Proposition d'ARC 2007-2008 :

SéSur : Sécurité et Surveillance dans les flots de données

Équipe AxIS	Équipe DREAM	Équipe KDD	Équipe TATOO
INRIA	IRISA	LGI2P/EMA	LIRMM
Sophia Antipolis	Rennes	Nîmes	Montpellier

Résumé

L'objectif de l'ARC SéSur est de réunir les compétences (conceptuelles et expérimentales) indispensables à l'étude de solutions efficaces pour sécuriser, surveiller et diagnostiquer les systèmes producteurs de données connues sous le nom de "**flots de données**" ou "**data streams**". Ces données présentent deux caractéristiques majeures : 1) elles sont les signes vitaux du système considéré et leur analyse est dans la plupart des cas une nécessité première et 2) elles sont produites à une vitesse et dans des quantités telles que la technologie actuelle ne permet pas de les traiter de façon satisfaisante.
Nous nous intéressons particulièrement à la surveillance de systèmes produisant des flots de données. La surveillance consiste dans ce cas à détecter dans ce flot de données des motifs caractéristiques du bon ou du mauvais fonctionnement du système. Jusqu'à présent de telles méthodes n'utilisaient que des ensembles de motifs fixés ou appris au préalable. Toutefois le cadre général des flots de données impose d'extraire et de détecter simultanément et "à la volée" les motifs synoymes de dysfonctionnement. Cette détection peut alors être le résultat de méthodes supervisées ou non supervisées.
Nous avons donc identifié deux thèmes à explorer dans le cadre de l'ARC SéSur :

1. Définir de nouvelles techniques d'extraction de connaissances capables de prendre en compte les caractéristiques nouvelles de ces données. Cela demande de revisiter l'ensemble des techniques existantes. Dans cette ARC, notre objectif sera d'étudier l'extraction de deux types de motifs temporels dans les flots de données : les motifs séquentiels et les chroniques qui sont des motifs contenant des événements non séquentiels mais reliés par des contraintes.

2. Proposer des solutions pour assurer la gestion et la maintenance de ces connaissances au fil du temps. En effet, compte tenu de la nature extrêmement dynamique des flots de données il est évident que l'évolution des connaissances extraites sera un défi pour ce projet. Notre objectif dans l'ARC SéSur est de proposer des méthodes de gestion des connaissances acquises sur le flot, en se basant sur des travaux préliminaires existants dans ce domaine chez les équipes concernées.

La pertinence et l'utilité de ces connaissances à des fins de sécurité, de surveillance et de diagnostic dans les systèmes ciblés sera également privilégiée. Dans ce but, les différents contextes applicatifs des partenaires impliqués dans cette proposition seront étudiés. Nous garderons donc à l'esprit des objectifs réalistes en terme d'applications potentielles, par la validation de nos propositions sur ces données réelles.
La recherche sur le thème général des flots de données se développe de manière importante dans le monde. En rassemblant des équipes actives dans le domaine, à l'INRIA et à proximité de l'INRIA, nous espérons obtenir des avancées dans des "niches" peu explorées pour l'instant, à savoir l'extraction de motifs temporels (séquentiels et non séquentiels) ainsi que l'évolution des motifs au cours du temps.

1 Contexte et motivations

La production rapide et massive de données sous la forme de flots s'est récemment révélée être une source de sujets de recherche majeurs. Ce phénomène de production, qui concerne un nombre grandissant de sources de données, se retrouve sous le nom de ń flots de données ż (ou ń data streams ż). Les flots de données peuvent être issus des données d'opérateurs téléphoniques, de la surveillance de patients dans les hôpitaux, de réseaux de capteurs (par exemple de consommation d'énergie électrique), des journaux d'usage de certains sites Web très fréquentés, du trafic IP, des transactions financières, des enchères en ligne, de procédés industriels ou encore du trafic routier urbain, etc.
Les flots de données nous confrontent à deux défis principaux :

1. Comment représenter un flot de manière fidèle sans le stocker exhaustivement ?
2. Comment extraire de la connaissance depuis un flot, sans le bloquer par des opérations coûteuses ?

Les deux questions sont fortement liées. D'un côté, les connaissances (motifs) extraites à partir d'un flot peuvent être une base de travail pour résumer ou représenter ce flot. D'un autre côté, le résumé d'un flot peut être utile pour y extraire de la connaissance de façon approximative mais fiable. Les méthodes et algorithmes traditionnels de gestion et de fouille des données statiques ne peuvent pas être appliquées directement sur les flots de données et de nouveaux paradigmes doivent être apportés. L'idée principale est qu'un flot ne peut pas être stocké mais doit être traité ń à la volée ż, que ce soit pour répondre à des requêtes ou pour exécuter une opération de fouille. Le traitement ń à la volée ż implique de mettre en place et de maintenir des résumés des données qui sont passées dans le flot. Ces résumés sont une représentation approximative du flot de données, qui permet d'approcher les résultats que l'on obtiendrait avec une requête ou un processus de fouille ń classiques ż .

1.1 La fouille de flots de données : de nouveaux défis

Dans les applications traditionnelles, le processus de fouille était prévu pour fonctionner sur des données stockées et statiques ou peu mises à jour. L'extraction de connaissances pouvait alors prendre des jours, des semaines ou même des mois, mais la nature statique des données ne perturbait pas le déroulement de l'extraction. Dans le cas des flots, on n'observe les données qu'une seule fois. Les caractéristiques des flots de données sont particulièrement contraignantes : données produites en continu et à une très grande vitesse, impossibilité de stocker les données et nécessité de les incorporer au modèle, même si elles ne seront plus rencontrées [10]. En raison du grand nombre d'applications concernées, la fouille des flots de données est devenu un sujet de recherche majeur [1, 6]. La fouille de flots de données pose deux défis principaux :

1. Les opérations traditionnelles de fouille sont inapplicables sur un flot de données. Les flots produisent des données en continu, très rapidement et de façon illimitée. Il est impossible d'utiliser des algorithmes traditionnels qui ont besoin de faire plusieurs passes sur les données. En prenant comme exemple l'extraction d'items ou de séquences fréquents, les principaux verrous à l'adaptation de méthodes traditionnelles sont : i) la technique ń générer-élaguer ż est inadaptée car l'étape de génération fait appel à des opérateurs de jointure, connus pour être typiquement bloquants car leur calcul nécessite de disposer de l'ensemble des données [3] ii) Les données ne peuvent être observées qu'une seule fois et iii) l'utilisation de la mémoire est limitée même si de nouveaux éléments continuent à être produits [20].
2. Le traitement exhaustif et exact des flots est impossible. La distribution des données change inévitablement dans le temps et l'utilisateur final est souvent plus intéressé par les changements récents (pour lesquels il veut une précision élevée) que par les changements plus éloignés (où une précision plus faible est satisfaisante) [8]. Par exemple, maintenir les items les plus fréquents

est récemment apparu comme étant un problème très intéressant [16, 17] pour des applications comme la détection d'intrusion dans les réseaux à fort trafic.

