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

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

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

WIT PRESS 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.

Printed in Great Britain by Athenaeum Press Ltd. Gateshead.

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 II: Techniques

Section 3: Neural networks and decision trees

Section 4: Link analysis

Section 5: Clustering and categorisation

Part III: Applications

Section 6: Consumer and strategic intelligence

Section 7: Applications in science, engineering and life sciences