

DATA MINING VII

DATA, TEXT AND WEB MINING AND THEIR BUSINESS APPLICATIONS

WIT*PRESS*

WIT Press publishes leading books in Science and Technology.

Visit our website for the current list of titles.

www.witpress.com

WIT*eLibrary*

Home of the Transactions of the Wessex Institute.

Papers presented at Data Mining 2006 are archived in the WIT eLibrary in volume 90 of WIT Transactions on Information and Communication Technologies (ISSN 1743-3517).

The WIT eLibrary provides the international scientific community with immediate and permanent access to individual papers presented at WIT conferences.
<http://library.witpress.com>

WIT Press publishes leading books in Science and Technology.
Visit our website for the current list of titles.

www.witpress.com

SEVENTH INTERNATIONAL CONFERENCE ON DATA MINING

DATA MINING VII

CONFERENCE CHAIRMEN

A. Zanasi

TEMIS Text Mining Solutions, Italy

C. A. Brebbia

Wessex Institute of Technology, UK

N. F. F. Ebecken

Federal University of Rio de Janeiro, Brazil

INTERNATIONAL SCIENTIFIC ADVISORY COMMITTEE

S. Ananyan

M. W. Berry

S. Bolasco

I. Caddy

C. Curotto

B. Drewes

A. G. Evsukoff

P. Giudici

M. Gottgroy

J. P. Lawler

B. Little

D. Malerba

P. J-S. Pan

A. Sidhu

S. Sirmakessis

D. Sitnikov

Organised by

Wessex Institute of Technology, UK

Czech Technical University, Czech Republic

Federal University of Rio de Janeiro, Brazil

Sponsored by

WIT Transactions on Information and Communication Technologies

WIT Transactions on Information and Communication Technologies

Transactions Editor

Carlos Brebbia
Wessex Institute of Technology
Ashurst Lodge, Ashurst
Southampton SO40 7AA, UK
Email: carlos@wessex.ac.uk

