ACTES DES JOURNEES

LA FOUILLE DANS LES DONNEES PAR LA METHODE D'ANALYSE STATISTIQUE IMPLICATIVE

Applications et traitement par C.H.I.C.

23-24 Juin 2000
à l'Institut Universitaire de Formation des Maîtres
de l'Académie de Caen

Edition des Actes coordonnée par Régis Gras et Marc Bailleul

AVANT-PROPOS

En Juin dernier, à l'Institut de Formation des Maîtres de Caen, co-organisateur avec l'Association de Recherche en Didactique des Mathématiques (A.R.D.M.), se sont déroulées des Journées internationales francophones sur le thème : « La fouille dans les données par la méthode d'analyse implicative ». Ces Journées font suite au colloque, « Méthodes d'Analyses de Données Multidimensionnelles », tenu en janvier 1995 dans le même I.U.F.M. et organisé, comme celui-ci, par Marc Bailleul, Maître de Conférences dans cet I.U.F.M. et par Régis Gras, Professeur Emérite à l'Ecole Polytechnique de l'Université de Nantes. Le thème des Journées, plus spécifique que celui du Colloque, était circonscrit cette fois aux travaux portant sur la méthode d'analyse implicative développée par R.Gras et ses élèves depuis vingt ans et contributive aux recherches sur la « Fouille dans les données » dans des domaines variés dont, bien sûr, la didactique des mathématiques. Ce thème a permis de rassembler des chercheurs de disciplines différentes et de plusieurs nationalités : Belge, Espagnole, Française, Italienne et Suisse.

Une "fouille" dans des données (encore appelée "Knowledge Discovery in Databases" ou encore "Data Mining" dans la littérature anglo-saxonne) part, en général, du croisement de sujets (ou objets) et de variables (propriétés ou attributs) binaires, ordinales ou numériques. L'analyse implicative vise l'extraction de connaissances, d'invariants, de règles non symétriques consistantes, du type « l'attribut a implique généralement l'attribut b » ou « quand a est choisi, on a tendance à choisir b ». Elle en mesure la qualité sur la base statistique d'un nombre significatif de contre-exemples où la règle n'est pas vérifiée et où certains déséquilibres cardinaux sont observés parmi les exemples et les contre-exemples à l'implication et à sa contraposée. Le logiciel, dénommé CHIC (Classification Hiérarchique Implicative et Cohésitive), développé maintenant par Raphaël Couturier, permet de :

- quantifier la significativité de ces nombres, la consistance de la règle associée, de classes de règles, la contribution des sujets ou de catégories de sujets à certaines règles,
- représenter, par un graphe, des chaînes de règles et, par une hiérarchie, des règles sur des règles que l'on appelle aussi méta-règles,
- supprimer, d'ajouter, de conjoindre ou de disjoindre des variables.

Au cours des deux Journées, des communications de chercheurs français ou étrangers ont rendu compte de résultats très significatifs obtenus dans différents domaines par une "fouille" dans leurs propres données. L'analyse implicative a permis de faire apparaître, de façon complémentaire à des méthodes factorielles ou taxonomiques, des règles sur la gestion des personnels dans l'entreprise, des règles sur l'attitude de consommateurs, des liaisons stables entre phénomènes et paramètres en biologie, des invariants en matière de psychologie d'apprentissage (didactique et sciences de l'éducation), des règles dans la représentation sociale, etc. Les conférences de Yves Kodratoff et Régis Gras ont fait le point respectif sur l'emploi des méthodes de fouille dans l'analyse de textes littéraires et sur la théorie implicative. Les débats, qui ont systématiquement accompagné chaque intervention, ont été très animés. Ils ont permis des échanges fructueux grâce à leur variété tant au niveau des contenus des applications présentées qu'à celui des approches et des méthodologies venues de champs scientifiques différents.

Au cours de trois ateliers, après un apprentissage du fonctionnement du logiciel CHIC et à partir d'un exemple (un fichier de données binaires), les participants ont pu effectuer les calculs liés à la théorie implicative, puis les valider par l'emploi du logiciel. Cette activité visait à démystifier les algorithmes de calculs opérés automatiquement et à permettre une

interprétation critique et plus maîtrisée des résultats obtenus au cours de l'utilisation de la méthode.

Au milieu des Journées, un repas normand, concocté par Marc Bailleul a permis d'apprécier les recettes et les produits régionaux, mais également a conduit à des échanges conviviaux entre tous les participants.

Les actes présents rassemblent les différentes formes de contributions. Mais l'intérêt exprimé par la communauté des chercheurs statisticiens pour ce type de méthode associée à ses applications doit permettre la publication d'un numéro spécial de la revue « Mathématiques et Sciences Humaines ».

