Studies in Computational Intelligence, Volume 406

Editor-in-Chief
Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Further volumes of this series can be found on our homepage: springer.com

Vol. 385. Alex Graves
Supervised Sequence Labelling with Recurrent Neural Networks, 2012
ISBN 978-3-642-24796-5

Vol. 386. Marek R. Ogiela and Lakhmi C. Jain (Eds.)
Computational Intelligence Paradigms in Advanced Pattern Classification, 2012
ISBN 978-3-642-24048-5

Vol. 387. David Alejandro Pelta, Natalio Krasnogor, Dan Dumitrescu, Camelia Chira, and Rodica Lung (Eds.)
Nature Inspired Cooperative Strategies for Optimization (NICSO 2011), 2011
ISBN 978-3-642-24093-5

Vol. 388. Tiansi Dong
Recognizing Variable Environments, 2012
ISBN 978-3-642-24057-7

Vol. 389. Patricia Melin
Modular Neural Networks and Type-2 Fuzzy Systems for Pattern Recognition, 2012
ISBN 978-3-642-24138-3

Vol. 390. Robert Bembenik, Lukasz Skonieczny, Henryk Rybiński, and Marek Niezgódka (Eds.)
Intelligent Tools for Building a Scientific Information Platform, 2012
ISBN 978-3-642-24808-5

Vol. 391. Herwig Unger, Kyandoghere Kyamaky, and Janusz Kacprzyk (Eds.)
Autonomous Systems: Developments and Trends, 2012
ISBN 978-3-642-24805-4

Vol. 392. Narendra Chauhan, Machavaram Kartikeyan, and Ankush Mittal
Soft Computing Methods for Microwave and Millimeter-Wave Design Problems, 2012
ISBN 978-3-642-25562-5

Vol. 393. Hung T. Nguyen, Vladik Kreinovich, Berlin Wu, and Gang Xiang
Computing Statistics under Interval and Fuzzy Uncertainty, 2012
ISBN 978-3-642-24904-4

Vol. 394. David A. Elizondo, Agusti Solanas, and Antoni Martínez-Ballesté (Eds.)
Computational Intelligence for Privacy and Security, 2012
ISBN 978-3-642-25236-5

Vol. 395. Srikanta Patnaik and Yeon-Mo Yang (Eds.)
Soft Computing Techniques in Vision Science, 2012
ISBN 978-3-642-25506-9

Vol. 396. Marielba Zacarias and José Valente de Oliveira (Eds.)
Human-Computer Interaction: The Agency Perspective, 2012
ISBN 978-3-642-25690-5

Vol. 397. Elena Nikolaevskaya, Alexandr Khimich, and Tamara Chistyakova
Programming with Multiple Precision, 2012
ISBN 978-3-642-25672-1

Vol. 398. Fabrice Guillet, Gilbert Ritschard, and Djamel Abdelkader Zighed (Eds.)
Advances in Knowledge Discovery and Management, 2012
ISBN 978-3-642-25837-4

Vol. 399. Kurosh Madani, António Dourado Correia, Agostinho Rosa, and Joaquim Filipe (Eds.)
Computational Intelligence, 2012
ISBN 978-3-642-27533-3

Vol. 400. Akira Hirose
Complex-Valued Neural Networks, 2012
ISBN 978-3-642-27631-6

Vol. 401. Piotr Lipiński and Konrad Świrski(Eds.)
Towards Modern Collaborative Knowledge, 2012
ISBN 978-3-642-27445-9

Vol. 402. Theodor Borangiu, Andre Thomas, and Damien Trentesaux (Eds.)
Service Orientation in Holonic and Multi-Agent Manufacturing Control, 2012
ISBN 978-3-642-27448-0

Vol. 403. Kit Yan Chan, Tharam S. Dillon, and C.K. Kwong
Computational Intelligence Techniques for New Product Design, 2012
ISBN 978-3-642-27475-6

Vol. 404. Ahmad Taher Azar (Ed.)
Modelling and Control of Dialysis Systems, 2012
ISBN 978-3-642-27457-2

Vol. 405. Ahmad Azar (Ed.)
Modeling and Control of Dialysis Systems, 2012
ISBN 978-3-642-27557-9

Vol. 406. Cai-Nicolas Ziegler
Mining for Strategic Competitive Intelligence, 2012
ISBN 978-3-642-27713-9