Editorial Board

P L Aguilar

University of Extremadura
Spain

J J Casares Long

Universidad de Santiago de Compostela
Spain

A Davies

University of Hertfordshire
UK

A De Montis

Universita di Cagliari
Italy

G K Egan

Monash University
Australia

D J Evans

Nottingham Trent University
UK

A Genco

University of Palermo
Italy

D Goulias

University of Maryland
USA

J Jaafar

UiTm
Malaysia

C-H Lai

University of Greenwich
UK

J Lourenco

Universidade do Minho
Portugal

N Milic-Frayling

Microsoft Research Ltd
UK

J Rao

Case Western Reserve University
USA

M P Bekakos

Democritus University of Thrace
Greece

M Costantino

Royal Bank of Scotland Financial Markets
UK

N J Dedios Mimbela

Universidad de Cordoba
Spain

G Deplano

Universita di Cagliari
Italy

K-H Elmer

Universitat Hannover
Germany

U Gabbert

Otto-von-Guericke Universitat Magdeburg
Germany

P Giudici

Universita di Pavia
Italy

A Gualtierotti

IDHEAP
Switzerland

T Kobayashi

University of Tokyo
Japan

G Loo

The University of Auckland
New Zealand

D Malerba

Universita degli Studi
Italy

G Nakhaeizadeh

DaimlerChrysler Research & Technology
Germany

J Roddick

Flinders University
Australia

F Rodrigues

Poly Institute of Porto
Portugal

R Turra

CINECA Interuniversity Computing Centre
Italy

H Westphal

University of Magdeburg
Germany

N Zhong

Maebashi Institute of Technology
Japan

W Schreiber

University of Alabama
USA

D Van den Poel

Ghent University
Belgium

J Yoon

Old Dominion University
USA

H G Zimmermann

Siemens AG
Germany

DATA MINING VII

DATA, TEXT AND WEB MINING
AND THEIR BUSINESS APPLICATIONS

Editors

A. Zanasi

TEMIS Text Mining Solutions, Italy

C. A. Brebbia

Wessex Institute of Technology, UK

N. F. F. Ebecken

Federal University of Rio de Janeiro, Brazil

WITPRESS Southampton, Boston



A. Zanasi

TEMIS Text Mining Solutions, Italy

C. A. Brebbia

Wessex Institute of Technology, UK

N. F. F. Ebecken

Federal University of Rio de Janeiro, Brazil

Published by

WIT Press

Ashurst Lodge, Ashurst, Southampton, SO40 7AA, UK

Tel: 44 (0) 238 029 3223; Fax: 44 (0) 238 029 2853

E-Mail: witpress@witpress.com

<http://www.witpress.com>

For USA, Canada and Mexico

WIT Press

25 Bridge Street, Billerica, MA 01821, USA

Tel: 978 667 5841; Fax: 978 667 7582

E-Mail: infousa@witpress.com

<http://www.witpress.com>

British Library Cataloguing-in-Publication Data

A Catalogue record for this book is available
from the British Library

ISBN: 1-84564-178-7

ISSN: 1746-4463 (print)

ISSN: 1743-3517 (on-line)

*The texts of the papers in this volume were set
individually by the authors or under their supervision.
Only minor corrections to the text may have been carried
out by the publisher.*

No responsibility is assumed by the Publisher, the Editors and Authors for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein.

© WIT Press 2006.

Printed in Great Britain by Cambridge Printing.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the Publisher.

Preface

This year the Data Mining Conference is held in the center of Europe: Prague, a symbolic door between West and East, North and South Europe. Prague is also a symbol of a new, unified Europe, a Europe, we hope, of peace, security and progress.

The Conference also, this year, remains faithful to the original idea of providing a platform to discuss theoretical and applicative aspects of data mining, with participants from all over the world, both from academia and from industry.

Its success is reflected in the many abstracts received, with participants coming from 17 countries, with a strong representation from USA, the Americas and Europe as well as many other continents, allowing a real multinational-multicultural exchange of experiences and ideas!

Last year we witnessed the explosion of interest in data mining applications to unstructured data and we forecasted¹ that, during the following year:

1. several applications dedicated to the analysis of unstructured content would have appeared;
2. the interest in unstructured data mining/text mining of researchers, OEM and system integrators working in sectors of information retrieval, semantic web, linguistics and knowledge management would have grown.

So, we were pleased during the last year to witness the appearance of IBM UIMA (Unstructured Information Management Architecture), launched by IBM with a group of sixteen worldwide partners, among them the big players of data/text mining: Clearforest, Inxight, SAS, SPSS and TEMIS, and the appearance of SAP TREX (Text retrieval and classification), containing text mining functions.

But probably our best forecasting was done again last year, when after having introduced the possibility of fighting terrorism using data and text mining capabilities, quoting the actions of CIA (which funded In-Q-Tel, a non-profit corporation designed to fuel private research, development and application in, amongst other things, data and text mining technologies²), and of the French government, which launched three initiatives, endowed with 250 million Euros, to support French start-ups working in this sector³. We wrote:

We expect that soon the European Union will decide to integrate these national interests in a common, European action.

Our expectation was written in March 2005 and printed in May. We were pleased that, in June, the first meeting of ESRAB-European Security Research Advisory Board was held, just having been nominated by the European Commission to advise

in starting the ESRP-European Security Research Program, already endowed with almost four billion Euros to fund the development of security research, with particular attention given to data and text mining!

For a discipline which is central to research and also to business, and which generates interest not only among software developers but also among large companies and government departments and agencies, it is important to look at the market and at its movements.

A current analysis⁴, of the growth rate of the text mining market, estimates it to be more than 22%, with a market share divided between organizations such as ClearForest, Inxight and Temis and the larger ones including IBM, SAP, SAS, SPSS and Google.