Nous tenons à remercier les différents organismes qui nous ont apporté un soutien à la tenue de ces Journées sur les plans scientifique, financier et logistique :

l'Association pour la Recherche en Didactique des Mathématiques (ARDM),

l'Institut Universitaire de Formation des Maîtres de l'Académie de Caen,

l'Ecole Polytechnique de l'Université de Nantes,

l'Institut de Recherche Informatique de Nantes (IRIN),

la Société PerformanSe, SA, Nantes,

la Ville de Caen,

l'International Association for Statistical Education (IASE),

ou leur parrainage scientifique *:*

la Société Francophone de Classification (SFC),

l'Université de Caen et

le Laboratoire de Recherche Informatique d'Orsay (LRI).

Marc Bailleul et Régis Gras, organisateurs des Journées

Comité scientifique et de programme

Marc BAILLEUL, IUFM de Caen,

Carmen BATANERO, Université de Grenade,

Henri BRIAND, Ecole Polytechnique de l'Université de Nantes et IRIN,

Edwin DIDAY, INRIA et CEREMADE Paris Dauphine,

Régis GRAS, Ecole Polytechnique de l'Université de Nantes et IRIN,

Yves KODRATOFF, CNRS, L.R.I. Université de Paris-Sud Orsay,

Pascale KUNTZ, Ecole Polytechnique de l'Université de Nantes et IRIN,

EduardoLACASTA, Université de Navarre, Pampelune,

Jean LEJEUNE, Université de Caen,

Israël-César LERMAN, IRISA, Université de Rennes,

Amedeo NAPOLI, LORIA, Université de Nancy,

Guy NOEL, Université de Mons,

Maria-Gabriella OTTAVIANI, Université La Sapienza, Rome,

Marie-Jeanne PERRIN, IUFM d'Amiens et DIDIREM Paris 7,

Jacques PHILIPPE, Ecole Polytechnique de l'Université de Nantes, IRIN et Société Performanse

Conférences

Régis Gras *(Université de Nantes)* : "Les fondements de l'analyse statistique implicative"
Yves Kodratoff *(CNRS et Université Orsay)* : "Les trois approches de la mesure de confirmation de l'implication : application à la fouille de textes"

Communications

Marc Bailleul *(IUFM de Caen)*: "Le mémoire professionnel, des réseaux de représentations fortement différenciés"
Antoine Bodin *IREM-Université de Besançon)*: " Apports de l'analyse implicative à l'exploitation des études à grande échelle"
Robin Gras, Régis Gras, Markus Müller, Ron D. Appel *(Université de Genève, Université de Nantes):* " Classification de protéines par un algorithme génétique et par l'implication statistique dans le cadre de l'identification automatique par "peptide mass fingerprinting""
Sylvie Guillaume, Ali Khenchaf *(Université de Nantes)*: "Règles ordinales : une généralisation des règles d'association. Application"
Eduardo Lacasta Zabalza *(Universidad de Navarre)*: "L'illusion graphique chez les élèves du secondaire"
Dominique Lahanier-Reuter *(Université de Lille III)*: "Algorithme de construction d'implications statistiques entre deux variables quantitatives. Application"
Patrick Leconte *(Université de Tours)* "Application à une recherche sur l'émergence et l'évolution des représentations sociales d'un outil de gestion des ressources humaines"
Rémi Lehn, Fabrice Guillet, Pascale Kuntz *(Université de Nantes)*: "Félix : un outil interactif d'aide à la fouille de connaissances s'appuyant sur l'intensité d'implication"
Guy Noël *(Université de Mons)*: "Des difficultés de l'initiation à l'algèbre"
Pilar Orüs Baguena *(Universidad de Castellon)* "Utilisation didactique des tableaux des données et du logiciel CHIC à l'école élémentaire"
Maria Gabriella Ottaviani, Silvia Zannoni *(Università di Roma "La Sapienza")* : "L'implication statistique et la didactique. L'utilisation d'un outil non symétrique d'analyse de données pour l'interprétation des résultats d'un test d'évaluation"
Jacques Philippé, T.Teusan, S. Baquedano, C.Bourcier *(Université de Nantes, ESC de Rouen, Performanse)* : L'analyse implicative dans un contexte d'extraction de connaissances pour la mise au point de systèmes d'aide à la décision en analyse des comportements"
Maria Polo, Michela Maggio *(Università di Cagliari)* : "L'implication statistique dans l'analyse du phénomène de l'abandon scolaire au niveau DEUG 1ère année en Italie"
Djamel A. Zighed *(Université Lyon 2)* : "Regroupement de modalités pour maximiser l'association de deux variables"