1.2 Surveillance et diagnostic par reconnaissance de motifs

Nous nous intéressons particulièrement à la surveillance et au diagnostic par reconnaissance de comportements caractéristiques du système observé. Cette technique fait partie des méthodes dites à *base de modèle*. Dans ce cas le modèle ne décrit pas de manière abstraite le fonctionnement du système mais fournit une description des différents comportements normaux ou anormaux du système. Dans le cas idéal les descriptions permettent de discriminer les différents états dans lequel peut se trouver le système.

La technique de surveillance varie selon que les comportements mémorisés dans le modèle sont normaux ou anormaux. Dans le premier cas il s'agit de vérifier qu'un ensemble d'indicateurs conservent des valeurs acceptables ou que des séquences d'événements indicatrices de bon fonctionnement se produisent à intervalles réguliers. Dans le second cas il s'agit principalement de reconnaître dans les données d'observation du système des motifs d'événements (des motifs séquentiels ou des chroniques) ou des ensembles de valeurs caractéristiques indiquant qu'une panne s'est produite. Dans ce cas une alarme est émise à l'attention de l'opérateur qui peut prendre les décisions de conduite qui s'imposent.

Dans la plupart des cas, de tels modèles doivent être construits de manière automatique du fait de leur complexité et de la nécessité de mises à jour régulières. Les techniques utilisées sont issues de l'apprentissage automatique ou de la fouille de données. L'extraction de séquences permet d'aquérir des chroniques ou des motifs significatifs selon divers critères liés à des contraintes temporelles ou d'apparition, par exemple. La différence essentielle entre les motifs et les chroniques est liée aux données manipulées. Dans le cas des chroniques, les données sont plus généralement issues de capteurs et se présentent sous la forme d'une longue séquence d'événements ininterrompue (e.g. des données d'électrocardiogramme issus d'un patient, ou des séquences de capteurs dans un contexte de supervision). Les motifs séquentiels, quant à eux, sont recherchés sur un très grand nombre de séquences qui sont bien plus courtes (e.g. les sessions de navigation des utilisateurs sur le mail, les traces d'accès à un serveur).

D'un point de vue conceptuel, cette méthode est proche du traitement de requêtes à des bases de données sur des flots (méthode *push*). Toutefois dans ces approches l'ensemble des requêtes (les motifs servant à interroger le flot) est fixe et l'accent est mis sur l'efficacité du processus d'interrogation "à la volée". Nous nous intéressons ici à la manière d'apprendre et de faire évoluer un tel ensemble de requêtes.

1.3 La sécurité et la surveillance dans les flots de données : applications

Nous donnons dans cette section quelques exemples d'applications permettant d'illustrer les liens existants entre le traitement des flots de données et les besoins en termes de sécurité. Ces liens sont assez naturels, dans la mesure ou les données produites par les systèmes générateurs de flots sont souvent le "pouls" de ces systèmes. S'assurer d'une analyse sûre et réactive de ces données permet de garantir le bon fonctionnement du système. Il peut s'agir de détecter des fraudes et réagir immédiatement, de détecter une attaque pirate sur un site en temps réel ou encore de détecter une arythmie cardiaque sur les données d'un électrocardiogramme.

Prenons tout d'abord l'exemple des usages du Web (actions effectuées par les utilisateurs sur un site internet). Il s'agit de données brutes et de bas niveau (date et heure de l'enregistrement, pages demandées, IP de l'utilisateur, etc.). L'analyste veut extraire, à partir de ces données, les changements,

les tendances, les motifs inhabituels et le tout avec un niveau de détails raisonnable. Il peut alors s'agir d'extraire des motifs exprimant que "Le trafic moyen venant du Japon, sur le site de l'INRIA Sophia Antipolis, dans les 15 dernières minutes est 40% plus élevé que dans les dernières 24 heures". Prenons maintenant le cas des attaques pirates sur un site. Ces attaques sont généralement à la recherche d'un ensemble de failles sur un site. Cet ensemble de failles est connu des pirates qui se communiquent des routines, contenant la série de requêtes à effectuer pour espérer faire s'écrouler un site Web. Le site de l'INRIA Sophia Antipolis est régulièrement la cible de ces attaques. Savoir analyser en temps réel le trafic du site et détecter des changement comme "Des URLs jusqu'ici rarement demandées sont l'objet de requêtes successives contenues dans une fenêtre temporelle très réduite avec une fréquence en augmentation de 80% ces 2 dernière minutes" peut permettre de réagir à de nouveaux types d'attaques (dont la technique ne serait pas encore connue des services informatiques) en temps réel et de prendre des dispositions immédiatement.

Un autre exemple se trouve dans les attaques de réseaux de type "Déni de Service" (DoS). Elles sont caractérisées par une tentative des attaquants de rendre impossible l'utilisation d'un service par ses utilisateurs légitimes. Il s'agit d'une des attaques internet les plus connues. Le mode d'opération de ces attaques consiste à consommer au maximum les ressources et la bande passante de la machine visée. Quand plusieurs attaquants se coordonnent pour lancer des attaquent de ce type, il s'agit d'une attaque DoS distribuée (DDoS). Une des formes de cette attaque est connue sous le nom de "*Email Bombing*". Il s'agit d'envoyer un nombre excessif de messages volumineux sur un ou plusieurs compte du site visé. L'*Email Bombing* est detecté quand le serveur du site victime est fortement ralenti, certainement en raison du grand nombre de messages à traiter. Une façon plus sécurisante de détecter cette attaque serait de surveiller le trafic du port 25/`SMTP`. Il s'agit alors de détecter les utilisations déviantes de ce port, afin d'exclure certaines adresses avant que le serveur ne se trouve saturé.

2 Problématique

Nous avons identifié deux thèmes majeurs à aborder dans le cadre de cette ARC, afin de répondre aux exigences posées par le contexte de l'analyse d'un flot à des fins sécuritaires. Tout d'abord nous aborderons les problèmes liés à de nouvelles techniques d'extraction de connaissances adaptées aux flots de données. Cela concerne à la fois les techniques issues de la fouille de données, mais aussi les techniques issues de l'apprentissage (supervisé ou non supervisé). En deuxième lieu, notre objectif est de considérer des problèmes liés à l'évolution des connaissances extraites. En effet, la nature extrêmement évolutive des données considérées va nous inciter à considérer les problèmes liés à l'évolution des connaissances acquises au fil du temps.
Nous avons cependant une troisième préoccupation : l'application de ces techniques sur des données réelles. Dans cette optique, la section 4 fera le point sur les données que nous envisageons de traiter afin de montrer l'utilité et l'efficacité des méthodes que nous pourrons développer dans un contexte de sécurité et de surveillance.

