2
	• Joab Michelle	15/10/2000	PR UM2
	• Poncelet Pascal	01/09/1999	PR UM2
• Sala Michel	10/04/2000	MCF UM1	
Doctorants	• Ayouni Sarah	23/02/2009 - 30/06/2012	
	• Di Jorio Lisa	01/10/2007 - 30/09/2010	
	• Rabatel Julien	14/01/2008 - 30/09/2011	
	• Low-Kam Cécile	01/10/2007 - 30/09/2010	
	• Saneifar Hassan	14/01/2008 - 30/09/2011	
	• Salle Paola	01/10/2007 - 30/09/2010	
	• Pitarch Yoann	12/06/2008 - 31/12/2011	
	• Li Haoyuan	01/10/2006 - 31/10/2009	

Associés	• Golbreich Christine	09/06/2008 - 09/06/2010 PU Univ. Versailles St Quentin
	• Melançon Guy	15/10/2000 - 31/12/2009 PU1 INRIA Bordeaux
	• Teisseire Maguelonne	01/09/1995 - 01/01/2011 DR2 CEMAGREF
Post-Docs	• Nin Guerrero Jordi	12/01/2009 - 30/04/2009
	• Chakkour Feirouz	01/11/2008 - 30/02/2009

## ● Publications (du 01/01/2005 au 31/08/2009)

	ACL	ACLN	ASCL	INV	ACTI	ACTN	COM	AFF	OS	OV	DO	AP	BV	Prix	Orga
2005-2009	27				69	31	2	4	11		4	17			3

## ● Publications (suite)

### *Thèses et HDR*

- 5 soutenances de thèse et 2 soutenances d'Habilitation à Diriger les Recherches
- Organisation de Conférences, workshops, challenges et sessions dans des conférences*
- 5ièmes Journées Entrepôts de Données et Analyses en lignes (EDA 2009), Montpellier, juin 2009
- Session « Decision and Health », International Symposium on Intelligent Decision Technologies (IDT'2009), Himeji, Japon, avril 2009
- Session « Fuzzy is Scalable: Managing Huge Databases Using Fuzzy Methods » (ACM/IEEE Int. Conference on Soft Computing as Transdisciplinary Science and Technology - CSTST'08), octobre 2008.
- Session « Mining Multidimensional Data », European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 08), Antwerp, Belgique, septembre 2008
- Challenge international « Discovery Challenge Data Mining and Classification Context », European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 07), Warsaw, Pologne, septembre 2007
- Session « Management of voluminous and Complex data: Data Warehouses, OLAP, Data Integration, Complex Data, Data Mining » du congrès IPMU, Paris, juillet 2006
- Workshop « Fouille de données temporelles » dans le cadre des Conférences Extraction et Gestion de Connaissances (EGC 2007, EGC 2008, EGC 2009).
- Workshop « Fouille de données d'opinion (FODOP) » Congrès INFORSID 08.

### *Articles récompensés*

> Meilleur article :

- 1 article pour la 9th Industrial Conference on Data Mining (ICDM 2008), Leipzig, Allemagne, December 2008
- > Articles sélectionnés comme meilleurs articles de la conférence
- 1 article pour European Conference on Machine Learning and Principles and

### *Practice of Knowledge Discovery in Databases (ECML/PKDD 08)*

- 4 articles pour les 9ièmes Journées Extraction et Gestion des Connaissances EGC 2009, Strasbourg, France
- 3 articles pour les 8ièmes Journées Extraction et Gestion des Connaissances EGC 2008, Sophia Antipolis, France, 2008
- 2 articles pour les 7ièmes Journées Extraction et Gestion des Connaissances EGC 2007, Namur, Belgique, 2007

### *Editions de Livres*

- « Scalable Fuzzy Algorithms for Data Management and Analysis: Methods and Design », IDEA Group Publisher, 2009.
- « Data Mining Patterns: New Methods and Applications », IDEA Group Publisher, August 2007, 307 page, ISBN-13 978-1599041629.
- « Successes and New Directions in Data Mining », IDEA Group Publisher, November 2007, 369 pages, ISBN-13 978-1599046457.

