Doru TANASA

Né le 30 Juin 1978
à Lugoj (Roumanie)

Résidence Les Primevères, Bât. Myosotis
112, Avenue Sainte Marguerite
06200 NICE

Tél: +33 (0)4 93 71 18 31
Mob: +33 (0)6 98 71 78 37
E-mail: dorutanasa@yahoo.com

Docteur en Informatique

spécialisé dans la Fouille de données

Plan du Curriculum Vitae

1 Formation et diplômes

2006 **Qualifié aux fonctions de maître de conférence** en 27e section (Informatique).

2001 – 2005 **Doctorat en Informatique,** Université de Nice Sophia Antipolis, France, Mention très honorable, thèse soutenue le **3 juin 2005**.
Travaux de recherche effectués dans l'équipe AxIS à l'INRIA Sophia Antipolis et dirigés par Brigitte TROUSSE.

Sujet : « *Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support* »**.**

Rapporteurs : Henri BRIAND et Fabrice GUILLET (Polytech'Nantes)
Patrick GALLINARI (Université Pierre et Marie Curie)
Osmar ZAÏANE (Université d'Alberta, Canada)

Jury :

Henri BRIAND	Professeur	Polytech'Nantes
Osmar ZAÏANE	Professeur Associé	Université d'Alberta, Canada
Djamel ZIGHED	Professeur	Université Lumière Lyon 2
Michel RUEHER	Professeur	Université de Nice Sophia Antipolis
Maguelonne TEISSEIRE	Maître de conférences	Université de Montpellier II
Brigitte TROUSSE	C.R. 1	INRIA Sophia Antipolis

2000 – 2001 **DEA d'Informatique** à l'Université de Nice Sophia Antipolis, France (classé 3ème/11).

Responsable du DEA : Professeur Emmanuel KOUNALIS.
Stage de DEA : dans l'équipe de recherche AxIS à l'INRIA Sophia Antipolis sous la coordination de Brigitte TROUSSE, de janvier 2001 à septembre 2001.
Sujet de stage : « Analyse et réutilisations de séquences temporelles. Application aux comportements d'internautes ».

1996 – 2000 **Maîtrise Informatique** à l'Université d'Ouest de Timisoara, Roumanie, Département Informatique, Faculté des Mathématiques (moyenne 9.67 sur 10).

1992 – 1996 **Baccalauréat en Sciences** au Lycée « Traian Lalescu », Resita, Roumanie (classe spéciale d'Informatique) (moyenne 9.44 sur 10).

2 Parcours professionnel

Nov. 2005 – Mai 2006 **Ingénieur de recherche** sur un contrat RNTL « EPIA » (« *Evolution d'un portail d'informations adaptatif* ») à l'INRIA Sophia Antipolis, Projet AxIS (*http://www-sop.inria.fr/axis/*), France

Sept.2004 – Déc. 2005 **Professeur d'informatique** (équivalent Moniteur) à l'Université Internationale de Monaco, (*http://www.monaco.edu/*), Monaco

Sept. 2001 – Juin 2005 **Doctorant** à l'INRIA Sophia Antipolis, Projet AxIS (*http://www-sop.inria.fr/axis/*), France

Janv. 2001 – Sept. 2001 **Stagiaire DEA** à l'INRIA Sophia Antipolis, Projet AxIS (*http://www-sop.inria.fr/axis/*), France

Mars 2000 – Sept. 2000 **Ingénieur de développement** (en Java, VB, Oracle - PL/SQL) pour IEngineer.com, Roumanie

Avr. 1999 – Mars 2000 **Ingénieur de développement** (en C++, Java) pour Optimal Solution (*http://www.optsol.at/*), Roumanie

Juin – Juil. et Oct. 1997 **Opérateur informatique** (saisie et formatage de données en SGML) pour DSS Romania (*http://www.dss.ro/*), Roumanie

3 Activités de recherche

3.1 Thèmes de recherche

Mes thèmes de recherche appartient au domaine de la fouille de données avec des contributions dans : le prétraitement des données Web, l'extraction des motifs séquentiels, en particulier l'extraction de motifs séquentiels à faible support. Dans la suite je donne une présentation détaillée de mes principaux résultats sur ces thèmes en commençant avec les deux contributions principales de ma thèse et continuant avec trois autres contributions dans le domaine de l'extraction des motifs séquentiels.

Méthodologie générale de prétraitement des logs Web (*thèse*)

Mots Clefs : *Web Usage Mining (WUM), fouille de données Web, journaux d'accès Web, méthodologie WUM, prétraitement WUM, WUM multi-sites*

Description : Le domaine du Web a connu une croissance formidable ces quinze dernières années tant dans le nombre de sites Web disponibles que dans le nombre d'utilisateurs de ces sites. Cette croissance a généré de très grandes masses de données relatives aux traces d'usage du Web par les internautes, celles-ci enregistrées dans des fichiers logs Web. De plus, les propriétaires de ces sites ont exprimé le besoin de mieux comprendre leurs visiteurs afin de mieux répondre à leurs attentes.

Le Web Usage Mining (WUM), domaine de recherche assez récent, correspond justement au processus d'extraction des connaissances à partir des données (ECD) appliqué aux données d'usage sur le Web. Il comporte trois étapes principales : le prétraitement des données, la découverte des schémas et l'analyse (ou l'interprétation) des résultats. Un processus WUM extrait des patrons de comportement à partir des données d'usage et, éventuellement, à partir d'informations disponibles sur le site (structure et contenu) et sur les utilisateurs du site (profils).