Nous exposons ici les problématiques de recherche liées aux aspects "extraction de connaissances" et "évolution de ces connaissances".

2.1 Vers de nouvelles techniques d'extraction de connaissances adaptées aux flots de données

Les contraintes inhérentes aux flots, e.g. données non statiques, volume, etc., sont telles qu'il devient indispensable d'adapter voir de reconsidérer complètement le processus d'extraction de connaissances. Dans le cadre des applications visées et en tenant compte des compétences des différents partenaires, le projet souhaite particulièrement aborder les problématiques suivantes :

- *Développer des algorithmes de fouille adaptés aux flots de données.* Il s'agit d'un grand défi qui implique de répondre aux types de questions suivantes : est il possible de faire un apprentissage à la volée ? Sur quelle fenêtre temporelle travailler ? Comment s'assurer que les motifs extraits sur une fenêtre temporelle correspondent à ceux qui seraient extraits si il était possible de stocker l'ensemble complet des données ? Comment repérer rapidement des séquences significatives sans pénaliser le transfert du flot ? Quel type de structure efficace mettre en oeuvre pour représenter les connaissances extraites ? Comment interroger le "passé" des connaissances apprises ?

- *Améliorer la précision des connaissances acquises par la prise en compte de la complexité des données.* Pour améliorer la précision des connaissances acquises une voie consiste à enrichir les données brutes avec les autres dimensions qui peuvent lui être associées (e.g. pour un internaute, outre sa navigation, de nombreuses informations peuvent être obtenues : pays, ville, type de machine, type de système ; pour un malade, ces informations pourraient concerner ses références médicales). Se posent alors les questions : comment introduire des informations complémentaires dans le processus de fouille de manière à diminuer l'imperfection des données ? Comment étendre les techniques d'extraction à des données complexes, en particulier multidimensionnelles ? Comment gérer l'imprécision dans les connaissances extraites ?

2.2 Faire évoluer les connaissances apprises au fil du temps

Dans le domaine de l'extraction de connaissances à partir de données statiques, il existe une séparation nette entre apprentissage et utilisation des connaissances : la base de connaissances est utilisée après l'apprentissage et elle est éventuellement réapprise si elle ne donne pas satisfaction. Lorsque l'on se place dans le cadre de la surveillance d'un système dynamique, les modèles (ensembles de motifs) sont extraits des données provenant de l'observation du système (du flot) et utilisés immédiatement pour le diagnostic. Les modèles extraits ne sont pas optimaux puisqu'appris sur un sous-ensemble des données d'observation (fenêtre temporelle ou résumé des données passées). Un premier problème est d'élaborer des indicateurs permettant d'évaluer en continu la qualité des données. Cette qualité peut être évaluée par un expert ou tout autre dispositif d'auto-évaluation du système lui-même, en particulier elle est liée à leur performance en détection d'anomalie. Une mauvaise qualité du modèle nécessite une révision du modèle mais se pose alors la question de la mise en œuvre de cette révision à partir de données d'apprentissage partielles et des performances obtenues en détection.
Par exemple, dans le cadre de la détection d'intrusion, les signatures d'attaques évoluent régulièrement : elles sont mises à jour quotidiennement et pourtant le nombre de faux positifs est considérable. Les experts n'ont que très peu de possibilités d'interagir avec le processus de détection d'attaques.
Dans le cadre de la surveillance médicale, la connaissance d'un patient particulier grandit au cours du temps. Il est important de pouvoir raffiner les connaissances extraites pour les adapter au patient et à son évolution. De la même manière, l'intégration de nouvelles variables de conduite dans le cas d'une application de sûreté de fonctionnement aura pour conséquence de remettre en cause certaines connaissances acquises sur les symptômes liés à une défaillance. L'adaptation des modèles est souhaitable pour améliorer la qualité et l'efficacité du diagnostic.
La prise en compte de cette évolution nécessite de répondre aux différentes questions suivantes : comment détecter l'inadéquation du modèle courant par rapport aux performances en détection? comment modifier le modèle appris de manière dynamique pour qu'il soit toujours représentatif ? L'apprentissage de modèles ou de motifs nécessite de préciser des paramètres, ceux-ci peuvent être adaptés pour une situation donnée mais comment et quand les faire évoluer lorsque de grandes modifications s'opèrent dans le comportement des données ? Comment considérer en même temps toutes les entrées (flots, experts, extérieurs) qui permettent de faire évoluer la base de connaissances ?

3 Propositions de recherche

Le principal objectif de cette ARC est d'aborder les questions évoquées plus haut 2 dans le cadre de l'extraction de motifs pour la surveillance et la sécurité. Plus précisément, nous souhaitons produire des méthodes d'extraction de connaissances qui soient capables de traiter les flots de manière réaliste et d'assurer le suivi de l'évolution de ces connaissances. Nous nous focaliserons plus particulièrement sur les techniques qui sont adaptées aux domaines d'applications visés et pour lesquelles les équipes disposent d'une expertise reconnue.

3.1 Extraction de connaissances dans les flots de données

Participants: AxiS (Inria), Dream (Inria), KDD (LGI2P/EMA Nimes), TATOO (LIRMM)

Dans le projet SéSur, nous souhaitons nous focaliser sur les approches d'extraction de motifs et sur les approches de classification (supervisée ou non) dans les flots de données. L'extraction de connaissances est dite supervisée lorsqu'il est possible de partager les données en entrée en un ensemble de séquences dans lesquelles le phénomène à caractériser est présent et un ensemble de séquences dans lesquelles on sait que le phénomène ne s'est pas produit. L'extraction non supervisée vise à extraire des schémas sans connaissance préalable. Des critères basés généralement sur la fréquence d'apparition des motifs extraits sont utilisés pour distinguer les motifs à retenir. Dans les deux cas, la qualité des données et leur complexité doivent être prises en compte (en pré-traitement ou pendant l'extraction). Nous commençons donc par présenter un axe qui consiste à étudier la prise en compte de la complexité et l'imperfection des données dans les approches supervisées et non supervisées, de manière à assurer la qualité des connaissances extraites.

Prise en compte de la complexité des données
Les données issues du monde réel sont souvent décrites au travers de différents attributs. De telles données sont dites *multidimensionnelles*, les dimensions pouvant même être munies de hiérarchies permettant de décrire les données à différents niveaux de granularité. De plus, les données sont souvent entachées d'imperfections, soit parce qu'elles sont incertaines, ou parce qu'elles sont imprécises. Ce phénomène est très fréquent dans les données manipulées dans les flots de données puisqu'elles proviennent la plupart du temps de capteurs renvoyant des informations imparfaites. Dans le contexte des flots de données, notre objectif est de proposer des méthodes permettant de considérer la complexité des données, tant sur le plan de leur aspect multidimensionnel que sur le plan de leur imperfection. Pour ce faire, nous proposons d'intégrer les travaux menés dans l'équipe TATOO liés au traitement de motifs séquentiels multidimensionnels et flous afin de les étendre au contexte des flots de données. Cet objectif est crucial dans le contexte des flots de données puisqu'il conditionne le fait que les méthodes proposées seront robustes et valides sur des données issues du monde réel. En effet, les méthodes classiques existantes pour le traitement des flots de données ignorent les imperfections et traitent donc des données très souvent biaisées.

