Advanced Information and Knowledge Processing

Peter Rausch
Alaa F. Sheta
Aladdin Ayesh *Editors*

Theory, Systems and Industrial Applications

Business Intelligence and Performance Management

Advanced Information and Knowledge Processing

For further volumes:
www.springer.com/series/4738

Peter Rausch • Alaa F. Sheta • Aladdin Ayesh
Editors

Business Intelligence and Performance Management

Theory, Systems and Industrial Applications

Springer

Editors

Peter Rausch
Department of Computer Science
Georg Simon Ohm University of Applied Sciences
Nuremberg, Germany

Aladdin Ayesh
Faculty of Technology
De Montfort University
Leicester, UK

Alaa F. Sheta
Department of Computer Science
Taif University
Taif, Saudi Arabia

ISSN 1610-3947   Advanced Information and Knowledge Processing
ISBN 978-1-4471-4865-4   ISBN 978-1-4471-4866-1 (eBook)
DOI 10.1007/978-1-4471-4866-1
Springer London Heidelberg New York Dordrecht

Library of Congress Control Number: 2013933013

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Editorial

Business Intelligence (BI) and Performance Management (PM) offer solutions to issues caused by the challenges of the 21st century, such as globalisation, volatile markets or technical progress. In all these challenges, handling growing volumes of data is a major issue that requires fast storage, reliable data access, intelligent retrieval of information and automated decision-making mechanisms, all provided at the highest level of service quality. BI and PM provide techniques to efficiently respond to the dynamic business. Selected aspects of both topics are discussed within this state-of-the-art volume. The contributing authors are leading academics and professionals representing various universities, research centres and companies worldwide. Their expertise covers multiple disciplines and industries that forms the pillars of BI and PM.

Figure 1 shows how this book is organised. The first part ("*Introduction*") contains a chapter written by Hans-Georg Kemper, Peter Rausch, and Henning Baars about general Business Intelligence (BI) and Performance Management (PM) concepts in which related terms and definitions are introduced. In the following four parts BI and PM are analysed from different points of views. Part II discusses aspects concerning the strategic level. Afterwards, applications contributing to business development are outlined. Subsequently, ideas about methodologies are given. The fifth part focuses on important aspects of the related technologies. Finally, ideas about further developments conclude this book.

Part II ("*BI/PM in Business Analytics, Strategy and Management*") begins with a chapter of Hans-Georg Kemper, Henning Baars, and Heiner Lasi. They present an integrated BI framework to close the gap between IT support for management and production. The next chapter, written by Peter Rausch and Michael Stumpf, gives insights into how to link the different management levels by means of BI and PM in the construction industry. Aspects of IT-based decision support on different management levels are discussed in the following chapter. Zafer-Korcan Görgülü and Stefan Pickl outline the integration of data mining and systems engineering as an integral part of the business strategy. Examples of clinical decision support and aviation management are outlined. To give the readers an idea how to introduce approaches supporting management instruments, Martin Kütz provides a guideline

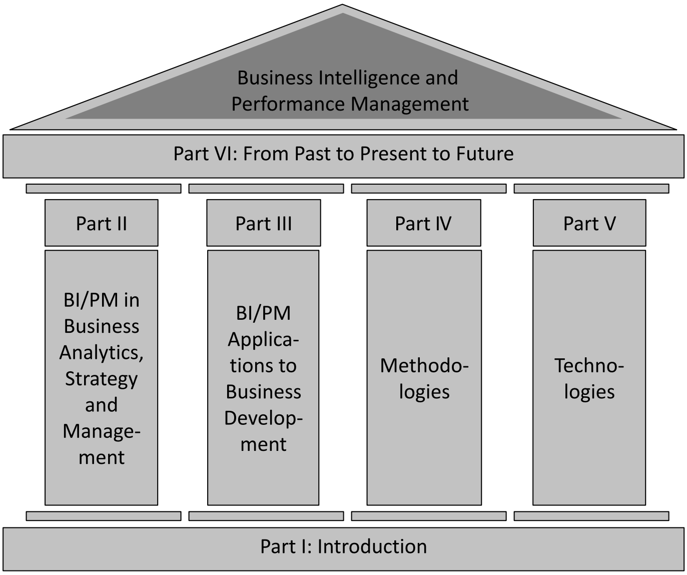

Business Intelligence and Performance Management
Part VI: From Past to Present to Future
Part II
BI/PM in Business Analytics, Strategy and Manage-ment
Part III
BI/PM Applica-tions to Business Develop-ment
Part IV
Methodo-logies
Part V
Techno-logies
Part I: Introduction

Fig. 1 How this book is organised

for the introduction of key performance indicators and scorecards. To illustrate his explanations, he relates his chapter to the example of IT Management.

Part III ("*BI/PM Applications to Business Development*") includes selected examples of BI and PM applications, which were taken from various domains. They are transferrable to different industries. In the chapter of Dieter Landes, Florian Otto, Sven Schumann, and Frank Schlottke data mining approaches to detect incidents in networks are presented. The insights are based on experiences, which were made in the insurance business. Security issues are also the focus of Rick Adderley's chapter. He explores the differences between the cross industry process for data mining and the national intelligence model using a self-organising map. The presented case study is based on experiences in the field of public services. Klaus Freyburger's chapter about business planning and support by means of IT-Systems gives a very useful overview of applications and solutions in the field of business planning. The following chapter, written by Hans Georg Zimmermann, Ralph Grothmann, and Hans-Jörg von Mettenheim, puts the focus on planning purchase decisions. The presented solutions are based on advanced neural networks.

Since the benefit of all applications strongly depends on the adequacy of methodologies in terms of the problem domain, Part IV ("*Methodologies*") includes selected approaches. As the readers could learn from Part III, time series processing can be very useful for planning problems. Thus, Daniela Pohl and Abdelhamid Bouchachia present a roadmap of online and offline methods which are applied in financial services and other industries. A major issue, of course, is coping with uncertainty. Hence, the following chapters are addressed to this subject. For instance, Peter Rausch and Birgit Jehle explore regression analyses to solve the issue of data supply for planning and budgeting processes under uncertainty. It also includes a fuzzy approach. The fuzzy set theory offers a rich approach that also can be used to solve other issues. Heinrich J. Rommelfanger presents a fuzzy approach to minimise the total cost in production and transportation planning. Alaa F. Sheta, Malik Braik, Ertan Öznergiz, Aladdin Ayesh, and Mehedi Masud combine fuzzy approaches with neural networks. Bringing these instruments together, they show how to improve steel making processes and how to model the dynamics of industrial processes. In addition to applying the methodologies, it is also very important to be able to measure efficiency of the deployed solutions. Martin Kütz presents useful approaches in his chapter and transfers them to IT organisations.

Part V ("*Technologies*") looks at BI and PM from another point of view. It puts the focus on technologies. Werner Schmidt's chapter explores business activity monitoring by means of a more or less loosely coupled combination of complex event processing functionality, process engines and dashboard applications. The idea is to introduce prerequisites for a continuous and simultaneous real-time monitoring. As another important issue, it is also necessary to process huge amounts of data. Therefore, Frederic Stahl, Mohamed Medhat Gaber, and Max Bramer analyse the aspect of scaling up data mining techniques to large datasets using parallel and distributed processing.

Part VI ("*From Past to Present to Future*") gives the readers an idea about further developments. Many trends are already included in the other parts. However, we have decided to highlight one important subject. The Editors are honoured to finish the book with William H. Inmon's chapter "*Evolution of Business Intelligence*". He spans the arch from the first simple reporting systems to today's evolved state of business intelligence and to the analysis of textual data.

The book describes selected theoretical aspects and presents practical solutions of the BI and PM area. The readers will get an excellent overview of how BI and PM are applied successfully to the challenges of the 21st century.

We hope that the readers will enjoy the book.

Nuremberg, Germany — Peter Rausch
Taif, Saudi Arabia — Alaa F. Sheta
Leicester, UK — Aladdin Ayesh

Acknowledgements

We are indebted to many people for their support and encouragement which was invaluable for the successful production of this book project. Firstly, we like to thank all authors for their excellent contributions and the wonderful cooperation during the review and publishing process. Also, we are deeply grateful to all referees for their valuable suggestions and their effort. Words can’t express how grateful we are to Michael Stumpf for contributing his great ideas and countless hours of his spare time to make this book a success. We appreciate his valuable feedback and the fruitful scientific discussions very much. Without Michael’s support we certainly wouldn’t have achieved this quality.