La quantité des données d'usage à analyser ainsi que leur faible qualité (en particulier l'absence de structuration) sont les principaux problèmes en WUM. Les algorithmes classiques de fouille de données appliqués sur ces données donnent généralement des résultats décevants en termes de pratiques des internautes (par exemple des patrons séquentiels évidents, dénués d'intérêt).

Dans ma thèse **[20]**, j'ai apporté une première contribution importante pour un processus WUM : une **méthodologie générale de prétraitement des logs Web**, domaine encore très peu abordé dans la littérature. L'originalité de la méthodologie de prétraitement proposée consiste dans le fait qu'elle prend en compte l'aspect **multi-sites** du WUM, indispensable pour appréhender les pratiques des internautes qui naviguent de façon transparente, par exemple, sur plusieurs sites Web d'une même organisation. Outre l'intégration des principaux travaux existants sur ce thème, j'ai proposé dans cette méthodologie quatre étapes distinctes : la fusion des fichiers logs, le nettoyage, la structuration et l'agrégation des données. En particulier, j'ai proposé plusieurs heuristiques pour le nettoyage des robots Web, le calcul des variables agrégées décrivant les sessions et les visites, ainsi que l'enregistrement de ces données dans un modèle relationnel. Plusieurs expérimentations ont été réalisées, montrant que la méthodologie proposée permet une forte réduction (jusqu'à 10 fois) du nombre des requêtes initiales et offre des logs structurés plus riches pour l'étape suivante de fouille de données.

Trois approches pour la découverte des motifs séquentiels de très faible support et application au Web Usage Mining (*thèse*)

Mots Clefs : *Web Usage Mining (WUM), fouille de données Web, fouille de données, méthodologie WUM, extraction de motifs séquentiels, support faible, classification non-supervisée, méthodologie divisive, boîte à outils WUM, Apriori-GST, AxisLogMiner*

Description : La deuxième contribution que j'ai apportée dans ma thèse **[20]** vise la découverte à partir d'un fichier log prétraité de grande taille, de comportements minoritaires correspondant à des motifs séquentiels de très faible support. Pour cela, j'ai proposé une méthodologie générale visant à

diviser le fichier log prétraité en sous-logs, se déclinant selon trois approches d'extraction de motifs séquentiels au support faible (**Séquentielle**, **Itérative** et **Hiérarchique**). Celles-ci ont été implémentées dans des **méthodes** concrètes **hybrides** mettant en jeu des algorithmes de classification et d'extraction de motifs séquentiels. Plusieurs expérimentations, réalisées sur des logs issus de sites académiques, m'ont permis de découvrir des motifs séquentiels intéressants ayant un support très faible, dont la découverte par un algorithme classique de type Apriori était impossible.
Enfin, j'ai proposé une **boîte à outils** appelée *AxisLogMiner*, qui supporte à la fois la méthodologie de prétraitement et, actuellement, deux méthodes concrètes hybrides pour la découverte des motifs séquentiels en WUM. Cette boîte à outils a donné lieu à de nombreux prétraitements de fichiers logs pour les besoins en recherche de l'équipe AxIS et de ses collaborateurs (sites Web de l'INRIA, de l'INRIA Sophia Antipolis et de l'Université de Recife, Brésil) et aussi à des expérimentations avec les méthodes implémentées.

GWUM : extraction des motifs séquentiels basée sur une généralisation des pages Web guidée par les usages

Mots cléfs : *Fouille des usage du Web, motifs séquentiels, classification*
Participants : Doru Tanasa, Florent Masseglia, Brigitte Trousse
Description : L'analyse des usages d'un site Web à partir d'une extraction de motifs est souvent limitée par le faible support de ces motifs. Cela est dû principalement à la grande diversité des pages et des comportements. Il est pourtant possible de regrouper la plupart des pages dans différentes catégories lors d'un pré-traitement. Travailler sur ces catégories, plutôt que sur les URLs, peut permettre de faire émerger certains comportements de manière "générique". Dans ce travail **[20,21]**, nous présentons une méthodologie originale d'analyse des usages du Web à partir d'une généralisation des URLs. Cette généralisation est réalisée à partir d'une catégorisation des URLs à l'aide d'informations extraites à partir de l'accès à ces pages par les internautes. Nous présentons ensuite une expérimentation relative à une généralisation des URLs basée sur des informations relatives aux accès aux pages : celle-ci permet de mettre en avant les changements de support des motifs extraits selon qu'ils sont obtenus avec ou sans généralisation.