### *Editions de numéro spéciaux*

- Numéro spécial « Mining Spatio-Temporal Data », Journal of Intelligent Information Systems (JIIS), Kluwer Academic Publishers, 2006
- Numéro spécial « Algorithmes pour la découverte de motifs dans les bases de données ». Revue I3, Cépadués Edition, 2007.
- Numéro spécial « Fouille de Données d'Opinions », Revue RNTI, Cépadués, à paraître en 2009.

### *Comités de programmes*

- Les membres de l'équipe TATOO participent régulièrement en tant que membres de comités de programme ou de relectures aux revues et conférences et prestigieuses du domaine (KDD, ICDM, PKDD, PAKDD, VLDB, IEEE TKDE, DMKD, DKE, IDA, DEXA, CIKM, IDEAS, SAC, CLA, FQAS...). En outre, TATOO participe la standardisation de OWL2, Web Ontology Language, en tant que membre du OWL W3C WG.

## ● Activités administratives significatives

- D. Héryn, Présidente Université Montpellier 2 (juin 2008)
- Vice-présidence du conseil scientifique de l'Université de Montpellier 2
- Direction du département Informatique et Gestion de Polytech' Montpellier
- Chargé de mission pour la FOAD (Formation Ouverte A Distance) de l'Université de Montpellier 1
- Membre du conseil de direction du pôle MIPS
- Responsabilité de la spécialité « Informatique Professionnelle » du Master IMS (Informatique, Mathématique et Statistique) de l'Université Montpellier 2
- Chargé de mission pour la mise en place du partenariat Université Montpellier 2/IBM Montpellier (Co-laboratoire, partenariats académiques)
- Membre nommé CNU 27
- Présidence du Conseil d'Orientation Scientifique du département « Informatique, Multimédia, TIC » de Transfert-LR, structure de transfert régional du Languedoc-Roussillon inscrite dans le Contrat Plan Etat Région
- Animation d'un pôle régional INTS, Intelligence Numérique et Technologies Sensibles regroupant scientifiques et industriels de la région Languedoc-Roussillon autour des thèmes de l'intelligence économique et de la sécurité globale des biens et des personnes

## ● Coopérations internationales

Philipps-Universität Marburg (Pr Eyke Hullermeir), Eindhoven Technical University (Dr Toon Calders), Simon Fraser University (Pr Jian Pei), Carleton University, Ottawa (Dr Babak Esfandiari), Institut Teknologi Bandung, Indonesia, Dept. Informatics Engineering (co-encadrement de thèse), Tunisia (co-encadrement de thèse), Aalborg University, Denmark (chercheur invité T.B. Pedersen), Pakistan, Malaisie, Indonésie (Projet STIC-ASIA EXPEDO), Prince of Songkla University, Thailand, Institut HEM, Maroc, Institut Fatronik Espagne (coencadrement de thèse)

## ● Coopérations nationales

Université Cergy-Pontoise, Université de Tours, Université d'Orsay (Projet STIC-ASIA EXPEDO), UPMC, LGI2PEMA, INRIA Sophia Antipolis, LIRIS, Université Antilles Guyane (CEREGMIA), EDF R&D, France Telecom

