



Revue des
Nouvelles
Technologies
de l' **I**nformation

Entreposage et Fouille de données

Rédacteurs invités : Omar BOUSSAID et Stéphane LALLICH

Année 2003 – Numéro 01

Revue des Nouvelles Technologies de l'Information
Sous la direction de Djamel A. Zighed et Gilles Venturini

Année 2003 – Numéro 1

Entreposage et fouille des données

Rédacteurs invités : Omar Boussaid, Stéphane Lallich
(Laboratoire ERIC, Université Lumière Lyon2)

LE MOT DES DIRECTEURS DE LA COLLECTION RNTI

Chères Lectrices, Chers Lecteurs,

La Revue des Nouvelles Technologies de l'Information a été créée en 2003 et a pour but de fournir à notre communauté scientifique un outil de communication de très grande qualité et ouvert à tous. Les thématiques générales couvertes par RNTI concernent tous les domaines liés à :

- l'Extraction de connaissances à partir des Données (ECD)
- la Fouille de données (FD)
- la Gestion des connaissances (GC)

A moyen terme, ces thématiques pourront se généraliser à d'autres domaines connexes (Web, ...).

RNTI publie uniquement des numéros spéciaux qui peuvent être de deux types :

- des actes de conférences sélectives garantissant une haute qualité des articles,
- des numéros à thème faisant l'objet d'un appel à communication. Chaque numéro à thème est édité par un ou plusieurs rédacteurs en chef invités. Un comité de programme spécifique est formé à cette occasion.

Pour connaître les informations de dernière minute sur cette collection, consulter des informations ou faire paraître des annonces, connaître les appels en cours, nous vous invitons à visiter le site consacré à RNTI et à prendre contact avec nous :

<http://www.antsearch.univ-tours.fr/rnti>

En outre, nous étudierons toutes vos propositions d'organisation de numéros spéciaux. Selon notre expérience passée, nous nous efforçons toujours de satisfaire à la fois les auteurs et les lecteurs.

Nous espérons que ce premier numéro vous donnera à tous entière satisfaction.

Djamel A. Zighed et Gilles Venturini.

PRÉFACE

Le laboratoire ERIC, Université Lumière Lyon 2, spécialisé dans l'extraction des connaissances à partir des données (ECD), a eu la charge d'organiser les 35èmes Journées de Statistique de la SFDS à Lyon du 2 au 6 juin 2003. Dans ce cadre, il a pris l'initiative d'organiser une session spéciale intitulée *ii* Entreposage et Fouille de données *ii*. Cette session a été l'occasion pour les chercheurs et les praticiens d'échanger des idées et de faire le point des récents développements relatifs à l'entreposage et à la fouille des données.

Les grandes masses de données hétérogènes dont disposent aujourd'hui les entreprises, ainsi que les flots de données circulant sur le réseau, posent des problèmes originaux de gestion, de fouille et de valorisation. L'extraction de connaissances à partir des données (Knowledge Data Discovery : KDD) apparue dans la communauté de l'intelligence artificielle, a pour but l'identification de structures inconnues, valides, et potentiellement exploitables dans les bases de données (U.M. Fayyad). L'ECD propose un cadre général dans lequel sont regroupées les méthodes qui permettent de faire face aux problèmes d'organisation des données et de leur exploitation, plus particulièrement :

- L'entreposage (data warehousing) a pour objet d'organiser les très grands volumes de données, de les structurer et de les préparer à l'analyse. Plus particulièrement, l'analyse en ligne (On Line Analytical Processing : OLAP) permet de résumer l'information et de naviguer dans les données.
- La fouille des données (data mining) a pour but d'extraire des connaissances à partir des données par des méthodes de structuration (apprentissage non supervisé) ou par des méthodes explicatives (apprentissage supervisé), une fois les données acquises et préparées.