Extraction de motifs séquentiels basée sur une indexation par arbre des suffixes généralisé : application aux navigations Web (*stage DEA*)

Mots clés : *comportement utilisateur*, *analyse des usages*, *sous-séquences fréquentes d'items*, *navigation*
Description : Dans le cadre de mon stage de DEA **[22]**, j'ai effectué des recherches sur l'indexation de séquences temporelles en utilisant un arbre des suffixes généralisé. Les séquences d'événements ont de nombreuses applications. Citons les alarmes dans un réseau de télécommunication, les données cliniques, les valeurs des actions sur le marché boursier, les sessions des utilisateurs d'un logiciel. Généralement on obtient ces séquences par l'observation des valeurs de paramètres d'un processus pendant une période de temps donnée. Mes recherches sur l'indexation de séquences temporelles s'intègrent dans celles sur les systèmes de recommandations du type Broadway menées dans l'équipe AxIS.
L'objectif de mon travail a été de permettre une recherche rapide de sessions similaires en termes de sous-séquences fréquentes. J'ai choisi d'indexer ces sessions avec un arbre des suffixes généralisé (*generalized suffix tree* – GST) qui permet la recherche d'un motif en $O(n)$ où n est la longueur du motif recherché, structure principalement utilisée pour l'indexation des chaînes de textes et des séquences génétiques. J'ai implémenté (en Java) l'algorithme appelé **Apriori-GST** qui utilisé cette méthode d'indexation de sessions. La structure de base de l'algorithme est similaire à celui d'Apriori. L'originalité de mon approche consiste dans l'utilisation de cet index (GST) pour l'indexation des séquences (sessions ou navigations Web). L'avantage de l'index GST dans ce cas est double :

1) l'incrémentalité qu'il apporte est très utile (on ne doit pas refaire l'index chaque fois que la base de données des sessions change, des nouvelles sessions sont ajoutées) et

2) l'algorithme Apriori-GST utilise l'index pour calculer le support d'une séquence lors de l'extraction des motifs ce qui augmente de manière significative les performances de cet algorithme par rapport à l'algorithme classique Apriori.

Extraction de motifs séquentiels à partir des profils d'expressions de gènes régulatrices

Mots clés : *motifs séquentiels, arbre de suffixes, Apriori-GST, puce Affymetrix, variations d'expression, puce à ADN, microarray, analyse de variations d'expressions, expression de gènes*
Participants : Doru Tanasa, Jesús López (Université d'Ulster, Irlande du Nord), Brigitte Trousse.
Description : Nous avons aussi appliqué l'algorithme Apriori-GST **[22]** dans le domaine de la bioinformatique **[5]**, lors d'une visite chez un de nos partenaires de l'action COST (« KnowlEST »), le groupe de recherche en bioinformatique de l'Université d'Ulster (Irlande du Nord).
Dans ce but, nous avons développé l'application GREPminer qui emploie Apriori-GST pour la découverte de motifs séquentiels fréquents à partir des séquences de variations d'expression de gènes régulatrices.
Etant donné l'arrivée de la technologie *microarray*, il est maintenant possible d'analyser l'expression d'un grand nombre de gènes simultanément. Dans l'article **[5]** nous nous sommes intéressés à appliquer nos méthodes d'extractions de motifs séquentiels aux données *microarray*. En particulier nous rapportons des résultats appliqués aux séquences d'expression de gènes associées au développement du cerveau de la souris. Ces données sont publiquement disponibles dans la base de données GEO (*http://www.ncbi.nlm.nih.gov/geo/*). L'intérêt biologique a été de comprendre la base moléculaire du développement neural du cerveau de la souris.
Les données *microarray* sont transformées en séquences de trois niveaux possibles d'exposition (e^+, e^0 ou e^-) et ces séquences sont indexées en utilisant GST. Un motif séquentiel extrait de ces données peut être vu, dans ce cas-ci, comme une sous-séquence de niveaux d'expression de gènes qui se produisent fréquemment. À partir des motifs extraits nous avons déduit l'hypothèse qu'il y avait une forte variation du niveau d'expression des gènes entre l'étape prénatale E18 et l'étape postnatale P7, variation qui devrait être étudiée plus en détail (sur plusieurs étapes).
L'outil GREPminer implémenté en Java nous a permis d'extraire et d'afficher ces séquences ainsi que le détail des gènes correspondants.
Les perspectives de ce travail concernent surtout l'interprétation génétique des phénomènes identifiés.

3.2 Perspectives de recherche

Mes perspectives de recherches concernent les extensions de la méthodologie de prétraitement WUM, les approches d'extraction de motifs séquentiels, ainsi que leurs applications éventuelles dans d'autres domaines comme, par exemple : la bioinformatique, l'analyse des usages des systèmes d'information, le Web sémantique, les données financières, etc.
Dans la suite, je décrit certaines recherches actuelles et futures.

Extension de la méthodologie de prétraitement WUM

J'envisage d'étendre la méthodologie de prétraitement proposée en **[20]** afin de :

- Prendre en compte la carte du site (sauvegardée dans une base de données ou dans un fichier XGMML) et utiliser éventuellement un système de versions (comme CVS) pour pallier les changements fréquents des sites Web qui nuiraient à la pertinence des résultats ;
- Traiter de manière différente les fichiers multimédia inclus dans les pages Web et ceux qui sont demandés par les utilisateurs ;
- Prendre en compte d'avantage d'informations sur les utilisateurs et sur le site Web, notamment dans le cas d'un site marchand où une page correspond à un produit ou à une catégorie de produits.

Nouvelles méthodes basées sur les trois approches proposées

J'envisage d'instancier les trois approches (Séquentielle, Itérative et Hiérarchique) que j'ai proposées dans ma thèse **[20]** dans d'autres méthodes d'extraction de motifs séquentiels en choisissant différents composants pour l'étape de classification et d'extraction de motifs.