PROGRAMME DES JOURNEES

Thèmes des Journées

Thème 1 : *Sur une application, étude des spécificités de la méthode implicative par rapport à d'autres méthodes exploratoires*
Thème 2 : *D'une problématique exploratoire à l'analyse par la méthode implicative*
Thème 3 : *Elaboration de nouveaux concepts d'association en réponse à une problématique exploratoire*

Vendredi 23 juin

9h-9h30 Accueil

9h30-10h Ouverture : **Marie-Jeanne Perrin**, Présidente de l'A.R.D.M., **Henri Briand**, Professeur à l'Ecole Polytechnique de l'Université de Nantes

10h-11h Conférence : **Régis Gras,** Université de Nantes, *"Les fondements de l'analyse statistique implicative"*

11h-11h15 Pause

11h15-12h30 T.P. CHIC : **Marc Bailleul**, IUFM de Caen, **Antoine Bodin,** IREM-Université de Besançon, **Raphaël Couturier,** IUT de Belfort, **Régis Gras**

14h30-15h30 Communications thème 1 :
Rémi Lehn, Fabrice Guillet et Pascale Kuntz., Ecole Polytechnique de l'Université de Nantes, *"Félix : un outil interactif d'aide à la fouille de connaissances s'appuyant sur l'intensité d'implication"*
Robin Gras, Markus Müller et **Ron D.Appel**, Université de Genève, **Régis Gras**, *« Classification de protéines par un algorithme génétique et par l'implication statistique dans le cadre de l'identification automatique par "peptide mass fingerprinting" »*

15h30-16h30 Communications thème 1 :
Pilar Orüs, Universidad de Castellon, *"Utilisation didactique des tableaux des données et du logiciel CHIC à l'école élémentaire"*
Patrick Leconte, Université de Tours, *"Application à une recherche sur l'émergence et l'évolution des représentations sociales d'un outil de gestion des ressources humaines"*

16h30-17h Pause

17h- 18h 30 Communications thèmes 1 et 2 :
Eduardo Lacasta, Universidad de Navarre, *"L'illusion graphique chez les élèves du secondaire"*
Guy Noël, Université de Mons, *"Des difficultés de l'initiation à l'algèbre"*
Jacques Philippé, T.Teusan, Serge Baquedano, C.Bourcier, Université de Nantes, ESC de Rouen, Société Performanse, *"L'analyse implicative dans un contexte d'extraction de connaissances pour la mise au point de systèmes d'aide à la décision en analyse des comportements"*

18h30-19h30 T.P. CHIC : **Marc Bailleul, Antoine Bodin, Raphaël Couturier, Régis Gras**

Samedi 24 juin

9h-10h Conférence :

Yves Kodratoff , CNRS et Université d'Orsay, *"Les trois approches de la mesure de confirmation de l'implication : application à la fouille de textes"*

10h-11h Communications thème 2 :

Maria-Gabriella Ottaviani et Silvia Zannoni., Università di Roma "La Sapienza", *"L'implication statistique et la didactique. L'utilisation d'un outil non symétrique d'analyse de données pour l'interprétation des résultats d'un test d'évaluation"*

Dominique Lahanier-Reuter, Université de Lille III, *"Algorithme de construction d'implications statistiques entre deux variables quantitatives. Application"*

11h-11h30 Pause

11h30-12h30 Communications thème 2 :

Marc Bailleul, IUFM de Caen,"*Mise en évidence de réseaux orientés de représentations dans deux études concernant des enseignants-stagiaires de l'IUFM"*

Arnaud Simon, Université de Nancy, *Analyse par l'implication statistique d'une enquête sur des données médicales*

14h30-16h Communications thème 3 :

Sylvie Guillaume et Ali Khenchaf., Université de Nantes, *"Règles ordinales : une généralisation des règles d'association. Application"*

Djamel Zighed, Université Lyon 2, *"Regroupement de modalités pour maximiser l'association de deux variables"*

Antoine Bodin, *" Apports de l'analyse implicative à l'exploitation des études à grande échelle"*

16h-16h15 Pause

16h15-17h15 T.P. CHIC : **Marc Bailleul, Antoine Bodin, Raphaël Couturier, Régis Gras**

17h15-17h45 Bilan-Clôture

Ouverture des Journées par Marie-Jeanne PERRIN, Présidente de l'Association pour la Recherche en Didactique des Mathématiques (A.R.D.M.)

C'est un grand plaisir pour moi de participer à l'ouverture de ces journées en tant que présidente de l'A.R.D.M. et aussi à titre personnel.