Last but not least thanks a lot to the Springer and the VTeX team, especially Helen Desmond, Ben Bishop and Edita Baronaitė for their great support.

Besides, Peter likes to thank his co-editors Alaa and Aladdin for their inspiring ideas. His sincerest thanks go to the lady of his heart, Birgit Jehle, for her encouragement, the feedbacks, her great ideas & suggestions, the scientific discussions, the valuable advices, the continuous support or just listening. It cannot be put into words how much she supported Peter and sacrificed her valuable time. Peter is also very grateful to Heinrich Rommelfanger and Patricia Brockmann who provided him with valuable feedback which helped so much to improve the quality of the chapters he was involved in. Additionally, Peter likes to express further greatest thanks for help and feedback to Carolin Theelke and to Caroline & Hans Benker. He dedicates this book to: Hannelore Wendt, Christian Amann, Thomas Hillerbrandt, Linda Kröber, Mr. Kuthning and an unknown lady from Italy. Without you this book wouldn’t exist!

Alaa would like to thank his co-editor Peter for his sincere and continue effort to make this book possible. He also wants to thank Aladdin for his great suggestion. Alaa would like to acknowledge his wife Rania for her support throughout his career and especially for the development of this book. She has been the source of motivation and the stimulus for me to improve my profession and move forward towards great achievements. Rania: I dedicate this book to you. I also thank my lovely children’s: Ahmed and Ayaa for their smart questions and wonderful smile. I hope one day they will read this book and know how much I love them. I cannot forget to thank my parents for their continuous praying. I want to dedicate this book to the sole of my best friend Ahmed Effat whom used to believe on me and tells that I can make great works.

Aladdin likes to thank his co-editors Peter and Alaa for their great efforts in realising this book, for their great friendship, and for their great display of professionalism. It has been a pleasure to work with them. He also likes to acknowledge his late parents for all they have done. They may have passed away but they are here in spirit through their great guidance, encouragement, and believing in him that made his career. In memory of Sabah and Saad.

Peter Rausch
Alaa F. Sheta
Aladdin Ayesh

Contents

Part I Introduction

1 Business Intelligence and Performance Management: Introduction . 3
Hans-Georg Kemper, Peter Rausch, and Henning Baars

Part II BI/PM in Business Analytics, Strategy and Management

2 An Integrated Business Intelligence Framework 13
Hans-Georg Kemper, Henning Baars, and Heiner Lasi

3 Linking the Operational, Tactical and Strategic Levels by Means of CPM: An Example in the Construction Industry 27
Peter Rausch and Michael Stumpf

4 Adaptive Business Intelligence: The Integration of Data Mining and Systems Engineering into an Advanced Decision Support as an Integral Part of the Business Strategy 43
Zafer-Korcan Görgülü and Stefan Pickl

5 How to Introduce KPIs and Scorecards in IT Management 59
Martin Kütz

Part III BI/PM Applications to Business Development

6 Identifying Suspicious Activities in Company Networks Through Data Mining and Visualization 75
Dieter Landes, Florian Otto, Sven Schumann, and Frank Schlottke

7 Exploring the Differences Between the Cross Industry Process for Data Mining and the National Intelligence Model Using a Self Organising Map Case study 91
Richard Adderley

8 Business Planning and Support by IT-Systems 107
Klaus Freyburger

9 Planning Purchase Decisions with Advanced Neural Networks 125
Hans Georg Zimmermann, Ralph Grothmann, and Hans-Jörg von Mettenheim

Part IV Methodologies

10 Financial Time Series Processing: A Roadmap of Online and Offline Methods 145
Daniela Pohl and Abdelhamid Bouchachia

11 Data Supply for Planning and Budgeting Processes under Uncertainty by Means of Regression Analyses 163
Peter Rausch and Birgit Jehle

12 Minimizing the Total Cost in Production and Transportation Planning—A Fuzzy Approach 179
Heinrich J. Rommelfanger

13 Design and Automation for Manufacturing Processes: An Intelligent Business Modeling Using Adaptive Neuro-Fuzzy Inference Systems 191
Alaa F. Sheta, Malik Braik, Ertan Öznergiz, Aladdin Ayesh, and Mehedi Masud

14 How to Measure Efficiency in IT Organizations 209
Martin Kütz

Part V Technologies

15 Business Activity Monitoring (BAM) 229
Werner Schmidt

16 Scaling up Data Mining Techniques to Large Datasets Using Parallel and Distributed Processing 243
Frederic Stahl, Mohamed Medhat Gaber, and Max Bramer

Part VI From Past to Present to Future

17 Evolution of Business Intelligence 263
W.H. Inmon

Contributors

Richard Adderley A E Solutions (BI) Ltd, Evesham, Worcestershire, UK

Aladdin Ayesh Faculty of Technology, De Montfort University, Leicester, UK

Henning Baars Chair of Information Systems I, University of Stuttgart, Stuttgart, Germany

Abdelhamid Bouchachia Bournemouth University, Bournemouth, UK

Malik Braik Electronic, Electrical and Computer Engineering Department, University of Birmingham, Birmingham, Edgbaston, UK

Max Bramer School of Computing, University of Portsmouth, Portsmouth, Hants, UK

Klaus Freyburger Hochschule Ludwigshafen am Rhein, Ludwigshafen, Germany

Mohamed Medhat Gaber School of Computing, University of Portsmouth, Portsmouth, Hants, UK

Zafer-Korcan Görgülü Institute for Theoretical Computer Science, Mathematics and Operations Research, Universität der Bundeswehr München, Neubiberg-München, Germany

Ralph Grothmann Siemens AG Munich, Corporate Technology, Munich, Germany

W.H. Inmon Inmon Consulting Services, Castle Rock, CO, USA

Birgit Jehle Noris Treuhand Unternehmensberatung GmbH, Nuremberg, Germany

Hans-Georg Kemper Chair of Information Systems I, University of Stuttgart, Stuttgart, Germany

Martin Kütz Fachbereich Informatik und Sprachen, Hochschule Anhalt, Köthen, Germany

Dieter Landes Coburg University of Applied Sciences and Arts, Coburg, Germany

Heiner Lasi Chair of Information Systems I, University of Stuttgart, Stuttgart, Germany

Mehedi Masud Department of Computer Science, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia

Florian Otto Coburg University of Applied Sciences and Arts, Coburg, Germany

Ertan Öznergiz Marine Engineering Operations Department, Faculty of Naval Architecture and Maritime, Yildiz Technical University, Istanbul, Turkey

Stefan Pickl Institute for Theoretical Computer Science, Mathematics and Operations Research, Universität der Bundeswehr München, Neubiberg-München, Germany

Daniela Pohl Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

Peter Rausch Department of Computer Science, Georg Simon Ohm University of Applied Sciences, Nuremberg, Germany

Heinrich J. Rommelfanger Faculty of Economics and Business Administration, Goethe University Frankfurt am Main, Schwalbach am Taunus, Germany

Frank Schlottke Applied Security, Stockstadt, Germany

Werner Schmidt University of Applied Sciences Ingolstadt, Ingolstadt, Germany

Sven Schumann HUK COBURG, Coburg, Germany

Alaa F. Sheta Department of Computer Science, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia

Frederic Stahl The School of Design, Engineering & Computing, Poole House, Bournemouth University, Poole, Dorset, UK

Michael Stumpf Department of Computer Science, Georg Simon Ohm University of Applied Sciences, Nuremberg, Germany

Hans-Jörg von Mettenheim Institut für Wirtschaftsinformatik, Leibniz Universität Hannover, Hannover, Germany

Hans Georg Zimmermann Siemens AG Munich, Corporate Technology, Munich, Germany

Part I
Introduction

Chapter 1
Business Intelligence and Performance Management: Introduction

Hans-Georg Kemper, Peter Rausch, and Henning Baars

Abstract Globalisation, volatile markets, legal changes and technical progress have an immense impact on business environments in most industries. More and more IT is deployed to manage the complexity. As a result, companies and organisations have to handle growing volumes of data which have become a valuable asset. The ability to benefit from this asset is increasingly essential for business success. Therefore, fast storage, reliable data access, intelligent information retrieval, and new decision-making mechanisms are required. Business Intelligence (BI) and Performance Management (PM) offer solutions to these challenges. Before important aspects of both topics are analysed from different points of view, this chapter gives an introduction to concepts and terms of BI and PM.