En choisissant de rapprocher les travaux dans ces domaines, nous avons voulu montrer une réelle complémentarité entre ces deux démarches. Ces méthodes se complètent et l'un des enjeux des années à venir est d'organiser leur synergie.

En outre, nous avons le plaisir d'accueillir dans ce numéro spécial différentes contributions sur l'application de la statistique à l'image. Nous avons reçu un total de 47 soumissions d'une excellente qualité globale. Compte tenu des contraintes d'exigence scientifique, nous n'avons retenu que les 24 meilleurs papiers. Nous remercions très vivement les membres du Comité de Lecture qui ont assuré la sélection des articles figurant dans ce numéro. Les articles ont été regroupés en fonction des thèmes qu'ils abordent : Entreposage des données, Fouille de données et Statistique pour l'image.

Omar Boussaid et Stéphane Lallich.

Comité de lecture du numéro :

M.A. Aufaure (INRIA)	A. Bar Hen (Université Aix-Marseille3)
F. Bentayeb (Université Lyon2)	P. Besse (UPS Toulouse)
F. Bouali (Université de Lille)	H. Briand (Université de Nantes)
J.-H. Chauchat (Université Lyon2)	J. Darmont (Université Lyon2)
X. Descombes (INRIA, Sophia-Antipolis)	G. Grégoire (IMAG, Grenoble)
M.S. Hacid (Université Lyon1)	A. Hillion (ENST Bretagne)
J.-M. Jolion (INSA Lyon)	Y. Kodratoff (LRI, Université Paris-Sud)
P. Kuntz (ENST de Bretagne)	M. Lamure (Université Lyon1)
D. Laurent (Université de Tours)	S. Miguet (Université Lyon2)
A. Morin (Université de Rennes)	F. Muhlenbach (Université Lyon2)
N. Nicoloyannis (Université Lyon2)	J.-M. Petit (Université de Clermont Ferrand)
X. Pennec (INRIA, Sophia-Antipolis)	P. Perez (Microsoft Research, Cambridge)
S. Rabaseda (Université Lyon2)	R. Rakotomalala (Université Lyon2)
G. Ritschard (Université de Genève)	D. Sarrut (Université Lyon2)
M. Schneider (Université de Clermont Ferrand)	M. Sebag (LRI, Université Paris-Sud)
M. Sebban (Université de Saint Etienne)	O. Teste (IRIT, Université de Toulouse)
L. Tougne (Université Lyon2)	C. Vangenot (EPFL Lausanne)
G. Venturini (Université de Tours)	L. Wehenkel (Université de Liège)

TABLE DES MATIÈRES

ENTREPOSAGE DE DONNÉES

Vers l'auto-administration des entrepôts de données, K. Aouiche, J. Darmont, L. Gruenwald	1
SMAIDoC : Un Système Multi-Agents pour l'Intégration des Données Complexes, F. Clerc, A. Duffoux, C. Rose, F. Bentayeb, O. Boussaid	13
Extension du modèle M3 aux évolutions temporelles dans les applications SOLAP, M. Miquel, A. Tchounikine	25
Utilisation de la théorie des sondages dans le cadre des OLAP, S. Goutier, V. Stéphan	37
Tirer profit des sources externes pour l'enrichissement des bases clients - Application du Data Mining prédictif aux bases EDF, C. Derquenne, S. Goutier, S. Lembo, V. Stéphan	49
Fusion numérique d'informations multi-sources et extraction de connaissances : Appli- cation à l'ingénierie du trafic, N. El Faouzi	61

