INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Équipe AXIS

Conception, Analyse et Amélioration de Systèmes d'Information dirigées par l'Usage

Sophia Antipolis

THÈME 3A

Table des matières

1. Composition de l'équipe **1**
2. Présentation et objectifs généraux **2**
2.1. Objectifs 2
3. Fondements scientifiques **2**
3.1. Sémantique et conception de systèmes d'informations hypertextes 2
3.2. ECD appliquées aux données d'usage 3
3.2.1. a) Sélection et transformation des structures de données 4
3.2.2. b) Extraction des règles d'associations 5
3.2.3. c) Découverte des motifs séquentiels 5
3.2.4. d) Recherche des structures classificatoires 5
3.2.5. e) Réutilisation dans l'analyse de l'usage. 6
3.3. Systèmes de recommandations personnalisées 6
3.4. Raisonnement à partir de cas 8
4. Domaines d'application **8**
4.1. Panorama 8
5. Logiciels **9**
5.1. Introduction 9
5.2. CLF - « Computer Language Factory » 9
5.3. Boite à outils de classification 10
5.4. CBR*Tools - Plate-forme objet en raisonnement à partir de cas 10
5.5. Broadway*Tools - Génération de systèmes de recommandations adaptatifs 11
5.6. Broadway-Web - Aide personnalisée à la navigation sur le Web 11
6. Résultats nouveaux **11**
6.1. Panorama 11
6.2. Aide à la vérification sémantique de sites Web 12
6.3. Sélection et transformation des structures de données 12
6.3.1. Extraction et construction de données agrégées 12
6.3.2. Forme Normale Symbolique appliquée à la transformation de données 13
6.3.3. Pré-traitement de logs HTTP multi-sites 13
6.3.4. Persistance et exploitation de logs comportementales des utilisateurs d'un site Web 13
6.3.5. Transformation de données spatiales 14
6.3.6. Améliorations de notre boite à outils de classification 14
6.3.6.1. Interfaces communes 14
6.3.6.2. Visualisation des cartes de Kohonen 14
6.4. Méthodes de partitionnement et validation 14
6.4.1. Classification non supervisée à partir de dissimilarités mesurant le lien entre représentations complexes des données 14
6.4.2. Classification non supervisée à partir d'un tableau de données intervalles 15
6.4.3. Algorithme des cartes topologiques auto-organisatrices et données symboliques 16
6.4.4. Méthode d'évaluation de la stabilité d'une classe 16
6.5. Extensions du modèle de la classification hiérarchique 16
6.5.1. La Structure 2-3 hiérarchique 16
6.5.2. Lien minimum pour l'Algorithme de Classification 2-3 hiérarchique 17
6.5.3. Etude des dissimilarités induites par les 2-3 hiérarchies 17
6.5.4. Complexité et implémentation de l'Algorithme de Classification 2-3 hiérarchique 17
6.6. Modélisation supervisée de données fonctionnelles par perceptron multi-couches 17
6.7. Analyse comparative de méthodes d'extraction de séquences 18

6.8. Analyse et utilisation de comportements visuels et non visuels 18
6.8.1. Conception et réalisation d'un plateforme générique d'expérimentation 18
6.8.2. Application de notre plateforme au contexte choisi 19
6.8.3. Expérimentation auprès de deux groupes de sujets 19
6.9. Réalisation d'un site CASA intégrant un service d'aide à la navigation 19
6.10. Extension et validation de Broadway*Tools et CBR*Tools 19
6.10.1. Broadway*tools, générateur de systèmes de recommandations perosnnalisées sur le Web 19
6.10.2. CBR*Tools plateforme objet pour la gestion de l'expérience 19
7. Contrats industriels 21
7.1. Contrats industriels 21
7.1.1. EDF : Classification de courbes et analyse de l'usage 21
7.1.2. EPIA un projet pré-compétitif RNTL 2002 21
8. Actions régionales, nationales et internationales 21
8.1. Actions régionales 21
8.2. Actions nationales 22
8.3. Actions européennes 22
8.3.1. Projet européen IST : ASSO 22
8.3.2. Réseaux européens 23
8.4. Actions internationales 23
9. Diffusion des résultats 23
9.1. Animation de la communauté scientifique 23
9.1.1. Revues 23
9.1.2. Comités de programme 24
9.1.3. Organisation de séminaires et conférences 24
9.1.4. Visites 24
9.1.5. Serveur interne Web 24
9.1.6. Divers 24
9.2. Formation 24
9.2.1. Enseignement universitaire 24
9.2.2. Thèses 25
9.2.3. Stages 26
9.3. Participation à des colloques, séminaires 27
10. Bibliographie 27

1. Composition de l'équipe

Responsable scientifique
Brigitte Trousse [CR Inria, UR Sophia Antipolis]

Responsable permanent
Yves Lechevallier [DR Inria, UR Rocquencourt]

Assistantes de projet
Sophie Honnorat [AI Inria, à 40 % dans le projet, UR Sophia Antipolis]
Sandrine Boute [TR Inria, à temps partiel dans le projet jusqu'au 1/10, UR Rocquencourt]
Christiane Demars [AI Inria, à temps partiel dans le projet à compter du 1/10, UR Rocquencourt]

Personnel Inria
Thierry Despeyroux [CR Inria, UR Sophia Antipolis puis UR Rocquencourt à compter du 01/09]
Florent Masseglia [CR Inria, depuis le 01/09, UR Sophia Antipolis]

Chercheur en détachement
Eric Guichard [Education Nationale (ex-chercheur ENS Ulm, à compter du 01/09, UR Sophia Antipolis]

Chercheur post-doctorant
Mohamed Semi Gaieb [UNSA, boursier ATER jusqu'au 31/08, puis INRIA à compter du 01/09, UR Sophia Antipolis]

Chercheurs doctorants
Hicham Behja [enseignant-chercheur, Université de Meknès, bourse réseau STIC-GL UNiversité de Casablanca (Maroc) et UR Sophia Antipolis]
Sergiu Chelcea [UNSA, 1/2 bourse région,à compter du 01/10, UR Sophia Antipolis]
Brieuc Conan-Guez [Université Paris IX Dauphine,1/2 ATER complémenté, UR Rocquencourt]
Aicha El Golli [Université Paris IX Dauphine, UR Rocquencourt]
Doru Tanasa [UNSA, UR Sophia Antipolis]

Ingénieur associé
Sébastien Simard [à compter du 01/09]

Collaborateurs extérieurs
Mireille Arnoux [MC, Université de Bretagne Occidentale, à compter du 1/05, UR Sophia Antipolis]
Patrice Bertrand [MC, Université de Paris IX Dauphine puis ENST Bretagne, UR Rocquencourt]
Marc Csernel [MC, Université de Paris IX Dauphine, UR Rocquencourt]
Fabrice Rossi [MC, Université de Paris IX Dauphine, UR Rocquencourt]

Chercheurs invités
Mireille Arnoux [MC, Université de Bretagne Occidentale, en congé sabbatique jusqu'au 30/04, UR Sophia Antipolis]
Lynne Billard [Université de Georgie, USA, UR Rocquencourt, juillet]
Francisco de Carvalho [université Federal de Pernambuco, Brésil,UR Rocquencourt, mai et septembre]
Jean-Paul Rasson [FUNDP à Namur, Belgique, UR Rocquencourt, janvier et juin]
Rosanna Verde [Professeur, Université de Naples II, UR Rocquencourt, juillet et septembre]

Stagiaires
Tarek Ait-Mohamed [DEA, Université Dauphine, du 01/04 au 31/07]
Fabien Benoit [ESSI, UNSA, à compter du 15/10]
Sergiu Chelcea [DEA, UNSA, du 21/01 au 30/09, UR Sophia Antipolis]
Nathalie Evan [ESSI, UNSA, jusqu'en avril]
Gentian Gusho [jusqu'au 31/09, UR Rocquencourt]
Vincent Giraudon [Cepun, du 07/01 au 31/08, UR Sophia Antipolis]
Laurent Jullien [DEA, Université Panthéon-Sornonne (Paris I), du 01/04 au 30/09]

Miha Jurca [à partir du 15/11, UR Sophia Antipolis]
Arnaud Santoni [ESSI, UNSA, à compter du 15/10]
Tao Wan [DEA,UVSQ, du 01/03 au 30/09]

2. Présentation et objectifs généraux

2.1. Objectifs

Mots clés : *système d'information*, *système de connaissances*, *site web*, *qualité*, *utilité*, *utilisabilité*, *design management*, *évaluation*, *filtrage collaboratif*, *intelligence artificielle*, *fouille de données*, *statistiques*, *classification*, *extraction des connaissances à partir de données (ECD)*, *génie logiciel*, *raisonnement à partir de cas*, *services adaptatifs*, *web sémantique*, *web service*, *vérification sémantique*, *analyse de l'usage*, *gestion de l'expérience*, *système de recommandation*, *adaptation à l'utilisateur*, *évolution*, *IHM*, *ergonomie*.

L'objectif d'AxIS est de concevoir des méthodes et des outils, dirigés par l'usage, pour l'aide à la conception et à l'analyse de systèmes d'informations et/ou de connaissances(SI). Bien qu'à court terme le projet s'oriente principalement sur les sites ou services Web, nous nous plaçons dans une optique globale de conception et d'évaluation de systèmes d'informations adaptatifs basés sur les standards du W3C. Par adaptatif, nous entendons à la fois des capacités d'adaptation à l'utilisateur voire de personnalisation et aussi des capacités d'apprentissage à partir d'une analyse de l'usage. Plus précisément, l'objectif du projet est

- de privilégier et d'anticiper dès la conception les problèmes liés à l'évolution du contenu d'un SI (architecture et documents) et ceux liés à l'usage (ainsi qu'à son évolution).
- d'aider le concepteur (par exemple l'éditeur, le Web-maître ou l'administrateur) d'un système d'informations à mieux prendre en compte l'utilisateur final.

Axis se propose de travailler selon deux points de vues centraux de la conception d'un SI que sont celui de l'*artefact* (visant à la fois le concepteur, l'éditeur ou le webmaster du SI) et celui de l'*usage*. Plus généralement nous visons à faciliter la gestion de points de vues [63]. Pour atteindre cet objectif, nous devons d'abord comprendre et formaliser la notion de système d'informations multi-vues. Les techniques mises en oeuvre sont à la confluence de disciplines différentes et complémentaires que sont l'intelligence artificielle (IA), l'extraction de connaissances à partir de données (ECD[1]) et le génie logiciel (GL).

Notre programme de recherche (cf. Fig. 1) s'articule autour des aspects statiques et dynamiques d'un SI et de deux thèmes transversaux et fédérateurs : a) aide à l'amélioration d'un SI par confrontation des aspects statiques et dynamiques et et b) capitalisation des connaissances liées a cette confrontation. Au niveau du développement logiciel, AxIS vise à la fois la définition de langages spécifiques (basés sur des ontologies relatives aux divers métiers visés) et la proposition de plates-formes logicielles pour l'aide à la spécification et l'évaluation de la qualité de SIs._

3. Fondements scientifiques

3.1. Sémantique et conception de systèmes d'informations hypertextes

Mots clés : *Web sémantique*, *aspects statiques d'un SI*, *vérification sémantqiue*, *typage*, *documents semi-structurés*, *spécification*, *maintenance*, *évolution*, *sites Web*, *sémantique formelle*, *génie logiciel*.

Concevoir et maintenir un système d'informations hypertexte comme un site Web est une tâche difficile. Il est beaucoup plus facile de trouver des informations inconsistantes qu'un site bien maintenu sur Internet. Notre but est d'étudier et de construire les outils nécessaires à la conception, à la production et à la maintenance de sites complexes et cohérents avec une approche pluri-diciplinaire (GL et IA).

[1] l'ECD fût introduit par Piatetsky-Shapiro en 1989 lors d'un workshop de la conférence IJCAI'89.

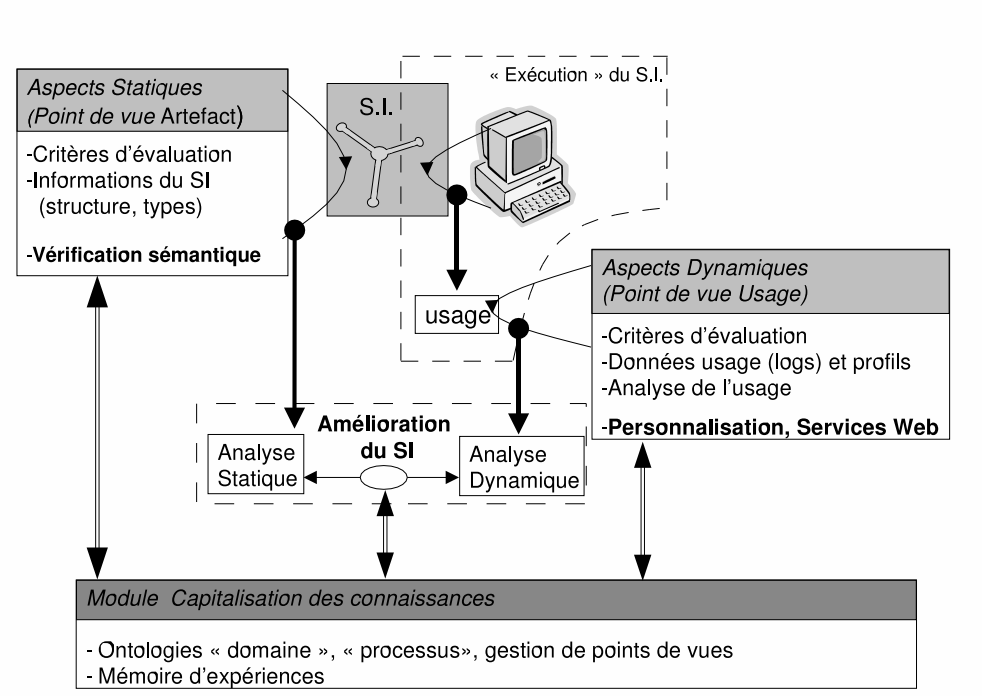

Figure 1. Problématique générale

Il existe un très fort parallèle entre un document structuré (tel qu'un site Web) et un programme, et le monde du Web est un très bon utilisateur d'idées développées il y a plusieurs années dans le monde du génie logiciel : la notion de syntaxe abstraite se retrouve dans un langage tel que XML et les DTD de même que l'idée de séparation entre structure et présentation concrète.

Jusqu'à présent, le monde du Web s'est principalement intéressé à la présentation des pages (HTML, CSS, XSL) et à la structure syntaxique du contenu des pages (XML), mais très peu à la sémantique des sites. Notons cependant les efforts du consortium W3C autour du « Web sémantique » (XML, RDF et « RDF schema ») ainsi que certains travaux de recherche issus de l'IA basés sur une approche ontologique comme WebMaster [64]. Notre approche diffère dans le fait que nous voulons exploiter plus loin le parallèle entre programmes et sites Web pour mieux aborder la sémantique formelle des sites.