Cai-Nicolas Ziegler

Mining for Strategic Competitive Intelligence

Foundations and Applications

Springer

Author
PD Dr. Cai-Nicolas Ziegler
PAYBACK GmbH
Albert-Ludwigs-Universität Freiburg i.Br.
München
Germany

ISSN 1860-949X
ISBN 978-3-642-27713-9
e-ISSN 1860-9503
e-ISBN 978-3-642-27714-6
DOI 10.1007/978-3-642-27714-6
Springer Heidelberg New York Dordrecht London
Library of Congress Control Number: 2012930495

Foreword

In recent years, progress in the analysis of business data by methods of data mining or machine learning has continued to be fast-paced. New computing paradigms like the map-reduce framework for fault tolerant large-scale distributed computing and the availability of vast computing resources without high setup costs through cloud computing now enable the analysis of sizes of data and complexities of models that have been unthinkable before. “Big Data”, “In Memory Analysis” and so forth are keywords describing these developments.

Accompanying this trend, a plethora of business applications based on data and derived analytic insights have been developed. Some business authors have coined the concept of “analytic enterprises”, whose key competitive advantage are not their products, services or customer base, but their analytic understanding and decision support facilities. Many, maybe most of these applications, serve immediate business goals such as increase in sales and thus could be counted towards the broad field of analytic marketing.

In his book, Cai-Nicolas Ziegler widens the scope to a much broader set of applications that aim to provide intelligence for many other aspects of companies, i.e., reputation of brands, the coverage of companies, products, or news events, and technological synergies between different units of large companies. At the same time the focus of these applications is on a much longer time frame for expecting pay-offs for investments in these types of analysis. The term “competitive strategic intelligence” very well describes this area of research.

For this type of analysis, evidence often is buried in complex data such as Web pages that mix relevant parts of text with lots of non-relevant information. Methods from different areas such as natural language processing, sentiment analysis, social network analysis, and, of course, machine learning, have to be adapted, developed and integrated to achieve these goals. As someone having many years of experience on both sides, the academic side and the business side, Cai-Nicolas Ziegler is the perfect candidate to translate between both spheres: to capture and describe relevant business problems, to boil them down to the analytical core problem involved, to solve this problem, as well as to translate the results back into the realm of business understanding.

This book provides the reader with a compact introduction into an exciting and complex field of research, starting with an overview of the basic methods and technologies, providing a summary of recent developments in this field, up to a sample of genuine research contributions for different problems and applications. This book could become a condensation point for researchers to start their own research from and for practitioners to get ideas about how to solve their own problems; to both ends it will serve a very good purpose.

Hildesheim, Lars Schmidt-Thieme
November 2011

Preface

The textbook at hand aims to provide an introduction to the use of automated methods for gathering competitive strategic intelligence. Hereby, the text does not describe a singleton research discipline in its own right, such as machine learning or Web mining. It rather contemplates an application scenario, namely the gathering of knowledge that is of paramount importance to organizations, e.g., companies and corporations.

To this end, the book first summarizes the range of research disciplines that contribute to addressing the issue, extracting from each those grains that are of utmost relevance to the depicted application scope. Moreover, the book presents systems that put these techniques to practical use (e.g., reputation monitoring platforms) and takes an inductive approach to define the *gestalt* of "mining for competitive strategic intelligence" by selecting major use cases that are laid out and explained in detail. These pieces form the *first part* of the book.

Each of those use cases is backed by a number of research papers, some of which are contained in its largely original version in the *second part* of the monograph. So, the book's structure is organized into three layers, with increasing detail: The first layer describes the foundations, the second lays out the use cases in a summarizing fashion, and the third then exposes the inner nucleus of these cases.

Most parts of the content have been drawn from the "Habilitationsschrift" (thesis for the post-doctoral lecture qualification) of the author. Some chapters have been significantly expanded while some papers from the original thesis have been left out in order to only include the ones that are seminal to the given context. Hereby, the contributions presented encompass more than five years of research work at various institutions, both in academia and industry.