## ● Contrats / Transferts et valorisation

- ANR MIDAS (2008-2011) « Etude, développement et démonstrateur de nouvelles méthodes de résumés de flux de données » avec l'ENST, CEREGMIA, INRIA, EDF R&D et FRANCE TELECOM R&D
- ANR RIAM « Ubiquitus » (2007-2008) « découverte, l'agrégation personnalisée, l'organisation ainsi que l'accès multi-terminaux, multi-formes et multi-usages à des services texte&audio » avec Netia et Nexwave Solutions
- Contrats de collaboration avec des sociétés nationales : France Telecom sur la visualisation de réseaux complexes (2006-2007) ; EDF Recherche&Développement sur la recherche de comportements atypiques dans des données multidimensionnelles (2006-2008) ; Groupe Européen d'Intérêt Economique SEMIDE (Système Euro-Méditerranéen d'Information sur les savoir-faire dans le Domaine de l'Eau (<http://www.semide.org>)) sur la construction et la propagation des annotations (2005-2007)
- Contrats régionaux : Projet Cartographie avec Expernova (2008-2010). « Cartographie de compétences dans le domaine scientifique ». Projet TextLog avec Satin (2009-2012) « Application de techniques de fouilles de textes pour analyser des logs ». Projet Capteurs avec Fatronik France (2009-2012) « Prévision de maintenance par analyse de comportement de capteurs » ; Projet Fraude avec Axiliance (2006-2007) « Détection de fraudes sur un réseau »
- Transferts de technologie régionaux avec des sociétés dans le cadre du Languedoc Roussillon Incubation : C6 sur la création et exploitation de réseaux sémantiques liés à des bases documentaires spécialisées; KEOSIA (2006) sur l'accompagnement de patients ; Airtist (2006) sur le ciblage des publicités pour des téléchargements de musique ; Phone Advance (2007) pour la détection de profils de consommation dans le domaine des cartes de fidélité dématérialisées ; Satin IP (2007) pour la mise en oeuvre d'une plate-forme logicielle de suivi de conception de blocs IP (projet incubé LRI, lauréat du concours OSEO émergence 2006 et du concours OSEO création&développement en 2007) ; Namae Concept (2008) sur la validation linguistiques en dépôt de noms (en collaboration avec l'équipe TAL) ; We Are Cloud (2008) sur l'analyse décisionnelle
- Projets Exploratoires Pluridisciplinaires (PEPS) : « GeneMining » (2008-2009) Apport des méthodes de recherche de motifs séquentiels pour exploiter des données issues des puces ADN - application à la maladie d'Alzheimer. Partenaire : MMDM (Université Montpellier 2) ; ST2I-SHS (2008-2009). « Langage, Mémoire et Alzheimer : une approche des maladies neurodégénératives fondées sur la densité des idées ». Partenaire : PRAXILING (UM3-CNRS)



## Extraction de connaissances dans les bases de données : motifs séquentiels et ontologies

### Résumé de la thématique

Motivées par des problèmes liés à l'informatique décisionnelle, les activités de recherche de l'équipe projet TATOO se focalisent sur des méthodes d'Extraction et de Représentation de Connaissances à partir de grandes bases de données. Notre objectif est d'étudier et de développer de nouvelles méthodes pour prendre en compte les besoins des nouvelles applications engendrés par les différents types de données. En outre, nous intégrons l'opérateur humain dans le processus d'extraction et proposons de l'aider et de l'outiller pour définir son champ de connaissance et qualifier les données à partir du champ des connaissances du domaine. Les travaux menés par le projet s'inscrivent dans la continuité des actions de recherche réalisées ces dernières années par les différents membres du projet et sont au cœur des préoccupations de la communauté nationale et internationale puisqu'ils concernent les différents axes suivants : (i) Fouille de données dans des bases de données complexes : données structurées, semi-structurées, multidimensionnelles, qualitatives et quantitatives, textuelles, entrepôts de données (données multidimensionnelles, agrégées, hiérarchisées), etc ; (ii) Fouille de données dans des bases de données dynamiques, i.e. dont les données sont exprimées sous la forme d'un flot de données continu, à grande vitesse ; (iii) Fouille de données approximatives et aide à la décision. Depuis ces dernières années, nous nous sommes intéressés à l'extraction de connaissances en utilisant à la fois des techniques d'apprentissage supervisées et non supervisées. Parmi ces dernières, nous avons, depuis 1996, une solide expérience dans

l'extraction de motifs séquentiels et, dans ce contexte, nos thématiques de recherche sont les suivantes :

- Algorithmes d'extraction de motifs, d'exceptions,
- Prise en compte de contraintes (par exemple temporelles pour affiner la connaissance extraite et offrir à l'utilisateur final la possibilité de considérer des contraintes du type : long terme, court terme),
- Approches incrémentales, temps réel, approximatives,
- Fouille de données d'entrepôts (résumés, règles d'association, motifs séquentiels, exceptions),
- Traitement des données multidimensionnelles, semi-structurées, arborescentes.

L'originalité des recherches menées se situe également dans la prise en compte de données incertaines et imprécises en utilisant par exemple la logique floue comme support théorique.