1.1 Current and Future Challenges

During the 21st century business environments have become more complex and dynamic than ever before. Companies operate in a world of change influenced by globalisation, volatile markets, legal changes, and technical progress. More and more IT is deployed to manage the complexity. As a result, growing volumes of data, for instance, provided by CRM systems, web shops or sensor technologies, have to be handled. Therefore, fast storage, reliable data access, intelligent information retrieval, and automated decision-making mechanisms, all provided at the highest level of service quality, are required. Successful enterprises are aware of these challenges and efficiently respond to the dynamic environment in which their business operates. Business Intelligence (BI) and Performance Management (PM) offer solutions to the challenges mentioned above and provide techniques to enable effective business change. The corresponding instruments allow transparency of processes

H.-G. Kemper · H. Baars
Chair of Information Systems I, University of Stuttgart, Keplerstr. 17, 70174 Stuttgart, Germany

P. Rausch (✉)
Department of Computer Science, Georg Simon Ohm University of Applied Sciences,
Kesslerplatz 12, 90489 Nuremberg, Germany
e-mail: perausch@prof-rausch.de

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_1, © Springer-Verlag London 2013

and their results on all management levels. Based on this information, action can be taken as fast as possible in the case of sudden market changes or critical developments. Meanwhile, companies in many industries, including technology suppliers, have realised this point and act as players in the fields of BI and PM. Thus, it is not surprising that the IT environment has changed dramatically over the last decades, both in terms of new business as well as in soft- and hardware requirements.

BI and PM are now well established and an important field for researchers as well as for professionals in all industries. Whether activities in this field are successful or not depends on certain prerequisites. It is important to address aspects from different points of view, which cover the following issues:

- BI/PM concepts to support business analytics, strategy and management,
- BI/PM applications to contribute to business development,
- methodologies
- and technologies.

Important aspects of the issues listed above will be covered by the following chapters of this book. Before we go into detail, it is important to have a common understanding what is meant by BI and PM. In the literature, the terms BI and PM are imprecise. A variety of definitions and interpretations for each term can be found. Therefore, in the next sections we will provide some basic definitions and outline the general concepts of BI and PM.[1] The chapter closes with a brief summary.

1.2 Business Intelligence (BI)

In 1996 the Gartner Group stated, "Data analysis, reporting, and query tools can help business users wade through a sea of data to synthesize valuable information from it—today these tools collectively fall into a category called 'Business Intelligence' " [1]. Consequently, leading companies in the field of management support environment adopted the term and subsumed all their tools for Data Warehouses (DWH), Data Marts (DM), Online-Analytical Processing (OLAP), data mining, etc. under the umbrella term Business Intelligence. Hence, in the early days the term BI was only used to describe the heterogeneous conglomerate of isolated tools, supporting various tasks of managers. It took years to establish a common understanding of BI in research and practice.

Meanwhile the various approaches merged into a common, rather inclusive understanding in the community that heavily focuses on aspects of integration and consistency. Based on this, BI is defined here as "an integrated, company-specific, IT-based total-approach for managerial decision support" [11].

Figure 1.1 shows a traditional BI architecture.

[1] The terms and definitions of this chapter are also valid for the other chapters. However, the editors didn't want the authors of other chapters to align to one single "truth". In case authors intend to use terms in a different way, the readers will find a reference, and the corresponding terms will be explained in context of the specific chapter.

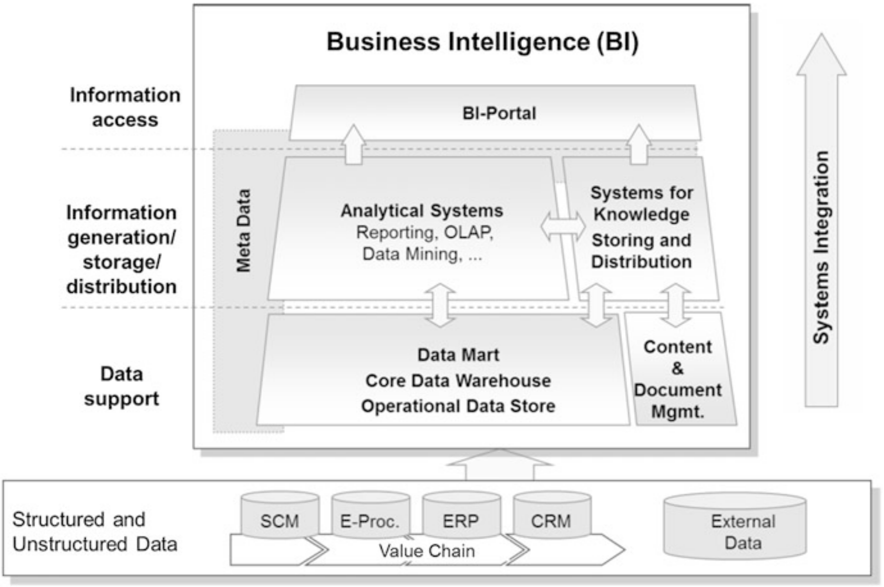


Fig. 1.1 Business Intelligence Architecture (adapted from [10])

From an application-oriented, logical perspective, a typical BI architecture consists of three layers [2]. These three layers are based on operational sources, like Supply Chain Management (SCM), E-Procurement Systems, Enterprise Resources Planning (ERP) Systems, Customer Relationship Management (CRM), and external sources. The systems of the Data Support Layer are fed by ETL processes (extraction-transformation-loading).

Data Support Layer The data support layer is responsible for storing transformed and harmonised, structured and unstructured data for decision support. Relevant data storing systems for unstructured data are document and content management systems. Structured data is stored in operational data stores (ODS), data warehouses (DWH) and data marts, whereas ODS are reservoirs for transactional data, which are often stored real-time without complex historisation routines. DWHs are data management systems for integrated, non-volatile, time-variant and subject-oriented data [9]. Bigger DWH "hub-and-spoke-architectures" have data marts, which are smaller data collections extracted from Core-DWHs, often based on multidimensional data models to support department-oriented ad-hoc reporting.

Information Generation, Storage, Distribution Layer This layer provides functionality to analyse structured data or unstructured content and supports the distribution of relevant knowledge. The analytical functionality of this layer includes OLAP and data mining, in addition to functionality to generate (interactive) business reports, ad-hoc analysis, or to implement performance management concepts, like the Balanced Scorecard or Value Driver Trees. For the distribution of knowledge, tools from Knowledge Management and CSCW domains are used, e.g. workflow support or tools for information retrieval.

Information Access Layer The information access layer offers the user convenient access to all relevant BI functions in an integrated environment—within the confines of defined user roles and user rights. Usually, the access layer is realised with some sort of portal software, which provides a harmonised graphical user interface.

In initial discussions on enterprise-wide BI approaches, practitioners and researchers propagated the development of a single enterprise-wide reservoir with harmonised data for decision support. It was argued that only in this way "a single point of truth" could be established, which is able to guarantee consistent support for all managerial decisions made in companies. In recent years, the debate became more controversial. It is now argued that BI approaches have to support heterogeneous decisions in strategic business units with often highly specific information needs. Besides, the field of BI is expanding and is being vitalised by new concepts and technologies in the area of data gathering and process support. Since the requirements of the primary and secondary business and production processes involved often differ fundamentally, it becomes clear that monolithic company solutions can not lead to satisfactory solutions. Modern Business Intelligence solutions therefore normally consist of a set of different interacting data storing systems, diverse ETL procedures, domain-specific data granularities, adequate analytic tools, and appropriate BI process models, in order to meet all of the challenges for effectively supporting processes [11].

1.3 Performance Management (PM)

Since a few years, the term Performance Management (PM), which partially overlaps with BI, attracts attention in science as well as in industry. According to an independent multivendor study of the Business Application Research Center (BARC) more than 80 % of all surveyed companies which were from different countries and industries have recognised the necessity to improve their PM processes. They claim a growing need for integrated technology platforms which support PM [3]. Subsequently, it is not surprising that the Gartner Group predicts a significant and growing demand for PM solutions [6].