L'évaluation de la qualité des motifs extraits est une étape essentielle que je voudrais aborder car c'est à travers cette étape qu'il sera possible de déterminer la méthode la mieux adaptée à un certain type ou domaine de données.

Interprétation et visualisation des résultats obtenus

Je suis aussi intéressé par la dernière étape d'un processus ECD, étape peu étudiée aujourd'hui car elle demande beaucoup d'investissement de la part de l'expert du domaine dans la conception de l'outil d'analyse. Dans la fouille des usages du Web, l'aide à l'interprétation à travers une visualisation est fortement souhaitable car à la fois les tailles des résultats et des données sont très importantes. J'envisage de poursuivre mes recherches sur les outils d'aide à la maintenance, re-conception et amélioration des sites Web.

3.3 Liste des publications[*]

Revue internationale avec comité de lecture (1)

[1] D. Tanasa, B. Trousse**.** *Advanced Data Preprocessing for Intersites Web Usage Mining*. IEEE Intelligent Systems, 19(2):59-65, March-April 2004

Revue internationale sur invitation (1)

[2] D. Tanasa, B. Trousse**.** *Data Preprocessing for WUM*. IEEE Potentials, 23(3):22-25, August-September 2004

Revue nationale avec comité de lecture (1)

[3] F. Masseglia, **D. Tanasa**, B. Trousse. *Diviser pour découvrir. Une méthode d'analyse du comportement de tous les utilisateurs d'un site Web*. In RSTI - Ingénierie des systèmes d'information (ISI), Vol. 9(1):61-83, 2004

Chapitre de livre (1)

[4] D. Tanasa, B. Trousse, F. Masseglia**.** *Mesures de l'internet*, chapitre *Fouille de données appliquée aux logs web : état de l'art sur le Web Usage Mining*, pages 126-143. Les Canadiens en Europe, 2004

Conférences internationales avec comité de sélection (4)

[5] D. Tanasa, J. Lopez, B. Trousse. *Extracting Sequential Patterns for Gene Regulatory Expressions Profiles*, In Knowledge Exploration in Life Science Informatics, KELSI 2004, Milan, Italy, November 2004. Proceedings, Vol. 3303:46-57 of LNCS, Springer-Verlag, 2004

[6] A. El Golli, B. Conan-Guez, F. Rossi, **D. Tanasa**, B. Trousse, Y. Lechevallier. *Les cartes topologiques auto-organisatrices pour l'analyse des fichiers Logs***.** In 11èmes Rencontres de la Société Francophone de Classification, Bordeaux, 8-10 Septembre, 2004

[7] F. Masseglia, **D. Tanasa**, B. Trousse. *Web Usage Mining: Sequential Pattern Extraction with a Very Low Support*. In Advanced Web Technologies and Applications: 6th Asia-Pacific Web Conference, APWeb 2004, Hangzhou, China, April 14-17. Proceedings, volume 3007 of LNCS, pages 513-522, 2004. Springer-Verlag

[*] Liste complète et ouvrages disponibles sur **:** **http://www-sop.inria.fr/axis/Publications/show.php?author=Doru_Tanasa**

[8] Y. Lechevallier, **D. Tanasa**, B. Trousse, R. Verde, *Classification automatique : Applications au Web Mining.* In Méthodes et Perspectives en Classification (10èmes Rencontres de la Société Francophone de Classification), Neuchâtel, Suisse, pages 157-160, Septembre 2003

Conférences nationales avec comité de sélection (3)

[9] D. Tanasa, B. Trousse, F. Masseglia. *Classer pour Découvrir : une nouvelle méthode d'analyse du comportement de tous les utilisateurs d'un site Web***.** Revue des Nouvelles Technologies de l'Information (RNTI), numéro spécial EGC'2004, 2:549-560, Janvier 2004

[10] F. Masseglia, **D. Tanasa**, B. Trousse. *Diviser pour Découvrir : une méthode d'analyse du comportement de Tous les utilisateurs d'un site Web*, BDA2003, Lyon, Octobre 2003

[11] D. Tanasa, B. Trousse. *Le prétraitement des fichiers log Web dans le Web Usage Mining Multi-sites*, In Journées Francophones de la Toile, Tours, Juillet 2003

Ateliers internationaux avec comité de sélection (5)

[12] S. Chelcea, A. Da Silva, Y. Lechevallier, **D. Tanasa**, B. Trousse. *Pre-Processing and Clustering Complex Data in E-Commerce Domain*. In Proceedings of the First International Workshop on Mining Complex Data 2005 (IEEE MCD'2005), held in conjunction with the Fifth IEEE International Conference on Data Mining (ICDM'05), Houston, Texas, 27 November 2005

[13] S. Chelcea, A. Da Silva, Y. Lechevallier, **D. Tanasa**, B. Trousse. *Benefits of InterSite Pre-Processing and Clustering Methods in E-Commerce Domain*, ECML/PKDD Discovery Challenge 2005, Porto, Portugal, October 2005

[14] M. Arnoux, Y. Lechevallier, **D. Tanasa**, B. Trousse, R. Verde. *Automatic Clustering for the Web Usage Mining*. In Proceedings of the 5th Intl. Workshop on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC03), Pages 54 - 66, Editura Mirton, Timisoara, October 2003

