

Information Extraction in Finance

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, the WIT electronic-library provides the international scientific community with immediate and permanent access to individual papers presented at WIT conferences. Visit the WIT eLibrary at

<http://library.witpress.com>

Advances in Management Information Series

Objectives of the Series

Information and Communications Technologies have experienced considerable advances in the last few years. The task of managing and analysing ever-increasing amounts of data requires the development of more efficient tools to keep pace with this growth.

This series presents advances in the theory and applications of Management Information. It covers an interdisciplinary field, bringing together techniques from applied mathematics, machine learning, natural language processing, data mining and data warehousing, as well as their applications to intelligence, knowledge management, marketing and social analysis. The majority of these applications are aimed at achieving a better understanding of the behaviour of people and organisations in order to enable decisions to be made in an informed manner. Each volume in the series covers a particular topic in detail.

The volumes cover the following fields:

Information
Information Retrieval
Intelligent Agents
Data Mining
Data Warehouse
Text Mining
Competitive Intelligence
Customer Relationship Management
Information Management
Knowledge Management

Series Editor

A. Zanasi
Security Research Advisor
ESRIF

Associate Editors

P.L. Aquilar

University of Extremadura
Spain

M. Costantino

Royal Bank of Scotland Financial
Markets
UK

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. Gultierotti

IDHEAP
Switzerland

J. Jaafar

UiTM
Malaysia

G. Loo

The University of Auckland
New Zealand

J. Lourenco

Universidade do Minho
Portugal

D. Malerba

Università degli Studi
UK

N. Milic-Frayling

Microsoft Research Ltd
UK

G. Nakhaeizadeh

DaimlerChrysler
Germany

P. Pan

National Kaohsiung University of
Applied Science
Taiwan

J. Rao

Case Western Reserve University
USA

D. Riaño

Universitat Rovira I Virgili
Spain

J. Roddick

Flinders University
Australia

F. Rodrigues

Poly Institute of Porto
Portugal

F. Rossi

DATAMAT
Germany

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

N. Zhong

Maebashi Institute of Technology
Japan

H.G. Zimmermann

Siemens AG
Germany

Information Extraction in Finance

Marco Costantino

Royal Bank of Scotland Financial Markets, UK

&

Paolo Coletti

Free University of Bolzano Bozen, Italy

WITPRESS Southampton, Boston



M. Costantino

Royal Bank of Scotland Financial Markets, UK

P. Coletti

Free University of Bolzano Bozen, Italy

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: 978-1-84564-146-7

Library of Congress Catalog Card Number: 2008924037

*The texts of the papers in this volume were set
individually by the authors or under their supervision.*

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. The Publisher does not necessarily endorse the ideas held, or views expressed by the Editors or Authors of the material contained in its publications.

© WIT Press 2008

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.

Contents

Preface	xi
Biographies	xiii
List of figures	xv
1 Financial information and investment decisions	1
2 Financial tools	11
2.1 Conventional quantitative tools	11
2.2 Artificial intelligence techniques	12
2.2.1 Semantic networks	14
2.2.2 Neural networks	16
2.2.3 Genetic algorithms	19
2.3 Qualitative tools	20
2.3.1 Expert systems	22
2.3.2 Natural language processing and information extraction	23
3 Traditional approaches on qualitative information	27
3.1 Reuters 3000 Xtra	27
3.2 Bloomberg	31
3.3 Other information systems	35
3.4 Weaknesses of traditional approaches	36
4 Natural language processing and information extraction	37
4.1 Information retrieval	37
4.1.1 Statistical and probabilistic approaches	38
4.1.2 Linguistic approaches	42

4.2	The TREC competitions	43
4.2.1	Tasks	44
4.2.2	Evaluation metrics	44
4.3	Information extraction	45
4.3.1	The scripts-frames systems	48
4.4	The MUC competitions	51
4.4.1	Evaluation of the MUC results	58
4.5	The MUC systems	60
4.5.1	New York University: Proteus	65
4.5.2	University of Sheffield: LaSIE	69
4.5.3	BBN technologies: PLUM	70
4.6	User-definable template interfaces	72
4.7	Conclusions	74