FOUILLE DE DONNÉES

Algorithme AntTree : Classification non supervisée par des fourmis artificielles, H. Azzag, N. Monmarché, M. Slimane, C. Guinot, G. Venturini	75
Classification Non Supervisée pour Données Catégorielles, P.-E. Jouve, N. Nicoloyanis	87
Partition BIC optimale de l'espace des prédicteurs, G. Ritschard	99
Traitement de données volumineuses par ensembles d'arbres aléatoires, P. Geurts	111
Critères d'évaluation des mesures de qualité des règles d'association, P. Lenca, P. Meyer, P. Picouet, B. Vaillant, S. Lallich	123
JEN : un algorithme efficace de construction de générateurs pour l'identification des règles d'association, A. Le Floc'h, C. Fisette, R. Missaoui, P. Valtchev, R. Godin	135
Des règles d'association ordinales aux règles d'association "classiques", S. Guillaume	147
Evaluation de la Résistance au Bruit de quelques Mesures Quantitatives, J. Azé, S. Guillaume, P. Castagliola	159
Des textes aux associations entre les concepts qu'ils contiennent, Y. Kodratoff, J. Azé, M. Roche, O. Matte-Tailliez	171

Apport des techniques de Text Mining pour la définition de caractéristiques clefs d'une mammographie, J. Clech, D. A. Zighed, A. Brémond	183
L'ADN en tant que texte : style et syntaxe, S. Lespinats, P. Deschavanne, A. Giron, B. Fertil	193
L'induction de graphes dans l'étude du Complexe Mycobacterium tuberculosis, G. Valétudie, S. Ferdinand, N. Rastogi, C. Sola	203

STATISTIQUE POUR L'IMAGE

Ondelettes et télédétection C. Charles, J.-P. Rasson	215
Une méthode générique pour la classification automatique d'images à partir des pixels, R. Marée, P. Geurts, L. Wehenkel	227
Analyse statistique de similarité dans une collection d'images M. Sarifuddin, R. Missaoui, J. Vaillancourt, Y. Hamouda, M. Zaremba	239
Analyse d'images multispectrales par diagonalisation simultanée de matrices de covariances, P. Courcoux, M.-F. Devaux , M. Séménou	251
Identification de Données Partiellement Occultées en RdF Statistique, B. Romaniuk, M. Desvignes, M. Revenu, M.-J. Deshayes	261
Une approche unifiée pour discriminer les tumeurs noires de la peau, C. Serruys, A. Giron, R. Viard, R. Triller, B. Fertil	273

INDEX DES AUTEURS

Aouiche K. 1
Azé J. 159, 171
Azzag H. 75
Bentayeb F. 13
Boussaid O. 13
Brémond A. 183
Castagliola P. 159
Charles C. 215
Clech J. 183
Clerc F. 13
Courcoux P. 251
Darmont J. 1
Derquenne C. 49
Deschavanne P. 193
Deshayes M.-J. 261
Desvignes M. 261
Devaux M.-F. 251
Duffoux A. 13
El Faouzi N. 61
Ferdinand S. 203
Fertil B. 193, 273
Fisette C. 135
Geurts P. 111, 227
Giron A. 193, 273
Godin R. 135
Goutier S. 37, 49
Gruenwald L. 1
Guillaume S. 147, 159
Guinot C. 75
Hamouda Y. 239
Jouve P.-E. 87
Kodratoff Y. 171
Lallich S. 123
Le Floc'h A. 135
Lembo S. 49
Lenca P. 123
Lepinat S. 193
Marée R. 227
Matte-Tailliez O. 171
Meyer P. 123
Miquel M. 25
Missaoui R. 135, 239
Monmarché N. 75
Nicoloyanis N. 87

Picouet P. 123
Rasson J.-P. 215
Rastogi N. 203
Revenu M. 261
Ritschard G. 99
Roche M. 171
Romaniuk B. 261
Rose C. 13
Sarifuddin M. 239
Séménou M. 251
Serruys C. 273
Slimane M. 75
Sola C. 203
Stéphan V. 37, 49
Tchounikine A. 25
Triller R. 273
Vaillancourt J. 239
Vaillant B. 123
Valétudie G. 203
Valtchev P. 135
Venturini G. 75
Viard R. 273
Wehenkel L. 227
Zaremba M. 239
Zighed D.A. 183