
DATA MINING VI

DATA MINING, TEXT 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 VI are archived in the

WIT eLibrary in volume 35 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.

Visit the WIT eLibrary at www.witpress.com.

Management Information Science

EDITORIAL BOARD

P L Aguilar

University of Extremadura
SPAIN

O Ciftcioglu

Delft University of Technology
THE NETHERLANDS

P Coupet

TEMIS
FRANCE

N J Dedios Mimbela

Universidad de Cordoba
SPAIN

A De Montis

Universita di Cagliari
ITALY

G Deplano

Universita di Cagliari
ITALY

P Giudici

Universita di Pavia
ITALY

D Goulias

University of Maryland
USA

A Gualtierotti

IDHEAP
SWITZERLAND

T V Hromadka II

California State University
USA

J Jaafar

UiTm
MALAYSIA

G Loo

The University of Auckland
NEW ZEALAND

J Lourenco

Universidade do Minho
PORTUGAL

D Malerba

Universita degli Studi
ITALY

G Nakhaeizadeh

Daimler Chrysler Research &
Technology.
GERMANY

P J-S Pan

National Kaohsiung University of
Applied Science
TAIWAN

J Rao

Case Western Reserve University
USA

D Riano

Universitat Rovira i Virgili
SPAIN

J Roddick

Flinders University
AUSTRALIA

F Rodrigues

Poly Institute of Porto
PORTUGAL

F Rossi

DATAMAT - Ingegneria dei Sistemi
S.p.A.
ITALY

D Sitnikov

Kharkov Academy of Culture
UKRAINE

R Turra

CINECA Interuniversity Computing
Centre
ITALY

D Van den Poel

Ghent University
BELGIUM

J Yoon

Old Dominion University
USA

A Zanasi

TEMIS Text Mining Solutions SA
ITALY

N Zhong

Maebashi Institute of Technology
JAPAN

HG Zimmermann

Siemens AG
GERMANY

SIXTH INTERNATIONAL CONFERENCE ON DATA MINING
DATA MINING VI

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

P. Giudici

M. Gottgroy

A. Gualtierotti

G. Lachtermacher

S. Larreina

D. Malerba

R. Maspons Bosch

N.M. Milic-Frayling

P.J-S. Pan

S.O. Rezende

M.F.R. Rodrigues

S. Sirmakessis

D. Sitnikov

R. Turra

D.E.N. Van den Poel

F. Wang

H.G. Zimmermann

Organised by

Wessex Institute of Technology, UK

DATA MINING VI

DATA MINING, TEXT 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: 1-84564-017-9

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 2005.

Printed in Great Britain by Athenaeum Press Ltd. Gateshead.

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 sixth Conference in the series dealing with Data Mining, Text Mining and their Business Applications took place in Skiathos, Greece, following the success of the fifth meeting held in Malaga, Spain. The Conference has become a traditional meeting for the international data mining community and provides a unique forum for practitioners coming from academia and industry from all over the world. The 2005 Conference has been a success, with a substantial increase in the number of participants and speakers as well as the countries represented. The quality of the meeting is demonstrated by the fact that it is considered the most popular in the world outside of the United States, according to the survey on the subject organized by Kdnuggets (www.kdnuggets.com). Twenty-five countries from all the continents are represented in the papers published in the book, offering a real multinational and multicultural range of experiences and ideas.

As already mentioned in the proceedings of the 2004 Data Mining book (Preface in *Data Mining 2004* – Eds. Zanasi, Brebbia, Ebecken, WIT Press, ISBN 1-85312-729-9), the most recent development has been the explosion of interest in data mining applications to unstructured data. This is reflected in a large increase in the number of papers dedicated to text mining in this book. Consequently it is not difficult to forecast that in the next months:

1. many applications dedicated to analysis of content coming from the billions of available web pages, newsgroups, emails, chat lines and message boards will appear;
2. the interest in unstructured data mining and text mining will grow amongst researchers, OEM and system integrators working in sectors such as information retrieval, semantic web, linguistics, knowledge management.

Hence, the decision taken last year of adding text mining in the title of our conference (*Data Mining, Text Mining and its Business Applications*), has proved to be correct seeing that the largest section of this book is the one dedicated to *text mining*, comprising about 25% of the contributions published.

Regarding *Business Applications*, we predicted as the most promising areas those regarding *National Security, Competitive Intelligence and Customer Relationship Management*. They can all be incorporated into the area of *Intelligence Analysis*, of especial interest for textual data mining and text mining applications. This is a topic that also takes into consideration *Consumer and Strategic Intelligence* applications.