L'ensemble de ces travaux a donné lieu à des thèses, de nombreux articles dans des revues et conférences internationales reconnues, et à des projets de transfert de technologie avec des sociétés régionales ou nationales.

### Des données de plus en plus complexes

L'un des challenges des années 90 était de proposer de nouveaux algorithmes d'extraction de connaissances

qui soient capables d'offrir à l'utilisateur des motifs, des modèles, des règles, des résumés de ces données. Des approches très efficaces ont été proposées et ont notamment permis de surmonter les difficultés de passage à l'échelle, de réduction d'espace de recherche, de prise en compte de contraintes dans le processus.

De nouveaux systèmes d'extraction ont fait alors leur apparition et sont de plus en plus utilisés pour aider le décideur.

Il faut aujourd'hui poursuivre cet objectif en considérant la complexité des données liées au développement de nouvelles applications. Il faut par exemple être capable de répondre aux questions suivantes :

- comment extraire de la connaissance à partir d'un ensemble de schéma XML ?

- comment analyser des données de gènes d'expression alors que les données sont à l'opposé de ce que les approches traditionnelles ont l'habitude de traiter ?

- comment tirer profit de données complexes et complémentaires pour améliorer la connaissance acquise ?

De manière plus générale, il faut examiner s'il est possible d'adapter les approches traditionnelles à ces nouveaux types de données ou s'il est indispensable de reconsidérer le processus dans son intégralité.

Ces dernières années, nos travaux ont porté sur des données de plus en plus complexes. Nous en décrivons



quelques-uns ci dessous.

Nos recherches menées sur des données complexes modélisées sous la forme de schémas XML offrent de nouvelles possibilités de génération de schémas médiateurs qui, utilisés dans un contexte de Web Sémantique, peuvent alors être interrogés de manière transparente par les différents utilisateurs et les requêtes sont ensuite propagées sur les sites sources. La recherche de structures typiques récurrentes dans des bases de données d'arbres a été étudiée

sous l'angle de l'optimisation

(EUSFLAT 05, EGC 06) ou de

l'approximation (Ouvrage

Fuzzy Logic and the

Semantic Web 05, IFSA07,

Fuzzy Sets and Systems

Journal 09, IDA Journal

09) et s'inscrivait

également dans le cadre

de la thèse de F. Del Razo

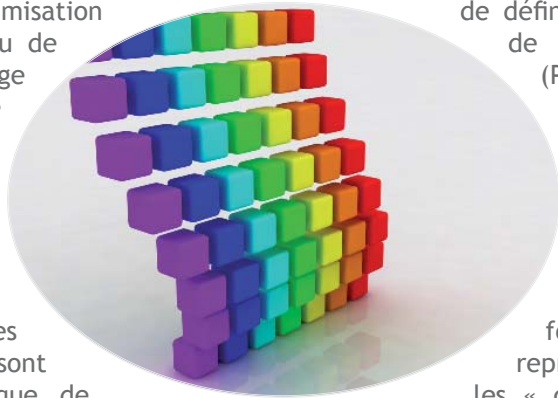
Lopez. Les données issues

de capteurs ou de logs sont

également un cas typique de

données complexes car les contenus

peuvent être très variables. Cette problématique est abordée dans le cadre de la thèse de J. Rabatel sur des données issues de capteurs localisés dans des trains afin d'anticiper les opérations de maintenance (ICDM 09) et de H. Saneifar pour analyser le contenu de fichier logs pour l'aide à la conception de semi-conducteurs (AusDM 2008, DEXA 2009). D'autres types de logs sont traités dans le cas du Web Usage Mining afin, par exemple, de rechercher les comportements courants des utilisateurs. Outre la définition d'algorithmes originaux pour extraire ces comportements, nous avons défini une approche qui offre également la possibilité d'extraire les périodes de temps les plus caractéristiques des usages (Revue Data Mining and Knowledge Discovery 07). De plus en plus d'entreprises, quelle que soit leur taille, sont maintenant munies d'un système d'archivage homogène de leurs données de production sous la forme