Why dedicating a monograph to the topic of "mining for strategic competitive intelligence", the reader may ask? The rationale can be found in the evolution of the World Wide Web:

With the emergence of the so-called "Web 2.0", the participation age has begun to flourish on the Internet: Everybody is able to overtly express his opinion in newsgroups and weblogs, which are best described as personal diaries open to the public. And these weblogs are being read by millions of others. Moreover, people contribute

side by side to the creation of massive knowledge structures like Wikipedia or ODP, the Open Directory Project. These represent collective efforts that would not have been conceivable by mere individuals or smaller groups. As such, the Web 2.0 has levelled the ground for the rise of new large-scale information sources and structures. And these sources of consumer-generated, mainly unstructured data (i.e., text) offer fresh new opportunities to be exploited:

Taking a closer look at them is particularly worthwhile from a *corporate* perspective, aiming to use its lush sprinkle in order to distill knowledge of strategic value for any given corporation or product. "Strategic knowledge" hereby refers to knowledge that can be utilized to obtain a better competitive position, e.g., by being able to more accurately target campaigns, or by receiving market feedback on new products more quickly so that product issues can be eradicated faster.

While these strategic insights were gathered before the advent of Web 2.0 already, they could not be gathered in an *automated* fashion. There was market research manually performed by people, there were surveys, opinion polls, and the like. Now it can be done at virtually no cost per unit. The information is out there – information generated by all of us – in a format digestible by machines. Here is the brave new world.

München, Cai-Nicolas Ziegler
November 2011

Acknowledgements

The core of research presented in this thesis has been conducted between 2005 and 2009, mainly during my post-Ph.D. time at the database group of the University of Freiburg, and (to the larger extent) during my tenure at Siemens Corporate Technology in Munich, the company's forefront research facility.

Before all, I wish to express my gratitude and thankfulness to Prof. Dr. Georg Lausen, whom I know since my days as Ph.D. student in Freiburg. He has supported me in my endeavor to write the thesis forming the foundation of this monograph – even beyond my university tenure – and has been much more to me than a mere supervisor. I value him also for his exceptional interpersonal skills and his trust in me. Moreover, I would like to thank my second supervisor, Prof. Dr. Georg Gottlob, whom I came to know through the Lixto project. Lixto is one of the best examples how university research can be turned into applicable practice.

I am also indebted to Hermann Friedrich, my former department head at Siemens AG Corporate Technology, Information & Communications, Knowledge Management (department names are long at Siemens): He gave me all the freedom of research I could think of and embraced my desire to keep on publishing in academic journals and conferences.

The coauthors of my research likewise deserve mentioning, namely Dr. Kai Simon (Averbis GmbH), Dr. Maximilian Viermetz, Dr. Stefan Jung, Michal Skubacz, Walter Kammergruber (all four at Siemens Corporate Technology), Christian Vögele (Telefónica O2), and Prof. Dr. Dietmar Seipel (University of Würzburg). Their input and discussions have been valuable and helped to shape my research's direction.

And, of course, there is my family, my parents Angelika & Klaus as well as my brother Chris, to whom I owe more than fits between the covers of this book. The list would not be complete without Miriam, in the same fashion as I would not be complete without her. She is my life.

To Miriam

Abbreviations

BLRT	Binomial Log-Likelihood Ratio Test
CGM	Consumer-generated Media
EM	Expectation Maximization
HMM	Hidden Markov Models
IDF	Inverse Document Frequency
IR	Information Retrieval
ME	Maximum Entropy
ML	Machine Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
ODP	Open Directory Project
OLAP	Online Analytical Processing
PCA	Principal Component Analysis
PMI	Point-wise Mutual Information
POS	Part-Of-Speech (tagging)
PSO	Particle Swarm Optimization
SNA	Social Network Analysis
SOM	Self-organizing Map
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TF	Term Frequency
VSM	Vector Space Model

Contents

1 Introduction 1
1.1 Motivation 1
1.2 Definition of Scope 2
1.3 Organization and Content 3
1.3.1 Chapter Outline 3
1.3.2 Publications 3

Part I Foundations and Use Cases

2 Research Foundations 7
2.1 Introduction 7
2.2 Information Retrieval and Search Technology 8
2.2.1 Vector Space Model Basics 8
2.2.1.1 Reducing Term Vector Dimensionality 10
2.2.1.2 Textual Search Based on VSM 12
2.2.2 On Search Engine Anatomy 12
2.2.2.1 Inverted Index Organization 13
2.2.2.2 Towards Semantic Search Engines 14
2.3 Data Mining 16
2.3.1 Supervised Learning 17
2.3.1.1 Decision Tree Induction 17
2.3.1.2 Neural Networks 18
2.3.1.3 Naïve-Bayes Classifiers 19
2.3.1.4 Support Vector Machines 20
2.3.2 Unsupervised Learning 20
2.3.2.1 *k*-Means Clustering 21
2.3.2.2 Expectation Maximization 22
2.3.2.3 Self-Organizing Maps 23
2.3.2.4 Density-Based Clustering 24
2.4 Natural Language Processing and Text Mining 26
2.4.1 Text-Based Classification and Clustering 27
2.4.2 Named Entity Recognition 27