Extraction non supervisée dans les flots de données.

Extraction de motifs séquentiels
Les motifs (ou connaissances) sont en général extraits en fonction de paramètres spécifiés par l'utilisateur : nombre d'occurrences d'un motif pour qu'il soit pertinent [26, 23], contraintes temporelles d'apparition entre événements [24], etc. Les techniques proposées jusqu'à présent considèrent un accès à la base dans son intégralité et nécessitent traditionnellement plusieurs parcours de cette base pour valider ou infirmer la présence de motifs candidats. Dans le cadre des flots de données, notre objectif est de reconsidérer ces approches pour éliminer ces jointures qui sont bloquantes par rapport au flot. De premières pistes ont été explorées par les partenaires impliqués dans ce projet pour l'apprentissage

non supervisé de motifs séquentiels dans les flots de données [33, 22]. En outre, le treillis des motifs extrait évolue constamment. Il peut donc être nécessaire de remettre en cause une partie de ce treillis et de le mettre à jour par *oubli* de certains motifs candidats.

Clustering dans les flots de données

1. Les travaux réalisés par l'équipe AxIS ont montré l'efficacité d'une approche basée sur une classification des séquences du flot de données, suivi par une extraction dans chaque cluster de la séquence qui le résume [22, 21]. Pour cette approche, une heuristique gloutonne a été définie afin d'affecter chaque nouvelle séquence dans une classe. Notre objectif est de proposer de nouvelles approches de classification de séquences dans les flots de données. Nous sommes en effet convaincus que l'extraction de motifs séquentiels dans les flots de données passe par des méthodes efficaces de classification qui permettent de diviser le problème et d'isoler d'éventuels individus susceptibles de provoquer un trop grand nombre de calculs (ce qui pourrait bloquer le flot).

2. Les techniques d'alignement de séquences permettent de proposer rapidement un résumé approximatif, mais fiable, d'un ensemble de séquences. Les techniques d'alignement existantes peuvent être appliquées mais elles ont été développées dans le cadre de données stockées et statiques. Notre objectif est de proposer une adaptation de ces techniques au contexte des flots de données, en considérant tout particulièrement que la précision des résultats doit rester aussi grande que possible. Ce dernier point est important, compte tenu du degré d'approximation déjà introduit par les techniques d'alignement appliquées aux données statiques.

Extraction supervisée dans les flots de données Les techniques proposées jusqu'à présent, telles les Bases de Données Inductives [32], pour extraire des chroniques ne permettent pas de prendre en compte des séquences très longues comme celles issues de flots de données. Ces techniques associent deux processus : la génération de motifs candidats et l'évaluation de leur pertinence. Ces deux points nécessitent des adaptations aux flots de données :

- la génération de motifs candidats ne peut plus procéder du plus général au plus spécifique car de nouveaux types d'objets peuvent apparaître dans les séquences et initialiser de nouveaux motifs complexes. Notre objectif est de proposer une méthode de gestion du treillis des motifs candidats qui évite de reconstruire ce treillis à l'apparition de nouvelles données. L'idée est de s'appuyer sur le treillis construit à l'instant $t - i$ et de ne réévaluer que le sous-treillis concerné par l'introduction du nouveau motif,

- l'évaluation de la pertinence des motifs candidats doit être réalisée sur une partie des données. Le support d'un motif calculé à un instant t peut être complètement différent de celui calculé à l'instant $t - i$. On se trouve alors confrontés aux mêmes problèmes que ceux rencontrés dans les approches non supervisées du paragraphe précédent. Notre objectif est d'étudier les techniques de résumés utilisées en extraction supervisée dans le cadre des flots de données et de les adapter au cas supervisé.

3.2 Faire évoluer les connaissances extraites d'un flot de données

Participants: Dream (Inria), KDD (LGI2P/EMA Nimes), AxiS (Inria), TATOO (LIRMM)

Gestion fine de l'historique des fréquents
La gestion de l'historique des motifs fréquents dans les flots de données est jusqu'ici basée sur le principe que "l'on est généralement plus intéressé par les changements récents que par les changements plus anciens". Les méthodes proposées ont alors eu pour objectif de stocker les historiques avec une granularité fine pour les événements récents et une granularité plus grande pour les événements anciens. C'est le cas pour les méthodes de fouille dans les flots proposées par Alice Marascu (AxIS) [22]

ou par Chedi Raissi (LGI2P/TATOO) [33]. Nous pensons que la gestion de l'historique ne doit pas être guidée par l'ancienneté des valeurs, mais plutôt par la variation de ces valeurs. Il s'agit donc de dire que "l'on est plus intéressé par les changements brusques ou marquants que par l'absence de changement". Nous proposons de développer des techniques de gestion des historiques, capables de s'adapter à l'évolution des fréquences demande d'étudier les méthodes de discrétisation. De plus, l'historique étant perpétuellement remis à jour, les échelles de valeurs vont changer (ce qui était considéré comme une évolution marquante devient monotone compte tenu des nouvelles valeurs).

Evaluation en ligne de la qualité des connaissances apprises
La qualité des connaissances apprises vis à vis du flot considéré est primordiale. Le problème principal provient du fait que les connaissances ont été apprises à partir d'un ensemble de données limité (une fenêtre temporelle ou un résumé). Il faut donc s'assurer que leur qualité reste suffisante vis à vis des données courantes et si besoin, les faire évoluer pour maintenir une qualité optimale. Ceci est d'autant plus important lorsque le système concerné évolue dans le temps (patient sous monitoring, procédé industriel dont les réglages évoluent, prise en compte de l'usure, etc.). Notre objectif est donc de proposer un ensemble d'indicateurs permettant d'évaluer la qualité en continu. Ceci peut se faire de différentes manières qu'il faudra comparer et éventuellement combiner : critères statistiques régulièrement vérifiés, réexamen de l'historique, prise en compte de la qualité de la détection ou de l'efficacité des décisions prises, confrontation à d'autres sources d'expertise, intervention d'un expert humain.