d'un entrepôt de données. Les données sont regroupées dans des cubes de données définis le long de plusieurs dimensions éventuellement avec des hiérarchies. Très souvent associés au concept de Business Intelligence (BI), ces outils permettent de donner aux décideurs un environnement de navigation et de décision adapté aux enjeux actuels. Cependant, les approches traditionnelles de fouille de données ne se focalisent que sur une seule dimension. Aussi, les travaux menés dans le cadre de la thèse de M. Plantevit ont permis de définir de nouvelles approches d'extraction

de motifs séquentiels multidimensionnels

(PKDD 05, DOLAP 06, DOLAP 07) et ont

notamment été appliquées dans le

cadre d'un projet de collaboration

avec EDF R&D. Enfin, quelque soit le

type de données, le constat que nous

avons pu faire ces dernières années,

notamment dans le cadre de nos projets

industriels, est que l'utilisateur n'est pas

forcément intéressé par les connaissances

représentant une majorité mais plutôt par

les « comportements surprenants ». Dans le

cadre de la thèse de Haoyuan Li, nous nous sommes

intéressés à l'extraction de motifs séquentiels qui

interviennent peu souvent mais qui sont surprenants par

rapport à la connaissance du domaine (ICDM 08, DEXA

08, IDA Journal 09). Par contre, les travaux menés dans

le cadre de la thèse de Cecile Low-Kam s'intéressent

à la découverte d'anomalies au sein de grandes bases

de séquences (ICMLA 08, EGC 09). La difficulté est de

différencier les comportements « différents » du bruit

possible au sein des données.

En s'intéressant à des bases de données textuelles, plusieurs propositions ont été réalisées. Tout d'abord, nous avons proposé une nouvelle approche d'indexation basée sur des co-occurrences, guidée par la structure des documents et contrôlée par une ontologie pour indexer de manière automatique de très grands corpus documentaires d'entreprises. De même, en partenariat

avec l'équipe TAL, l'originalité de nos propositions de classification ou de clustering est d'une part de permettre à des documents d'appartenir à différents clusters (parfois à l'aide d'approche floue) mais également d'améliorer la pertinence des classements en considérant l'ordre des mots. Ces techniques sont par exemple utilisées pour effectuer du filtrage par le contenu de documents et même pour faciliter la détection de fraudes sur les réseaux. De manière à résumer et à adapter le contenu des documents textuels, nous nous sommes également intéressées à de nouvelles approches de cartographie automatique de documents (résumé) et à la proposition de nouveaux documents adaptés aux comportements des utilisateurs (profiling). Ces dernières années nous avons étendu nos travaux à la détection d'opinions dans des données issues du Web (blogs, forums, ...). Nous proposons ainsi de nouvelles approches dont l'originalité est d'apprendre automatiquement les adjectifs pertinents pour un domaine d'application particulier (Dawak 08, Edition revue RNTI 09).

## L'opérateur humain de plus en plus impliqué

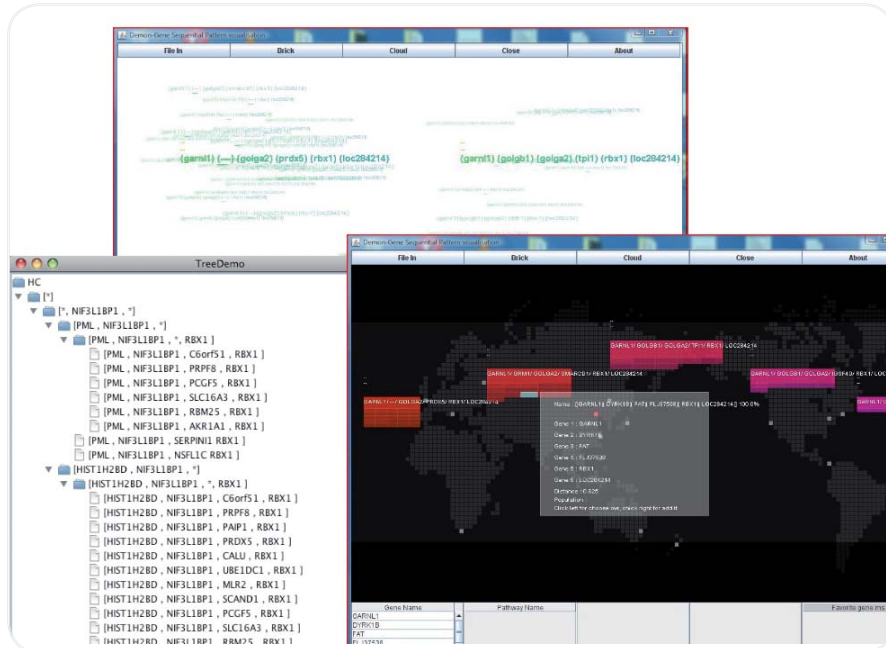
L'un des constats importants de ces dernières années est que pour obtenir des connaissances actionnables il est indispensable d'intégrer l'opérateur humain au centre du processus d'extraction. En effet, les outils existent mais ne correspondent pas forcément à ce que souhaite l'utilisateur. Il devient donc indispensable d'intégrer le plus tôt possible l'opérateur humain dans le processus d'extraction.