[15] 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

[16] D. Tanasa, B. Trousse. *Web Access pattern Discoevry and Analysis based on Page Classification and on Indexing Sessions with a Generalised Index Tree.* In Proceedings of SYNASC'01, Timisoara, Romania, October, 2001

Ateliers nationaux avec comité de sélection (3)

[17] S. Chelcea, A. Da Silva, Y. Lechevallier, **D. Tanasa**, Brigitte Trousse. *Prétraitement et classification de données complexes dans le domaine du commerce électronique*. In Atelier N°6: *Fouille de Données Complexes dans un processus d'extraction de connaissances*, EGC 2006, Pages 51 - 64, Lille, 17 Janvier 2006

[18] Y. Lechevallier, F. Masseglia, **D. Tanasa**, Brigitte Trousse. *Techniques des généralisation des URLs pour l'analyse des usages du Web*. In Atelier N°6: *Fouille de Données Complexes dans un processus d'extraction de connaissances*, EGC 2006, Pages 51 - 64, Lille, 17 Janvier 2006

[19] M. Arnoux, Y. Lechevallier, **D. Tanasa**, B. Trousse, R. Verde. *Classification automatique à partir de logs Web et de connaissances sur le site***.** In Atelier Fouille de Données Complexes à EGC'04, Clermont-Ferrand, pages 59-72, Janvier 2004

Thèse

[20] D. Tanasa. *Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support***.** Thèse de doctorat, Université de Nice Sophia Antipolis, 3 Juin 2005

Poster (1)

[21] D. Tanasa, B. Trousse. *Découverte et analyse de patrons d'accès Web*. In Poster papers in Extraction et de Gestion des Connaissances (EGC'02), pages 424-425, Janvier 2002

Mémoire de DEA

[22] D. Tanasa. *Analyse et réutilisations de séquences temporelles. Application aux comportements d'internautes*. Mémoire de DEA d'Informatique, Université de Nice Sophia Antipolis (UNSA), Juillet, 2001

Articles soumis (3)

[23] D. Tanasa, F. Masseglia, B. Trousse. *GWUM : une généralisation des pages Web guidée par les usages*. INFORSID 2006

[24] D. Tanasa, F. Masseglia, B. Trousse**.** *GWUM: Usage-driven Web Page Generalization*, DEXA 2006

[25] D. Tanasa, F. Masseglia, B. Trousse**.** *Encyclopedia of Data Warehousing and Mining*, 2nd Ed. Chapitre *Mining Generalized Web Data for Discovering Usage Patterns*, 2008

3.4 Participation à la communauté scientifique

Relectures d'articles

- **2004 – 2005**: IEEE Transactions on Knowledge and Data Engineering (6)
- **2004**: INFORSID 2004 (3)

Participation à des projets de recherche

- Participation au projet COLOR de l'INRIA, **e-Mimetic** (2005 – en cours) avec le LIRMM et l'Université du Sud Toulon-Var, « *Indicateurs de pertinence de moteur de recherche basé sur l'analyse mimétique* », Responsable du site Web *http://axis.inria.fr/e-mimetic*
- Participation au niveau européen à l'action COST **KnowlEST** (*Knowledge Exploration in Science and Technolgy*)

Participation à des groupes de travail nationaux et locaux

- Participation au groupe national de travail « **Fouille de Données Complexes** » créé par D. Zighed (2003 – en cours)
- Participation à l'action spécifique du CNRS « **Mesures de flux Internet** » (2003 – 2004), coordonnée par Ludovic Lebart et Valérie Beaudouin, Responsable du site Web *http://egsh.enst.fr/AS_fluxinternet/*
- Participation aux séminaires du **Laboratoire des usages de Sophia Antipolis**, Groupement d'Intérêt Scientifique entre le CNRS (STIC, SHS), le GET, l'INRIA et l'UNSA (2002).

3.5 Exposés scientifiques

Au niveau international

- KELSI2004 – Knowledge Exploration in Life Science Informatics, Milano, Italie
- APWEB2004 – The Sixth Asian Pacific Web Conference, Hangzhou, Chine
- SYNASC Workshops (*International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*) (2001, 2003), Timisoara, Roumanie
- SFC2003 (*Rencontres de la Société Francophone de Classification*), Neuchâtel, Suisse
- ICDM2002 (*International Conference on Data Mining*), Maebashi, Japon

Au niveau national

- Atelier FDC-EGC (*Fouille de Données Complexes dans un processus d'extraction de connaissances*) (2006), Lille, France
- EGC (*Extraction et Gestion des Connaissances*) (2002, 2004), France
- BDA2003 (*Bases de Données Avancées*), Lyon, France
- JFC2003 (*Journées Francophone de la Toile*), Tours, France

4 Activités pédagogiques

4.1 Synthèse des enseignements effectués entre 2001-2006

Matière	Resp./ Co-Resp.	Etablissement, Filière	Volume horaire (heures eq. TD)		
			Cours	TD	TP
Web Usage Mining en SAS (projet étudiant tuteuré)	Co-R puis R	IUT Menton, LP SID		138	
Programmation Java		Ecole Polytechnique de l'Université de Nice Sophia Antipolis (Polytech'Nice-Sophia), ESSI 1			145
Techniques de Programmation	R	Polytech'Nice-Sophia, ITII	270		
Algorithmique	R	Polytech'Nice-Sophia, ITII	36		
Data Mining et Web Mining		Ecole Doctorale STIC, Master	4,5		
Computer-based Systems*	R	Université Internationale de Monaco (IUM)	135		
Principles of Programming*	R	IUM	202,5		
Management Information Systems*	Co-R	IUM	33		
Business Analysis & Systems Design*	R	IUM	67,5		
Total : 1031,5 heures			**748,5**	**138**	**145**