5 LOLITA and IE-expert systems 75

5.1	Introduction and scope	75
5.2	Architecture of the system	76
5.2.1	The semantic network SemNet	77
5.2.2	Syntactic analysis	79
5.2.3	Analysis of meaning	81
5.2.4	Inference	82
5.2.5	Generation	82
5.3	Information extraction	83
5.3.1	Types of slots	85
5.4	Templates available in the system	86
5.4.1	Concept-based templates	86
5.4.2	Summary templates	87
5.4.3	Hyper-templates	88
5.4.4	The information to be extracted	88
5.4.5	Financial templates	90
5.5	Implementation of the financial templates	94
5.5.1	Prototypes	95
5.5.2	Domain-specific knowledge	95
5.5.3	Unification	95
5.6	The takeover financial template	96
5.6.1	The takeover main-event	96
5.6.2	The <i>takeover</i> template slots	100
5.7	Performance	107
5.8	Integration with elementised news systems	109
5.9	The IE-expert system	110
5.9.1	Implementation	112
5.9.2	The expert system component	113

6	Conclusions	115
A	Other MUC systems	117
A.1	Hasten	117
A.2	Alembic	119
A.3	NLToolset	120
A.4	Oki	120
A.5	IE ²	121
A.6	New York University: MUC-6 system	122
A.7	SIFT	123
A.8	TASC	123
A.9	New York University: MENE	124
A.10	FACILE	125
B	Recent systems	129
B.1	University of Utah system	129
B.1.1	Architecture	130
B.1.2	Evaluation and results	131
B.2	University of Sheffield pattern extraction	131
B.3	GATE framework	133
B.3.1	ANNIE	134
B.3.2	Architecture	135
B.3.3	Processing resource modules	138
B.4	Ontotext Lab: KIM	139
B.4.1	Ontology and knowledge base	139
B.4.2	Architecture	140
B.4.3	Evaluation and results	142
B.5	LoLo	143
B.5.1	Architecture	143
B.5.2	Evaluation and results	144
B.6	University of Wisconsin's systems	144
B.7	Elie	146
C	Other systems	149
C.1	Assentor	149
C.1.1	Introduction and scope	149
C.1.2	Architecture and performance	150
C.2	NewsInEssence	151
C.2.1	Introduction and scope	151
C.2.2	Architecture of the system	151
C.2.3	Implementation	153

C.3	Newsblaster	154
C.3.1	Introduction and scope	154
C.3.2	Centrifuser	155
C.3.3	System description	157
C.4	TermExtractor	158
Bibliography		159
Index		169

Preface

This book analyzes the state of the art of applied research in a challenging field: natural language understanding of financial news. Currently, thanks to the worldwide technological spreading, stock market traders are overwhelmed with financial information, both numerical and textual that has to be analyzed quickly in order to react before market conditions change again. While there are several well-known numerical techniques for quantitative data, textual information is usually manually examined investing a lot of precious human time. This book shows how information extraction (IE) can be successfully applied to this task, at the same time speeding up the process and freeing the trader from this workload.

The book's main focus has therefore a double identity: finance, especially intraday trading with large amounts of news arriving at a too fast pace to be examined manually, and IE, especially real-time analysis of predetermined events. Both sectors bring new problems and innovative techniques that are overviewed through many examples.

We start with an historical introduction to the first IE systems built in the 80s for the TREC competitions and then to the most promising approaches of MUC competitions, both statistical and rule-based, some of which lead to the development of the most interesting techniques in use today. Then we present recent systems, with a particular focus on their mixing of statistical and rule-based strategies. Finally, we show in deep detail the LOLITA system, together with its application IE-expert, two good examples of how IE and an expert systems can be applied to financial news analysis. Moreover, we introduce systems for other tasks, from which new ideas can be borrowed into this sector.

Biographies

Marco Costantino (marco.costantino@advanced-finance.com) achieved BSc and BA in Economics and Business Administration at the University of Trento, Italy. He then obtained a PhD in Computer Science at the University of Durham, UK. He worked in technology, analytics, and trading at JPMorgan equity derivatives in London and New York. He is currently head of equity derivatives quantitative development technology at Royal Bank of Scotland Financial Markets.

Paolo Coletti (paolo.coletti@advanced-finance.com) achieved BS and MSc in Mathematics at the University of Trento, Italy. After his PhD in Applied Mathematics in the field of computational fluid dynamics, he worked as a researcher in the field of natural language understanding. He is currently professor of Computer Science and Information Processing at the School of Economics and Management at the University of Bolzano, Italy.