Evolution des connaissances apprises en fonction du flot de données
Les connaissances apprises (en particulier lorsqu'elles sont utilisées dans un contexte de détection) doivent être constamment ajustées pour améliorer, d'une part, la qualité de représentation qu'elles donnent du flot et, d'autre, part la qualité de détection qui en dépend. L'objectif est d'utiliser les résultats obtenus en détection et surveillance pour faire évoluer l'ensemble des chroniques. Il faudra identifier les chroniques responsables d'une baisse de qualité, et proposer des modifications, qui pourront nécessiter des changements des contraintes temporelles, l'ajout ou la suppression d'événements, et même l'ajout ou la suppression des chroniques elles-même. Un autre point sera de prévoir des indicateurs pour contrôler l'effet positif de ces modifications. Ces indicateurs pourront probablement être proches de ceux proposés dans le paragraphe précédent.

4 Validation des recherches dans différents contextes applicatifs

Les méthodes qui seront developpées dans le cadre de cette ARC ont pour objectif l'extraction de connaissances dans les flots de données, la gestion de ces connaissances et de leur évolution dans le temps. L'objectif de ce projet est également de privilégier la qualité de ces connaissances et leur pertinence pour des besoins de sécurité. Dans cette optique nous avons l'intention de valider les méthodes qui seront proposées et développées, sur des données issues du monde réel. Toutes les données décrites par la suite seront systématiquement anonymisées pour des besoins de confidentialité.

Données cardiaques : ECG, pression, etc.
En unité de soins intensifs les patients sont "monitorés" : un ensemble d'électrodes placées sur diverses parties du corps enregistre les signaux électriques correspondant à l'activité cardiaque (trois électrodes permettent la reconstruction de l'ECG classique sur douze voies) et des catheters enregistrent la pression artérielle en continus. L'ensemble des informations obtenues par ces voies correspond à un flot multidimensionnel. Celui-ci doit être ensuite prétraité par des algorithmes de traitement de signal et être analysé en ligne pour détecter d'éventuels troubles cardiaques.
Les prothèses cardiaques enregistrent de plus en plus d'informations qui peuvent ensuite être utilisées pour évaluer la qualité de l'activité d'un patient et adapter le programme de stimulation en con-

séquence. Les données ne peuvent être stockées telles quelles et sont agrégées selon des granularités différentes (24 heures ou 30 jours, par exemple). Par ailleurs, la mémoire de ces prothèses est limitée et les données nouvelles écrasent les anciennes. Ces cararactéristiques présentent de nombreuses similarités avec les flots de données : vision partielle des données, nécessité de les résumer, analyse pour les résumer au mieux.
Même si les contraintes en espace mémoire sont bien différentes dans ces deux cas, on retrouve dans ces deux applications les problématiques de fouille de données sur des flots. Partant d'un ensemble de chroniques (qu'on supposera au départ appris hors-ligne), il s'agit d'évaluer en-ligne la qualité de cet ensemble de motifs et de l'adapter afin qu'il corresponde au mieux au patient surveillé et à son état courant, tout en garantissant une qualité de détection.

Données d'une société éditrice de solutions de sécurité sur le Web
Bee Ware est une société éditrice de solutions de sécurité applicative web. L'un des défis de la sécurité applicative est de protéger aussi bien des attaques connues qu'inconnues. C'est à la fois un objectif technologique, rendu nécessaire par la diversité applicative, et une contrainte d'exploitation, afin d'éviter toutes les tâches fastidieuses de mise à jour. Aucune technologie n'a su à ce jour apporter une solution acceptable contre des attaques non référencées. Seul un oeil expert et humain s'avère capable de s'orienter dans la diversité de ce trafic et d'identifier les requêtes suspectes. Il a appris et fonctionne par analogie. C'est un oeil expert à la fois des technologies applicatives et des malversations potentielles qu'elles impliquent. Cette capacité à catégoriser rapidement l'information est l'une des composantes de l'intelligence humaine, la capacité d'apprendre en est une autre. C'est en essayant de reproduire ces capacités que les nouveaux systèmes permettront d'analyser et de classifier le trafic, de détecter et de bloquer toutes les tentatives d'attaques.
Afin de développer son programme de recherche et développement, la société Bee Ware recherche des partenariats lui permettant de tester de nouvelles solutions d'apprentissage. Pour cela elle dispose d'un ensemble de données d'attaques telles que :

- Fichiers logs de webserver : Logs générés par les utilisateurs accédant à un site WEB. Ceci concerne les utilisateurs normaux ainsi que ceux ayant un comportement anormal. Dans un souci de confidentialité, seule une partie des informations est stockée (pas de données postées, pas de header http, ...).

- Données simulées : SimFlux est un outil crée par BeeWare. Il est capable de générer des requêtes HTTP randomisées (dans le cadre d'un modèle décrit en XML), d'insérer des attaques dans ces requêtes, de les jouer sur un serveur WEB ou de les écrire dans un format maîtrisé. Cet outil contient un module de génération de flux SOAP et WEBDAV.

- Pattern d'attaques : Pour l'entraînement et la validation des moteurs de sécurité, nous avons constitué une base d'attaque classifiée. Cette base comprend les grandes classes d'attaques et les variations à l'intérieur de ces classes.

- Attaques référencées De nombreux sites sur internet recensent les failles découvertes dans les applications (webmail, forum, ...). Malgré l'absence d'unification, de formalisation et de classification il est possibles d'en collecter quelques une et d'obtenir ainsi des signatures d'attaques existantes.

- Génération : session d'attaques générées par des scanners applicatifs

Entre les outils disponibles en interne et ceux en téléchargement libre sur internet il est possible de rassembler 5 ou 6 scanners de vulnérabilité Web. Ces scanners font en général des attaques variées, peuvent tester des techniques d'évasion, et les plus pertinents effectuent une phase de reconnaissance puis une phase de recherche avant d'effectuer la phase d'attaque.

Logs d'un serveur mail
En accord avec le service informatique de l'Inria Sophia-Antipolis, nous disposerons de l'ensemble des opérations effectuées sur le serveur mail de sophia (toujours anonymisées). Les traces disponibles contiennent pour chaque jour les informations suivantes (entre autres) :

- Adresse mail de l'emetteur du mail
- Taille du mail
- Identifiant du mail dans le système
- Relais utilisé pour ce mail
- Destinataire du mail
- Nombre de destinataires du mail

Notre objectif est d'appliquer les techniques développées en matière d'extraction de connaissances et d'évolution de ces connaissances afin de proposer des alarmes du type "*Un motif est en train d'émerger à une grande vitesse. Ce motif concerne X adresse d'émetteurs, qui envoient des mails de 4 Mo à une seule adresse destinataire* ". Ce type d'alarme, une fois mise en place sur un serveur réel, permettrait l'apprentissage, puis la détection de comportements d'attaque. Cela permettrait également de prendre rapidement les mesures nécessaires (cela peut aller du simple message d'alarme au blocage des adresses concernées afin d'éviter que le serveur ne s'écroule).