*Cours en anglais

4.2 Description détaillée des enseignements effectués par matière

Web Usage Mining en SAS, Université de Nice Sophia Antipolis, IUT Menton, LP SID, 2001 – 2004, 2005 – 2006

Responsables : Brigitte Trousse (2001 – 2002) et Doru Tanasa (2001 – 2004, 2005 – 2006),

Description : J'ai encadré le projet tuteuré des étudiants de la Licence Professionnel « Statistique et Informatique Décisionnelle ». Le projet consiste dans l'analyse statistique avancée des fichiers logs des sites Web INRIA en utilisant des logiciels tel que SAS et Microsoft SQL Serveur. Les étudiants ont dû dans une première étape structurer les données et les stocker dans une base de données Microsoft SQL Serveur. Ensuite, ils ont analysé les différents objets de cette base (sessions, navigations, utilisateurs) avec des techniques statistiques telles que : analyse en composantes principales, règles d'association, clustering, etc.

Programmation Java, Polytech'Nice-Sophia (ESSI), 1ère année, 2001 – 2003

Responsable : Peter Sander (sander@essi.fr)

Description : J'ai assuré les TPs de Java pour le cours de Programmation Java. A travers ce module les étudiants prennent connaissance de la programmation objet en Java. J'ai aussi participé à l'évaluation des étudiants car certains des TPs étaient notés.

Techniques de Programmation, Polytech'Nice-Sophia (ESINSA), ITII, 2001 – 2006

Responsable : Doru Tanasa

Description : Le module fait partie d'une formation ITII (Institut des techniques d'ingénieur de l'industrie) qui est une formation par alternance pour les ingénieurs (de BAC+2 à BAC+5). Je suis le responsable du cours depuis la première promotion des étudiants (2001). J'ai mis en place ce

cours qui inclut des notions sur les types de données, les tableaux, les pointeurs et les fichiers. Les étudiants doivent aussi réaliser un projet comptant pour la note finale. J'ai préparé les supports, les sujets des TPs et des projets ainsi que les sujets d'examen.

Algorithmique, Polytech'Nice-Sophia (ESINSA), ITII, 2004 – 2006

Responsable : Doru Tanasa

Description : Depuis 2004 je suis le responsable de ce module pour la formation ITII. Le programme que j'ai mis en place (3 cours) inclut : les algorithmes de tri, la récursivité et la librairie graphique *libsx*.

Data Mining et Web Mining, Ecole Doctorale STIC de l'Université de Nice Sophia Antipolis, Master PMLT, 2004 – 2006

Responsables : Martine Collard (mcollard@unice.fr) et Brigitte Trousse

Description : Je suis intervenu dans ce module optionnel du Master PLMT pour présenter le domaine du Web Usage Mining et les techniques de pré-traitement et d'extraction de motifs séquentiels à faible support.

Computer-based Systems, Université Internationale de Monaco (IUM), 1ère année, 2004 – 2005

Responsable : Doru Tanasa

Description : L'IUM est une école de commerce en langue anglaise située à Monaco. Les étudiants de l'IUM proviennent principalement d'Europe mais aussi des Etats-Unis, d'Australie, d'Asie. Afin de financer ma quatrième année de doctorat j'y ai travaillé en tant que professeur d'informatique (équivalent moniteur) délivrant plusieurs cours dont celui-ci sur les « Systèmes basés sur les ordinateurs ».

Le cours est une introduction au système d'exploitation Windows et aux outils de bureautique de Microsoft Office (Word, Excel, PowerPoint). Par ailleurs, j'ai présenté aussi une courte histoire de l'informatique, un cours sur les virus et la sécurité. J'ai été le responsable de deux groupes d'étudiants (d'environ 20 personnes chacun) et j'ai préparé les supports de cours, les devoirs et les corrections ainsi que les 2 examens (intermédiaire et final).

Principles of Programming, Université Internationale de Monaco (IUM), 2ème année, 2004 – 2006

Responsable : Doru Tanasa

Description : Ce cours est destiné aux étudiants ayant une certaine expérience en informatique, mais aucune expérience précédente en programmation. Le cours est conçu pour enseigner des pratiques de programmation applicables à n'importe quel environnement de programmation, mais avec une attention spécifique sur le langage de programmation Visual Basic sous Windows. Comme pour le cours précèdent, j'ai été le responsable de 2 groupes en 2004 – 2005 et d'un groupe en 2005 – 2006. Pour promouvoir ce cours les étudiants ont dû en plus des examens (2) réaliser un projet en binôme sous Visual Basic.

Management Information Systems, Université Internationale de Monaco (IUM), 4ème année, 2004 – 2006

Responsables : William Ligthfoot (wlightfoot@monaco.edu) et Doru Tanasa

Description : Le cours des systèmes d'information présente aux étudiants les principales notions des NTIC (hardware, software, réseaux, sécurité, SGBD), l'intégration d'un processus économique avec la technologie, par exemple, dans le E-Commerce. Le cours a été enseigné conjointement avec le professeur W. Lightfoot, responsable du programme MBA de l'université.