Pour répondre à cette problématique, nos travaux ont abordé différentes approches pour d'une part prendre en compte le plus rapidement possible les contraintes de l'utilisateur mais surtout pour intégrer ce dernier tout au long du processus. La prise en compte des contraintes des utilisateurs permet de pouvoir offrir de

nouvelles connaissances plus adaptées aux besoins de décideurs. Cependant, l'une des difficultés rencontrées dans les approches traditionnelles est que ces contraintes sont souvent trop rigides et donc difficiles à mettre en œuvre. Dans le cadre de la thèse de C. Fiot, en offrant plus de flexibilité dans la spécification des contraintes nous avons permis d'offrir non seulement des connaissances adaptés aux besoins des utilisateurs mais nous avons également apporté des solutions à la problématique des données manquantes (TIME 07, IEEE Transactions on Fuzzy Systems 07, FuzzIEEE 08, DASFAA 08). Dans le cadre des travaux sur les données liées à la santé comme les données associées au suivi de patient ou les données transcriptomiques issues de puces à ADN, nous avons pu constater que les connaissances extraites étaient difficilement utilisables car trop nombreuses. Ainsi, dans le cadre des thèses de P. Salle, nous avons proposé de nouveaux algorithmes afin d'intégrer plus tôt dans le processus les connaissances expertes. Ceci offre d'une part l'avantage de réduire l'espace de

recherche et donc les temps de traitements associés et, d'autre part, d'offrir des connaissances plus en adéquation par rapport aux attentes des biologistes et médecins partenaires (CSTST 08, FuzzIEEE'09, AIME 09). En outre, un outil de visualisation a été réalisé dans le cadre du PEPS ST2I "GeneMining" pour permettre de parcourir et valider les connaissances extraites (motifs séquentiels de gènes) :

L'utilisation d'ontologies offre la possibilité de se focaliser sur les connaissances importantes d'un domaine et donc d'extraire des connaissances utiles. Ainsi, les travaux menés dans le cadre de la thèse de L. Abrouk se sont intéressés à la définition et à la mise à jour de telles ontologies (BDA 07). Enfin les travaux menés dans la thèse de L. Di Jorio ont permis de montrer que l'utilisation de motifs séquentiels offrait une nouvelle approche prometteuse pour générer de manière automatique de nouvelles ontologies (ODBASE 08).



Ces dernières années nous avons également étudié d'autres types de contraintes liées notamment aux nouvelles normes européennes et à la loi HIPAA au Etats-Unis sur la protection de la vie privée. Dans ce cas, la contrainte imposée par les institutions, et donc au final par l'utilisateur, est de garantir que l'application d'approches d'extraction de connaissances ne permet pas d'avoir accès à ses données privées. Les travaux menés dans ce domaine ont permis de définir de nouvelles approches de préservation de la vie privée pour les motifs séquentiels et pour la détection de fraudes dans des réseaux (CIKM 06, BDA 06, Revue ISI 07, EGC 09).