Logs d'accès Web
Nous disposons des enregistrements concernant les logs d'accès Web au site de l'Inria Sophia-Antipolis sur les 15 derniers mois. Les équipes AxIS, TATOO et KDD sont habituées à travailler sur ce type de données, en les considérant de manière statique ou incrémentale [27, 25, 26]. Notre objectif est de considérer désormais ces données sous forme de flot continu. Cela nous permettra de vérifier la qualité des connaissances extraites et de les comparer à celles obtenues par une analyse statique. De plus, l'équipe AxIS a déjà extrait (par des techniques d'extraction de motifs séquentiels à très faible support sur des données statiques) des attaques pirates sur le site de l'Inria Sophia-Antipolis contenues dans ces logs [26]. Nous serons donc interessés de voir si de nouvelles attaques seront détectées dans les conditions difficiles imposées par les flots de données, afin de montrer notre capacité, à l'issue de ce projet, de les détecter en temps réel.

5 Collaborations et synergie entre les équipes

L'objectif de SéSur est de favoriser les synergies et les échanges entre des Centres de Recherche INRIA et des Laboratoires qui s'intéressent à la problématique de la fouille de données dans les flots de données. Plus précisément, nous souhaitons profiter des expériences complémentaires pour être à même de proposer de nouvelles approches et de valider leur utilité et leur efficacité afin d'améliorer la sécurité et la surveillance des systèmes observés.

Afin d'assurer la collaboration et l'échange d'informations entre les différentes équipes constituant ce projet, et constituer un noyau de compétences sur le thème étudié, nous envisageons trois principales actions :

- Durant la première partie du projet (6 mois), un des objectifs est de rédiger un état de l'art tirant parti des compétences et des connaissances complémentaires de chacune des équipes. En particulier, une attention particulière sera apportée à présenter de manière unifiée les approches d'extraction de motifs utilisées par les différents projets qu'elles soient de type non supervisé ou supervisé.

- Nous prévoyons de mettre en place une boîte à outils, rassemblant les logiciels communs et permettant le partage des logiciels existants ou développés par les équipes.

- Nous prévoyons d'échanger nos données d'expérimentations afin d'avoir un terrain d'application commun, en particulier lors de l'étape validation.
- Des réunions régulières sont prévues (voir le budget) afin d'enrichir la réflexion de chacune des équipes du projet et assurer un avancement cohérent de nos travaux.
- Un espace de travail collaboratif (forge) propre au projet SéSur sera mis en place afin de réunir les informations sur l'avancement du projet, de mettre en commun les compte-rendus de réunions, les articles publiés, les données et toute information susceptible d'aider au bon avancement du projet.

6 Positionnement

Il n'existe pas à l'heure actuelle (et à notre connaissance) de projet en France abordant cette thématique, en dépit des enjeux qui lui sont associés. Les systèmes producteurs de flots de données sont en effet de plus en plus nombreux et les techniques permettant de les sécuriser et de les surveiller sont encore à développer, ce qui est l'objectif de cette ARC. Les quelques travaux existant en France sont issus de la communauté bases de données et portent sur l'aspect gestion des données et réponse aux requêtes. La principale originalité de notre proposition est de se focaliser sur l'aspect extraction de motifs temporels à partir des flots de données, ces motifs étant ensuite utilisés dans le cadre de la surveillance en ligne des systèmes producteurs. Les méthodes de fouille de données et d'apprentissage développées dans chacun des projets concernés par cette ARC sont à ce jour confrontées à des problématiques nouvelles face aux flots de données. Les méthodes développées par AxIS, le LGI2P ou le LIRMM, doivent en effet être étendues à de longues séquences comme le sont les données d'ECG ou les logs d'opérateurs téléphoniques qu'exploite le projet DREAM. D'un autre côté, le projet DREAM possède une expertise en matière d'extraction de chroniques pertinentes, mais les méthodes développées dans ce contexte sont jusqu'ici étudiées pour des cas de données statiques. Le passage au contexte des flots de données demandera donc un travail commun de tous les partenaires afin de converger vers des méthodes à la fois réalistes, permettant l'extraction de connaissances pertinentes et utiles, et afin également de valider nos travaux dans un contexte applicatif réel.
Au niveau international, la problématique de l'analyse des flots de données est de plus en plus abordée comme en témoigne les nombreux workshops ou sessions organisés sur ce thème par des conférences prestigieuses (KDD, SIGMOD, VLDB, PKDD, ICDM, SAC). Il existe également de nouveaux projets qui s'intéressent soit à l'extraction soit à la détection dans les flots de données. Le projet le plus proche de nos problématiques est MAIDS (Mining Alarming Incidents in Data Streams) de l'Université Illinois Urbana-Champaign (http://maids.ncsa.uiuc.edu/about/index.html) qui a été initié en 2003 et considère l'analyse des flots de données à des fins de détection. Même si les objectifs restent les mêmes, SéSur diffère de Maids sur un certain nombre d'aspects. En particulier, nous souhaitons faire collaborer les approches d'extraction de motifs et de chroniques. En proposant ces deux types d'approches nous sommes alors à même de répondre à deux grandes catégories d'applications : celles qui considèrent une séquence longue et continue (e.g. électrocardiogramme, supervision) et celles qui considèrent un ensemble de longues séquences (e.g. site Web). Un autre point est l'aspect évolution des connaissances qui n'est jusqu'à présent pas abordé alors que les conséquences sur des applications réelles sont nombreuses.

7 Les partenaires

Équipe AxIS – Inria Sophia Antipolis

L'équipe AxIS est basée sur une approche multi-disciplinaire pour l'analyse, la conception et l'amélioration des systèmes d'information. Les techniques développées font appel à des compétences en intelligence

artificielle, fouille de données ou encore génie logiciel. Ces compétences complémentaires sont nécesaires pour atteindre nos objectifs principaux: 1) développer ds méthodes et outils pour aider à la fois à l'analyse d'un SI et à celle de son utilisation; 2) aider à la validation, la maintenance et l'amélioration de SIs. Actuellement nous travaillons sur les SIs basés sur le Web et les NTIC. Nos objectif en termes d'analyse des usages sont liés à des données volumineuses mais stockées et statiques. Dans le cadre de cette ARC, les techniques de fouille de données seront particulièrement présentes, dans le but de les faire progresser et de les adapter au contexte des flots de données. L'objectif pour AxIS est, entre autres, d'améliorer l'aspect pertinence des connaissances extraites et leur utilité dans un contexte applicatif réel avec des contraintes fortes comme celles exigées par une application à caractère sécuritaire.