Il existe (au moins) deux sens au terme « sémantique » : ce peut être l'étude scientifique du sens des unités linguistiques, mais aussi l'étude de propositions d'une théorie déductive du point de vue de leur vérité ou de leur fausseté. C'est à cette dernière définition que nous voulons nous référer en formalisant une partie du contenu des systèmes d'information.

Nous pouvons d'ores et déjà distinguer les « `aspects statiques` » d'un site qui peut être vu comme un ensemble de contraintes globales (pas seulement syntaxiques, mais aussi sémantiques et dépendantes du contexte) qui doivent être vérifiées et la « `aspects dynamiques` » qui prend en compte la navigation d'un utilisateur dans un site et rejoint donc l'analyse des usages. Pour les aspects dynamiques, nous pensons à moyen terme, formaliser les notions de qualité et de fiabilité d'un site en faisant un parallèle avec la notion de preuve de programme.

3.2. ECD appliquées aux données d'usage

Mots clés : *analyse des usages*, *web usage mining*, *entrepot de données*, *fouille de données*, *sous-séquences fréquentes*, *classification*.

Rappelons en Fig. 2 les quatre étapes du processus ECD. a) L'étape de **sélection des données** vise tout d'abord à extraire d'un entrepôt de données ainsi constitué les ensembles d'informations utiles aux méthodes de fouille de données. b) L'étape de **transformation des données** concerne quant à elle l'utilisation de "parseurs" construisant les tableaux de données directement utilisables par les algorithmes de l'ECD. c) Les techniques de fouille de données utilisées peuvent être **l'extraction de règles d'association**, **la découverte des motifs séquentiels**, **la recherche de structures classificatoires**.d) Enfin la dernière étape est de permettre une **réutilisation dans l'analyse de l'usage** des résultats obtenus par les techniques de fouille de données.

Les recherches en ECD appliquées aux données d'usage sont motivées par un double but : augmenter les usages d'un SI ou améliorer le SI en confrontant les informations structurelles du SI aux résultats de l'analyse de l'usage.

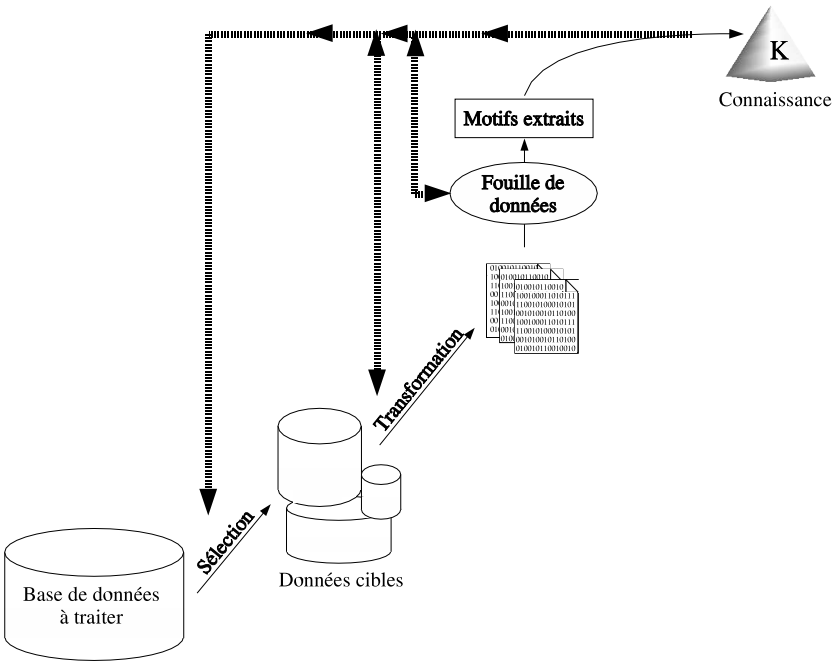

Figure 2. Les différentes étapes du processus d'ECD

3.2.1. a) Sélection et transformation des structures de données

Les méthodes ECD envisagées s'appuieront sur la notion de session dont la représentation peut être realisée avec l'aide d'un modèle tabulaire (items), d'un modèle des règles d'association (séquences d'items) ou enfin d'un modèle de graphe. Cette notion de *session* permet d'intervenir au bon niveau dans l'extraction de connaissances à partir des fichiers logs. Notre objectif est que nos outils de pré-analyse puissent construire des résumés et générer des statistiques sur ces résumés. A ce niveau de formalisation nous pouvons, introduire des règles et des graphes, définir des structures hiérarchiques sur les variables, extraire des séquences temporelles et donc, constituer de nouveaux types de données en utilisant des méthodes d'ECD.

Enfin, comme les méthodes d'analyse viennent de divers domaines de recherches (Analyse de données, statistique, Data Mining, Intelligence artificielle,...), une transformation des données en entrée et en sortie de ces programmes est nécessaire et sera contrôlée par des traducteurs. Les données en entrée seront issues des bases de données ou bien d'un fichier d'un format standard (XML) ou d'un format propriétaire.

Nous insistons sur l'importance de cette étape dans le processus d'extraction des connaissances.

3.2.2. b) Extraction des règles d'associations

L'objectif de nos outils de pré-analyse ou opérateurs de généralisation définis dans le paragraphe précédent est de construire des résumés et de générer des statistiques sur ces résumés. A ce stade de la formalisation, nous pouvons introduire des règles et des graphes, définir des structures hiérarchiques sur les variables, extraire des séquences temporelles et donc, constituer de nouveaux types de données en utilisant des méthodes de recherche d'ensembles fréquents ou des règles d'associations.

Ces méthodes ont été introduites en 1993 par R. Agrawal, T. Imielinski et A. Swami, chercheurs en bases de données au Centre de Recherche IBM d'Almaden. Elles sont aujourd'hui disponibles dans les logiciels du marché dits de « data mining » (Intelligent Miner d'IBM ou Entreprise Miner de SAS), essentiellement dans le domaine du commerce électronique.

Notre approche s'appuiera sur des travaux réalisées dans le domaine des opérateurs de généralisation et de construction de données agrégées. Ces résumés pourront être intégrés dans un mécanisme de recommandation pour l'aide à l'utilisateur.

Nous proposons d'adapter les méthodes de recherche d'ensembles fréquents ou des règles d'associations au Web Usage Mining. On pourra s'inspirer des méthodes utilisées dans le cadre du génome qui présente quelques similarités avec notre problématique. Si l'objectif de l'analyse peut se formuler dans un cadre décisionnel alors des classifieurs pourront identifier des groupes d'usage basés sur les règles extraites.

3.2.3. c) Découverte des motifs séquentiels

La connaissance de l'utilisateur permet une recherche des motifs séquentiels qui sont des règles d'associations entre sessions ordonnées dans le temps. Les résultats obtenus par les algorithmes d'extraction classiques ne sont pas suffisamment précis si l'on souhaite analyser de manière détaillée le comportement des utilisateurs ou des clients au cours du temps. Les possibilités offertes par l'analyse d'un log nous semblent dépasser le cadre de l'utilisation qui en est faire à l'heure actuelle. En effet nous pouvons envisager de travailler sur la qualité et la pertinence des résultats obtenus de différentes manières en prenant en compte la temporalité des résultats, en proposant une classification sur les résultats (pour mieux cibler les catégories de population) en considérant une information textuelle (pour filtrer avant d'analyser) ou encore en prenant en compte la structure du site lors de l'extraction de motifs.

3.2.4. d) Recherche des structures classificatoires

De plus l'intégration du niveau *session utilisateur* améliore la qualité de l'information à analyser et permet d'utiliser des méthodes de classification ayant une modélisation sous-jacente plus complexe. Partant de ces informations un découpage de notre population en groupes homogènes rend la modélisation de notre problème plus facile. L'objectif de cette classification est de pouvoir comparer les parcours des usagers dans le site de référence. L'extraction et l'interprétation de certains comportements types peut aider, d'une part le webmaster à restructurer son site et d'autre part les futurs usagers du site à rechercher une information.

En fonction des objectifs, nous proposons d'élaborer une méthode de classification bien adaptée. Cette adaptation sera exprimée sous la forme d'une optimisation d'un critère sous-jacent aux objectifs. Par exemple quand l'aide à la conception d'un site se résume en la mise en correspondance d'informations issues de l'architecture du site et de l'analyse des sessions d'usagers, alors la classification croisée peut être une bonne approche car elle permet de construire simultanément des classes de sessions et des classes de pages (rubriques).

Une étape de *validation* est nécessaire pour évaluer si la structure en classes déterminée par l'algorithme représente bien le jeu de données étudié. Il s'agit alors de tester si la valeur d'un critère d'adéquation entre la structure en classes obtenues et les données est significativement différente des valeurs prises par ce critère sur des jeux de données simulés sous une hypothèse d'absence de structure.

Une autre difficulté résulte de la taille de plus en plus conséquente de l'ensemble des données traitées. Dans ce cas, les techniques usuelles de la modélisation statistique ne peuvent être appliquées sans précaution, car sur un ensemble suffisamment grand d'indicateurs statistiques calculés pour un jeu de données sans structure, il est très probable que certains d'entre eux aient une valeur significative. Dans ce cas, notre approche consiste à réaliser une structuration de l'espace de représentation par un processus classificatoire puis de réaliser une

modélisation sur chaque classe obtenue. Cette modélisation est d'autant plus facile que la classe est homogène. Cette recherche de classes homogènes est maintenant indispensable dans l'analyse de grands ensembles de données pour la production de métadonnées dont le rôle est, non seulement, de "qualifier" la donnée mais surtout de guider son traitement.

D'autre part, le modèle d'indexation de situations comportementales[51] utilisé dans nos systèmes de recommandations nous permet, dans un site donné, d'extraire - soit on-line soit off-line à partir des logs - des comportements jugés intéressants par l'analyste puis de les indexer.

Ces travaux s'inscrivent dans une analyse off-line de l'usage et feront appel à des méthodes d'analyse exploratoire de données et des méthodes numériques de classification. Ceci peut prendre la forme d'une analyse statistique et d'une classification des sessions ou de profils utilisateurs, et mener à un mécanisme de recommandation afin d'aider l'utilisateur.

Nous proposons de développer une vaste famille de méthodes de classification afin de répondre aux problèmes de l'analyse de l'usage et d'introduire systématiquement dans ces méthodes une étape de validation afin de pouvoir avoir de bons indicateurs de comparaison. Nous nous proposons de fusionner les approches IA et AD dans le cadre supervisé afin de développer une boîte à outils contenant un ensemble de méthodes de classification dédiées à l'analyse de l'usage.

3.2.5. e) Réutilisation dans l'analyse de l'usage.

Ce thème vise à réutiliser un résultat d'analyse précédente dans l'analyse courante : nous envisageons à court terme un premier travail relatif à une approche incrémentale de découverte de motifs séquentiels et à plus long terme un deuxième relatif à une approche basée sur des techniques de raisonnement à partir de cas.

A l'heure actuelle des algorithmes très rapides ont été développés pour rechercher efficacement des dépendances entre attributs (algorithme de recherche de règles d'associations) ou des dépendances entre comportements (algorithme de recherche de motifs séquentiels) dans de grandes bases de données.

Malheureusement, même si ces algorithmes sont très efficaces, ils prennent, selon la taille de la base de données, entre plusieurs minutes et plusieurs jours pour extraire des informations pertinentes et utiles. Aussi, la variation des paramètres offerts à l'utilisateur (support minimal et confiance), nécessite de relancer les algorithmes sans tenir compte des résultats précédents. De la même manière lorsque de nouvelles données sont ajoutées ou supprimées de la base, il est souvent nécessaire de relancer le processus d'extraction pour maintenir la connaissance extraite. Etant donnée la taille des données manipulées il est alors indispensable de proposer une approche à la fois interactive (variation des paramètres) et incrémentale (variation des données de la base) pour répondre le plus rapidement possible aux besoins de l'utilisateur final. Cette problématique est à l'heure actuelle reconnue comme un problème de recherche ouvert dans le cadre du Data Mining et même si quelques propositions existent, elles ne sont malheureusement pas satisfaisantes car elles ne permettent que de répondre partiellement à la problématique.

3.3. Systèmes de recommandations personnalisées

Mots clés : *Web*, *hypermedia*, *fouille de données*, *extraction de connaissances à partir de données*, *KDD*, *raisonnement à partir de cas*, *CBR filtrage collaboratif*, *calcul de recommandations*.

L'objectif d'un système de recommandations est d'aider les utilisateurs à faire leurs choix dans un domaine où ils disposent de peu d'informations pour trier et évaluer les alternatives possibles [61][60][58].

Un système de recommandations peut être décomposé en trois entités de base (cf figure 3) : le groupe d'agents *producteurs* de recommandations, le module de *calcul de recommandations* et le groupe de *consommateurs* des recommandations.

Un défi majeur dans le domaine de la conception de systèmes de recommandations est le suivant :

> Comment produire des recommandations *personnalisées* et de haute *qualité* tout en *minimisant l'effort* requis de la part des producteurs et des consommateurs.

Deux grandes approches complémentaires sont proposées dans la littérature : 1) l'approche basée sur le contenu et fondée sur l'apprentissage automatique de profils utilisateurs et 2) l'approche dite de filtrage

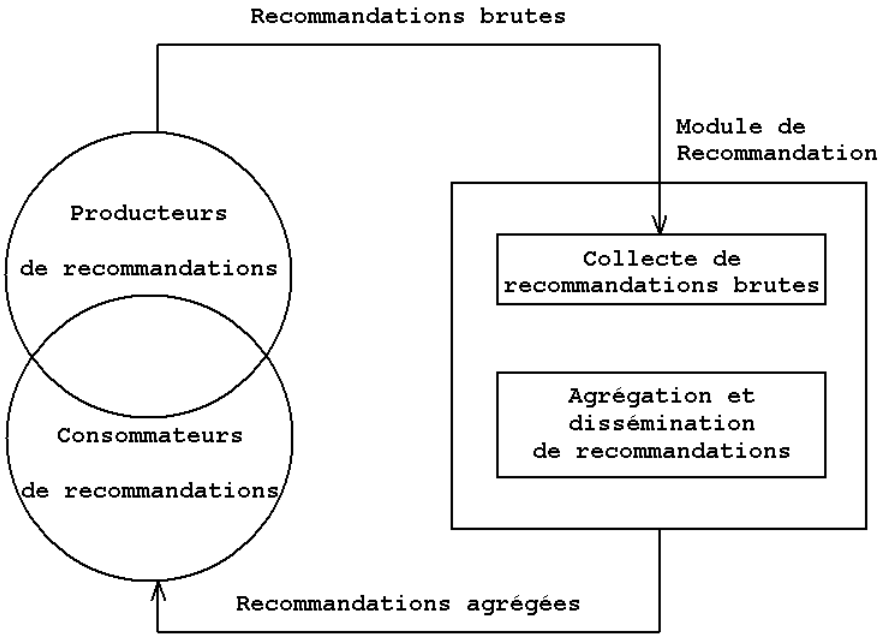

Figure 3. Architecture générale d'un système de recommandation

collaboratif fondée sur des techniques de fouille de données. Le profil utilisateur est une structure de données qui décrit les centres d'intérêts d'un utilisateur dans l'espace des objets à recommander. Celui-ci est une structure construite dans la première approche ou donnée dans la seconde par l'utilisateur . ce profil est utilisée soit pour filtrer les objets disponibles (on parle alors de filtrage basé sur le contenu), soit pour recommander à l'utilisateur ce qui a satisfait d'autres utilisateurs ayant un profil similaire (on parle alors de filtrage collaboratif) [60].