## Des données disponibles de plus en plus rapidement

Les dernières avancées technologiques nous confrontent à un autre problème important : comment appréhender le fait que les données soient disponibles de plus en plus rapidement ? A l'heure actuelle, quel que soit l'algorithme de fouille utilisé, l'obtention des résultats peut être très long (plusieurs heures, jours) dès que l'on manipule de grandes bases. Si l'on considère par exemple des sites de e-commerce où plusieurs Gigaoctets de données peuvent arriver par heure, la connaissance que l'on va extraire n'est déjà plus représentative des données de la base. Le problème à résoudre dans ce cas est : sachant que les données ne peuvent plus être stockées, est-il possible d'extraire de la connaissance à la volée ? La conséquence immédiate est que les approches traditionnelles ne sont plus du tout adaptées et qu'il est indispensable de proposer de nouvelles techniques.

Ces techniques doivent, comme précédemment, permettre d'extraire rapidement la connaissance (comment gérer des données qui arrivent de manière continue, e.g. données de capteurs, données boursières, news, ...) mais également proposer à l'utilisateur une certaine flexibilité dans les connaissances extraites. Dans le cadre de ces travaux, nous nous sommes intéressés à la détection de motifs séquentiels dans des flots de données. Motivés par des systèmes aussi dynamiques que les systèmes P2P, nous avons proposé une nouvelle approche basée sur des algorithmes génétiques (EGC 06, AINA 06) pour extraire des comportements d'utilisateurs dans des systèmes P2P. Dans un cadre plus général d'extraction de motifs, nos derniers travaux se sont intéressés à de nouvelles approches basées sur des techniques d'échantillonnage basées sur l'optimisation du choix du rappel et de la précision (CIKM 05, MLDM 05, Revue Pattern Recognition 07, Revue IDA 06) ou de nouvelles techniques de représentation des

connaissances approximatives (Revue JIS 07, BDA 05, IEEE IS 06). Nos travaux réalisés dans le cadre de la thèse de C. Raïssi ont également permis d'aborder la notion de représentations condensées pour les flots (Revue Data Mining and Knowledge Discovery 08) et de nouvelles approches d'échantillonnages basées sur des réservoirs (ICDM 07). Nos travaux actuels, qui s'inscrivent dans le cadre de la thèse de Y. Pitarch (BDA 08, FINA 09) et du projet d'ANR MIDAS, s'intéressent à la possibilité de gérer des données multidimensionnelles disponibles sous la forme de flots et proposent de nouvelles approches de création de résumés tenant compte des hiérarchies disponibles.

## Perspectives

Si le stockage des données augmente exponentiellement dans toutes les industries, les domaines de l'Environnement et de la Santé ne font pas exception. Pour ce dernier, des milliers de milliards de dossiers médicaux sont stockés chaque année dans le monde entier. Malheureusement, cette précieuse source de données n'est pas ou peu exploitée et les challenges sont multiples : Comment faire émerger de la connaissance pertinente pour les décideurs en charge des établissements de soins (e.g. recherche de tendances, analyse multi-dimensionnelle...) ? Comment soutenir la pratique médicale quotidienne des professionnels de santé dans leurs tâches de suivi (e.g. détection d'anomalies), de diagnostic (e.g. recherche de profils similaires) ? Comment soutenir les activités de recherche des praticiens en leur permettant d'avoir une vue par spécialité sur les données de santé (e.g. traitement de données semi structurées).

Dans le domaine de l'Environnement, la problématique est similaire et nécessite l'analyse de très grandes masses de données complexes et hétérogènes. Il suffit de citer la gestion des ressources en eau, le maintien de la biodiversité ou les changements climatiques... Dans les années à venir, nous nous intéresserons plus

particulièrement à la recherche et au développement de nouvelles méthodes d'extraction de connaissances avec pour objectif des applications innovantes dans le domaine de la Santé et de l'Environnement.

Ces travaux seront menés tant d'un point de vue fondamental que du point de vue de ces applications, et seront en particulier menés dans le contexte du partenariat fort en cours de création avec IBM Montpellier. Dans ce contexte, nous nous attacherons à étudier la fouille de données utilisant les paradigmes du calcul haute performance (HPC) ainsi que les nouvelles méthodes innovantes de Business Intelligence, en particulier pour la définition d'outils d'entrepôts et de fouille de données temps réel.