Unfortunately, there is no clear definition of PM and its different variants. In the literature, a huge variety of PM definitions can be found, for example [4, 8, 12–14, 16]. Sharma states that PM is based on "the process of assessing progress towards achieving predetermined goals". It involves "the relevant communication and action on the progress achieved against these predetermined goals" [16]. In contrast to that, Lebas [14] calls PM a philosophy. It is obvious that these definitions differ in scope. Both remain a little bit imprecise. Geishecker's and

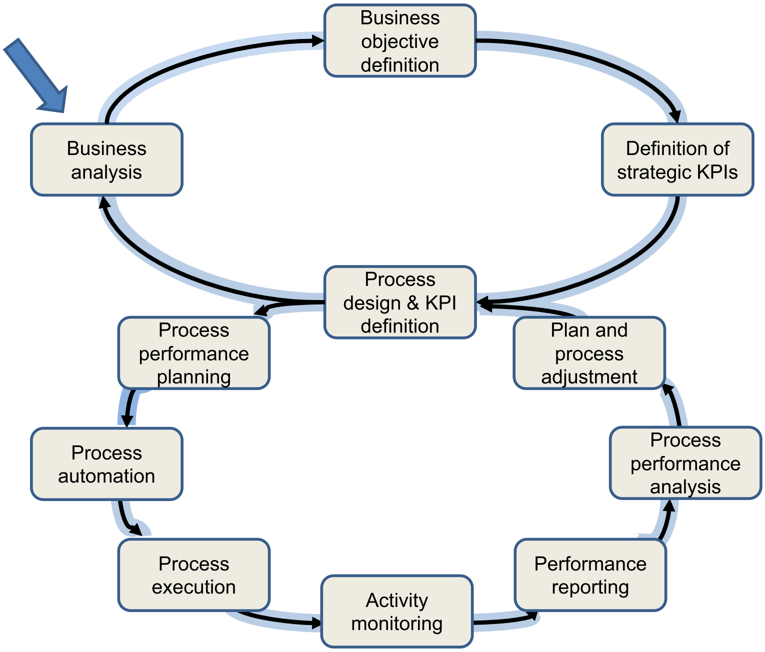

Business objective definition
Business analysis
Definition of strategic KPIs
Process design & KPI definition
Process performance planning
Plan and process adjustment
Process automation
Process performance analysis
Process execution
Activity monitoring
Performance reporting

Fig. 1.2 Closed-loop approach (adapted from [5, 15])

Rayner's interpretation is more precise. They define PM as methodologies, metrics, processes and systems which are used to monitor and manage business performance [7]. These aspects are all covered by this book. Therefore, we will use this definition in the following chapters. If the focus is set on PM in the context of enterprises, the terms Corporate Performance Management (CPM) or Enterprise Performance Management (EPM) can be used. They are subsets of PM. The term CPM is widespread in science as well as in industry. EPM is used by well-known software companies, such as Oracle or SAP. Because EPM and CPM exclude public institutions and non-profit organisations by definition, the term Business Performance Management (BPM) can be used in a more general context.

The basic idea of PM and its variants, as illustrated in Fig. 1.2, is a closed-loop approach. It helps to bridge the gap between the strategic and the operational levels by means of at least two linked loops. In contrast to "conventional" BI, which is a little bit more focused on technology, process-orientation is a significant attribute of PM.

While the operational level deals with aspects of monitoring, controlling and the optimisation of work processes, the strategic level defines business objectives and strategic Key Performance Indicators (KPIs). The starting point is the analysis of the business and the subsequent definition of business objectives. Based on the business objectives, strategic KPIs are derived. They influence the process design and the definition of process-oriented indicators [5]. It is important to align processes with strategic KPIs by defining operational KPIs. Operational KPIs periodically quantify the performance on the operational level. Compared to strategic KPIs their aggregation level is lower.

On the operational level, process performance has to be planned. In the case of automated processes, process execution can be monitored by business activity monitoring (BAM) tools. Of course, it is also possible to collect or add data manually. The data collected, which is processed by performance reporting tools, allows analyses of the process performance in terms of the objectives. The key data from the operational layer is monitored and analysed on a regular basis. The analyses make the actual performance transparent. Indicators, such as “average operational hours per day”, are compared with planned values in order to identify possible issues in process execution [15]. If the benefits overcome the effort, for example, if it is very important to recognise manufacturing problems as early as possible, real-time data monitoring can be desirable. BI components, as well as PM tools, can be used for further analyses. Abnormalities in the indicators denote issues which, for instance, can be caused by inefficient processes or exceptional market fluctuations. As a result, action can be taken, such as changing the process or revising the goals. In the ideal case, potential problems are avoided and identified before they arise. The positive or negative effects of those adjustments are measured in the next iteration, and a new cycle starts [17].

Of course, the results which are achieved on the operational level have an impact on the strategic level. Aggregated data is used to analyse business or rather the corresponding strategic KPIs, regularly. Deviations of current parameters from target values can indicate alignment problems on the operational level or an inadequate business strategy [15]. As a consequence, for instance, an adjustment of business objectives on the strategic level can be triggered. This can result in a complete redesign of business processes. The impacts on the operations are measured and analysed again by means of figures. Thus, the loop is closed and the strategic level is linked to the operational level.

As already mentioned in the last section, the information generation, storage, and distribution layers of BI architectures include functionality to implement PM concepts and its components, such as Balanced Scorecards. In this context, PM can be seen as an extension of BI. While BI applications are focused on the automated collection of data and the analyses by means of tools, such as data mining or OLAP, PM focuses on the process of systematic monitoring and on the control of business objectives on different management levels. The intention is to achieve sustainable success by means of continuous process improvements in terms of the company’s strategy.

1.4 Summary and Outlook

BI and PM offer a rich set of concepts and tools to efficiently master the challenges which are caused by the dynamic environment of companies and organisations. The successful application of BI and PM requires a common understanding of all of the parties involved. Due to a huge variety of interpretations, this chapter provided basic ideas of BI and PM concepts. First, the field of BI was outlined. The three layers of a typical BI architecture, which are based on operational sources, were explained. The application-oriented view includes

- the data support layer,
- the information generation, storage, and distribution layer,
- and the information access layer.

Afterwards, different interpretations of PM were analysed. It was shown that PM is based on the idea of the closed-loop approach. Closed-loops can be established on different management levels and should be linked to achieve full transparency of processes and their results. This enables companies and organisations to respond quickly to current developments. The process-oriented view of PM contributes to continuous improvements in terms of the strategic goals, and the holistic approach is an important requirement for sustainable success.

By means of integrated BI and PM components, the complete information chain, ranging from data supply to decision-making, can be supported and automated to a great extent. Thus, the growing amount of data can be processed efficiently. They are a valuable asset for companies and organisations. The ability to benefit from this asset is more and more essential for business success in competitive environments.

The following chapters of this book cover important aspects of this challenge, illustrated by many examples of industrial applications. BI and PM will be analysed from different points of view, and strategic concepts, business applications, methodologies, and technologies will be explored.

References

1. Anandarajan, M., Anandarajan, A., Srinivasan, C.A.: Business Intelligence Techniques. Springer, Berlin (2004)
2. Baars, H., Kemper, H.G.: Management support with structured and unstructured data—an integrated business intelligence framework. Inf. Syst. Manag. **25**(2), 132–148 (2008)
3. BARC: Performance Management—Aktuelle Herausforderungen und Perspektiven (2009). Available via http://www.barc.de/de/marktforschung/research-ergebnisse/performance-management.html. Accessed 7 Sep 2012
4. Becker, D., Brunner, J., Bühler, M., Hildebrandt, J., Zaich, R.: Value-Based Performance Management. Gabler, Wiesbaden (1999)
5. Dinter, B., Bucher, T.: Business performance management. In: Chamoni, P., Gluchowski, P. (eds.) Analytische Informationssysteme, 3rd edn., pp. 23–50. Springer, Berlin (2006)
6. Eddy, N.: BI, Performance management software market surpassed $12B in 2011. Available via http://www.eweek.com/c/a/IT-Management/Business-Intelligence-Performance-Management-Software-Market-Surpassed-12-Billion-in-2011/. Accessed 7 Sep 2012
7. Geishecker, L., Rayner, N.: Corporate performance management: BI collides with ERP. Research note SPA-14-9282, Gartner, Inc., December 17 (2001)
8. Hoffmann, O.: Performance management. Diss., Bern et al. (1999)
9. Inmon, W.H.: Building the Data Warehouse, 4th edn. Wiley, New York (2005)
10. Kemper, H.G., Baars, H.: Business Intelligence und Competitive Intelligence. HMD, Prax. Wirtsch.inform. **43**(247), 7–20 (2006)