Dans l'action AxIS, nous poursuivons le développement d'une approche hybride de calcul de recommandations basée sur l'analyse du contenu visité et centrée fouille de données où les comportements passés d'un **groupe** d'utilisateurs sont utilisés pour calculer les recommandations (cf. filtrage collaboratif). La plupart des autres approches fondées sur la fouille de données sont principalement des approches statistiques où l'ordre d'occurrence d'événements dans l'historique n'est pas pris en compte lors du calcul de recommandations. Citons comme exemple, dans le domaine d'aide à la navigation sur le Web, le système FootPrints [66] et le système de Yan et al [67].

Les problèmes difficiles pour la mise en œuvre de notre approche concernent les aspects suivants :

1. fournir des techniques d'identification et d'extraction de comportements pertinents (i.e. des enseignements ou des cas) à partir des données brutes des historiques,
2. définir des méthodes et des techniques de mesure de similarités entre comportements,
3. définir des techniques d'inférence de recommandations personnalisées à partir des comportements pertinents passés identifiés (ou à partir des cas remémorés).

Nous étudions l'ensemble des trois problèmes ci-dessus en explorant la possibilité d'application des techniques RàPC et plus généralement de l'ECD.

Nous étudions la classe de systémes de recommandations s'appuyant sur une **réutilisation d'expériences passées** issues d'un groupe d'utilisateurs utilisant des techniques de raisonnement à partir de cas (RàPC). Dans cette classe de systèmes, nous privilégions les deux types de systèmes suivants :

- ceux dont le calcul de recommandations se base sur une réutilisation d'expériences d'un groupe d'utilisateurs **recherchant de l'information** dans un système d'informations hypertexte comme le Web ou dans un site internet/intranet. Ces systèmes visent une **aide personnalisée** à l'activité de recherche d'informations ;
- ceux dont le calcul de recommandations s'appuie sur une **réutilisation d'expériences** passées issues d'**experts** en vue de fournir un aide à des activités de **conception**.

3.4. Raisonnement à partir de cas

Mots clés : *gestion de l'expérience*, *réutilisation d'expériences passées*, *indexation*, *raisonnement à partir de cas case-based reasoning*, *séquences temporelles*.

Glossaire

Raisonnement à partir de Cas (RàPC). Se dit d'une approche de résolution de problèmes basée sur la réutilisation par analogie d'expériences passées appelées « cas ». Un cas est généralement indexé pour permettre de le retrouver suivant certaines caractéristiques pertinentes et discriminantes, appelées « indices » ; ces indices déterminent dans quelle situation (ou contexte) un cas peut être de nouveau réutilisé.

Le raisonnement à partir de cas [59] se décompose habituellement en quatre phases principales [45][57] :

1. une phase de recherche de cas ayant des similarités (i.e. des indices similaires) avec le problème courant,
2. une phase de réutilisation, permettant de construire une solution au problème courant en se basant sur les cas identifiés dans la phase précédente,
3. une phase de révision de la solution qui permet d'affiner cette dernière grâce à un processus d'évaluation,
4. une phase d'apprentissage, dont le but est de mettre à jour les éléments du raisonnement en prenant en compte l'expérience qui vient d'être réalisée et qui pourra ainsi être utilisée pour les raisonnements futurs.

Les problèmes difficiles en RàPC sont très fréquemment liés : à la définition et la représentation d'un cas, l'organisation de la base de cas, les diverses indexations utilisées et la définition de « bonnes » mesures de similarités pour la recherche de cas, le lien recherche-adaptation de cas (le meilleur cas étant le cas le plus facilement adaptable), la définition d'une stratégie d'adaptation à partir du (ou des) cas retrouvé(s), l'apprentissage de nouveaux indices, etc.
Nous poursuivons l'évaluation de nos résultats en RàPC en particulier notre modèle d'indexation par situation comportementale, notre plate-forme orientée objet CBR*Tools et boîte à outils Broadway*Tools. De plus, nous étudions plus particulièrement des techniques d'indexation des sessions et algorithmes de recherche de sous-séquences fréquentes d'items pour l'analyse en ligne et hors ligne de l'usage des internautes.

4. Domaines d'application

4.1. Panorama

Mots clés : *système d'information*, *Web*, *multimédia*, *commerce électronique*, *e-marketing*, *e-CRM*, *santé*, *ingénierie*, *éducation*, *environnement*, *télécommunications*, *conception de sites Web*, *conception de services adaptatifs*, *recherche d'information*, *analyse des usages*, *analyse des logs*, *personnalisation*.

Le projet vise tout domaine applicatif portant sur la conception, l'évaluation et l'amélioration d'un système hypermédia d'informations de grande taille où la prise en compte de l'utilisateur final est primordiale. A

court terme, nous nous focalisons sur des sites Web (internet, intranet) ou parties de sites disposant d'une des caractéristiques suivantes :

1. Présence ou intégration souhaitée de services d'aide à la recherche collaborative d'informations et de personnalisation (ranking, filtrage, ajout de liens, etc.) ;
2. Fréquente évolution du contenu générant de nombreux problèmes de maintenance : citons, par exemple,

   – Site témoignant des activités d'un groupe de personnes, celui-ci pouvant être une institution (ex, Inria), une entreprise, une communauté scientifique, un réseau européen sur internet ou intranet, etc.
   – Site indexant un ensemble vaste de productions (documents, produits) issues du Web ou d'une organisation (entreprise) selon un critère thématique par exemple : citons les moteurs de recherche (Yahoo, Voila), les guides internet pour des cibles données (Educadoc de FT) ou portails (communautés scientifiques).

3. Interprétation de la satisfaction utilisateur (selon le point de vue concepteur) ou satisfaction explicite par les utilisateurs : citons, à titre d'exemples, les sites commerciaux ou les sites e-learning mais aussi les moteurs de recherche.

En fait, nous nous intéressons (cf. rapport d'activités Inria 2001 de l'action AID) aux points suivants :

- Vérification sémantique d'un système d'informations,
- Analyse des usages d'un système d'information (internet, intranet),
- Reconception d'un système d'informations basée sur une analuyse de l'usage,
- Recherche collaborative d'informations sur le web.

Enfin, il est à noter, que d'autres domaines (santé, transports, etc.) peuvent être soumis à l'étude dans la mesure où ils offrent un cadre expérimental de validation de nos travaux de recherche en ECD et en réutilisation d'expériences en gestion d'historiques : ce type d'approche peut être jugée pertinente dans des applications mal résolues en automatique de type contrôle (par exemple nutrition de plantes sous serres, commande en robotique etc.).

5. Logiciels

5.1. Introduction

Mots clés : *informatique décisionnelle*, *service Web*, *personnalisation*, *gestion de l'expérience*, *raisonnement à partir de cas*, *plate-forme orientée objet*, *réutilisation*, *UML*, *patron de conception*, *système de recommandations*, *filtrage collaboratif*, *coopération*, *argumentation*, *Web*, *langages de spécification*, *sémantique formelle*, *CLF*, *CBR*Tools*, *Broadway*Tools*, *Java*, *Prolog*.

Les logiciels de l'équipe sont conçus pour la plupart avec l'atelier de conception objet *Rational Rose* et réalisés avec le langage de programmation Java.

5.2. CLF - « Computer Language Factory »

Mots clés : *langages de spécification*, *sémantique*, *Centaur*, *Prolog*.

Participant : Thierry Despeyroux [correspondant].

CLF, développé par Thierry Despeyroux (dans l'ex-projet CROAP) est un ensemble d'outils et de formalismes de spécification de la syntaxe et la sémantique de langages informatisées. CLF propose actuellement les langages AS [47] (Abstract Syntax) et CS (Concrete Syntax).

Une partie de CLF a été directement adaptée au monde Prolog comme une extension des DCG (« Definite Clause Grammars »). Par rapport aux DCG, notons une plus grande facilité d'expression due à la possibilité de récursions gauches et à la compilation de certaines règles de grammaire permettant d'utiliser certaines optimisations très importantes des compilateurs Prolog (indexation de clauses). De plus l'analyseur ainsi réalisé contient une méthode générique permettant de faire le lien entre les occurrences du terme construit et les positions textuelles dans le texte source, une caractéristique qui manque à la plupart des générateurs d'analyseurs lexicaux.

Cette extension des DCG a été en 2001 utilisée pour contruire un analyseur XML utilisé pour nos travaux sur la vérification sémantique de systèmes d'informations écrits en XML (cf. section 6.2).

5.3. Boite à outils de classification

Mots clés : *classification automatique*, *classes*, *visualistaion*, *cartes de Kohonen*.

Participants : Yves Lechevallier [correspondant], Marc Csernel [correspondant], Brigitte Trousse.

Cette boite à outils écrite en C++ (sous linux et windows) permet de regrouper les outils et les méthodes de classification développés dans l'équipe et utilise notre bibliothèque de transformation et d'exploitation des données développée par M. Csernel, proposant ainsi une interface commune entre méthodes. Cette boite à outils permet à chacun d'intégrer facilement sa méthode de classification, de la tester et de la comparer avec les autres méthodes.

Cette année deux interfaces utilisateur ont été réalisées pour permettre un usage soit via une application C++ soit via un serveur intranet Web..

5.4. CBR*Tools - Plate-forme objet en raisonnement à partir de cas

Mots clés : *raisonnement à partir de cas*, *plate-forme objet*, *composants logiciels*, *réutilisation*, *UML*, *patron de conception*.

Participants : Sémi Gaieb, Sébastien Siémard, Brigitte Trousse [correspondante].

CBR*Tools est une plate-forme à objets développée dans l'équipe depuis 97 pour faciliter le développement d'applications nécessitant des techniques de raisonnement à partir de cas.

CBR*Tools [51] [54] est une plate-forme à objets (ou « object-oriented framework » [55][48]) en RàPC, qui offre un ensemble de classes abstraites modélisant les principaux concepts nécessaires pour développer une application intégrant des techniques de raisonnement à partir de cas : cas, bases de cas, index, mesures de similarité, contrôle du raisonnement. Elle offre également un ensemble de classes concrètes qui implantent de nombreuses méthodes classiques (indexation par plus proches voisins, indexation par Kd-tree [65], indexation par prototypes [52], indexation basée sur une approche neuronale, mesures de similarités standards). CBR*Tools comporte actuellement plus de 200 classes avec notamment deux grands groupes : le package *core* pour le fonctionnement de base et le package *time* pour la gestion spécifique des situations comportementales. La programmation d'une nouvelle application se fait par spécialisation de classes existantes, par composition d'objets ou en utilisant les paramètres des classes existantes.

CBR*Tools vise tout particulièrement des domaines d'application nécessitant une réutilisation de cas devant être indexés par des situations comportementales.

CBR*tools a été installée à France télécom (R&D) à Lannion en 1998 et 2000 dans le cadre de `Broadway-Web` et `educaid` (FT-CTI) et a été utilisée dans le cadre d'un contrat XRCE-INRIA (98). Une documentation sur le Web est accessible à l'adresse suivante :
http://www-sop.inria.fr/axis/cbrtools/manual/.

La plate-forme CBR*Tools a été évaluée via la conception et la réalisation de cinq applications (Broadway-Web, educaid, BeCKB, Broadway-Predict, RA2001). Nous avons montré que, pour chaque application, l'expertise approfondie nécessaire pour utiliser CBR*Tools ne concerne que 20% à 40% des points d'ouverture validant ainsi l'aide apportée par notre plate-forme tant sur la modélisation que sur l'implantation, grâce à la réutilisation de son architecture abstraite et de ses composants (index, similarité).

5.5. Broadway*Tools - Génération de systèmes de recommandations adaptatifs

Mots clés : *agent personnalisé*, *boîte à outils*, *système distribué*, *programmation asynchrone*, *composants logiciels*, *réutilisation*, *services adaptatif*, *personnalisation*.

Participants : Sémi Gaieb, Sébastien Siemard, Brigitte Trousse [correspondante].

Broadway*Tools est une boîte à outils pour faciliter la réalisation de systèmes de recommandations Web adaptatifs pour l'aide à la recherche d'informations sur le Web ou dans un site internet/intranet. Cette boîte à outils offre actuellement différents serveurs dont un serveur de calcul de recommandations basé sur notre modèle d'indexation comportementale pour l'observation des sessions utilisateurs et sur la réutilisation de sessions passées d'un groupe d'utilisateur. Un système de recommandations réalisé avec Broadway*tools observe les navigations de différents utilisateurs et récolte les évaluations et les annotations de ces utilisateurs pour établir une liste de recommandations pertinentes (documents Web, mots clés, etc.).

Cette boîte à outils a été utilisée pour deux systèmes d'aide à la navigation :
Broadway-Web (ex Broadway-V1) pour l'aide à la navigation sur le Web et Broadway-educaid pour l'aide à la navigation dans un clone d'un site de France Télécom.

5.6. Broadway-Web - Aide personnalisée à la navigation sur le Web

Mots clés : *filtrage collaboratif*, *système de recommandations*, *aide à la navigation*, *Web*, *analyse réutilisation*, *comportements utilisateurs*, *profils utilisateurs*.

Participants : Sergiu Chelcea, Semi Gaieb, Brigitte Trousse [correspondante].

Broadway-Web[2] [53] est un assistant pour la navigation sur le Web réutilisant les navigations passées d'un groupe d'utilisateurs. Broadway-Web observe les navigations de différents utilisateurs et récolte les évaluations et les annotations de ces utilisateurs pour établir une liste de documents pertinents.

Broadway-Web (http://www-sop.inria.fr/axis/broadway/), issu de la thèse de M. Jaczynski (1998), est un serveur HTTP utilisé comme proxy : il est inséré entre le navigateur et le reste du web et il intercepte ainsi toutes les demandes de documents pour le protocole HTTP. Broadway est alors capable d'observer les différentes navigations des utilisateurs en enregistrant notamment : les adresses des documents visités, un ensemble de mots clefs issus de l'analyse automatique des pages HTML et les évaluations des documents. Durant une navigation, Broadway peut afficher un ensemble de documents qu'il conseille suivant l'état courant de la navigation, et permet aux utilisateurs d'évaluer ou d'annoter les documents traversés grâce à une barre d'outils insérée dynamiquement dans les pages HTML visualisées. Broadway intègre le serveur HTTP Jigsaw du W3C programmé en Java et utilise notre plate-forme CBR*Tools pour implanter le système de raisonnement à partir de cas permettant la réutilisation des navigations passées.