En cette année mondiale des mathématiques, les colloques fleurissent et ce début d'été est particulièrement chargé : entre la mi-juin et la mi-juillet, j'ai compté au moins quatre ou cinq colloques en France qui intéressent au premier chef la didactique des mathématiques et il en est de même un peu partout dans le monde. Abondance de biens ne nuit pas dit-on. Peut-être en ce qui concerne la participation. Si les courriers électroniques se déplacent quasi instantanément, malgré les moyens de transport rapides, nul de dispose encore du don d'ubiquité et les jours n'ont toujours que 24 heures, ce qui fait que nous n'avons pas toujours la possibilité d'assister à toutes les rencontres qui nous auraient intéressés. Mais ces colloques donneront lieu à des écrits, des productions que chacun pourra ensuite étudier à loisir, ce qui permettra des retombées bien au delà de cette année.

Le but de l'Association pour la Recherche en Didactique des Mathématiques est de favoriser le développement et le rayonnement de la recherche en didactique des mathématiques. Le soutien à des colloques est un moyen d'atteindre ce but, venant compléter les actions régulières que sont le séminaire national, les Écoles d'été et bien sûr la publication de la revue Recherches en didactique des mathématiques avec la collection associée d'ouvrages. C'est donc tout naturellement que l'association soutient plusieurs des manifestations qui concernent cette année la didactique des mathématiques et je m'en réjouis.

Mais si, parmi tous ces colloques, il m'est particulièrement agréable d'inaugurer celui-ci, c'est qu'un des membres fondateurs et un des plus actifs de notre association y joue un rôle central. Je veux parler bien sûr de Régis Gras qui est à la fois co-organisateur de ces journées, conférencier, et surtout à l'origine de la méthode que nous allons y étudier.

En didactique des mathématiques, on étudie des phénomènes complexes où interviennent des quantités de facteurs et il est bien difficile d'organiser ces facteurs, surtout si on ne veut pas se contenter de réponses en oui / non , et qu'on veut non seulement des corrélations mais aussi des dépendances entre variables. La méthode mise au point par Régis Gras et par les chercheurs qui ont travaillé avec lui – plusieurs sont présents ici – répondait à un grand besoin de la didactique des mathématiques. La conception d'un logiciel qui la met en pratique dote les didacticiens d'un outil à la fois performant et convivial, comme le montrent beaucoup de travaux. Le colloque qui s'est tenu ici même en 1995 en est le témoin. Mais ce que montrent de plus les présentes journées, à travers les multiples applications qui en sont faites, c'est que le domaine d'application de la méthode d'implication statistique dépasse largement la didactique des mathématiques. Je n'en dirai pas plus. Il y a ici beaucoup de gens plus compétents que moi pour en parler. Je tenais cependant à remercier et féliciter Régis pour son travail fructueux pour la didactique des mathématiques.

J'ai une autre raison de me réjouir de ces journées, c'est leur tenue à l'I.U.F.M. de Caen que je remercie particulièrement de son accueil, une nouvelle fois, au nom de l'A.R.D.M.. J'ai déjà parlé du colloque qui s'est tenu en 1995, mais il me faut encore mentionner les Écoles d'été de didactique des mathématiques qui se sont tenues à Houlgate en 1997 et 1999 et pour lesquelles l'I.U.F.M. de Caen a apporté un soutien logistique important. A l'I.U.F.M. de Caen, je voudrais particulièrement remercier notre collègue Marc Bailleul qui, dans chacune de ces manifestations, a fait preuve d'une efficacité et d'un dévouement

remarquables. Il est aussi l'un de ceux qui ont travaillé au développement de la méthode d'analyse implicative.

Je suis très heureuse que ces journées se tiennent dans un I.U.F.M. car il est important de montrer qu'une recherche de qualité peut et doit se développer dans les I.U.F.M. C'est un point qui me tient particulièrement à cœur et que nous essayons de mettre en pratique à l'I.U.F.M. Nord-Pas-de-Calais où j'exerce.

Il me reste à espérer que vous trouverez dans ce colloque l'occasion d'enrichir vos connaissances et vos propres recherches. Je souhaite à chacun un fructueux travail.

Marie-Jeanne Perrin-Glorian
Professeur à l'Université d'Artois,
Directeur d'études à l'I.U.FM. Nord-Pas-de-Calais
Présidente de l'A.R.D.M.

Ouverture des Journées par Henri BRIAND

GESTION DE LA CONNAISSANCE DANS L'ENTREPRISE ET ANALYSE IMPLICATIVE

Plus de 40% des 1000 entreprises sélectionnées par FORTUNE possèdent un spécialiste de la gestion de la connaissance qui est responsable de la création d'une infrastructure et d'un environnement culturel pour le partage de la connaissance.