11. Kemper, H.G., Baars, H., Mehanna, W.: Business Intelligence – Grundlagen und praktische Anwendungen, 3rd edn. Vieweg, Wiesbaden (2010)
12. Klingebiel, N.: Performance Measurement. Gabler, Wiesbaden (1999)
13. Krause, O.: Performance Management, Eine Stakeholder-Nutzen-orientierte und Geschäftsprozessbasierte Methode. Gabler, Wiesbaden (2005)
14. Lebas, M.: Performance measurement and performance management. Int. J. Prod. Econ. **9**(41), 23–36 (1995)
15. Melchert, F., Winter, R., Klesse, M.: The enabling role of information technology for business performance management. In: Proceedings of the 2004 IFIP International Conference on Decision Support Systems, Prato, pp. 535–546 (2004)
16. Sharma, S.K.: Human Resource Management: A Strategic Approach to Employment. Global India Publications, New Dehli (2009)
17. White, C.: Closed-loop business intelligence: reality or simply another buzzword? Available via http://www.b-eye-network.com/view/10275. Accessed 03/09/2011

Part II
BI/PM in Business Analytics, Strategy and Management

Chapter 2
An Integrated Business Intelligence Framework

Closing the Gap Between IT Support for Management and for Production

Hans-Georg Kemper, Henning Baars, and Heiner Lasi

Abstract Information Technology (IT) support in the manufacturing sector has reached a watershed with digital components beginning to permeate all products and processes. The classical divide between “technical” IT and “business” IT begins to blend more and more. Data from design, manufacturing, product use, service, and support is made available across the complete product lifecycle and supply chain. This goes hand in hand with the diffusion of sensor and identification technology and the availability of relevant information streams on the customer side—leading to unprecedented amounts of data. The challenge is to purposefully apply emerging BI concepts for a comprehensive decision support that integrates product and shop floor design phases, the steering and design of operational industrial processes, as well as big and unstructured data sources. This chapter brings those pieces together in order to derive an integrated framework for management and decision support in the manufacturing sector.

2.1 A New Role for Business Intelligence in the Manufacturing Sector

Globalization, scarcity of natural resources, complexity, and the powerful rise of the BRICS economies are the biggest challenges for the leading industrialized countries. For these nations, the major tasks for the next 20 years will be securing versatile production capabilities, resource efficient engineering environments, and a consequent time-to-market delivery of highly sophisticated industrial products [1].

In order to cope with these challenges, engineers are concentrating their research activities on complex concepts like the “Digital Factory” or “Intelligent Production Systems” as well as on introducing a variety of systems for steering and controlling their specific, production oriented operational processes. The main objective of these measures is to fully digitalize and integrate all processes of the product lifecycle and across supply chains [1]. In these contexts, large volumes of data are

H.-G. Kemper (✉) · H. Baars · H. Lasi
Chair of Information Systems I, University of Stuttgart, Stuttgart, Germany
e-mail: kemper@wi.uni-stuttgart.de

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_2, © Springer-Verlag London 2013

generated and stored within the IT infrastructures that support engineering, production, and logistics. The integration of this technical-oriented data with management support information, however, is still unsatisfactory. An integrated strategic, administrative, and operational control and a comprehensive managerial decision support still promises relevant untapped business potential. This article focuses on this topic. It extends and adapts the BI framework by [28] that has been introduced in Chap. 1 and derives an integrated framework for closing the gap between management- and production-oriented IT support.

2.2 Reshaping the BI Toolset

The more comprehensive the BI-based decision support becomes and the closer it is linked to the actual (and in the realm of manufacturing: *physical*) business processes, the more questions arise regarding requirements for an augmentation of classical BI-systems. Required are pertinent components and concepts for defining the interplay between the evolving BI landscape and existing operational application systems. Additionally, striving for a detailed understanding of processes leads not only to an ever increasing volume of data of both structured and unstructured nature but also to volatile use profiles and workloads.

In the following, existing concepts dealing with these developments are introduced. These are later contrasted with available systems for the support of the product lifecycle in the manufacturing sector.

2.2.1 Operational BI and BI and Business Process Management

The diffusion of BI into operational and tactical management layers has been discussed under the label "Operational BI" (OpBI) [17, 38]. The term OpBI is problematic because it does not clearly distinguish between the realm of BI and that of operational systems. In fact, some examples given by vendors appear to be rather manifestations of an insufficient operational support than of an innate need for new BI applications. If there already is a mature IT landscape in place—as in the manufacturing industry—the claim of a better operational decision support needs to be thoroughly substantiated [32]. This does not mean that OpBI is without merits. BI-technologies come into play when they can exert their strengths: Integrating large volumes of data from various sources, refining them for the purposes of decision support, and presenting the results in a comprehensive fashion.

This is also why OpBI is so closely related to the connection between business process management and BI—an area where the aspect of integration clearly comes into focus. There are various facets of this, which are covered in different, partly overlapping concepts [27].

A widespread example for viable OpBI is the area of Business Application Monitoring (BAM). In this case, data from various sources is combined in near-real-time to process-level key performance indicators (KPIs) and visualized via operational dashboards (e.g. on the status of open orders, delivery processes etc.). BAM applications are often embedded in broader concepts for Business Process and Business Performance Management, which aim at providing a consistent base of indicators across process steps and managerial levels [22, 42].

An approach that goes beyond the mere presentation of refined data is "Process-centric BI". Here, next to data, analytic *functionality* is embedded into operational systems in order to enable operational staff to conduct analysis on operational data [11]. The term "Embedded BI" goes even further. It denotes the application of BI functionality to process data from local repositories [30]. In this case, however, the specific contribution of a BI system is not obvious.

While the discussed OpBI concepts are directed towards an inclusion into running processes, *Business Process Intelligence* (BPI) has a more strategic momentum. BPI is concerned with the analysis of data on process instances for purposes of uncovering and optimizing the underlying process structures and models [16, 23, 38]. An example for a BPI application is process mining where operational log files are used for the extraction, enrichment, and evaluation of as-is-process-structures [44]. Another option for BPI is tailoring existing BI analysis tools (OLAP, reporting) [9, 14]. This, however, makes it necessary to extract, store, and handle data on the *process logic* rather than just on the *process results*, i.e. the order of activities and the related constrains need to be traceable. The concepts developed for this include the introduction of a respective "Process DWH" that is designed for such an analysis [47, 50]. Examples for relevant sources of process data in the realm of manufacturing are the Manufacturing Execution Systems (MES) or systems which allow an automatic tracing of objects, e.g. based on RFID technology (cf. Sect. 2.3).

2.2.2 Big Data, Cloud BI, and In-Memory BI

Collecting relevant data for the in-depth analysis of processes and activities on the operational layer leads to data repositories with sizes beyond those of ordinary DWHs. The relevant data can come in various forms—structured machine and sensor data, semi-structured reports form quality testing, feedback e-mails from customers, product evaluations on web pages, discussions in social networks, etc. Performance bottlenecks have always been an issue in BI that required an arsenal of strategies on multiple levels [10]. Nowadays, however, data volumes reach a level that classical relational technologies cannot efficiently handle anymore. This topic is currently summarized under the rather unspecific term *Big Data* [26, 37]. It can be dealt with in various ways, which can in parts also be applied in combination.

One approach, particularly suited for conglomerates of semi-, and unstructured data ("polystructured data"), is to apply database technologies that relax the strict scheme requirements of relational data bases as a trade-off for a better distribution of the data processing tasks and a higher query performance ("NoSQL"—not only

SQL). Examples include key-value stores, document stores, and extensible record stores [12, 45]. Contemporary NoSQL BigData repositories are particularly suited for parallelizing data aggregation and analysis task and for utilizing large clusters of computing infrastructure. While their eventual role in the domain of BI remains yet unclear, Big Data stores seem to be particularly interesting as *data sources* and as components for pre-processing the semi- or unstructured contents residing within those sources. Their applicability as full-scale replacements for a business-oriented DWH is limited however, as they are *by design* not meant to guarantee full consistency at all points of time ("BASE" model—Basically Available, Soft state, Eventual consistency).

A second strategy for dealing with large data sets that is intensively discussed is to apply "In-Memory data base" solutions. In-Memory solutions are tailored for handling larger volumes of data in the higher layers of the memory hierarchy, i.e. Random Access Memory (RAM), processor cache, and processor registers. Combined with pertinent data structures (e.g. a column-based instead of row-based storage of data base tables) this can lead to significant gains in query performance, e.g. in OLAP solutions [40, 41]. Implementations can particularly be found in specific DWH and/or OLAP appliances. The suitability for OpBI solutions is palpable—which leads some authors to the conclusion that in the future managerial and operational enterprise systems will rest upon a (re)unified data socket that is realized in an in-memory fashion [40]. While such a scenario is most probably only viable in a limited set of environments, the assumptions illustrate the increasing overlap between the operational and the managerial systems and the relation to questions of performance.