Broadway-Web (ex Broadway-V1) a été installée à France Télécom (R&D) à Lannion en 98 et a fait l'objet d'une démonstration aux huitèmes et neuvièmes rencontres INRIA-Industrie (novembre 1998 et décembre 2001) ainsi qu'à lm conférence TTNL'02 à Sophia Antipolis en novembre 2002 (cf. 5) à Sophia Antipolis..

6. Résultats nouveaux

6.1. Panorama

Outre nos résultats sur les aspects statiques d'un SI (cf. 6.2), les résultats sur les aspects dynamiques sont ordonnés en fonction des étapes du processus d'ECD : a) sélection et transformation des données (cf. 6.3), b) fouille de données, (cf. 6.4, 6.5, 6.6, 6.7) et utilisation des motifs extraits pour des systèmes de recommandations basés sur l'analyse de l'usage(cf. sections 6.8, 6.9 et 6.10). En cette année de création d'AxIS, certains de nos résultats sont liés à nos objectifs applicatifs, d'autres sont plus théoriques.

[2] Brodway - « BROwsing ADvisor reusing pathWAYs »

6.2. Aide à la vérification sémantique de sites Web

Mots clés : *sémantique*, *sites Web*, *services adaptatifs*, *Web Semantics*, *approches formelles*, *sémantique naturelle*, *typage*, *vérification*, *Centaur*, *CLF*, *adaptation à l'utilisateur*, *personnalisation*.

Participants : Thierry Despeyroux, Brigitte Trousse.

Nous avons poursuivi notre évaluation de l'apport d'une sémantique de style sémantique naturelle [56] pour la spécification et la vérification d'une page Web, voire d'un site Web.

Notre motivation au niveau du « Web Sémantique » concerne principalement l'aide à la spécification et à la vérification de sites Web. Très peu de travaux abordent la vérification sémantique de sites, que ce soit avec des techniques issues de l'IA ou du génie logiciel : citons l'un d'eux (en IA) WebMaster [64].

Notre approche est inspirée de travaux précédents en sémantique des langages de programmation, traçant un parallèle entre la syntaxe des langages de programmation et la structure des sites Web (ou de documents semi-structurés) et entre la sémantique des programmes et la sémantique des sites Web, appliquant des notions de types et de règles sémantiques aux documents présents sur le Web.

En 2001, nous nous étions d'une part affranchis du système Centaur utilisé précédemment, d'autre part nous nous étions intéressés à des données réelles (comme des pages Web du site de l'Inria et aussi des données d'analyse générées en XML à partir d'une base de données).

Si on rapproche la vérification de systèmes d'information de la notion de compilation de programmes, apparaît rapidement la nécessité d'exprimer des règles de dépendance comme celles qui apparaissent dans un " makefile ". Cependant, la complexité de la structure d'un système d'information nous a fait toucher les limites d'un programme tel que " make " : manipulation d'un grand nombre de fichiers et de répertoires, délocalisation, utilisation de données accessibles par leur URL etc.

Cette année, suite à la réalisation de maquettes en utilisant Prolog et Make, nous avons commencé le design d'un langage de spécification permettant d'exprimer des règles de contraintes sémantiques. Bien que cachant en grande partie l'esprit "sémantique naturelle", qui est sans doute trop hermétique pour l'utilisation que nous voulons en faire, notre but est de générer un code exécutable qui reste bien, lui, dans l'esprit de la sémantique naturelle.

Ce langage de spécification peut être vu comme une extension de XML, avec en particulier des variables logiques. Le développement d'un analyseur pour ce langage est largement facilité par le fait que nous disposons d'un analyseur pour XML construit en utilisant CLF.

Nos travaux ont fait l'objet d'une présentation lors des journées de l'action STIC CNRS Web sémantique [24].

6.3. Sélection et transformation des structures de données

Participants : Mireille Arnoux, Marc Csernel, [F. A. T. De Carvalho], Aicha El Golli, Nathalie Evan, Sémi Gaieb, Miha Jurca, Yves Lechevallier, Doru Tanasa, Brigitte Trousse, [Rosanna Verde].

6.3.1. Extraction et construction de données agrégées

Mots clés : *généralisation*, *données agrégées*.

En analyse de données, la méthode de généralisation permet d'agréger les informations d'une base de données en décrivant des concepts sous-jacents aux données. Elle est non seulement un outil descriptif pour l'utilisateur mais aussi une étape intermédiaire permettant d'autres analyses sur ces concepts. Ayant un ensemble d'individus G de Ω, ensemble de la population, le but de la généralisation est de construire une bonne représentation de G par un vecteur multidimensionnel résumant toutes les descriptions des individus (réel, binaire, catégories ou modalités...). Une méthode déjà adoptée consiste à associer à ce vecteur un poids et un scalaire résumant la dispersion de ces individus. Une seconde approche proposée dans le cadre de l'analyse de données symbolique, permet de résumer un ensemble d'observations par une description symbolique (intervalle, distribution, ...).

Cette année nous avons développé un nouvel opérateur de généralisation [25]. Celui-ci construit un ensemble de classes homogènes qui peuvent être modèlisées sous la forme d'objets symboliques. L'objectif

est d'extraire d'un ensemble de bases de données relationnelles, en utilisant l'opérateur de généralisation proposé par Véronique Stéphan dans sa thèse, un ensemble d'assertions décrivant les concepts à analyser. Si la description de certains concepts est assez hétérogènes alors nous proposons une méthode de classification divise afin améliorer la qualité de cette description. Cette nouvelle description est formalisée par une disjonction d'assertions. Cette approche a été introduite dans DB2SO utilisé par le logiciel d'analyse de données symboliques SODAS dans le cadre du projet européen ASSO. Ce dernier permet de créer des objets symboliques à partir des bases de données relationnelles. Dans la première version de ce logiciel, DB2SO permettait de regrouper les observations appartenants au même groupe et les généraliser par une description symbolique qu'on appelle "assertion". Certaines de ces assertions incluent des observations atypiques. Dans la nouvelle version des améliorations ont été effectuées. Ces améliorations permettent de généraliser chaque groupe par une disjonction d'assertions, ce qui a permis d'avoir des descriptions plus homogènes et de meilleure qualité. Pour cela, une étape de décomposition a été ajoutée au processus de généralisation. Cette décomposition appliquée à chaque groupe est basée sur un algorithme de classification divisive.

6.3.2. Forme Normale Symbolique appliquée à la transformation de données

Mots clés : *distance*, *espace de description*, *Forme Normale Symbolique*.

La Forme Normale Symbolique (FNS), inspirée de la 3ème Forme Normale des base de données relationnelles, consiste à factoriser les descriptions des objets symboliques selon les contraintes exprimées par des règles entre les variables de telle façon que seulement la partie cohérente de ces objets soit représentée.

Une nouvelle version de notre bibliothèque de transformation et manipulation de données (écrite en C++) [21][22] a été mise au point par M. Csernel avec l'aide des étudiants de l'université de Recife en collaboration avec F.A.T. de Carvalho (cf. projet CLADIS). Cette nouvelle version intègre la possibilité de travailler avec des données intervalles.

Un premier essai, réalisé en collaboration avec Rosanna Verde, de traduction manuelle au format F.N.S d'un tableau de données intervalles nous a permis d'évaluer la possibilité d'introduire cette nouvelle version dans les programmes actuellement développés dans ASSO.

6.3.3. Pré-traitement de logs HTTP multi-sites

Mots clés : *logs HTTP*, *fusion de données*, *multi sites*, *Web*.

Nous avons établi une méthodologie pour le pré-traitement des fichiers logs Web [31] avant d'appliquer des méthodes de fouille de données sur ces logs issus de sites différents. Le pré-traitement consiste dans le filtrage et nettoyage de ces logs dans le but d'éliminer les requêtes pour les images, les requêtes provenant des robots Web et former en suite des sessions de navigations. Les données ont été rendus anonymes pour permettre en suite l'analyse par des étudiants du STID. Nous avons implémenté cette méthode sous forme des scripts Perl que nous aons appliquée sur trois mois de logs Web (Novembre 01 - Janvier 02).

6.3.4. Persistance et exploitation de logs comportementales des utilisateurs d'un site Web

Mots clés : *persistance*, *modèle relationnel*, *logs*, *Web*.

Cette année nous avons revisité le modèle relationnel de `logaudience` issue d'une première capitalisation des diverses modélisations des sessions des internautes naviguant sur le Web ou dans un site Web réalisées ces dernières années (contrat FT CTI, prédiction du comportement utilisateur, etc.) lors de la conception de systèmes de recommandations.

Cette année fut l'occasion dans le cadre de notre projet Colors `e-behaviour` (cf. 6.8) de revisiter notre modèle relationnel de notre boite à outils `logaudience` de sorte 1) à bien séparer les données brutes des données d'usage en vue d'une analyse et 2) à y intégrer les traces des comportements visuels des internautes. Ce travail s'est fait dans le cadre du projet DESS de Nathalie Evan (2001-2002) (resps : M. Arnoux et S. Gaieb) s'intitulant : "Persistance et exploitation des données comportementales des utilisateurs d'un site web" [35] Celui-ci s'intégrait dans le développement d'un système de recommandation adaptatif basé sur l'approche Broadway. Un effort particulier a été fat sur la structuration et le traitement des données relatives aux comportements visuels et non visuels des internauets. Le suivi des données comportementales était assez

fin car enrichi par un suivi visuel (dispositif Eye tracking) ; de plus il était important de collecter des données historisées et aussi de les agréger notamment par rapport aux sessions des utilisateurs. Nous avons donc choisi de construire un entrepôt de données. Les faits essentiels correspondent aux accès aux pages web avec pour dimensions principales la session, le temps et la page. La dimension page peut être considérée sous l'angle de plusieurs hiérarchies selon une organisation physique ou logique ; nous avons retenu tout d'abord la vision qui correspond à un annuaire thématique avec un parcours arborescent illustré par l'annuaire des rapports INRIA de 2001 (cf. 4).

6.3.5. Transformation de données spatiales

Mots clés : *Analyse des données symbolique*, *fouille de données spatiales*.

L'encadrement du stage de DEA de Tao Wan a été réalisé conjointement par Karine Zetouni de l'UVSQ et Yves Lechevallier [44]. La fouille de données spatiales (dite **data mining spatial**) répond à un besoin réel de nombreuses applications en permettant de tirer profit de la disponibilité croissante de données localisées et de leur richesse potentielle. Il se caractérise par l'introduction des relations spatiales dans l'analyse.

L'objectif de ce stage est de réaliser une étude bibliographique sur le sujet de la fouille de données spatiales et d'introduire les spécificités des données spatiales dans les méthodes d'arbre de décision.

Cette étude s'est appuyée sur une base de données correspondant à un échantillon représentatif d'accidents routiers (plus de 10 000 accidents) fournie par l'INRETS.

Durant ce stage une conversion du schéma de la base de données relationnelles des accidents routiers en un tableau de données symboliques a été faite. Le logiciel SODAS a été utilisé pour réaliser une analyse de ce tableau.

6.3.6. Améliorations de notre boite à outils de classification

Cette année les améliorations de notre boite à outils de la classification ont principalement été sur les deux parties suivantes :

6.3.6.1. Interfaces communes

Le travail a consisté à regrouper les outils et les méthodes de classification (utilisant notre bibliothèque de transformation et d'exploitation des données développées dans AxIS via une interface commune. Ainsi chacun peut intégrer facilement sa méthode de classification, la tester et la comparer avec les autres méthodes. Actuellement deux interfaces ont été réalisées : une en C++ et une sur le Web (stage de M. Jurca) offrant ainsi un serveur intranet dans l'équipe.

6.3.6.2. Visualisation des cartes de Kohonen

La représentation "classique" des cartes de Kohonen consiste en un ensemble de U cellules disposées en réseau. La structure du réseau est libre mais généralement une grille est choisie comme structure de visualisation. Durant son stage de DEA, Tarek Ait-Mohamed [33] (resp : Y. Lechevallier) a proposé trois nouvelles formes de description des cellules et une première version écrite en Java. Ces descriptions ont été construites à partir du vecteur de pondération associé à chaque cellule de la carte.

6.4. Méthodes de partitionnement et validation

Participants : Patrice Bertrand, [Marie Chavent], Marc Csernel, [F. A. T. De Carvalho], Achia El Golli, Yves Lechevallier, [Rosanna Verde].

Mots clés : *clustering*, *données agrégées*, *distances*, *intervalles*, *distance de Hausdorff*, *cartes topologiques*, *données symboliques*, *stabilité*, *validation*, *analyse de données symboliques*.

6.4.1. Classification non supervisée à partir de dissimilarités mesurant le lien entre représentations complexes des données

Nous avons travaillé sur une approche classificatoire [23] dont l'objectif est l'obtention d'un partitionnement d'un grand nombre d'objets en un nombre réduit de classes homogènes à partir d'un tableau de dissimilarités calculé sur ces objets. L'algorithme choisi est issu de l'algorithme des Nuées Dynamiques sur un tableau

de dissimilarités dont le critère de classification est basé sur la somme des dissimilarités entre les individus appartenant à la même classe. L'algorithme proposé fait décroître ce critère.

Le choix d'une mesure de proximité est nécessaire : aussi, durant le projet CLADIS, nous avons étudié diverses mesures de proximité, utilisables sur un tableau de descriptions d'objets ayant une structure complexe.

Lors du calcul de la mesure de proximité il est nécessaire de tenir compte de la variabilité, (liée aux valeurs observées sur chaque variable) et de la connaissance du domaine exprimée par des règles (dépendances entre variables). Nous les avons modélisées en créant deux types de dépendances, l'une est basée sur la structure hiérarchique, l'autre est logique. Concernant cette famille d'indices, nous proposons deux approches :

- Les indices de la première approche utilisent pour chaque variable une fonction de comparaison suivie d'une fonction d'agrégation. La fonction de comparaison, utilisant des opérateurs symboliques (jonction, conjonction), est basée sur la différence de contenu et de position. La fonction d'agrégation s'inspire de la métrique de Minkowsky. La prise en compte des contraintes se fait par la pondération de chacune des valeurs ;
- Les indices de la seconde approche n'utilisent que la fonction de comparaison. La comparaison entre une paire d'objets est réalisée globalement par une fonction qui utilise des opérateurs symboliques et des mesures positives.

Le problème majeur de toutes ces approches est l'aspect combinatoire du calcul lors de la prise en compte des dépendances logiques. Il est linéaire en fonction du nombre de variables et, malheureusement, exponentiel en fonction du nombre de règles. Pour depasser cette difficulté, nous proposons un approche par décomposition des descriptions symboliques selon la Forme Normale Symbolique (cf. 6.3) que nous sommes en train d'étudier.

Un prototype de cette méthode a été réalisé dans le cadre du projet CLADIS entre l'INRIA et le CNPq.

6.4.2. Classification non supervisée à partir d'un tableau de données intervalles

Nous avons étudié [15][27][32] trois nouvelles approches classificatoires d'un tableau de données à structure complexe. Deux approches sont applicables aux tableaux d'intervalles, la dernière peut être appliquée aux tableaux issus d'un ensemble de variables multivaluées.