It is not by coincidence that at an important meeting sponsored by the Central Intelligence Agency and dedicated to Intelligence, it was underlined that: “Analytic communities are continually challenged by the need to analyze massive volumes, velocities and varieties of multilingual and multimedia data. This situation occurs in multiple disciplines including HUMINT, SIGINT, IMINT, MASINT, GEOINT, and OSINT. This takes place in multiple domains including, but not limited, to terrorism, politics, economics, chemical, nuclear, and biological weapons of mass destruction, information assurance, science and technology and industry analysis” supporting “the management of knowledge, information and data sources”! Given the popularity of Data Mining, it is not surprising that discussions about the need of fighting terrorism using data and text mining capabilities started a political debate, fearing the risk that this represents to data privacy and democracy itself.

The strategy, nevertheless, has been already designed and it will not change: CIA launched In-Q-Tel, a non profit corporation designed to promote private research, development and application in, among the other, data and text mining technologies; the French government (the more active among the European governments in the sector of the so called “intelligence economique”) appointed Alain Juillet, coming from DGSE, as economic intelligence responsible to the Prime Minister’s office, and then launched three initiatives to support start ups working in this sector, considered strategic to French national interests. We expect that the European Union will soon decide to integrate these national interests in a common, European action.

For a discipline which is central to research, but also to business and which interests not only software developers, but also large companies and government departments and agencies, it is important to define its market size, even if this is a difficult task. A recently published synthesis of the current size of the data and text mining market estimates it to be about \$ 6 billions in 2006 (A. Zanasi – Preface in *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, WIT Press, 2005, ISBN 1-85312-995-X), a large sum which assures the involvement in this area of all the most important system integration companies in the world.

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

The next Data Mining 2006 Conference will take place in Prague, Czech Republic on 11-13 July 2006.

The Editors
Skiathos, 2005

Contents

Part I: Methodological Approaches

Section 1: Data mining

Learning networks for tornado forecasting: a Bayesian perspective
T. B. Trafalis, B. Santosa & M. B. Richman 5

Outlier detection based on projection-based ordering
A. Shojaie & P.-N. Tan 15

On extending F-measure and G-mean metrics to multi-class problems
R. P. Espindola & N. F. F. Ebecken 25

Multivariate interdependent discretization in discovering the
best correlated attribute
S. Chao & Y. P. Li 35

Estimation and extension of the Stochastic Schemata Exploiter
T. Maruyama & E. Kita 45

Decision making on operational data: a remote approach to
distributed data monitoring
V. Benson 55

A multi-strategy approach for mining multimedia data repositories
H. L. Viktor & E. Paquet 63

Section 2: Text mining

A multi-criteria decision making approach in feature selection for enhancing
text categorization
S. Doan & S. Horiguchi 77

Multilingual text mining <i>F. Neri</i>	89
The protein ontology project: structured vocabularies for proteins <i>A. S. Sidhu, T. S. Dillon, B. S. Sidhu & E. Chang</i>	95
Text mining for stock movement predictions: a Malaysian perspective <i>Y.-C. Phung</i>	103
Medical communication quality in the Italian pharmaceutical industry: measurement and analysis by ‘NOOS’ <i>P. Mariani & G. Ventre</i>	113
A comparison of two algorithms for discovering repeated word sequences <i>R. Tesar, D. Fiala, F. Rousselot & K. Jezek</i>	121
A genetic algorithm for text mining <i>G. Desjardins, R. Godin & R. Proulx</i>	133
The process of sensemaking on the telework virtual community using text mining <i>L. Giuliano & G. La Rocca</i>	143
Knowledge discovery in large text databases using the MST algorithm <i>V. Romanov & E. Pantileeva</i>	153
Textual document pre-processing and feature extraction in OLEX <i>R. Curia, M. Ettorre, L. Gallucci, S. Iritano & P. Rullo</i>	163
Naive rule induction for text classification based on key-phrases <i>N. N. Karanikolas & C. Skourlas</i>	175
Renovation of terms adjustment and effective model combination impact on information retrieval performance <i>A. Kasam & H.-C. Kwon</i>	183
Linguistic summaries on small screens <i>E. D’Avanzo & T. Kuflik</i>	195
Rule discovery in Web-based educational systems using Grammar-Based Genetic Programming <i>C. Romero, S. Ventura, C. Hervás & P. González</i>	205