An alternative to a high-end in-house BI infrastructure is the import of services based on Cloud Computing approaches, i.e. internet-based services that can ideally be deployed in an ad-hoc manner, scaled dynamically with changing demand based on virtualization technologies, and be used in a pay-per-use model [39]. The application of Cloud Computing approaches to the domain of BI ("Cloud BI") can be an answer to issues of volatile workloads and of unpredictable requirements on the information generation and access layer [4, 46]. One source of such requirements is the unpredictable demand for BI on mobile devices ("Mobile BI")—where the rapid succession of innovation cycles quickly renders investments in specific components worthless (among others: mobile clients for various platforms, user and device management, security settings, etc.) [5]. The subject of mobile BI also gains relevance with the trend towards OpBI—an *in-process* decision often goes along with the need for an *on-site* decision, e.g. on the premises of the customer, in a distribution center, or at the shop floor.

2.3 Source Systems for BI in the Manufacturing Sector—Developments

The level of IT support in manufacturing is currently taken to a new level. This development can be broken down into three interdependent trends: First, activities across the product lifecycle are increasingly connected via digital networks. Second, identification and sensor technology is increasingly embedded into the physical environment and attached to objects ranging from transportation equipment, material, Work-In-Progress (WIP), up to machines, vehicles, and buildings [20]. Third, there is an increasing amount of semi- and unstructured data available for analysis (cf. Sect. 2.3). All this provides an increasing foundation of interrelated data that can be utilized for decision and management support. Injecting this data into BI systems is fruitful from two perspectives: First, integrating data on technical processes and business outcomes enables a more purposeful planning and steering at operational and tactical level (OpBI). Second, it allows for the provision of in-depth insights that can be used for strategic decisions.

The following sections detail the developments regarding the IT support within the product lifecycle, the relevance of sensor and identification technologies, and the role of semi- and unstructured data sources.

2.3.1 IT Systems Within the Product Lifecycle

Industrial companies are characterized by developing, designing, and manufacturing physical goods. While technological leaders create more and more complex products that are sold in bundles with non-material extensions like services and maintenance, the actual products are still of a *physical* nature and require intricate development and manufacturing processes in which expertise from several domains needs to be brought together [18, 35]. For example, technical goods are regularly composed of mechanical engineering, software- and electronic-based components, as well as fluid or electric power modules. Each of those domains comes with specific tasks and needs specific IT support. Three examples illustrate this: CAD systems, simulation, and production control. As for CAD, mechanical design, electronic design, as well as the fabric layout planning all apply Computer Aided Design (CAD) systems to build Digital Mock Ups (DMUs) [48]. The concrete functionality and data models of those CAD systems, however, strongly vary depending on the tasks they have to fulfill. This has the consequence that industrial businesses use several types (and brands) of CAD systems in parallel, often one per domain. Another example is simulation: The development and manufacturing of high-end products contains specific tasks like finite element simulation for strength calculation purposes, the simulation of product functionality or manufacturing planning. Each of those tasks is supported by its own specific IT-system [19]. In consequence, industrial businesses use a broad variety of heterogeneous IT systems. A third example is production control: Manufacturing is increasingly digitalized with numeric control systems, digital actors and sensors [3]. Specific steering and control tasks lead to specific IT systems. For example manufacturing execution system (MES) are increasingly used for the collection of machine and sensor data for right time control and steering tasks [31]. Those systems have to fit to the kind of manufacturing processes and tools. Therefore, even here industrial businesses apply separate MES in different manufacturing environments.

Taking into account that most industrial businesses act as global players, the IT infrastructure regularly gets more heterogeneous with different plants brining in their own IT systems depending on their size and functions.

In summary, the points mentioned above are leading to heterogeneity of the IT system landscape. And so far, business-oriented systems like ERP and CRM systems have not been considered: Even in medium-sized industrial businesses, it is common to find a large number of respective IT systems. This results in relevant product, process, and machine data being distributed across industrial businesses. It is indispensable for a holistic decision support purpose to collect and semantically integrate this data.

2.3.2 Identification and Sensor Technologies

Embedded, wirelessly interconnected, and mobile IT components that jointly provide new types of IT services have been discussed under the heading of "Ubiquitous Computing" (UC) for quite a while among scholars [36, 49]. It was the attention that has been given to the technology of "Radio Frequency Identification" that eventually propelled the diffusion of viable business applications of UC [20, 25]. RFID is applied in a variety of applications, yard management, theft prevention for tools, up to tracking product flows in production. A crucial development has been the diffusion of standards, esp. the "Electronic Product Code" (EPC) family of standards which not only covers codes and physical interfaces, but middleware platforms, and services for data exchange across enterprise borders [15, 24]. Originating in the retail sector, EPC is also increasingly applied in the manufacturing industry. Beyond its initial focus on identification, RFID can be augmented by sensor technologies, measuring environmental states such as temperature, humidity, acceleration, strain etc.

Direct effects of the application of RFID and sensor technology result from the automation of data capturing activities and encompass cost savings, faster availability of data, data quality improvements, and an avoidance of various mistakes and inefficiencies that result from erroneous manual data input. More interesting are the information effects that are an indirect consequence of the real-time data availability and the potentially higher resolution of automatic measurements of the presence, the identity, and/or the state of objects. This can even enable complete new ways of conducting processes or designing products (transformation effects) [7].

From the view of decision support, the new data enables not only a real-time process steering but also an *ex-post analysis* of process instances—particularly if objects are identified on item level [6, 13]. Among others, this allows process analysis on shop floor level, e.g. if WIPs or transportation material (cases, pallets, containers etc.) are tagged with pertinent RFID transponders and tracked with systems like MES. In case of EPC-based applications, the definition of a globally unique identifier like the EPC code even fosters data integration and analysis across enterprise borders. This is of particular interest in the realm of SCM, e.g. for pinpointing root causes of loss, faults and damages, for identifying and analyzing routing options, or for evaluating lot sizes or transport modalities.

Another relevant development towards a UC manufacturing environment with a BI impact results from the increasing degree of network-attached and IT-controlled "smart" machines and the trend to collect, distribute, and archive machine data in digital form.

UC data can enter the realm of BI either indirectly via operational systems (material management systems, warehouse management systems, SCM, ERP, MES, PPS, etc.). Or it can bypass this layer by being fed more or less directly into the DWH environments (after going through basic filtering and data transportation steps with specific "edge ware" and middleware). Either way, UC data can become a rich source for insights for both the steering and iterative adjustments of processes as well as for the design of new ones.

2.3.3 Unstructured Data

The discussed digitization of the shop floor leads to a large amount of structured data, e.g. sensor data [3]. However, numerous sources are not as structured and therefore not readily processable by BI applications. Examples include reports, emails or plain text documentations. Even the results from BI-based analyses are usually at some point translated into an unstructured form (e.g., a PDF file) for purposes of distribution or archival—the handling of these procedures is still considered unsatisfactory in many larger organizations [2]. Even more challenging are non-text representations of information, e.g. pictures from optical sensors or drawings, which are also needed for decision support, esp. within engineering tasks.

This leads to the requirement of coupling "classical" BI infrastructures for management support with systems that are specifically designed to handle, refine, and analyze semi- and unstructured data. In general, the semi- and unstructured data is either integrated into the information access layer (e.g. by the means of interlinked documents), integrated into the data support and information generation layer by processing (existing or extracted) meta data or distributed via components from the domain of knowledge management for knowledge storage and distribution [3].

2.4 Extending the Scope of Integrated Decision Support

In the following section, business scenarios are presented that highlight the potential of integrating the data sources discussed in section three—under consideration of the concepts and technologies introduced in section two. This leads to a BI with a much broader scope: First, product, process, and shop floor design phases are explicitly considered—necessitating dedicated product and process DWHs. Second, process steering and management become part of BI, which leads to components for OpBI and BPI. Third, large and unstructured data sources are considered in more analysis scenarios.

2.4.1 Including Product and Shop Floor Design Phases

Within the industrial product creation process and the subsequent phases of product usage and recycling there are several decisions with strategic implications. For this reason it is advisable to devote attention to the decision support within the product lifecycle. The following two exemplary management tasks will explain the typical decisions contained in the product lifecycle and the resulting information demand that needs to be covered by BI.