Ces trois approches utilisent des algorithmes de classification de type Nuées Dynamiques optimisant un critère lié aux distances entre les objets à classer ou bien mesurant l'adéquation entre un ensemble de prototypes (noyaux, centroides) et une partition de ces d'objets, les prototypes étant une modélisation d'une classe de cette partition.

Dans la première approche, les prototypes sont des éléments de l'espace de représentation des objets à classer c'est-à-dire un vecteur dont les coordonnées sont des intervalles. La distance entre un prototype et un individu est basée sur la distance de Hausdorff entre deux vecteurs d'intervalles. La distance de Hausdorff entre deux intervalles ξ_k^j et ξ_l^j est égale à :

$$\delta_j(\xi_k^j, \xi_l^j) = max\{|\alpha_k^j - \alpha_l^j|, |\beta_k^j - \beta_l^j|\}$$

La distance d entre deux vecteurs d'intervalles $\xi_k = (\xi_k^1, ..., \xi_k^p)$ et ξ_l est alors une combinaison des distances de Hausdorff $\delta_1, ..., \delta_p$ d'où :

$$\begin{array}{rcl} d & : & \Omega \times \Omega \longrightarrow \Re^+ \\ & & (k,l) \longrightarrow d(\xi_k, \xi_l) = (\sum_{j=1}^{p} max\{|\alpha_k^j - \alpha_l^j|, |\beta_k^j - \beta_l^j|\}^2)^{1/2} \end{array}$$

Dans la seconde approche, les prototypes sont des vecteurs de distributions calculés à partir de systèmes de pondérations associés à un ensemble d'intervalles, dits *intervalles élémentaires*. Dans ce cas on parlera de *prototypes généralisés*.

Dans la troisième approche, le prototype généralisé et les individus ne sont plus représentables dans le même espace de description. La mesure de comparaison utilisée n'est donc pas une dissimilarité mais une fonction de comparaison (" matching "). Cette fonction est constituée de deux composantes car elle intègre non seulement l'écart entre deux intervalles mais aussi le système de pondération de ces deux intervalles.

Un prototype comprenant ces trois approches a été réalisé dans le cadre du projet CLADIS entre l'INRIA et le CNPq et du projet européen ASSO.

6.4.3. Algorithme des cartes topologiques auto-organisatrices et données symboliques

L'algorithme des cartes topologiques auto-organisatrices (SOM : *Self Organising Map*), introduit par Kohonen, est un outil structurant les relations entre classes via un réseau de neurones, et réduisant la dimension des données initiales tout en préservant, au moins partiellement, la topologie de l'espace des variables. Il permet également de visualiser les données dans un espace de dimension faible, généralement égale à 2. Ces propriétés constituent des avantages considérables par rapport aux algorithmes de partitionnement, surtout en phase d'exploration. Les cartes topologiques réalisent une interface entre les méthodes de classification et les techniques de réduction des données [41].

L'algorithme des cartes topologiques proposé par Kohonen est un procédé d'auto-organisation qui cherche à projeter des données représentées dans un espace de grande dimension, dans un espace de faible dimension. En fin d'apprentissage, le but du réseau est de reproduire sur la carte de sortie les corrélations présentes dans les données d'entrées.

Cette année, dans le cadre de la thèse d'Aicha El Golli, nous avons proposé une nouvelle version « batch » de l'algorithme des cartes topologiques qui prend en entrée des tableaux de dissimilarités. Dans cette version les prototypes ne sont plus recalculés à chaque fois qu'on présente une observation mais après une phase d'affectation de l'ensemble d'apprentissage. C'est une version non stochastique de l'algorithme des cartes topologiques. La méthode ainsi proposée répond aux deux objectifs suivants : 1) traiter aussi bien les données classiques que les données symboliques, 2) fournir une interprétation symbolique des neurones et donc des classes obtenues. Chaque neurone sera caractérisé par un prototype qui est modélisé par un objet symbolique.

6.4.4. Méthode d'évaluation de la stabilité d'une classe

Dans le cadre du contrat européen ASSO, nous avions déjà proposé une méthode d'évaluation de la stabilité d'une classe (arbitraire) générée par une méthode de partitionnement. La stabilité d'une classe est évaluée à partir de deux statistiques qui estiment le degré de stabilité inhérent à l'isolation et à la compacité de cette classe. Les valeurs prises par ces statistiques sont évaluées à l'aide d'un test de Monte-Carlo, l'hypothèse nulle de ce test étant l'absence de structure en classe de l'ensemble des données.

Cette année, P. Bertrand et Y Lechevallier ont défini une méthodologie à suivre pour interpréter les valeurs prises par ces statistiques de stabilité dans le cas de données symboliques. Dans le cas es données symboliques la principale difficulté provient du fait que l'hypothèse nulle utilise l'enveloppe convexe calculée sur la représentation dans un espace euclidien des données (i.e. des représentations sous forme d'intervalles, de lois de probabilité discrètes non métriques, de valeurs ensemblistes, ...) et qu'il se produit souvent une explosion combinatoire qui est liée au nombre de coordonnées à prendre en compte.

6.5. Extensions du modèle de la classification hiérarchique

Mots clés : *hiérarchie*, *clustering*, *CAH*, *lien minimum*, *complexité*.

Participants : Patrice Bertrand, Sergiu Chelcea, Gentian Guscho, Laurent Jullien, Brigitte Trousse.

6.5.1. La Structure 2-3 hiérarchique

P. Bertrand s'est intéressé à un type particulier de classification pyramidale : la *classification 2-3 hiérarchique* [14][13]. Plus précisément, une collection $\mathcal{C}$ de parties non vides d'un ensemble fini E est appelée 2-3 hiérarchique si pour chaque partie $X \in \mathcal{C}$, il existe au plus une partie $Y \in \mathcal{C}$ telle que $X \cap Y \notin \{\varnothing, X, Y\}$, autrement dit si pour tout $X \in \mathcal{C}$, il n'existe pas plus d'un élément $Y \in \mathcal{C}$ tel que $X \cap Y$, $X - Y$ et $Y - X$ ne soient pas vides. Nous avons déterminé quelques propriétés des collections 2-3 hiérarchiques, en particulier

nous avons montré que ce sont des collections d'intervalles pour au moins un ordre total défini sur E, et que le nombre maximal d'éléments (non réduits à des singletons) d'une collection 2-3 hiérarchique est égale à $\lfloor \frac{3}{2}(\mid E \mid -1) \rfloor$.
Nous avons également proposé le principe d'un algorithme de *Classification Ascendante 2-3 Hiérarchique (2-3 CAH)* qui est naturellement associé à ce type de structure et qui généralise l'algorithme bien connu de la CAH. Finalement, nous avons prouvé une extension de la bijection de Johnson-Benzécri (entre hiérarchies indicées et ultramétriques) au cas des 2-3 hiérarchies indicées au sens large.

6.5.2. Lien minimum pour l'Algorithme de Classification 2-3 hiérarchique

Le stage de Laurent Jullien (avril - septembre 2002) (responsable : P. Bertrand) a porté essentiellement sur l'étude des propriétés de l'algorithme de 2-3 CAH avec le lien simple. Il a été montré notamment qu'en introduisant un nouvel indice du lien simple (appelé indice du "double lien" minimum"), l'algorithme de 2-3 CAH conduit à une 2-3 hiérarchie indicée au sens large qui induit une 2-3 ultramétrique qui est inférieure à la dissimilarité initiale et supérieure à l'ultramétrique sous-dominante.

6.5.3. Etude des dissimilarités induites par les 2-3 hiérarchies

Le stage de Gentian Guscho [37] (responsable : P. Bertarnd) avait pour objectif d'étudier la structure de classification 2-3 hiérarchique, en examinant plus particulièremement les dissimilarités appelée 2-3 ultramétriques, qui sont induites par ce type de classifications. L'étude avait aussi pour objectif de situer les 2-3 ultramétriques dans le cadre général du treillis des dissimilarités définies sur un meme ensemble de données. On a ainsi a pu mettre en évidence l'existence d'une sous-dominante pour une classe particulière de 2-3 ultramétriques.

6.5.4. Complexité et implémentation de l'Algorithme de Classification 2-3 hiérarchique

Dans le cadre de l'encadrement de stage de DEA de Sergiu Chelcea [34][18] (responsables : P. Bertrand et B. Trousse), nous avons étudié le nouvel algorithme de classification, appelé *Classification Ascendante 2-3 Hiérarchique*, proposé par Patrice Bertrand dans Bertrand (2002), et nous l'avons implémenté et intégré dans la boite à outils de Raisonnement à Partir de Cas (RàPC), CBR*Tools, développée à l'INRIA. L'étude théorique de la classification 2-3 hiérarchique a révélé quelques propriétés concernant ce nouvel algorithme, et une reformulation de l'algorithme a permis de proposer un algorithme en $\mathcal{O}(n^2 \log n)$ au lieu de $\mathcal{O}(n^3)$. Une modification de la phase de fusion a aussi été proposée afin d'obtenir un indiçage strict de la structure de classes générées après chaque fusion de deux classes. L'algorithme a été testé dans le cadre d'une application de RàPC pour la détermination de facteurs de risque d'assurance des voitures." Actuellement nous nous focaliserons sur l'utilsiationd e cet algorithme pour la classification de comportement utilisateurs sur le Web.

6.6. Modélisation supervisée de données fonctionnelles par perceptron multi-couches

Mots clés : *analyse de données fonctionnelles*, *perceptron multi-couches*, *réseau de neurones*, *approximation universelle*, *consistance*.

Participants : Brieuc Conan-Guez, Yves Lechevallier, Fabrice Rossi.

L'Analyse de Données Fonctionnelles (ADF) est une extension de l'analyse de données traditionnelles à des données fonctionnelles. Dans cette approche, chaque individu est décrit par une ou plusieurs fonctions réelles, plutot que par un vecteur de R^n. L'avantage majeur de l'ADF sur l'approche multivariée est de prendre en compte de manière naturelle la régularité des fonctions étudiées.

Dans le cadre de l'ADF, on s'est intéressé à l'extension du Perceptron Multi-Couches (PMC)[46] à des données fonctionnelles : le Perceptron Multi-Couches Fonctionnel (PMCF). On a montré [30][29][19][28][20] que ce nouveau modèle partageait avec le PMC deux propriétés théoriques importantes : le PMCF est un approximateur universel, l'estimation des paramètres du modèle est consistante (ce résultat reste valable dans le cas d'une connaissance discrète des fonctions).

Lors de l'année écoulée, deux nouvelles extensions du PMCF ont été proposées :

a) la première extension s'appuie sur une phase préliminaire de projection des fonctions d'entrée sur une base de fonctions choisie au préalable (B-spline, séries de Fourier, etc). La représentation régularisée obtenue est alors soumise au PMCF. L'avantage de cette méthode par rapport à l'approche précédemment proposée (traitement direct des fonctions) est que le PMCF calcule sa sortie grâce à une entrée partiellement débruitée : l'apprentissage du modèle est donc facilité.

D'un point de vue théorique, on a montré que le PMCF basé sur une étape préalable de projection est un approximateur universel et que l'estimation des paramètres est consistante. D'un point de vue pratique, diverses simulations ont été effectuées afin de comparer les deux approches (approche directe/approche par projection),

b) la seconde extension du PMCF proposée consiste à modifier le modèle afin d'obtenir une sortie fonctionnelle. Ce type de modèle est très intéressant car il permet de modéliser les processus fonctionnels à temps discret. On a montré que ce modèle est un approximateur universel, et que de plus, sous l'hypothèse d'indépendance des individus, l'estimation des paramètres est consistante.

6.7. Analyse comparative de méthodes d'extraction de séquences

Mots clés : *analyse de séquences*, *motifs séquentiels*, *sous-séquences fréquenetes*.

Participants : Doru Tanasa, Florent Masséglia, Brigitte Trousse.

On peut distinguer deux types de séquences fréquentes en termes de fouille de données : les sous-séquences fréquentes et les motifs séquentiels. Nous avons récemment implémenté une nouvelle version de l'index « Arbres de Suffixes » (cf.. STIndex[62]) plus performante du point de vue « occupation de la mémoire ». Cette étape va nous permettre de lancer une étude comparative des méthodes d'extraction de séquences fréquentes sur deux plans :

- La qualité des résultats (en comparant la pertinence des séquences obtenues par une méthode d'extraction de sous-séquences ou de motifs séquentiels).
- La rapidité d'extraction.

Pour amorcer ces travaux de comparaison différentes méthodes d'extraction de séquences ont déjà été appliquées sur des données de type access log.

6.8. Analyse et utilisation de comportements visuels et non visuels

Mots clés : *eye-tracking*, *analyse*, *comportements*, *navigation*.

Participants : Mireille Arnoux, [Thierry Baccino, Térésa Colombi, Sylvain Denis, LPEQ UNSA], Nathalie Evan, Sémi Gaieb, Vincent Giraudon, Brigitte Trousse.

Le but du projet COLOR e-Behaviour, projet commun avec le LPEQ de l'UNSA
http://www-sop.inria.fr/COLOR/2001/e-BEHAVIOUR.html est de proposer une plate-forme pour l'analyse et l'utilisation des comportements visuels et non visuels des internautes dans un système d'aide à la navigation dans un annuaire thématique. L'originalité de ce projet est la prise en compte par le système de recommandations, du parcours visuel des utilisateurs sur l'écran. La " saisie " par un système d'« Eye-Tracking » et l'analyse des données relatives à ce parcours se font en collaboration avec le LPEQ/UNSA. Les travaux menés sont tripler :

6.8.1. Conception et réalisation d'un plateforme générique d'expérimentation

La plate-forme comporte une partie statique et une partie dynamique. La partie statique (2001-2002) est composée de l'annuaire thématique sous forme de pages HTML et d'un ensemble d'informations sur cet annuaire enregistré dans une base de données (MySQL). Cette partie est générée à partir de spécifications XML et à l'aide d'un script Perl.

La partie dynamique est composée d'un recommandeur basé sur l'approche Broadway et de sessions utilisateurs (navigations effectuées dans l'annuaire thématique) enregistrées dans la base de données. Deux

recommandeurs ont été conçus et développés basés sur les composants logiciels de Broadway*tools relatifs à l'aide à la navigation dans un annuaire thématique (cf. l'application réalisée « educaid » en 2000 pour FT) : le premier raisonnant sur les comportements non visuels et le deuxième raisonnant sur les comportements visuels corrélés aux comportements non visuels des internautes.

6.8.2. Application de notre plateforme au contexte choisi

Nous avons ensuite utilisé notre plateforme générique avec le système d'eye-tracking utilisé par le LPEQ de l'UNSA et avec un clone de la partie « rapports d'activités 2001 » du site Web de l'Inria : ce clone a été généré à partir des fichiers XML des rapports d'activités.