Participants :

- Yves Lechevallier – DR / Inria Rocquencourt
- Alice Marascu – Doctorante (sujet : "Extraction de Motifs Séquentiels dans les flots de données").
- Florent Masseglia – CR
- Brigitte Trousse – CR

Équipe DREAM – IRISA Rennes

Le projet DREAM a pour thème de recherche principal l'aide à la surveillance et au diagnostic de systèmes ou activités complexes évoluant dans le temps. Les domaines d'application sont la santé, les télécommunications et l'environnement. Plus précisément, DREAM étudie l'utilisation de techniques de type "reconnaissance de chroniques" pour le diagnostic et la prédiction de pannes de systèmes dynamiques, ainsi que pour la surveillance et le diagnostic des pathologies cardiaques de patients.
Motivé par l'acquisition des chroniques, un autre sujet de recherche de l'équipe est l'acquisition de modèles temporels basés sur des chroniques par des techniques d'apprentissage symbolique et plus récemment de fouille de données de type relationnel. Nous nous sommes, pour l'instant, focalisés sur des données d'apprentissage statiques. L'extension des techniques à des données dynamiques telles que celles fournies par des flots de données est essentiel. Cet aspect est en particulier présent dans des applications industrielles que nous abordons telles que la sécurité dans les réseaux de télécommunications. Le travail commun avec les partenaires de cette ARC devrait nous permettre d'aborder cette problématique dans les meilleures conditions, en particulier en renforçant notre compétence en fouille de données.

Participants :

- Marie-Odile Cordier – Professeure / Université de Rennes 1
- René Quiniou – CR Inria
- Alexandre Vautier – Doctorant (sujet : "Intégration de contraintes temporelles dans les bases de données inductives. Application à la détection d'intrusion.")

L'équipe KDD LGI2P/EMA Nîmes

L'un des thème de recherche du LGI2P concerne les risques et la décision. Dans ce cadre, il s'intéresse plus particulièrement au suivi et au contrôle d'activités complexes au cours du temps. Parmi les dernières domaines d'applications traités par le centre, nous pouvons citer la santé (détection de chutes de personnes agées), le transport, les détection de fraudes dans le cas d'opérateur téléphonique, ou encore la supervision de procédés continus dynamiques. Parmi les sujets de recherche abordés, l'équipe s'intéresse aux techniques de fouilles associées à la classification, au clustering et à l'extraction de séquences. En ce qui concerne ce dernier point, les premières approches proposées se sont intéressées à des bases de données statiques puis à la possibilité de considérer l'ajout de nouvelles données. Depuis

cette année, nous nous intéressons à l'apprentissage de ces séquences dans un cadre de flots de données. La participation à ce projet nous offrira la possibilité d'étendre nos propositions à la prise en compte de données de plus en plus complexes et à renforcer nos compétences en considérant également un apprentissage supervisé.

Participants :

- Gérard Dray – Enseignant Chercheur
- Jacky Montmain – Chercheur CEA
- Pascal Poncelet – Professeur
- Chedy Raissi – Doctorant (sujet : "Détection de motifs séquentiels dans les data streams")

Équipe TATOO – LIRMM Montpellier

Les travaux menés au sein du projet Ingénierie des Données et des Connaissances du LIRMM dans le contexte de l'extraction de connaissances portent sur la fouille de données dans les bases de données complexes (e.g. données structurées, semi structurées, multidimensionnelles, qualitatives et quantitatives, textuelles etc.) et dynamiques, sur la fouille de données approximatives et l'aide à la décision. L'un des domaines d'application du projet concerne l'aide au diagnostic et à la prévention de pannes dans un système dynamique de capteurs solaires. Les travaux récents traitent de la recherche de motifs séquentiels multidimensionnels comme une approche réellement générique applicable aussi bien aux cubes de données qu'aux données enrichies, de motifs séquentiels flous comme aide à la prise en charge des valeurs manquantes. La participation à ce groupe de recherche nous permettra d'affiner nos connaissances dans le domaine de l'apprentissage supervisé. En outre, elle permettra de favoriser l'émergence de solutions permettant de considérer les données issues du monde réel dans la gestion des flots de données sans perte d'information. Ces données, souvent définies sur plusieurs dimensions et entachées d'erreurs et d'imprécisions, sont en effet la plupart du temps difficilement appréhendables par les méthodes classiques qui les réduisent en des données simples, ce qui entraîne des lacunes dans la connaissance produite.

Participants :

- Céline Fiot – Doctorante (sujet : "Motifs séquentiels et prise en compte des données manquantes ou incomplètes")
- Anne Laurent – MCF
- Marc Plantevit – Doctorant (sujet : "Fouille de données pour les bases de données multidimensionnelles")
- Maguelonne Teisseire – MCF

8 Budget demandé

La coordination de cette ARC sera assurée par Florent Masseglia (AxIS Sophia Antipolis). Le budget demandé pour cette proposition est de 135 Keuros et vise à financer :

- 2 post-docs de 12 mois en co-encadrement
- 4 stages de 5 mois sur les quatre sites
- Missions (4 X 2,5 Keuros par an) : les déplacements entre les sites (Sophia-Antipolis, Rennes, Nîmes et Montpellier) des partenaires et des personnels recrutés, l'organisation des réunions ainsi que la participation à des conférences liées à cette ARC.

- Invitations - séminaires (10 Keuros) : la prise en charge de visites de chercheurs invités lors de nos réunions pleinières ou lors de l'organisation d'un séminaire à la fin de la deuxième année. Nous envisageons dès à présent d'inviter :
  - Dr. Raffaele Perego, HPC Lab, Istituto di Scienza e Tecnologie (Pise, Italie).
  - Pr. Salvatore Orlando, Università Ca' Foscari di Venezia (Italie).
  - Georges Hébrail (ENST Paris).

Voici une synthèse du budget demandé pour cette ARC :

Année	Post-doc	Stage	Mission	Total
2007	37,5 Keuros	15 Keuros	10 Keuros	62,5 Keuros
2008	37,5 Keuros	15 Keuros	20 Keuros	72,5 Keuros
Total	75 Keuros	30 Keuros	30 Keuros	135 Keuros

References

[1] The MAIDS Project, http://maids.ncsa.uiuc.edu.

[2] *VLDB 2002, Proceedings of 28th International Conference on Very Large Data Bases, August 20-23, 2002, Hong Kong, China*. Morgan Kaufmann, 2002.

[3] *Next Generation Data Mining*, chapter Mining frequent patterns in data streams at multiple time granularities. MIT Press, 2003, 2003.

[4] *Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management, New Orleans, Louisiana, USA, November 2-8, 2003*. ACM, 2003.

[5] P.A. Laur ans R. Nock, J.-E. Symphor, and P. Poncelet. On the Estimation of Frequent Itemsets for Data Streams: Theory and Experiments. In *Proceedings of the 14th ACM Conference on Information and Knowledge Management (CIKM 2005)*, Bremen, Germany, October 2005.

[6] Arvind Arasu, Brian Babcock, Shivnath Babu, Mayur Datar, Keith Ito, Rajeev Motwani, Itaru Nishizawa, Utkarsh Srivastava, Dilys Thomas, Rohit Varma, and Jennifer Widom. Stream: The stanford stream data manager. *IEEE Data Eng. Bull.*, 26(1):19–26, 2003.