Management of Engineering Regulation and Standardization

The management of engineering regulation and standardization is usually part of the role of *Knowledge Engineers* (KE). KEs e.g. have to deal with identifying relevant engineering knowledge, acquiring that knowledge, and encoding it as input for knowledge or expert systems, construction rules, (construction) scripts, or templates. A primary task of these engineers is the acquisition and association of (fragmented) information in order to regulate construction with the objectives of coming to a holistic view on the relevant information [21, 43] and to implement a permanent active learning organization [29, 34]. These are prerequisites for the support of a "design for X" approach (e.g. design for assembly, design for logistics, or design to standards). As most business strategies require more than one "design for X" commitment (e.g. simultaneously demanding design for cost, design for quality, and design for assembly), these commitments are very often conflicting: Reaching a higher quality level (design for quality) requires a trade-off with the reduction of costs (design for cost). The KEs therefore have to figure out the impacts of changes across different commitments, if possible based on historical data. Examples of the data that needs to be collected and integrated for this includes actual geometric data and its history of changes (as stored in CAD systems), data on actual and historic assemblies, e.g. with respect to the reuse of parts (stored in PDM/PLM systems), non-financial KPIs (e.g. timeliness) from different manufacturing sites (mostly extracted from MES), plan, actual, and historical data about resources in production (from PPS-systems), or budged and actual financial key performance indicators from ERP systems.

Maturity Stage Level Management

Maturity Stage Level Management (MSLM) is important in the context of manufacturing engineering, quality management and lifecycle management. A core task of MSLM again is the acquisition and association of (fragmented) information—here aiming at reporting a certain maturity stage concerning key figures (e.g. warranty costs, loss claims) under consideration of different views (e.g. manufacturing site, region of use, and kind of defect), areas of responsibility (e.g. part managers, module managers, project managers), and hierarchical levels. The goal of MSL managers is to permanently enhance the maturity stage of products [33]. The information demand of MSL managers includes actual and historical data from different business units as well as heterogeneous external data sources (e.g. market data). Examples include geometric and feature data, a history of changes (from CAD systems), actual and historic assemblies (from PDM/PLM systems), quality data (from quality systems), documents and spreadsheets, plan, actual as well as historic financial data (from ERP systems), data concerning customer satisfaction, e.g. complaints (from CRM systems), as well as external data form retailers, service or repair shops about failures and repairs.

There are many tasks in the product lifecycle with similar characteristics like product and project management, management of product variants, or manufacturing engineering management: They all require information from the whole lifecycle and the plethora of applied systems contained within.

2.4.2 Including Process Steering and Management

Forerunners in applying both BI-based process management approaches in industrial environments can be found in the areas of logistics and SCM. This is not surprising, given the fact that the core concepts of those functions are characterized by an overarching process view. A process DWH, potentially filled automatically by UC technologies, can be used both for tasks of steering and of analyzing. A relevant complication for such scenarios comes from the cross-border nature of many of those scenarios and the need to quickly include and exclude partners, react to changing transportation and inventory strategies, consider modification of business models, and temporary demand for advanced analytic functionality (data mining, simulation, predictive analytics). From this point of view, logistics and SCM also illustrate the potential of Cloud BI [4].

Steering of Product Flows

Providing comprehensive information on product flows is a task that is heavily characterized by integrating and aggregating data from a variety of involved partners (manufacturers, second, third, and fourth party logistics providers, wholesalers, retailers) and their respective systems. Ideally, the status of a supply network (e.g. inventories, number, location, and status of moving goods and vehicles, service levels, throughput times etc.) can be accessed without manual or semi-manual data capturing and integration effort in adequate accuracy, correctness, and timeliness. This way it becomes possible to react to unexpected events (e.g. unavailable routes, losses and damages etc.) and to find solutions (e.g. alternative routes, re-directing oversupplies to retail outlets that face an Out-of-Stock situation etc.). Such scenarios are both relevant for the internal logistics of a single enterprise as for complete Supply Networks [6, 8, 13]. These are also examples for Operational BI.

Analyzing Process Structures

Interlinked with the (ad-hoc) steering is the step of uncovering patterns behind already observed events and of pinpointing root causes of reoccurring issues—a prime example for the application of BPI solutions (cf. Sect. 2.2.1).

Applications include the identification of problematic product configurations that lead to problems during transportation, identifying transportation routes that can be linked to quality impairments, or bottlenecks causing decreasing cycle times. This type of BPI application requires not only pertinent analysis tools but also data on both process logic (for tracing back problems) and on business results (for evaluating the problem impact). Here, a higher granularity of the data corresponds with the ability to adequately narrow down cause-effect-relationships. Again, scenarios can be found both in internal (production) logistics (where esp. an MES can act as a rich source of relevant and interconnected process data) and in broader SCM approaches (which, however, requires object traceability, e.g. based on RFID technologies) [8, 32].

2.4.3 Including Big and Unstructured Data

Decision support within various tasks in industrial businesses typically needs to consider both economical as well as technical aspects—with the latter often coming in extremely detailed and high-dimensional form (e.g. geometric data). Usually, the respective types of analyses also require the consideration of information that is only available in a semi-structured or unstructured form, e.g. service reports that sketch technical and geometric specifications in quality protocols or technical drawings. A comprehensive framework for BI in the manufacturing sector therefore needs to include both: an integrated presentation interface to connect structured and unstructured data as well as analytics of structured descriptions (meta data) to unstructured files. Given the sheer volume of the resulting data repositories and the need to also include Internet data for the reconciliation of decisions with customer and market trends, the potential of an inclusion of In-Memory and Big Data technologies (possibly Cloud-based) is salient.

Decision Support Based on Data from PDM and PLM Systems

Many PDM and PLM systems are based on meta data centric file systems. They handle drawings, DMUs, or reports as *files* that are managed by their meta data. Diving deeper into the processes of the digital firm, increasingly often simulation and virtual reality methods are used for testing and enhancing product features. Those methods typically generate large quantities of data—with large portions being unstructured or semi-structured. Typical analytics regarding PDM data, simulation data, and data from the customer side are explorations on the degree of customer satisfaction. Therefore, web analytics are increasingly used in social networks and use groups.

Optical Methods in Manufacturing and Quality Monitoring

The same applies to more and more widespread optical methods in manufacturing and quality monitoring. Process monitoring based on high resolution image processing leads to quickly increasing volumes of both structured and unstructured data, which both have to be integrated in decision support concepts to conduct lifecycle oriented root cause analysis e.g. for analyzing

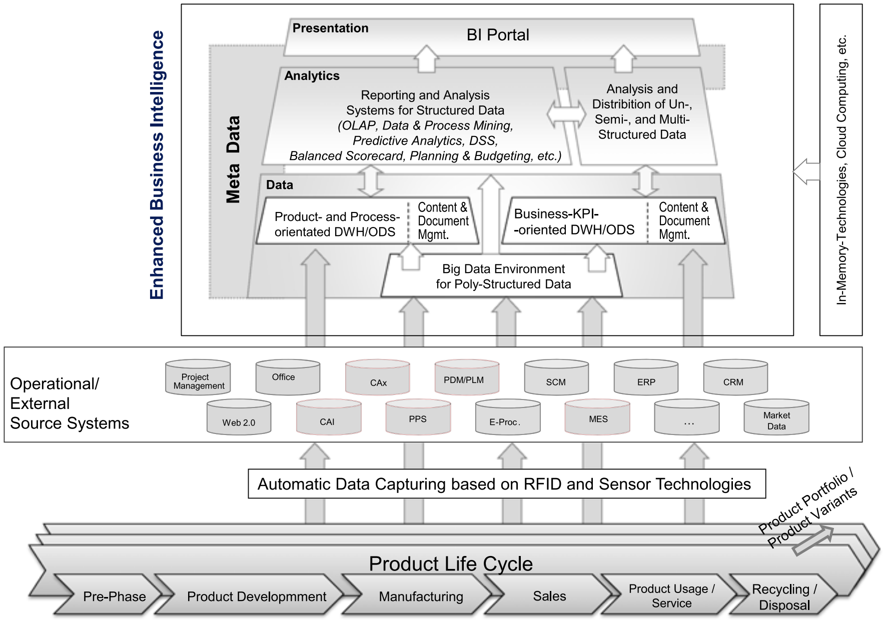

Enhanced Business Intelligence
Presentation
BI Portal
Meta Data
Analytics
Reporting and Analysis Systems for Structured Data
(OLAP, Data & Process Mining, Predictive Analytics, DSS, Balanced Scorecard, Planning & Budgeting, etc.)
Analysis and Distribution of Un-, Semi-, and Multi-Structured Data
Data
Product- and Process-orientated DWH/ODS
Content & Document Mgmt.
Business-KPI--oriented DWH/ODS
Content & Document Mgmt.
Big Data Environment for Poly-Structured Data
In-Memory-Technologies, Cloud Computing, etc.
Operational/ External Source Systems
Project Management
Office
CAx
PDM/PLM
SCM
ERP
CRM
Web 2.0
CAI
PPS
E-Proc.
MES
...
Market Data
Automatic Data Capturing based on RFID and Sensor Technologies
Product Portfolio / Product Variants
Product Life Cycle
Pre-Phase
Product Developmment
Manufacturing
Sales
Product Usage / Service
Recycling / Disposal

Fig. 2.1 Enhanced business intelligence architecture

failure driven warranty cost in maturity stage management or engineering management.