6.8.3. Expérimentation auprès de deux groupes de sujets

V. Giraudon dans le cadre de son stage de DESS Ergontic co-encadré par T. Colombi et B. Trousse a conçu et mené une expérimentation visant à valider l'aide apportée par le premier recommandeur (selon leur niveau d'expertise sur le web) et à étudier l'impact de l'introduction de ce recommandeur sur le comportment des utilisateurs [36]. Une expérimentation a été menée avec deux groupes de 12 sujets devant effectuer 8 taches prédéfinies, un groupe sans aide et un avec aide. Une analyse statistique a permis de mettre en évidence l'aide apportée par le recommandeur et l'impact sur le comportement visuels des utilisateurs du groupe « avec aide ». Cette expérimentation se poursuit en 2003 avec le deuxième recommandeur et surtout une analyse comparative des deux recommandeurs.

6.9. Réalisation d'un site CASA intégrant un service d'aide à la navigation

Mots clés : *transports*, *service adaptatif*, *recommandations*, *perosnnalisation*, *site Web*, *CASA*.

Participants : Sergiu Chelcea, Sémi Gaieb, [Georges Gallais], Brigitte Trousse.

Dans le cadre de l'encadrement de stage de DEA de Sergiu Chelcea [18],, nous avons étudié l'intéret d'utiliser une approche de recommandations perosnnalisées sur une maquette de site Web dédié à la la CASA (Communauté d'Agglomération de Sophia Antipolis) permettant ainsi de faire de sliens transverses utiles pour les citoyens ou touristes.
Ce travail a été effectué en collaboration avec Georges Gallais et Patrick Rives de l'action VISA.

6.10. Extension et validation de Broadway*Tools et CBR*Tools

Mots clés : *classification automatique*, *classes*, *visualistaion*, *cartes de Kohonen*.

Participants : Tarek Ait-Mohamed, Sergiu Chelcea, Sémi Gaieb, Achia El Golli, Yves Lechevallier, Sébastien Simard, Brigitte Trousse.

Lors du développement des projets e-Behaviour (site RA2001) et recommandations dans le domaine des transports (site CASA), les biliotèques CBR*Tools et Broadway*Tools ont du subir quelques évolutions ponctuelles qui ont entrainé une légère reconception de nos anciennes applications de Broadway*Tools, Broadway-Web et Educaid.

*6.10.1. Broadway*tools, générateur de systèmes de recommandations perosnnalisées sur le Web*

La modification principale a été la factorisation d'un ensemble de composants logiciels communs de Educaid et RA, qui a été placé dans Broadway*Tools De plus, une procédure d'installation automatique de l'ensmeble des applications faites avec Broadway*Tools a été mise en place.

*6.10.2. CBR*Tools plateforme objet pour la gestion de l'expérience*

Cbr*Tools a bénéficié de l'ajout de nouveaux algorithmes concernant la partie indexation de cas :

- 2-3 Hierachical Ascending Classification [34],
- Neuron Index [12] grace à une nouvelle version du code réalisé par Attila Benedek.

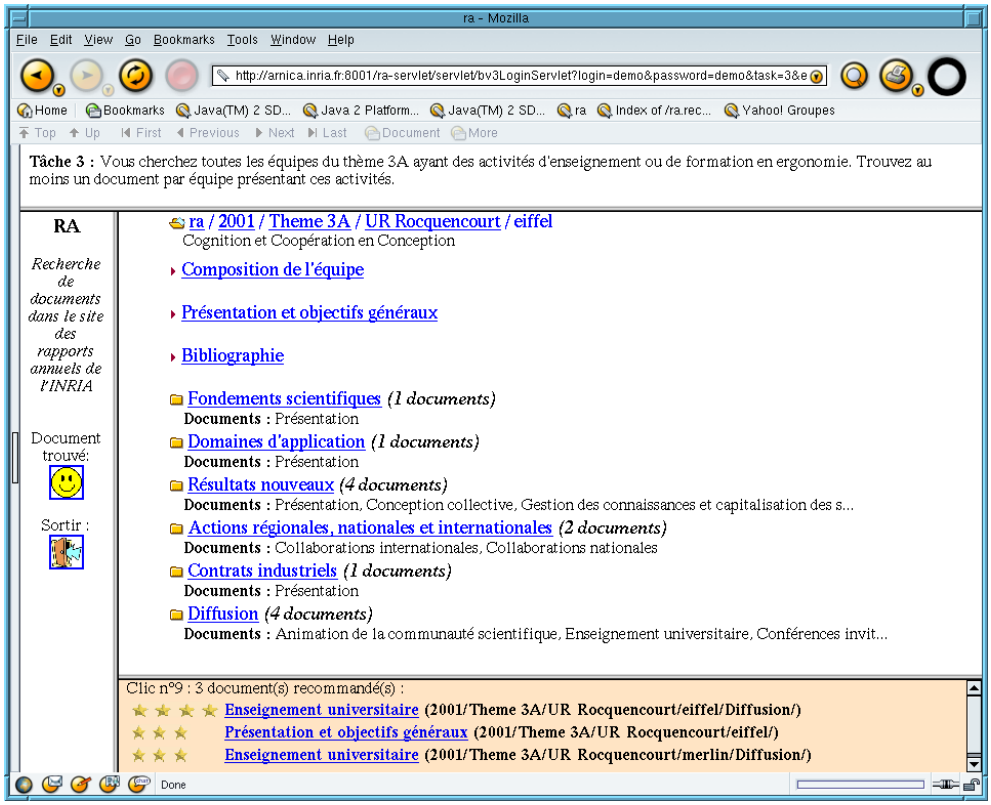

Figure 4. Recommandations personnalisées dans le cadre du RA 2001

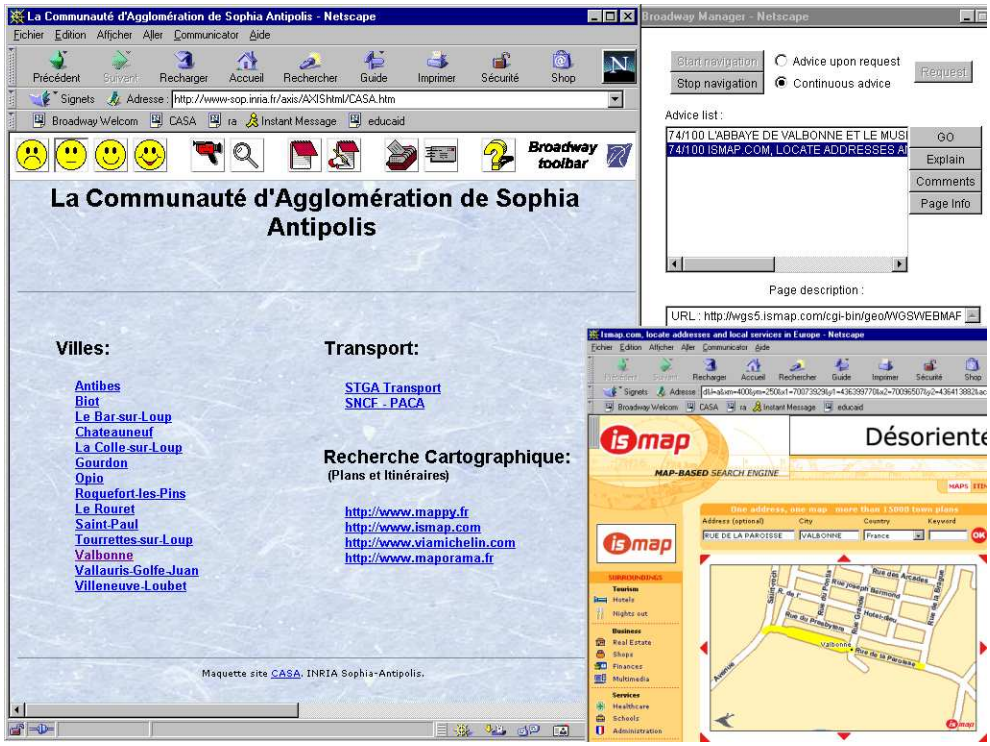

Figure 5. Recommandations personnalisées dans le cadre de la CASA

7. Contrats industriels

7.1. Contrats industriels

7.1.1. EDF : Classification de courbes et analyse de l'usage

Cette année un contrat d'association a été signé avec la DER de l'EDF pour une durée de deux ans. L'objectif est de continuer notre collaboration sur la classification de courbes et, en particulier, des courbes de consommation des clients EDF et d'initialiser une nouvelle collaboration sur l'analyse des fichiers log. Cette année notre participation a été uniquement sur l'analyse des courbes. Les courbes obtenues auprès d'EDF sont essentiellement des courbes journalières, établies à partir de mesures de consommation horaire.

Dans le cadre des courbes de consommation, les cartes de Kohonen offrent la possibilité de visualiser de manière immédiate la structure d'une population de courbes dont l'effectif peut dépasser le millier. Les motivations d'une carte de Kohonen rejoignent alors celles de l'analyse en composantes principales en permettant de visualiser le contenu des classes mais en conservant la structure classificatoire ce qui permet d'explorer et d'interpréter les différentes parties du nuage de données.

D'une manière générale, les courbes sont bien adaptées à la classification par les cartes de Kohonen car leur représentation plane est immédiatement compréhensible par n'importe quel utilisateur.

7.1.2. EPIA un projet pré-compétitif RNTL 2002

Le projet EPIA « Evolution d'un Portail d'Informations Adaptatif » a été labélisé à l »appel d'offre RNTL 2002. Les partenaires de ce projet sont : Dalkia, Mediapps et Inria (AxIS Sophia Antipolis et Rocquencourt).

8. Actions régionales, nationales et internationales

8.1. Actions régionales

Du fait de la bilocalisation d'AxIS, nos actions concernent deux régions :

- Action Colors (2001-2002) : le projet **e-Behaviour** (cf. section 6.8) est une coopération entre notre équipe et le LPEQ (UNSA). http://www-sop.inria.fr/COLOR/2001/e-BEHAVIOUR.html sur l'analyse des comportements visuels et non visuels des inernautes sur un site Web [36].
- Collaboration avec G. Gallais et P. Rives de l'action VISA (Inria Sophia Antipolis) et avec M. Riveill de l'équipe Rainbow (I3S Sophia Antipolis) sur le thème de l'adaptation de services dans le domaine des transports (co'encadrement d'un stage de DEA et réponses à divers appels d'offre RNRT et PREDIT).
- Laboratoire des Usages, CNRT, Sophia Antipolis (http://www-sop.inria.fr/axis/Labo_des_usages.pdf), laboratoire structuré sous forme d'un GIS visant à éclairer les principales pistes d'une recherche sur les usages des TIC et d'une recherche technologique inspirée par ces usages (B. Trousse, F. Masséglia et E. Guichard). B. Trousse est membre du comité scientifique (CS) de ce laboratoire et membre suppléant du comité de groupement. Participation à la première réunion du CS le 22 octobre 2002 à Sophia Antipolis (B. Trousse et E. Guichard en tant qu'invité).
  Participation de B. Trousse à plusieurs réunions réunissant des représentants des membres fondateurs (C. Charbit ENST,C. Guéguen GET et R. Aréna UNSA&CNRS) visant la préparation des textes de présentation du laboratoire au CS en collboration avec) [38].
- Laboratoire PRISM : Co-encadrement d'un stage de DEA avec Karine Zeitouni sur le Data mining spatial par ap
- Laboratoire LISE-CEREMADE de l'université Paris IX Dauphine : Activité de recherche avec Fabrice Rossi autour l'approche fonctionnelle dans les méthodes de classification non supervisées et principalement les réseaux de neurones [30][29][28][19][20].

8.2. Actions nationales

- GDR-PRC I3 - Groupement de Recherches « Information - Interaction - Intelligence » du CNRS (F. Masséglia).
- participation à la journée GafoDonnées le 5(après-midi) Novembre. : F. Masséglia.
- Laboratoire MAB de l'université de Bordeaux : Développement de la méthode de classification divisive DIV et actions de recherche sur les méthodes de classification sur des données intervalles et d'outils d'aide à l'interprétation de classes ou de partitions avec Marie Chavent [15][16][17]
- Université de Technologie de Compiégne : Participation à l'écriture d'un chapitre d'un livre sur l'analyse des données chez Hermès, l'éditeur scientifique de ce livre étant Gérard Govaert.

8.3. Actions européennes

8.3.1. Projet européen IST : ASSO

http://www.info.fundp.ac.be/asso/index.html

La composante parisienne d'AxIS participe à un projet communautaire de recherche en statistique, le projet ASSO (IST-2000-25161) démarré début 2001 jusqu'à la fin 2003, qui est géré par Eurostat. Ce projet appartient au programme DOSIS (Development Of Statistical Information Systems) qui doit servir de catalyseur, au niveau européen, en réunissant les fournisseurs et utilisateurs de données, autour de thèmes et d'enjeux communs (conception des systèmes d'information, collecte des données, ainsi que leur traitement, analyse et stockage). Il veut favoriser l'émergence de nouveaux modèles, de nouvelles techniques, de nouveaux logiciels. La participation de l'INRIA [40] comprend trois parties :

- la normalisation [40], sur le plan syntaxique, des structures classificatoires de données numérique-symboliques et mise sous format XML du tableau de données symboliques et des metadonnées associées à ce tableau,
- coordination [40][43][42] du développement des méthodes de classification et participation à la réalisation de SCLUST et DCLUST,et proposition d'une méthode d'évaluation de la stabilité d'une classe générée par une méthode de partitionnement

- modification du module d'extraction de données agrégées (DB2SO) [40].

La première partie comprend la définition d'un langage commun de représentation de connaissances sous forme XML. Ce langage a deux objectifs :

- permettre aux divers algorithmes présents dans ASSO de communiquer entre eux,
- disposer d'un outil de formulation de la connaissance permettant aux utilisateurs de transmettre leurs tableaux de données aux algorithmes de ASSO.

Au cours de cette année, la bibliothéque C++ associée au langage de représentation des connaissances et des metadonnées a été réalisée. Ce travail permet de contrôler la syntaxe et la sémantique de fichiers XML qui servent à assurer les communications entre les différentes méthodes du projet.

La seconde partie est la fourniture d'un programme de classification automatique SCLUST qui permet de réaliser un partitionnement des données de type intervalle. Un prototype a été réalisé.

La dernière partie concerne la modification du module DB2SO suivant les spécifications données par ASSO.

8.3.2. Réseaux européens

- IST Ontoweb : l'action AxIS participe au réseau OntoWeb (Ontology-based Information Exchange for Multilingual Electronic Commerce and Information Integration) qui a été proposé en 2000 à l'intiative de Dieter Fensel (Division of Mathematics & Computer Science, Vrije Universiteit Amsterdam). Participation : T. Despeyroux et B. Trousse.
- COST Action 282 (2001-2005) : « Knowledge Exploration in Science and Technology ». B. Trouse a été nommé en 2002 représentante pour la France avec A. Mille (ERIC, Lyon) au « Management Committee ».
  http://cost.cordis.lu/src/action_detail.cfm?action=282.