J'ai participé à l'élaboration et la présentation des supports du cours, ainsi qu'à l'évaluation des étudiants.

Business Analysis & Systems Design, Université Internationale de Monaco (IUM), 1er année, 2004 – 2006

Responsable : Doru Tanasa

Description : Ce cours que j'ai mis en place est une introduction à l'utilisation des systèmes d'information pour l'aide à la décision dans les organisations. Le contenu du cours inclut des concepts de *hardware* et *software*, les applications, et des concepts de télécommunication. Pendant le cours nous avons examiné et expliqué :

- la nature de l'information et son utilisation dans la prise de décision,
- le rôle du système d'information dans la stratégie d'organisation,
- la manière dont l'information est organisée, stockée et traitée par la technologie moderne de l'information,
- les développements récents en technologie de l'information,
- la façon dont les développements dans les réseaux et l'Internet ont eu un impact sur les affaires et
- le processus d'analyse et de conception d'un système d'information économique.

En tant que responsable du cours j'ai préparé et présenté les supports, préparé et corrigé les contrôles continus (4) et les examens (2).

4.3 Encadrement

J'ai été le responsable du projet de fin d'études d'Arnaud Santoni, étudiant de l'EPU Nice-Sophia (ESSI 3), projet effectué à compter du 15.10.2002 et jusqu'au 01.04.2003 sur le sujet « *Construction de la structure logique d'un site Web en XML* ». Il a travailler au développement de l'application WebLogic qui permet d'extraire la structure d'un site Web dans une base de données ou dans un fichier XGMML.

5 Logiciels développés

5.1 Sommaire des logiciels développés

No	Nom	Employeur	Période	Langages et Techniques utilisées
1	Structured Decision Tree Designer	Optimal Solution	04/1999 – 03/2000	Java et C/C++
2	Address Module	Optimal Solution	04/1999 – 03/2000	Visual C++ 6.0, MFC
3	AxISLogMiner Preprocessing	INRIA	Pendant ma thèse, entre : 10/2001 – 06/2005	Java, Perl, Scripts Shell
4	Cluster & Discover	INRIA		Java
5	Apriori-GST/GREPminer	INRIA		Java

5.2 Description des logiciels développés

Structured Decision Tree Designer

Cette application est à présent commercialisée par la société Optimal Solution (Vienne, Autriche) : *http://www.optsol.at/site_en/products/ods/dTree.html*, dans le cadre de son système d'aide à la décision appelé « Optimal Decision System » : *http://www.optsol.at/site_en/products/ods/index.html*.

L'application permet la construction, visualisation, manipulation et validation des arbres de décision ainsi que la transformation d'une table de décision dans un arbre de décision. L'application inclut aussi la possibilité de générer du code C++ à partir des arbres de décision. En effet, une fois l'arbre de décision construit, l'utilisateur a la possibilité d'exporter du code C/C++ fonctionnel en utilisant éventuellement ses propres fichiers de définitions (.h). A travers le mécanisme JNI, le code généré peut être débugué avec l'application Java.

Address Module

J'ai travaillé sur le module d'administration des adresses des clients et des fournisseurs pour une application de gestion, en réalisant les tâches suivantes :

- Réalisation de l'interface graphique en utilisant les classes MFC de Microsoft;
- L'ajout/modification/suppression d'une adresse dans la base de données des adresses ;
- La visualisation des adresses avec la possibilité de trier par différents critères ;
- La personnalisation des informations affichées (en colonnes) dans un format similaire avec l'explorateur de Windows.

AxISLogMiner Preprocessing

L'application *AxISLogMiner Preprocessing* (*http://www-sop.inria.fr/axis/axislogminer/*) fait partie d'une plateforme spécialement créée pour la fouille des données d'usage du Web, appelée *AxISLogMiner*.

Cette plateforme a été développée pendant ma thèse **[20]** à l'INRIA Sophia Antipolis et comprends plusieurs modules écrits en Perl, des scripts Shell et des packages Java. Elle représente l'implémentation des méthodologies de prétraitement et analyse de données log Web proposé dans ma thèse de doctorat.

L'application est destinée aux systèmes Linux/Unix mais son portage en Windows est réalisable en utilisant éventuellement la librairie cygwin.

AxISLogMiner Preprocessing est un ensemble de programmes Perl et de scripts Shell qui offre la possibilité de prétraiter des fichiers log Web de grande taille (plusieurs GO) et provenant de plusieurs sites Web partenaires en vue d'une analyse Data Mining. Les données sont structurées et stockées dans une base de données MySQL à l'aide d'un module Java.

L'application est en phase de finalisation en vue de la rendre disponible en version libre (licence GPL).

Cluster & Discover

L'outil *Cluster & Discover* fait partie de la boîte à outils *AxisLogMiner* et implémente la méthode de même nom proposée dans ma thèse **[20]**. L'outil est implémenté en Java est il fait appel à la méthode PSP pour l'extraction des motifs fréquents.