Traditionnellement, le savoir de l'entreprise résidait en son sein (documentation, base de données internes, compétences des employés). Ce savoir est de plus en plus réparti. Il réside de plus en plus en dehors des entreprises (banques d'information, documentation électronique). L'accroissement des débits des réseaux rend possible cet accès à l'information dans le monde entier (les spécialistes du domaine pensent que le terabit -mille milliards de bits par seconde seront bientôt atteints). Le volume de données double tous les 20 mois. L'évolution de la technologie informatique permet aujourd'hui et dans le futur de stocker ces données (Master Card On Line dispose d'un entrepôt de données de 1,2 teraoctets) et de les traiter avec efficacité (les mémoires centrales de micros sont composées de centaines de mégaoctets et la fréquence de certaines puces dépassent le gigahertz).

Ainsi, aujourd'hui dans les entreprises, l'objectif est de mémoriser les données créées par les transactions effectuées dans les entreprises dans un entrepôt de données. Il s'agit aussi de mettre sous forme accessible par la machine la documentation et la capitalisation de la connaissance dans l'entreprise. Ce dernier objectif peut être réalisé en créant des bases de données sur les "leçons apprises" et sur les "meilleures pratiques" d'un spécialiste. Les entreprises puisent aussi leurs connaissances à l'extérieur de l'entreprise en consultant des données structurées (banque de données) ou semi-structurées (documents textuels par exemple) souvent accessibles sur le Web.

Le traitement de données provenant de différentes sources suppose une interopérabilité sémantique notamment réalisée par la conception d'ontologies qui définissent le vocabulaire partagé et utilisé dans les systèmes de gestion de la connaissance. L'accroissement du volume de données pose le problème de leur utilisation efficace. Pour ce faire, les employés des entreprises doivent disposer d'outils qui cherchent la connaissance à partir des données (extraction de connaissances, algorithme de data mining).

L'extraction de connaissance dans les données (Data mining, Kowledge Discovery in data) a pour but de filtrer l'information pertinente pour une entreprise à partir des données qui sont à sa disposition.

La première source de données est l'entreprise elle-même. L'activité courante génère des données qui sont stockées dans la base de données. Ces données sont ensuite sélectionnées et réorganisées dans un entrepôt de données qui mémorise toutes les données utiles à la prise de décision.

La seconde source de données est constituée par l'ensemble des informations extérieures à l'entreprise et qui lui sont accessibles (banques de données consultées par l'entreprise, données disponibles sur la Toile).

Le volume de données pouvant être utilisé par une entreprise est considérable et son taux de croissance très important. Pour éviter d'être submergé par ce raz de marée de données, il est stratégique pour une entreprise de développer des méthodes et des outils qui sont capables de discerner dans ce magma de données, les connaissances pertinentes qui seront intégrées dans le système d'aide à la décision : c'est l'objectif de l'extraction de connaissances à partir des données (ECD).

D'autres connaissance "tacites" existent notamment chez les acteurs de l'entreprise et doivent être explicitées et représentées. C'est l'ensemble de ces connaissances qui constitue la base de connaissances qui doit être mémorisée et accessible par les différentes personnes de l'entreprise.

Parmi les algorithmes d'extraction des connaissances, la *méthode d'analyse statistique implicative* a pour but de définir des liens de causalité entre des variables, de les représenter et de mesurer la qualité de ce lien de causalité. Cette méthode d'analyse de données, au centre du thème qui nous réunit aujourd'hui et demain, se place donc comme un des outils performants pour participer à l'extraction de connaissances à partir des données, non seulement dans les domaines scientifiques fréquentés par les universitaires et dans le domaine de l'éducation, mais également dans celui de l'entreprise. Ces connaissances s'expriment en termes de règles d'association dont l'interprétation, puis la maîtrise ne peuvent que favoriser les prises optimales de décision.

C'est à partir des recherches menées par Régis Gras que des développements ont été effectués par des chercheurs de l'Ecole Polytechnique de l'Université de Nantes sur la mesure de qualité de la connaissance, sur l'amélioration d'algorithmes fondée sur l'intensité d'implication. L'acceptation de nombreux papiers dans des congrès internationaux montre la pertinence et l'opérationnalité des solutions offertes par la méthode implicative.

Henri BRIAND
Professeur à l'Ecole Polytechnique de l'Université de Nantes
Responsable de l'équipe C.I.D. (Connaissances – Informations - Données)
à l'Institut de Recherche Informatique de Nantes (I.R.I.N.)