8.4. Actions internationales

- Université de Naples II(Italie) : Le 6 et 7 mai 2002 Marc Csernel a été invité pour deux conférences sur la forme normale symbolique et sur l'extraction de conaissances d'une base de données sous forme d'objets symboliques.
- Coopération Inria-CNPq : Yves lechevallier et Marc Csernel ont été invités à Recife dans le cadre du projet CLADIS. Le projet CLADIS se termine cette année aussi nous avons présenté les travaux de recherche, réalisés durant ce projet, dans le cadre de la conférence des sociétés brésiliennes d'intelligence artificielle (SBIA) et des réseaux de neurones (SBRN) [23].
- Facultés Universitaires Notre-Dame de la Paix à Namur (Belgique) : Durant le mois de février et pour une durée d'une semaine Yves Lechevallier a été invité à l'université de Namur, dans le cadre des "questions spéciales", pour réaliser un cours sur les méthodes de classification et de classement. Suite à cette invitation une collaboration sur la détermination du bon nombre de classes a été initialisée [26].
- Projet LIAMA : Ce projet s'est terminé cette année. Une publication [11] a été réalisée à partir d'une partie des résultats de ce projet.

9. Diffusion des résultats

9.1. Animation de la communauté scientifique

9.1.1. Revues

B. Trousse est membre du comité de rédaction de deux revues nationales :

- Revue d'Intelligence Artificielle RIA publiée par Hermès qui diffuse les résultats de la communauté francophone d'IA (rédacteur en chef : M. Pomerol)

- Revue électronique I3 du GDR-I3 (rédacteurs en chef : C. Garbay et H. Prade) : http://www.Revue-I3.org/

Y. Lechevalier est

- l'un des rédacteurs du « Journal of Symbolic Data Analysis (revue électronique http://www.jsda.unina2.it ;
- co-éditeur en 2002 avec L. Tricot (CNAM) des numéros 28 et 29 de la Revue de Modulad, périodique semestriel du Club Modulad.

E. Guichard a été relecteur de deux articles pour la Revue Internationale de Géomatique.

9.1.2. Comités de programme

Sont membres de comité de programme :

- au niveau national : EGC'03(Lyon) Extraction et Gestion des Connaissances, SFdS'03 (Lyon) Journées de la Société Française de Statistique : Y. Lechevallier, Inforsid'03 (F. Masséglia)
- au niveau international : ECCBR'02 European Conference of Case-Based Reasonng (UK) : B. Trousse

9.1.3. Organisation de séminaires et conférences

- Colloque « Mesures de l'internet » (2002) : E. Guichard (responsable), S. Honnorat et B. Trousse en relation avec D. Sergeant et M-H Zeitoun du service BMC.
  URL : http ://ww-sop.inria.fr/axis/cmi/

9.1.4. Visites

L'action AxIS a reçu cette année plusieurs visites :

- Thierry Baccino, Sylvain Denis et Térésa Colombi de l'UNSA (LPEQ de l'UNSA) dans le cadre du projet e-behaviour (Colors) ;
- Claude Guéguen dans le cadre de la constitution du laboratoire des usages sous forme d'un GIS,
- Jean-Paul Rasson, FUNDP, Namur, en janvier et en juin.
- Rosanna Verde, Université de Capoue, Italie en juillet et septembre ;

Dans le cadre d'accords de coopération, nous avons accueilli les chercheurs étrangers suivants :

- Franscico A. T. de Carvalho, Université de Récife, Brésil en mai et octobre ;
- Huiwen Wang, BUUA, Pekin, Chine en septembre ;

9.1.5. Serveur interne Web

L'action AxIS a mis au point une première version de son site Web permettant d'accéder à un certain nombre d'informations relatives à nos recherches et en particulier à notre plate-forme objet CBR*Tools de partage et de réutilisation d'expériences et aux deux systèmes Broadway et Hermès : http://www-sop.inria.fr/axis/.

9.1.6. Divers

T. Despeyroux est président de l'AGOS, membre titulaire réélu en 2002 de la commission technique paritaire (CTP) ainsi que du conseil d'administration de l'INRIA en tant que représentant du personnel.

9.2. Formation

9.2.1. Enseignement universitaire

Nous avons participé cette année aux enseignements suivants : La partie sophipolitaine du projet fait partie actuellement de plusieurs équipes enseignantes à l'UNSA et à l'Université de Montpellier UNSA

- DEA Informatique (resp Mr Kounalis) à l'UNSA Sophia Antipolis :

  - Cours sur les *frameworks objet* (4h) dans le cours du tronc commun (TC4) « Langages et Modèles à objets » (resp M. Blay) : B. Trousse
  - Coresponsables d'un module optionnel (O15) sur *Sémantique et Conception de sites Web* (30h) dans le cadre des modules optionnels du DEA Informatique proposé depuis 1999 : T. Despeyroux et B. Trousse. (réalisé en 1999)
- DESS « Ergonomie et NTIC » (resps T. Baccino et J. Araszkiewiez) à l'UNSA : responsable de deux cours sur la *conception de sites Web évolutifs* et sur la *conception de sites/services adaptatifs d'aide à la recherche d'informations* : B. Trousse.
- Licence professionnelle Franco-italienne STID « Statistiques et Informatique Décisionnelle » (resp. J. Lemaire) à l'UNSA, Menton :

  - Encadrement d'un projet (100h/étudiant, 22 étudiants, 54h TD encadrées) sur *l'analyse de logs HTTP dans un contexte multi-sites Web* (50h) sur les deux proposés : D. Tanasa, B. Trousse (resp.). Le projet consistait dans l'analyse (avec l'outil SAS) des logs INRIA pré-traitées (cf. 5.1).
- ESINSA à l'UNSA, Sophia Antipolis : cours de C pour la formation ITII (36 h en juillet) : D. Tanasa.
- ESSI à l'UNSA, Sophia Antipolis : TP (40h) de *Java* : D. Tanasa..
- IUT GTR (Génie des Télécommunications et Réseaux) à l'UNSA, Sophia Antipolis : TP (140h) de *Réseaux* : S. Gaieb.

La partie Rocquencourt a la responsabilité des cours suivants et accueille chaque année environ plusieurs stagiaires de DEA :

- DEA « Modélisation et traitement des données et des connaissances » (resp : S. Pinson) de l'université Paris IX-Dauphine (6h) : Cours sur *Analyse des connaissances numériques et symboliques* : Y. Lechevallier.
- DESS « Mathématiques appliquées et sciences économiques » (MASE) de l'université Paris IX-Dauphine : Cours (18h) sur les *méthodes neuronales en classification* Y. Lechevallier.
- DESS « Ingénierie de la Décision » (resp : S. Pinson) de l'université Paris IX-Dauphine : Cours sur le *Traitement des enquêtes* : P. Bertrand.
- ISUP de l'Université de Paris 6 : Cours sur les *méthodes de classification et de classement* (30h) : Y. Lechevallier.
- ENSAE : Cours sur le Data Mining (12h) : Y. Lechevallier.

9.2.2. Thèses

AxIS est équipe d'accueil de doctorants et stagiaires de DEA de la formation doctorale STIC de Nice-Sophia Antipolis ainsi que de l'Université de Paris IX Dauphine.
Thèses soutenues dans AxIS[3] :

1. **Brieuc Conan-Guez**, "Modélisation supervisée de données fonctionnelles par perceptron multi-couches Université de Paris Dauphine (18 décembre 2002).
2. **Marc Csernel**, "La Forme Normale Symbolique » Uniiversité de Paris Dauphine (19 décembre 2002) (directeurs : E. Diday et F. de A.T. De Carvalho).

Thèse en cours :

[3] Notons également la thèse de E. Guichard soutenue en octobre 2002 [49]

1. **Aicha El Golli** (démarrage fin 2001), Cartes topologiques et modèles statistiques : application à la classification de données symboliques, Université de Paris IX Dauphine (directeur de thèse : E. Diday).
2. **Doru Tanasa**, (démarrage fin 2001), « Trace et analyse de l'usage pour l'aide à la reconception d'un site Web », Université de Nice-Sophia Antipolis (directeur de thèse : Brigitte Trousse).
3. **Sergiu Chelcea**, (démarrage fin 2002), « Classification de profils utilisateurs d'un site Web », Université de Nice-Sophia Antipolis (co-directeurs de thèse : Jacques Lemaire et Brigitte Trousse).
4. **Hicham Behja**, (démarrage fin 2002), « Gestion de points de vues multiples dans l'analyse d'un observatoire sur le Web », Université de Casablanca, co-encadrants de thèse : Aziz Marzark et Brigitte Trousse). cette thèse s'inscrit dans le projet STIC GL de la coopération France-Maroc(2002-2005).

Jurys de thèse :
Yves Lechevallier a été membre de 3 thèses : celles de Brieuc Conan-Guez et de Marc Csernel d'AxIS à l'université Paris-Dauphine au mois de décembre etcelle de Florent Domenach sur "Structures latticielles, correspondances de Galois contraintes et classification symbolique" à l'université Paris-I.

Yves Lechevallier a été rapporteur des quatre thèses suivantes :

- d'Aurélien de Reyniès "Classification et Discrimination en analyse de données symboliques" en novembre à l'université de Paris-Dauphine,
- de Vincent Bertholet "Apports de l'estimation de densité aux arbres de discriminatioon" en mai aux facultés universitaires notre-dame de la paix à Namur (Belgique),
- de Jean-Pascal Aboa "Méthodes de segmentation sur un tableau de variables aléatoires" en mars à l'université de Paris-Dauphine,
- de Céline Robardet "Contribution à la classification non supervisée : proposition d'une méthode de bi-partitionnement" en mai à l'université de Lyon-1.

9.2.3. Stages

Nous avons accueilli dix stahiaires qui ont travaillé sur les projets et stages suivants (ordre chronologique) :

1. **Nathalie Evan**, étudiante de l'ESSI (UNSA, 2001-2002), a effectudée son projet [35] ; à compter de mi-octobre 2001 jusqu'à fin mars 2002 (responsables : M. Arnoux et S. Gaieb) sur les aspects trace des comportements visuels et non visuels dans une base de données relationnelle dans le cadre du projet e-behaviour ;
2. **Sergiu Chelcea** [34] (DEA Université de Nice Sophia Antipolis) sur une étude théorique et pratique de l'algorithme de 2-3 hiérarchies appliqué à la classification de comportements d'internautes sur une maquette de site de la CASA (resp : P. Bertrand et B. Trousse) ; .
3. **Gentian Gusho** [37] (DEA MIASH, Université de Paris 1), (resp : P. Bertrand) sur « Etude d'une classe particulière de robinsoniennes » ;
4. **Laurent Jullien** [39](DEA MIASH, Université de Paris 1). (resp : P. Bertrand) sur « 2-3 Hiérarchies, 2-3 Ultramétriques, algorithme 2-3 CAH » ;
5. **Tarek Ait-Mohamed** [33], étudiant de DEA de l'Université de Paris IX Dauphine ;
6. **Vincent Giraudon** [36], étudiant DESS Ergontic (UNSA) ;
7. **Tao Wan** [44], étudiant DEA UVSQ ;
8. **Miha Jurca**, stagiaire roumain à compter du 15/11sur l'intégration des méthodes de classification développés par la composante parisienne dans CBR*Tools et dans le développement d'un serveur Web intranet relativement à notre savoir faire en classification (encadrement : M. Csernel, Y. Lechevallier et B. Trousse) ;
9. **Fabien Benoit**, étudiant de l'ESSI (UNSA 2002-2003), effectue son projet à compter du 28/10/02 au 28/04/03 (resp. F. Masséglia). L'objectif de ce stage est de fournir une implémentation de la méthode SPAM [50]. Cette méthode d'extraction de motifs séquentiels est basée sur l'exploitation

de bitmaps. De plus nous pouvons ajouter aux résultats attendus, une étude du comportement de cet algorithme en fonction de la structure des données explorées ;

10. **Arnaud Satoni**, étudiant de l'ESSI (UNSA 2002-2003), effectue son projet à compter du 15.10.2002 jusqu'au 01.04.2003 (resp. D. Tanasa) sur la "Construction de la structure logique d'un iste Web en XML".

9.3. Participation à des colloques, séminaires

Nous avons présenté nos travaux de recherche dans des séminaires et conférences internationales. Voir la bibliographie de cette année pour plus de détails.
Outre ces conférences ou séminaires, nous avons participé

- Séminaire du SAMOS (Paris I PANTHEON - SORBONNE) en mars 2002 : B/ Conan-Guez.
- Séminaire du LISE-CEREMADE (Université Paris-IX Dauphine)en mars 2002 : T. Lechevallier.
- 2ème COST Management Committee Meeting à Lyon le 20 juillet : B. Trousse
- "Workshop on Symbolic Data Analysis" de l'IFCS2002 à Cracovie (Pologne) et du SBRN'02 et SBIA'02 au Pernambuco (Brésil) : Y. Lechevallier.
- ECAI'02 European Conference on Artificial intelligence, Lyon, juillet : B. Trousse.
- Journée d'accueil de l'UR Sophia le 4 Octobre 2002 : E. Guichard, F. Masséglia et S. Siemard.
- Journée GafoDonnées le 5(après-midi) novembre : F. Masséglia.
- Journée prospective du RNRT pour définir ses futurs programmes de recherche et appels d'offre (7 nov) : E. Guichard.
- RNTL'02 : présentation d'un poster sur notre projet RNTL EPIA (labélisé en 2002) aux journées RNTL organisées à Toulouse du 23-25 octobre.
- TTNL'02 : présentation d'une démonstration sur notre approche de calcul de recommandations personnalisées dans un site web en relation avec les transports. Illustration au niveau de la CASA (Communauté d'agglomération de Sophia Antipolis) : S. Chelcea, S. Gaieb et B. Trousse.
- Assises du GDR I3 les 4 et 5 Décembre (Nancy) : F. Masséglia.

10. Bibliographie

Bibliographie de référence

[1] P. BERTRAND, M. JANOWITZ. *The k-weak Hierarchies : An Extension of the Weak Hierarchical Clustering Structure.* in « Discrete Applied Maths », North-Holland, 1999.

[2] *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data.* éditeurs H. H. BOCK, E. DIDAY., Studies in Classification, Data Analysis and Knowledge Organisation, Springer-Verlag, 1999.

[3] G. CARAUX, Y. LECHEVALLIER. *Règles de décision de Bayes et méthodes statistiques de discrimination.* in « Revue d'intellignece artificielle », numéro 2-3, volume 10, 1996, pages 219-284.

[4] M. CHAVENT. *A monothetic clustering method.* in « Pattern Recognition Letters », 1999, pages 989-996.

[5] M. CSERNEL. *On the complexity of computation with symbolic objects using domain knowlege.* in « New Advances in Data Science and Classification », Springer-Verlag, 1998, pages 85-90.

[6] T. DESPEYROUX, B. TROUSSE. *Web sites and Semantics.* in « HYPERTEXT'01, the twelfth ACM Conference on Hypertext and Hypermedia, Aarhus, Danemark », pages 239-240, août, 2001.

[7] M. JACZYNSKI. *Modèle et plate-forme à objets pour l'indexation des cas par situation comportementale : application à l'assistance à la navigation sur le Web.* thèse de doctorat, Université de Nice Sophia-Antipolis, Sophia-Antipolis, décembre, 1998.