A Conference such as this can only succeed as a team effort, so we want to thank the International Scientific Advisory Committee for their excellent work in reviewing the papers as well as their invaluable input and advice.

The next Data Mining conference will take place in the New Forest, UK in 2007. For further information visit www.wessex.ac.uk, or contact rgreen@wessex.ac.uk

The Editors
Prague, July 2006

¹ Preface in *Data Mining 2005* – Eds. Zanasi, Brebbia, Ebecken – WIT Press

² John Gannon-*The Strategic Use of Open-Source Information* –in *Intelligence and the National Security Strategist: Enduring Issues and Challenges*-Sherman Kent Center for Intelligence Studies National War College National Defense University (2004)

³ *Guerre économique: la France passe à l'offensive* in “Le quotidien de l'Expansion” – 11-03-2005

⁴ A. Weissman, Apex Partner – An investor's Perspective on Information Retrieval – Information Intelligence Summit, Phoenix, AZ - April, 10-13, 2006

Contents

Section 1: Data preparation

Nonlinear dimensionality reduction of large datasets for data exploration <i>V. Tomenko & V. Popov</i>	3
Text preparation through extended tokenization <i>M. Hassler & G. Fliedl</i>	13

Section 2: Clustering technologies

A method for association rule quality evaluation based on information theory <i>D. Sinikov, E. Titova & O. Ryabov</i>	25
K-means algorithm and its application for clustering companies listed in Zhejiang province <i>Y. Qian</i>	35
Fuzzy geo-processing for characterization of social groups: an application to a Brazilian mid-size city <i>G. R. A. Gonzalez, A. G. Evsukoff, R. C. Pinto, A. P. B. Sobral & J. A. Silva</i>	45
Dynamic classification: economic welfare growth in the EU during 1995–2004 <i>I. Gertsbakh & I. Yatskiv</i>	53
Cluster analysis of 3D seismic data for oil and gas exploration <i>D. R. S. Moraes, R. P. Espindola, A. G. Evsukoff & N. F. F. Ebecken</i>	63
Clustering of time series using a similarity between segments and bands determined by patterns of technical analysis <i>R. Basagoiti & E. Juaristi</i>	71

The CLUSTER3 system for goal-oriented conceptual clustering: method and preliminary results <i>W. D. Seeman & R. S. Michalski</i>	81
---	----

Section 3: Categorisation methods

A neural-networks associative classification method for association rule mining <i>P. Sermswatsri & C. Srisa-an</i>	93
---	----

An efficient Bayesian network approach for discovering interesting patterns <i>R. Malhas & Z. Al Aghbari</i>	103
---	-----

Kernel Discriminant Analysis and information complexity: advanced models for micro-data mining and micro-marketing solutions <i>C. Liberati & F. Camillo</i>	115
---	-----

**Section 4: MS SQL Server data mining
(Special session by C. L. Curotto and N. F. F. Ebecken)**

Mining cross-predicting stochastic ARMA time series in SQL server 2005 <i>B. Thiesson & J. Lind</i>	125
--	-----

Stability analysis of time series forecasting with ART models <i>A. Bocharov, D. Chickering & D. Heckerman</i>	141
---	-----

Multi-relational data mining in Microsoft SQL Server 2005 <i>C. L. Curotto & N. F. F. Ebecken & H. Blockeel</i>	151
--	-----

Electrical thunderstorm nowcasting using lightning data mining <i>C. A. M. Vasconcellos, C. L. Curotto, C. Benetti, F. Sato & L. C. Pinheiro</i>	161
---	-----

Section 5: Text mining

Computational system for the textual processing of industrial patents <i>G. M. Caputo & N. F. F. Ebecken</i>	169
---	-----

Using text mining to understand the call center customers' claims <i>G. M. Caputo, V. M. Bastos & N. F. F. Ebecken</i>	177
---	-----

A neural-based text summarization system <i>S. P. Yong, A. I. Z. Abidin & Y. Y. Chen</i>	185
---	-----