Dans une première étape, les navigations des utilisateurs (sequences des pages Web) sont groupées dans des sous-logs à l'aide d'un algorithme de classification neuronal développé dans l'équipe AxIS et modifié pour l'occasion. Ensuite, sur ces sous-logs, l'utilisateur a le choix d'appliquer ou non (en fonction de son contenu) un processus d'extraction de motifs fréquents à l'aide de l'agorithme PSP développé par Florent Masseglia.

Les résultats sont présentés dans l'interface graphique et consistent dans des sous-séquences de navigations communes à au moins *x*% des utilisateurs.

Apriori-GST/GREPminer

Implémenté aussi en Java, l'algorithme *Apriori-GST* **[22]** permet d'extraire des motifs séquentiels à partir des séquences d'items. L'outil *GREPminer* emploie cet algorithme tout en ajoutant une interface graphique qui facilite la sélection, l'extraction et la visualisation des motifs de variation fréquents d'expression de gènes. Il prend en entrée des données *microarray* prétraitées et l'utilisateur choisit un support minimum pour les motifs extraits.
L'outil *GREPminer* est issu d'une collaboration avec le groupe de recherche en bioinformatique de l'Université d'Ulster en juillet 2004 et a été développé suite à mon visite dans ce laboratoire (visite financée par le projet européen KnowlEST).

6 Compétences

Sept années de développement logiciel dans les langages Java, C/C++, Perl sous Windows et Linux
Langages de Programmation : Java, C/C++, Perl, VB, SQL, PL/SQL
Bases de Données : MySQL, Oracle, Microsoft SQL Server, SAS, Microsoft Access
Systèmes d'exploitation : Windows (9x, NT, 2K, XP), Linux
Domaines de compétences : Fouille de données Web (usage, contenu, structure), Fouille de données, Extraction de Connaissances à partir de Données (ECD), Extraction des motifs séquentiels, Bases de données, Technologies Web, Web Sémantique, Algorithmes, Bioinformatique, Systèmes d'information

7 Participations aux concours

- *Concours de Programmation ACM*, phase Europe de Sud-Est – avec l'équipe de l'Université d'Ouest de Timisoara (1998, 1997), Bucarest, Roumanie
- *Concours National d'Informatique* (3ème place) avec l'équipe de l'Université d'Ouest de Timisoara (1996), Iasi, Roumanie

8 Langues

Roumain – Langue maternelle
Français – Bilingue (diplôme DALF)
Anglais – Courant (300h de cours en anglais à l'Université Internationale de Monaco)
Italien – Intermédiaire (lu, parlé)
Russe – Notions

9 Références

Dr. Brigitte TROUSSE
Chargée de Recherche 1, Responsable du Projet Axis, Inria Sophia Antipolis
Adresse : Inria, AxIS, 2004 Route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex, France
E-mail : Brigitte.Trousse@inria.fr **Tel :** 04 92 38 77 45 **Fax :** 04 92 38 77 55

Dr. Yves LECHEVALLIER
Directeur de Recherche, Responsable Scientifique du Projet AxIS, INRIA Rocquencourt
Adresse : INRIA-Rocquencourt, 78153 Le Chesnay cedex
E-mail : Yves.Lechevallier@inria.fr **Tel :** 01 39 63 54 34 **Fax :** 01 39 63 58 92

Prof. Dr. Jacques LEMAIRE
Professeur à l'IUT Nice - Côte d'Azur, Département STID
Adresse : 58, Chemin du collège, 06500 - Menton
E-mail : Jacques.Lemaire@unice.fr **Tél.:** 04 93 28 66 86 **Fax :** 04 93 28 66 81

Prof. Dr. Robert VIANI
Directeur des Études ITII, EPU Nice-Sophia
Adresse : Parc de Sophia Antipolis, 1645,route des Lucioles - F - 06410 Biot
E-mail : viani@unice.fr **Tél :** 04 92 38 85 52 **Fax** : 04 92 38 85 01

Prof. Dr. Antonella PATRAS
Doyen Adjoint, Directeur des Programmes Undergraduate, Université Internationale de Monaco
Adresse : International University Of Monaco, 2, Rue Du Prince Hereditaire Albert, 98000 Monaco
E-mail : apatras@monaco.edu **Tel :** +377 97 986 986 **Fax :** +377 92 052 830

10 Travaux sélectionnés

Les publications suivantes seront envoyées en cas de sélection pour l'audition :

D. Tanasa, B. Trousse**.** *Advanced Data Preprocessing for Intersites Web Usage Mining*. IEEE Intelligent Systems, 19(2):59-65, March-April 2004

D. Tanasa. *Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support***.** Thèse de doctorat, Université de Nice Sophia Antipolis, 3 Juin 2005.

D. Tanasa, J. Lopez, B. Trousse. *Extracting Sequential Patterns for Gene Regulatory Expressions Profiles*, In Knowledge Exploration in Life Science Informatics, KELSI 2004, Milan, Italy, November 2004. Proceedings, Vol. 3303:46-57 of LNCS, Springer-Verlag, 2004

D. Tanasa, B. Trousse, F. Masseglia. *Classer pour Découvrir : une nouvelle méthode d'analyse du comportement de tous les utilisateurs d'un site Web***.** Revue des Nouvelles Technologies de l'Information (RNTI), numéro spécial EGC'2004, 2:549-560, Janvier 2004

D. Tanasa, F. Masseglia, B. Trousse**.** *GWUM: Usage-driven Web Page Generalization*, DEXA 2006 (*article soumis*)