[8] F. MASSEGLIA. *Algorithmes et Applications Pour l'Extraction de Motifs Séquentiels Dans le Domaine de la Fouille de Données : de l'Incrémental au Temps Réel.* thèse de doctorat, Universite de Versailles St-Quentin en Yvelines, France, January, 2002.

[9] B. TROUSSE. *Vers des outils d'aide à la conception coopérative : « Design Groupware ».* éditeurs J.-M. FOUET., in « Connaissances et savoir-faire en entreprise - Intégration et capitalisation », Hermes, Paris, 1997, chapitre 17, pages 317-341.

[10] B. TROUSSE. *Viewpoint Management for Cooperative Design.* in « Proceedings of the IEEE Computationnal Engineering in Systems Applications (CESA'98) », UCIS - Ecole Centrale de Lille - CD-Rom, éditeurs M. K. P. BORNE, A. E. KAMEL., april, 1998.

Articles et chapitres de livre

[11] H. WANG, F. LINGHUI, Y. LECHEVALLIER. *Disaster Pattern of flood and waterlog in Poyang lake analysis.* in « Rivista di Statistica Applicata », 2002.

Communications à des congrès, colloques, etc.

[12] A. BENEDEK, B. TROUSSE. *Adaptation of Self-Organizing Maps for CBR case indexing.* in « Proceeding of the 4th International Workshopon Symbolic and Numeric Algorithms for Scientific Computing (toappear) », pages 31-45, Timisoara, Romania, October, 2002.

[13] P. BERTRAND. *Les 2-3 hiérarchies : une structure de classification pyramidale parcimonieuse.* in « Société Francophones de Classification », SFC2002, pages 143-148, Toulouse, France, septembre, 2002.

[14] P. BERTRAND. *Set Systems for Which Each Set Properly Intersects at Most One Other Set - Application to Pyramidal Clustering.* in « IFCS2002, Classification, Clustering, and Data Analysis », IFCS2002, éditeurs K. JAJUGA, A. SOKOLOWSKI., pages 38-39, Cracow, Poland, juillet, 2002.

[15] M. CHAVENT, Y. LECHEVALLIER. *Dynamical Clustering of Interval Data Optimization of an Adequacy Criterion Based on Haussdorff Distance.* in « Classification, Clustering, and Data Analysis », Springer, éditeurs K. JAJUGA, A. SOKOLOWSKI, H. H. BOCK., pages 53-60, Berlin, Germany, juillet, 2002, , aussi dans les actes d'IFCS2002, Poland,.

[16] M. CHAVENT, Y. LECHEVALLIER, R. VERDE. *Symbolic Clustering Interpretation and Visualisation.* in « Between Data Science and Everyday Web Pratice », GfKl2002, pages 65-66, Mannheim, juillet, 2002.

[17] M. CHAVENT, Y. LECHEVALLIER, R. VERDE. *Symbolic Clustering Interpretation and Visualization.* in « Between Data Science and Everyday Web Pratice », GfKl2002, pages 65-66, Mannheim, Germany, juillet, 2002.

[18] S. CHELCEA, P. BERTRAND, B. TROUSSE. *Hierarchical Ascendant Clustering : theoritical study and*

implementation. in « Proceeding of the 4th International Workshop on Symbolic and Numeric Algorithms for Scientific Computing », Timisoara, Romania, October, 2002.

[19] B. CONAN-GUEZ, F. ROSSI. *Approche Régularisée du Traitement de Données fonctionnelles par un perceptron multi-couches.* in « Société Francophones de Classification », SFC2002, pages 169-172, Toulouse, France, septembre, 2002.

[20] B. CONAN-GUEZ, F. ROSSI. *Multi-Layer Perceptrons for Functional Data Analysis : a Projection Based Approach.* in « ICANN 2002 », pages 667-672, Madrid, Spain, aout, 2002.

[21] M. CSERNEL, F. DE A. T. DE CARVALHO. *Modelling Memory Requirement with Normal Symbolic Form.* in « Classification, Clustering, and Data Analysis », Springer, éditeurs K. JAJUGA, A. SOKOLOWSKI, H. H. BOCK., pages 289-296, Berlin, Germany, juillet, 2002, aussi dans les actes d'IFCS2002, Poland.

[22] M. CSERNEL, F. A. T. DE CARVALHO. *On Memory Requirement with Normal Symbolic Form.* in « Exploratory Data Analysis in Empirical Research », Springer, éditeurs M. SCHWAIGER, O. OPITZ., pages 22-30, Berlin, 2002.

[23] F. A. T. DE CARVALHO, M. CSERNEL, Y. LECHEVALLIER. *Clustering constrained symbolic data based on distance functions.* in « Workshop on Symbolic Data Analysis », SBRN'02 et SBIA'02, Pernambuco, novembre, 2002.

[24] T. DESPEYROUX, B. TROUSSE. *De la sémantique des langages de programmation à la vérification sémantique des sites Web.* in « Journées scientifiques de l'action spécifique Web Sémantiques », CNRS, Paris, October, 2002, http://www.lalic.paris4.sorbonne.fr/stic/octobre/programme0209.html.

[25] A. E. GOLLI. *Bases de données relationnelles : construction d'objets symboliques.* in « Extraction des connaissances et apprentissage », EGC2002, éditeurs D. HÉRIN, D. A. ZIGHED., pages 422-423, Montpellier, France, janvier, 2002, (poster paper).

[26] A. HARDY, P. LALLEMAND, Y. LECHEVALLIER. *Détermination du nombre de classes pour la méthode de classification symbolique SCLUST.* in « Société Francophones de Classification », SFC2002, pages 221-224, Toulouse, France, septembre, 2002.

[27] Y. LECHEVALLIER, R. VERDE. *Classification des données multi-valuées et interprétation des classes.* in « Société Francophones de Classification », SFC2002, pages 243-246, Toulouse, France, septembr<e, 2002.

[28] F. ROSSI, B. CONAN-GUEZ, F. FLEURET. *Functional Data Analysis With Multi Layer Perceptrons.* in « IJCNN (part of WCCI) proceeding », IJCNN2002, pages 2843-2848, Honolulu, Hawaii, mai, 2002.

[29] F. ROSSI, B. CONAN-GUEZ. *Modélisation supervisée de données fonctionnelles par perceptron multi-couches.* in « Société Francophones de Classification », SFC2002, pages 93-100, Toulouse, France, septembre, 2002.

[30] F. ROSSI, B. CONAN-GUEZ. *Multi-layer Perceptron on Interval Data.* in « Classification, Clustering, and Data Analysis », Springer, éditeurs K. JAJUGA, A. SOKOLOWSKI, H. H. BOCK., pages 427-436, Berlin, Germany, juillet, 2002, , aussi dans les actes d'IFCS2002, Poland.

[31] D. TANASA. *Lessons from a Web Usage Mining Intersites Experiment.* in « First International Workshop on Data Cleaning and Preprocessing », ICDM02, Maebashi, Japon, 9 December, 2002.

[32] R. VERDE, Y. LECHEVALLIER. *Analysis of Merovingian Tombs on the Basis of Beads by Aid of a symbolic clustering algorithm.* in « Between Data Science and Everyday Web Pratice », GfKl2002, pages 185-186, Mannheim, Germany, juillet, 2002.

Divers

[33] T. AIT-MOHAMED. *Visusualisation des cartes de Kohonen.* rapport technique, DEA, Paris-Dauphine, 2002.

[34] S. CHELCEA. *Etude, implémentation et évaluation d'un algorithme de classification pour des systèmes de recommandations dans le domaine du tourisme.* rapport technique, Université de Nice Sophia Antipolis, 2002, DEA d'Informatique.

[35] N. EVAN. *Persistance et exploitation des données comportementales des utilisateurs d'un site web.* rapport technique, Université de Nice Sophia Antipolis, 2002, projet DESS Informatique (ESSI).

[36] V. GIRAUDON. *Projet e-Behaviour : Analyse des comportements visuels et non visuels.* rapport technique, DESS Ergontic, UNSA, 2002, 70 pages.

[37] G. GUSHO. *Etude d'une classe particulière de robinsoniennes.* rapport technique, Université Panthéon-Sorbonne (Paris 1) et Ecole des Hautes Etudes en Sciences Sociales (EHESS), 2002, DEA.

[38] C. GUÉGUEN, B. TROUSSE. *Plateformes, services et usags colectifs des STIC.* Document interne Laboratoire des usages (GIS) Sophia Antipolis, juin, 2002.

[39] L. JULLIEN. *2-3 Hiérarchies, 2-3 Ultramétriques, algorithme 2-3 CAH.* rapport technique, Université Panthéon-Sorbonne (Paris 1) et Ecole des Hautes Etudes en Sciences Sociales (EHESS), 2002, Stage de DEA.

[40] Y. LECHEVALLIER, M. CSERNEL, P. BERTRAND. *INRIA Activity Report.* IST-2000-25161 INRIA, septembre, 2002.

[41] Y. LECHEVALLIER. *Construction de super-classes à partir de la carte de Kohonen, indicateurs de qualité de cette carte.* Séminaire ERIC, Lyon, avril, 2002, http ://www-sop.inria.fr/axis/talks/02eric/.

[42] Y. LECHEVALLIER. *Scientific technical report for WP6.* IST-2000-25161 WP6/D1.1, juin, 2002.

[43] Y. LECHEVALLIER. *WP6 Data Sets.* IST-2000-25161 WP6/0001, janvier, 2002.

[44] T. WAN. *Data mining spatial par apprentissage symbolique.* rapport technique, DEA, UVSQ, 2002.

Bibliographie générale

[45] A. AAMODT, E. PLAZA. *Case-Based Reasoning : Foundational Issues, Methodological Variations, and System Approaches.* in « The European Journal oo Artificial Intelligence », numéro 1, volume 7, 1994, pages

39-59.

[46] C. BISHOP. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[47] T. DESPEYROUX. *AS, for Abstract Syntax - Manual - V1.0.* rapport technique, numéro 197, Inria, septembre, 1996, http://www.inria.fr/rrrt/rt-0197.html.

[48] M. E. FAYAD, D. C. SCHMIDT. *Object-Oriented Application Frameworks.* in « Communication of the ACM », numéro 10, volume 40, 1997, pages 32-38.

[49] E. GUICHARD. *L'internet : mesures des appropriations d'une technique intellectuelle.* thèse de sciences sociales, option sciences de l'information et de la communication, École des hautes études en sciences sociales, october, 2002.

[50] T. Y. J. AYRES, J. FLANNICK. *Sequantial PAttern Mining using A Bitmap Representation.* in « Actes des journées Bases de Données Avancées (BDA'02) », 2002.

[51] M. JACZYNSKI. *Modèle et plate-forme à objets pour l'indexation des cas par situation comportementale : application à l'assistance à la navigation sur le Web.* thèse de doctorat, université de Nice-Sophia Antipolis, Sophia-Antipolis, décembre, 1998.

[52] M. JACZYNSKI, B. TROUSSE. *Fuzzy Logic for the Retrieval Step of a Case-Based Reasoner.* in « Second European Workshop on Case-Based Reasoning (EWCBR'94) », pages 313-320, Chantilly, 1994.

[53] M. JACZYNSKI, B. TROUSSE. *Broadway : a Case-based System for Cooperative Information Browsing on the World-Wide-Web.* in « Collaboration between Human and artificial Societies. Coordination and Agent-based distributed Computing », LNAI Series, éditeurs J. A. PADGET., pages 264-283, 1999.

[54] M. JACZYNSKI, B. TROUSSE. *Patrons de conception dans la modélisation d'une plate-forme po ur le raisonnement à partir de cas.* in « Revue l'Objet », numéro 2, volume 5, 1999, Numéro Spécial sur les patterns orientés objets, D. Rieu et J-P. Giraudon (guest editors).

[55] R. JOHNSON, B. FOOTE. *Designing Reusable Classes.* in « Journal of Object-oriented programming », numéro 2, volume 1, 1988, pages 22-35.

[56] G. KAHN. *Natural Semantics.* in « Proceedings of STACS'87 », Lecture Notes in Computer Science n 247, Springer-Verlag, Berlin, février, 1987, http://www.inria.fr/rrrt/rr-0601.html, aussi Rapport de Recherche de l'INRIA Sophia Antipolis N 601.

[57] J. KOLODNER. *Case-Based Reasoning.* Morgan Kaufmann Publishers, 1993.

[58] J. A. KONSTANT, B. N. MILLER, D. MALTZ, J. L. HERLOCKER, L. R. GORDON, J. RIEDL. *GroupLens : Applying collaborative filtering to usenet news.* in « Communications of the ACM », numéro 3, volume 40, 1997, pages 77-87.

[59] A. NAPOLI, A. MILLE, M. JACZYNSKI, B. TROUSSE, ALII. *Aspects du raisonnement à partir de cas.* in

« Actes des 6 èmes journées nationales PRC-GDR Intelligence Artificielle », hermes, Paris, éditeurs S. PESTY, P. SIEGEL., pages 261-288, mars, 1997.

[60] P. RESNICK, H. R. VARIAN. *Recommender systems.* in « Communications of the ACM », numéro 3, volume 40, 1997, pages 56-58.

[61] U. SHARDANAND, P. MAES. *Social Information Filtering : Algorithms for Automating Word of mouth.* in « CHI'95 : Mosaic of creativity », ACM, pages 210-217, Denver, Colorado, mai, 1995.

[62] D. TANASA, B. TROUSSE. *Web Access pattern Discovery and Analysis based on Page Classification and on Indexing Sessions with a Generalised Index Tree.* in « SYNASC 2001, Timisoara, Roumanie », pages 62-72, octobre, 2001.

[63] B. TROUSSE, M. JACZYNSKI, R. KANAWATI. *Towareds a fraemwork for building collaborative information seraching systems.* in « Proceedings of the second european conferenve on digital libraries (ECDL'98) », série LNCS, Springer, septembre, 1998, (Poster session).

[64] F. VAN HARMELEN, J. VAN DER MEER. *WebMaster : Knowledge-based Verification of Web-pages.* in « Twelfth International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems IEA/AIE'99 », série Lecture Notes in Artificial Intelligence, Springer Verlag, 1999.

[65] S. WESS, K. ALTHOFF, G. DERWAND. *Using K-d Trees to Improve the Retrieval Step in Case-Based Reasoning.* in « Lecture Notes in Artificial Intelligence, Topics in Case-Based Reasoning », Springer-Verlag, éditeurs S. WESS, K. ALTHOFF, M. M. RICHTER., pages 167-181, 1994.

[66] A. WEXELBLAT, P. MAES. *Footprints : Visualizing histories for web browsing.* in « Proceedings of RIAO'97, Computer Assisted Information Retrieval on the Internet », Montreal, 1997.

[67] T. YAN, M. JACOBSEN, H. GARCIA-MOLINA, U. DAYAL. *From user access patterns to dynamic hypertext linking.* in « Computer Network and ISDN systems », volume 28, mai, 1996, pages 1007-1014, (proceedings of the 5th international WWW conference).