Analysis and development of latent semantic indexing techniques for information retrieval <i>M. Bottello</i>	193
--	-----

Section 6: Web mining

High performance environment for knowledge discovering in Portuguese language texts in the Web <i>V. M. Bastos & N. F. F. Ebecken</i>	205
---	-----

On the relationship between click rate and relevance for search engines <i>K. Ali & C. C. Chang</i>	213
---	-----

Selecting clickstream data mining plans using a case-based reasoning application <i>C. Wanzeller & O. Belo</i>	223
--	-----

Web page recommendation using a stochastic process model <i>B. J. Park, W. Choi & S. H. Noh</i>	233
--	-----

A new algorithm to measure relevance among Web pages <i>M. S. Sadi, M. M. H. Rahman & S. Horiguchi</i>	243
---	-----

A Web Mining process for e-Knowledge services <i>M. Castellano, F. Fiorino, F. Arcieri, V. Summo & G. Bellone de Grecis</i>	253
--	-----

Section 7: Customer relationship management

A study of customer relationship management (CRM) on apparel European Web sites <i>J. Lawler, P. Vandepuette & D. Anderson</i>	267
--	-----

Maximum resolution dichotomy for customer relations management <i>J. K. Ho</i>	279
---	-----

Telco churn analysis classification using a wavelet and RBF approach <i>Á. M. Cister & N. F. F. Ebecken</i>	289
--	-----

Section 8: Applications in science and engineering

Protein Ontology Project: 2006 updates <i>A. S. Sidhu, T. S. Dillon, B. S. Sidhu & E. Chang</i>	301
Evidence-based medicine: data mining and pharmacoepidemiology research <i>B. B. Little, R. A. Weideman, K. C. Kelly & B. Cryer</i>	307
SEQUEST: mining frequent subsequences using DMA-Strips <i>H. Tan, T. S. Dillon, F. Hadzic & E. Chang</i>	315

Section 9: Applications in business, industry and government

Traceability in the food-sector: the state of the art in a North Eastern Italian region <i>A. Payaro & S. Busetto</i>	329
A data mining approach to support the development of long-term load forecasting <i>M. R. Maia, K. de Oliveira Gonçalves Veloso, M. T. Okamoto, A. dos Santos Rigueira, G. M. Tavares, Á. M. Cister, M. A. F. Zarur, F. T. de Souza, G. S. Terra, A. G. Evsukoff & N. F. F. Ebecken</i>	339
Corporate bankruptcy prediction using data mining techniques <i>M. F. Santos, P. Cortez, J. Pereira & H. Quintela</i>	349
A Text Mining based content gathering system as strategic support for SMEs <i>N. Baldini, F. Neri & M. Perrone</i>	359
Intelligent analysis tools for computer based assessments <i>T. Filatov & V. Popov</i>	369
Information visualization for the taking of decisions <i>S. Larreina, S. Hernando & D. Grisaleña</i>	379

Section 10: Information systems, strategies and methodologies

Challenges in developing a cost-effective data warehouse for a tertiary institution in a developing country <i>A. Nazir & T. McDonald</i>	389
---	-----

A proposal of information system architecture for public transport <i>C. García, G. Padrón, F. Alayón & J. Caraballo</i>	399
Local nulls in summarised mobile and distributed databases <i>D. Chan & J. F. Roddick</i>	407
A Semantic Web Portal to construction knowledge exchange <i>M. Argüello, A. El-Hasia & M. Lees</i>	417
Ontological support to knowledge management in a hydrogeological information system <i>M. T. Pazienza, M. Pennacchiotti & A. Stellato</i>	429
Enterprise Intelligence Platform in the airline industry <i>G. Dragosavac, A. Viljoen & C. Badenhorst</i>	441
Improvement of generation change on SSE algorithm <i>T. Maruyama & E. Kita</i>	451
Author index	459