## Quelques publications significatives

C. Fiot, F. Masseglia, A. Laurent and M. Teisseire «Evolution Patterns and Gradual Trends» In International Journal of Intelligent Systems. 2010.

D. H. Li, A. Laurent, P. Poncelet and M. Roche. «On Unexpected Phrases in Text Documents ». Intelligent Data Analysis (IDA) Journal, Vol. 14, N. 1, 2010.

M. Plantevit, Y.W Choong, A. Laurent, D. Laurent and M. Teisseire. «Mining Multidimensional and Multiple-Level Sequential Patterns». A paraître dans ACM Transaction on Knowledge Discovery from Data (ACM TKDD), 2009

M. Teisseire, S. Bringay et A. Laurent. Actes des 5èmes Journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA'09). Numéro spécial de la Revue des nouvelles technologies de l'information (RNTI). Editions Cépaduès. 2009.

D. H. Li, A. Laurent and P. Poncelet. « WebUser: Mining Unexpected Web Usage». International Journal of

Business Intelligence and Data Mining, 2009.  
F. De Razo Lopez, A. Laurent, P. Poncelet and M. Teisseire. «FTMnodes: Fuzzy Tree Mining Based on Partial Inclusion». Fuzzy Sets and Systems Journal, 2009.

M. Plantevit, A. Laurent et M. Teisseire. «OLAP-Sequential Mining : Summarizing Trends from Historical Multidimensional Data using Closed Multidimensional Sequential Patterns». Annals of Information Systems, special issue in New Trends in Data Warehousing and Data Analysis, 2009.

H. Saneifar, S. Bonniol, A. Laurent, P. Poncelet, M. Roche. «Terminology Extraction from Log Files». In Proceedings of the 20th International Conference on Database and Expert Systems Applications (DEXA 2009), September 2009, Linz, Austria

G. Singh, F. Masseglia, C. Fiot, A. Marascu and P. Poncelet. «Data Mining for Intrusion Detection : from Outliers to True Intrusions». In Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2009), April 2009, Bangkok, Thailand.

C. Raïssi, T. Calders and P. Poncelet. «Mining Conjunctive Sequential Patterns». Data Mining and Knowledge Discovery, Springer Verlag (Selected paper from ECML/PKDD 08), Vol. 17, N. 1, August 2008.

F. Masseglia, P. Poncelet and M. Teisseire. «Efficient Mining of Sequential Patterns with Time Constraints: Reducing the Combinations», Expert Systems With Applications, Vol. 40, N. 3, 2008.

F. Masseglia, P. Poncelet, M. Teisseire and A. Marascu. «Web Usage Mining: Extracting Unexpected Periods from Web Logs». Data Mining and Knowledge Discovery, Springer Verlag, Vol. 16, N. 1, February 2008, pp. 39-65.

C. Fiot, A. Laurent and M. Teisseire. «Fuzzy sequential pattern mining in Incomplete Databases». Mathware

---

and *Soft Computing Journal*, Vol 15, No 1, pp. , 41-59, 2008.

P.A. Laur, J.E. Symphor, R. Nock and P. Poncelet. «Mining Evolving Data Streams for Frequent Patterns». *Patterns Recognition*, Elsevier, Vol. 40, N. 2, 2007, pp. 492-503.

C. Raïssi, P. Poncelet and M. Teisseire. «Towards a New Approach for Mining Maximal Frequent Itemsets over Data Stream». *Journal of Intelligent Information Systems*, Springer, Vol. 28, N. 1, 2007, pp. 23-36.

C. Raïssi and P. Poncelet. «Sampling for Sequential Pattern Mining: From Static Databases to Data Streams». *Proceedings of the IEEE International Conference on Data Mining (ICDM 07)*, Omaha NB, USA, October 2007.

F. Masegla, P. Poncelet and M. Teisseire (Editors). «Successes and New Directions in Data Mining», IDEA Group Publisher, November 2007, ISBN-13 978-1599046457.

P. Poncelet, F. Masegla and M. Teisseire (Editors). «Data Mining Patterns: New Methods and Applications», IDEA Group Publisher, August 2007, ISBN-13 978-1599041629.