[7] G. Carrault, M.-O. Cordier, R. Quiniou, and F. Wang. Temporal abstraction and inductive logic programming for arrhyhtmia recognition from electrocardiograms. *Artificial Intelligence in Medicine*, 28:231–263, 2003.

[8] Yixin Chen, Guozhu Dong, Jiawei Han, Jian Pei, Benjamin W. Wah, and Jianyong Wang. Online analytical processing stream data: Is it feasible? In *DMKD*, 2002.

[9] Marie-Odile Cordier and René Quiniou. Apprentissage relationnel de motifs temporels. In *Atelier Extraction de motifs temporels pour la détection en ligne de situations critiques à EGC 2005*, RNTI. Cépaduès, 2005.

[10] Pedro Domingos and Geoff Hulten. Catching up with the data: Research issues in mining data streams. In *DMKD*, 2001.

[11] A. Evsukoff, S. Gentil, and J. Montmain. Fuzzy reasoning in co-operative supervision systems. *Control Eng. Practice*, 8:389–407, 2000.

[12] C. Fiot, G. Dray, A. Laurent, and M. Teisseire. A la recherche des motifs séquentiels flous. In *Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA 04)*, pages 131–138, Nantes, France, 2004.

[13] Elisa Fromont, Marie-Odile Cordier, and René Quiniou. Learning from multi source data. In *PKDD'04 (Knowledge Discovery in Databases)*, volume 3202 of *Lecture Notes in Artificial Intelligence*, Pise, Italie, 2004. Springer.

[14] Elisa Fromont, René Quiniou, and Marie-Odile Cordier. Learning rules from multisource data for cardiac monitoring. In E. Keravnou S. Miksch, J. Hunter, editor, *AIME'05 (Artificial Intelligence in Medicine)*, pages 484–493, Aberdeen, Scotland, 2005. Springer.

[15] M. Jaczynski. *Scheme and Object-Oriented Framework for case Indexing By Behavioural Situations: Application in Assisted Web Browsing.* PhD thesis, Doctorat Thesis of the University of Sophia-Antipolis (in french), December 1998.

[16] Cheqing Jin, Weining Qian, Chaofeng Sha, Jeffrey Xu Yu, and Aoying Zhou. Dynamically maintaining frequent items over a data stream. In *CIKM* [4], pages 287–294.

[17] Richard M. Karp, Scott Shenker, and Christos H. Papadimitriou. A simple algorithm for finding frequent elements in streams and bags. *ACM Trans. Database Syst.*, 28:51–55, 2003.

[18] P.A. Laur, J.E. Symphor, R. Nock, and P. Poncelet. Mining Sequential Patterns on Data Streams: A Near-Optimal Statistical Approach. In *Proceedings of the 2nd International Workshop on Knowledge Discovery from Data Streams (KDDS 2005) In conjunction with ECML-PKDD2005 The 16th European Conference on Machine Learning (ECML) and The 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Porto, Portugal, October 2005.

[19] A. Laurent, P. Poncelet, and M. Teisseire. *Fuzzy Data Mining for the Semantic Web: Building XML Mediator Schemas.* In Fuzzy Logic and the Semantic Web. Elsevier, To appear.

[20] Gurmeet Singh Manku and Rajeev Motwani. Approximate frequency counts over data streams. In *VLDB* [2], pages 346–357.

[21] A. Marascu and F. Masseglia. Mining data streams for frequent sequences extraction. In *Proceedings of the first IEEE Workshop on Mining Complex Data (MCD'05). Held in conjunction with ICDM'05*, Houston, USA, 2005.

[22] Alice Marascu and Florent Masseglia. Mining sequential patterns from data streams: a centroid approach. *Journal of Intelligent Information Systems (JIIS).*, 27(3):291–307, November 2006.

[23] F. Masseglia, F. Cathala, and P. Poncelet. The PSP Approach for Mining Sequential Patterns. In *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery*, Nantes, France, September 1998.

[24] F. Masseglia, P. Poncelet, and M. Teisseire. Extraction efficace de motifs séquentiels : le prétraitement des données. In *Actes des Journées Bases de Données Avancées (BDA'99)*, Bordeaux, France, Octobre 1999.

[25] F. Masseglia, P. Poncelet, and M. Teisseire. Incremental mining of sequential patterns in large databases. *Data an Knowledge (DKE) Journal*, 46(1):97–121, July 2003.

[26] F. Masseglia, D. Tanasa, and B. Trousse. Web usage mining: Sequential pattern extraction with a very low support. In *Proceedings of the 6th Advanced Web Technologies and Applications: 6th Asia-Pacific Web Conference (APWeb 2004)*, pages 513–522, Hangzhou, China, 2004.

[27] F. Masseglia, M. Teisseire, and P. Poncelet. HDM: A Client/Server/Engine Architecture for Real Time Web Usage Mining. *Knowledge and Information Systems (KAIS) Journal*, 5:439–465, November 2003.

[28] F. Masseglia, M. Teisseire, and P. Poncelet. *Sequential Pattern Mining: A Survey on Issues and Approaches*. In Encyclopedia of Data Warehousing and Mining. Information Science Publishing, 2005.

[29] J. Montmain. Des modèles pour la supervision. In *Habilitation à Diriger des Recherches*, Institut National Polytechnique de Grenoble, 2000.

[30] J. Montmain and S. Gentil. Dynamical causal model diagnostic reasoning for online technical process supervision. *Automatica*, 36:1137–1152, 2000.

[31] M. Plantevit, Y.W. Choong, A. Laurent, D. Laurent, and M. Teisseire. M2SP: Mining Sequential Patterns Among Several Dimensions. In *proceedings of PKDD'05: Principles and Practice of Knowledge Discovery in Databases*, pages 205–216, Porto, Portugal, October 2005.

[32] Luc De Raedt. A perspective on inductive databases. *SIGKDD Explorations*, 4(2):69–77, 2002.

[33] C. Raissi, P. Poncelet, and M. Teisseire. Need for speed: Mining sequential pattens in data streams. In *Actes des 21ièmes Journées Bases de Données Avancées (BDA 2005)*, Saint Malo, France, 2005.

[34] Alexandre Vautier, Marie-Odile Cordier, and René Quiniou. An inductive database for mining temporal patterns in event sequences. In Leslie Pack Kaelbling and Alessandro Saffiotti, editors, *Proceedings of IJCAI-05 (International Joint Conference on Artificial Intelligence)*, Edinburgh, 2005.

[35] Alexandre Vautier, Marie-Odile Cordier, and René Quiniou. An inductive database for mining temporal patterns in event sequences (long version). In *PKDD (Principles and Practice of Knowledge Discovery in Databases) - Workshop mining spatio-temporal data*, Porto, Portugal, 2005.