Part II: Techniques

Section 3: Neural networks and decision trees

Multi-relational data mining in Microsoft® SQL Server™ <i>C. L. Curotto & N. F. F. Ebecken</i>	219
Neural network models for the development and evaluation of new fuels <i>G. S. Terra, A. G. Evsukoff, N. F. F. Ebecken, R. A. B. Sá & R. M. C. F. Silva</i>	229
CC4.5: cost-sensitive decision tree pruning <i>J. Cai, J. Durkin & Q. Cai</i>	239
Cooling Growing Grid: an incremental self-organizing neural network for data exploration <i>V. Tomenko & V. Popov</i>	247
Pleiotropic microarray gene expression data: advanced tandem multivariate data mining <i>B. B. Little, E. Barner & A. T. Dobson</i>	257
A decision tree classifier for vehicle failure isolation <i>N. Charkaoui, B. Dubuisson, C. Ambroise & S. Millemann</i>	265

Section 4: Link analysis

A method for generating aggregated associations between discrete data features <i>E. Tiova, D. Sitnikov, O. Ryabov, B. D'Cruz & O. Romanenko</i>	277
X3-Miner: mining patterns from an XML database <i>H. Tan, T. S. Dillon, L. Feng, E. Chang & F. Hadzic</i>	287

Section 5: Clustering and categorisation

Mining association rules from qualitative and quantitative clustering <i>A. Salazar, J. Gosalbez & I. Bosch</i>	299
HyperClustering: from the digital divide to a GRID e-workspace <i>M. Vafopoulos, V. Aggelis & A. Platis</i>	311
DEA implementation and clustering analysis using the K-Means algorithm <i>C. A. A. Lemos, M. P. E. Lins & N. F. F. Ebecken</i>	321

A hybrid method to categorize HTML documents <i>M. Khordad, M. Shamsfard & F. Kazemeyni</i>	331
--	-----

Part III: Applications

Section 6: Consumer and strategic intelligence

Application of technology prospective to business sectorial studies <i>S. Larreina & S. Hernando</i>	345
---	-----

Discovering common interests and problems to improve working conditions at a large company <i>M. C. S. Lopes, M. Onoda, V. M. Bastos & N. F. F. Ebecken</i>	353
---	-----

The use of knowledge discovery techniques for behavioural scoring <i>N. Meeus, J. Huysmans, B. Baesens, J. Vanthienen & M. Vandebroek</i>	361
--	-----

Providing database encryption as a scalable enterprise infrastructure service <i>U. T. Mattsson</i>	371
--	-----

National Security and threat awareness <i>F. Ghioni</i>	381
--	-----

Measuring user satisfaction with intelligent agents: an exploratory study <i>F. Fourati</i>	389
--	-----

A clustering approach for knowledge discovery in database marketing <i>M. F. Santos, P. Cortez, H. Quintela & F. Pinto</i>	399
---	-----

Section 7: Applications in science, engineering and life sciences

Evaluation of clinical prediction rules using a convergence of knowledge-driven and data-driven methods: a semio-fuzzy approach <i>M. Kwiatkowska, N. T. Ayas & F. Ryan</i>	411
---	-----

Evolving neural networks to flow cytometric data interpretation <i>G. C. Pereira, A. Bonomo & N. F. F. Ebecken</i>	421
---	-----

Classification algorithms and analyzing the functionality of protein families <i>L. Gao & D. K. Y. Chiu</i>	431
--	-----

Mining GPS logs to augment location models <i>M. Saraee & S. Yamaner</i>	445
---	-----

An adaptive Bayesian classification for real-time image analysis in real-time particle monitoring for polymer film manufacturing <i>K. Torabi, S. Sayad & S. T. Balke</i>	455
--	-----

Section 8: Applications in business, industry and government

Mining effective design solutions based on a model-driven approach <i>T. Katsimpa, S. Sirmakessis, A. Tsakalidis & G. Tzimas</i>	463
---	-----

Application of fuzzy models and neural models in financial time series <i>M. A. F. Allemão, A. G. Evsukoff & N. F. F. Ebecken</i>	475
--	-----

E-commerce models for banks' profitability <i>V. Aggelis</i>	485
---	-----

Sarbanes-Oxley, Basel II, and data mining opportunities in compliance systems <i>G. Allen</i>	497
--	-----

Survival data mining in the telecommunications industries: usefulness and complications <i>Z. Mohammed & D. Kotze</i>	505
--	-----

Data mining methods in a metrics-deprived inventory transactions environment <i>E. A. Beardslee & T. B. Trafalis</i>	513
---	-----

Ecological mining – a case study on dam water quality <i>M. F. Santos, P. Cortez, H. Quintela, J. Neves, H. Vicente & J. Arteiro</i>	523
---	-----

Improving effectiveness of Web sites using incremental data mining over clickstreams <i>F. Cavalcanti & O. Belo</i>	533
--	-----

Data mining education for external auditors <i>A. M. Young</i>	543
---	-----

Author Index	549
---------------------------	-----