2.4.3 Keyword Extraction 28
2.4.4 Identification of Word Collocations 29
2.5 Social Network Analysis 29
2.5.1 SNA in the Context of Competitive Strategic Intelligence 30
2.5.2 Network Centrality Measures 31
2.5.2.1 Degree Centrality 32
2.5.2.2 Closeness Centrality 32
2.5.2.3 Betweenness Centrality 33
2.5.2.4 Eigenvector Centrality 34
2.5.3 Statistical Properties of Social Networks 36
2.5.3.1 On Random Graphs 37
2.5.3.2 Small-World Models 37
2.5.3.3 Preferential Attachment Models 38
2.5.4 Conclusion 39
2.6 Sentiment Detection 40
2.6.1 Issues and Challenges 41
2.6.1.1 Impact of Context on Polarity 41
2.6.1.2 Topic Identification 42
2.6.1.3 Language, Grammar, and Idiomatic Expressions 42
2.6.2 Sentiment Detection Techniques 44
2.6.2.1 Machine Learning and Lexical Approaches 44
2.6.2.2 Enhancing Sentiment Detection through NLP Approaches 46
2.6.3 Going beyond Research 48
2.6.4 Conclusion and Outlook 48

3 Competitive Intelligence Capturing Systems 51
3.1 Introduction 51
3.2 Common Platform Architecture 54
3.2.1 Document Collection Layer 54
3.2.2 Analysis and Semantic Annotation Layer 54
3.2.3 Reporting and User Interface 55
3.3 Typical Analysis Tasks 55
3.4 Survey of Existing Systems 58

4 Research Topics and Applications 63
4.1 Introduction 63
4.2 Trend and Topic Analysis 63
4.2.1 Reputation Monitoring Platform 64
4.2.1.1 Analysis Services 65
4.2.1.2 Research Contributions 69
4.2.2 Exploratory Customer Feedback Analysis 69
4.2.2.1 Platform Components 70

4.2.2.2 Conclusion 72
4.2.3 Temporal Topic Evolution 72
4.2.4 Analysis of Social Tagging Behavior 73
4.3 Automated News Extraction 74
4.3.1 Particle Swarm Optimization for Learning 78
4.3.1.1 PSO Training Procedure 78
4.3.1.2 Evaluation and Application 80
4.3.2 Learning Based on Classification Models 80
4.3.2.1 Block Labelling 81
4.3.2.2 Classifier Training and Evaluation 82
4.3.2.3 Feature Impact Assessment 83
4.4 Discovery of Technology Synergies 84
4.4.1 Semantic Similarity between Named Entities 85
4.4.1.1 Use of Background Knowledge 85
4.4.1.2 Creating Semantic Fingerprints 86
4.4.1.3 Computing Fingerprint Similarity 87
4.4.1.4 Training and Evaluation 87
4.4.2 Web 2.0 Leverage in Finding Technology Synergies 88
4.4.2.1 Characteristics of Synergetic Technologies 89
4.4.2.2 Building Classifiers 90
4.4.2.3 Empirical Evaluation 91
4.5 Concluding Remarks 93

5 Conclusion 95
5.1 Summary 95
5.2 Outlook and Future Directions 96

References 99

Part II Selected Publications

6 Research Papers – Section 4.2 109
Towards Automated Reputation and Brand Monitoring on the Web 111
Mining and Exploring Customer Feedback Using Language Models and Treemaps 121

7 Research Papers – Section 4.3 135
Content Extraction from News Pages Using Particle Swarm Optimization 137
Distilling Informative Content from HTML News Pages Using Machine Learning Classifiers 151

8 Research Papers – Section 4.4 167
Automatic Computation of Semantic Proximity Using Taxonomic Knowledge 169
Leveraging Sources of Collective Wisdom on the Web for Discovering Technology Synergies 189