2.5 An Integrated Framework

Summarizing the previous insights, BI can unlock benefits in the manufacturing sector by bringing together technical and business-oriented information in a comprehensive decision support. This is on the one hand relevant at the operational and tactical level for the design and steering of processes. It is on the other hand of interest from a strategic perspective, as a holistic and in-depths view on these contents enables to uncover new decision options and a better understanding of the potentials and limitations of certain decisions.

This requires changes to traditional BI architectures. A resulting architecture framework is depicted in Fig. 2.1.

In more detail:

Data Support Layer

To support the decisions mentioned above, data has to be extracted both out of business and technical systems. For example, product feature data can be found in DMUs, while process logic information can e.g. be gained from MES, SCM, or directly gathered from UC systems. Within the enhanced framework, the data sources therefore encompass IT systems from the complete product lifecycle. Due to the need to store additional data formats, structures, and models with distinctive use and access profiles, the data support layer is extended by additional data warehouses, esp. for product oriented and technical data (product DWHs) and for data on process logic as required by BPI applications (process DWHs). Some of the relevant data directly streams in from sensors on the shop floor. This, as well as unstructured data from inside and outside the enterprise, leads to real time and Big Data requirements that complement the data support layer.

Information Generation, Storage, and Distribution Layer

The Information Generation, Storage, and distribution layer needs to include tools that are capable of analyzing the newly categories of data (polystructured data, process data)—leading to the need to build connections to NoSQL and Big Data components in the Data Support Layer. This goes along with the requirement to design pertinent data models.

Infrastructure Options

As they open up or prohibit application options, new infrastructural options (e.g. In-Memory or Cloud Computing) for coping with the aggravated performance requirements and the increasing volatility of the solutions, need to be included in the framework.

As this overview indicates, an industry-specific BI architecture that incorporates and adapts various new trends in BI and consequent might unfold competitive advantages. However, as many of the discussed concepts are still evolving and so far only implemented selectively and rudimentarily, the architecture framework can only function as a starting point that cannot replace a comprehensive company-specific evaluation. Furthermore, many open questions need to be addressed on the research side as well, e.g. questions on how to balance out trade-offs when choosing between Cloud-based Big Data services and in-house In-Memory solutions, or when comparing the use of established operational systems (like PLM or MES systems) and integrated DWH based solutions for various decision scenarios. Furthermore, a full-fledged integrated product DWH brings various challenges regarding the integrated data models, the data visualization, and the interplay with the process and “classical” business KPI DWHs.

However, tackling these issues might be highly valuable—particularly for enterprises in turbulent, complex, and global environments. Here, the capability to come to a thorough understanding of the business and to respond in-time to unexpected challenges can have consequences for the sustained survival of the enterprise.

References

1. ActionPlanT: ICT for manufacturing—the ActionPlanT roadmap for manufacturing 2.0 (2012). Available via http://www.actionplant-project.eu/public/documents/roadmap.pdf. Accessed 25th July 2012
2. Alter, A.: Business intelligence—are your BI systems making you smarter? CIO Insight **05/2003**, 77–85 (2003)

3. Baars, H., Kemper, H.G.: Management support with structured and unstructured data—an integrated business intelligence framework. Inf. Syst. Manag. **25**(2), 132–148 (2008)
4. Baars, H., Kemper, H.G.: Ubiquitous computing—an application domain for business intelligence in the cloud? In: Proceedings of the 17th Americas Conference on Information Systems (AMCIS), USA (2011)
5. Baars, H., Qie, L.: BI in the Cloud – Die Cloud als neuer Ansatz zur Erhöhung der BI-Agilität? BI Spektrum **7**(2), 26–29 (2012)
6. Baars, H., Sun, X.: Multidimensional analysis of RFID data in logistics. In: Proceedings of the 42th Hawaii International Conference on System Sciences (HICSS-42), USA (2009)
7. Baars, H., Gille, D., Strüker, J.: Evaluation of RFID applications for logistics: a framework for identifying, forecasting and assessing benefits. Eur. J. Inf. Syst. (EJIS) **18**(6), 578–591 (2009)
8. Baars, H., Kemper, H.G., Lasi, H., Siegel, M.: Combining RFID technology and business intelligence for supply chain optimization—scenarios for retail logistics. In: Proceedings of the 41th Hawaii International Conference on System Sciences (HICSS-41), USA (2008)
9. Bottani, E., Bertolini, M., Montanari, R., Volpi, A.: RFID-enabled business intelligence modules for supply chain optimization. Int. J. Technol.: Res. Appl. **1**(4), 253–278 (2009)
10. Brinkmann, A., Effert, S., Heidebuer, M., Vodisek, M., Baars, H.: An integrated architecture for business intelligence support from application down to storage. In: Proceedings of the 3rd International Workshop on Storage Network Architecture and Parallel I/Os, Saint Louis, USA (2005)
11. Bucher, T., Gericke, A.: Process-centric business intelligence. Bus. Process. Manag. J. **15**(3), 408–429 (2009)
12. Cattell, R.: Scalable SQL and NoSQL data stores. SIGMOD Rec. **39**(4), 12–27 (2010)
13. Cho, D.Y.: Ubiquitous data warehouse—integrating RFID with multidimensional online analysis. In: Proceedings of the San Diego International Systems Conference, San Diego (2005)
14. Chow, H.K.H., Choy, K.L., Lee, W.B., Chan, F.T.S.: Design of a knowledge-based logistics strategy system. Expert Syst. Appl. **29**, 272–290 (2005)
15. Curtin, J., Kauffman, R.J., Riggins, F.J.: Making the 'Most' out of RFID technology: a research agenda for the study of the adoption, usage and impact of RFID. Inf. Technol. Manag. **8**(2), 87–110 (2007)
16. Dayal, U., Hsu, M., Ladin, R.: Business process coordination: state of the art, trends, and open issues. In: Proceedings of the 27th International Conference on Very Large Data Bases (VLDB), Italy (2001)
17. Eckerson, W.E.: Best practices in operational BI—converging analytical and operational processes. In: TDWI best practice report, 3rd quarter 2007 (2007)
18. Ehrlenspiel, K.: Integrierte Produktentwicklung, 3rd edn. Hanser, München (2007)
19. Eigner, M., Stelzer, R.: Product Lifecycle Management, 2nd edn. Springer, Heidelberg (2009)
20. Fleisch, E.: What is the internet of things? An economic perspective. In: Auto-ID Labs white paper (WP-BIZAPP-053), Auto-ID Labs, St. Gallen (2010)
21. Giannakakis, T., Vosniakos, G.C.: Sheet metal cutting and piercing operations planning and tools configuration by an expert system. Int. J. Adv. Manuf. Technol. **36**(7–8), 658–670 (2008)
22. Golfarelli, M., Rizzi, S., Cella, I.: Beyond data warehousing: what's next in business intelligence? In: Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP, USA (2004)
23. Grigoria, D., Casatib, F., Castellanosb, M., Dayalb, U., Sayalb, M., Shan, S.C.: Business process intelligence. Comput. Ind. **53**, 321–343 (2004)
24. GS1: EPCglobal standards (2012). Available via http://www.gs1.org/gsmp/kc/epcglobal. Accessed 25 July 2012
25. Günther, O., Kletti, W., Kurbach, U.: RFID in Manufacturing. Springer, Heidelberg (2008)
26. Jacobs, A.: The pathologies of big data. Commun. ACM **52**(8), 36–44 (2009)
27. Kemper, H.G., Baars, H.: From data warehouses to transformation hubs—a conceptual architecture. In: Proceedings of the 17th European Conference on Information Systems (ECIS), Italy (2009)

28. Kemper, H.G., Baars, H., Mehanna, W.: Business Intelligence – Grundlagen und praktische Anwendungen, 3rd edn. Vieweg, Wiesbaden (2010)
29. Kendal, S.: An Introduction to Knowledge Engineering. Springer, London (2007)
30. Klawans, B.: Embedded or conventional BI—determining the right combination of BI for your business. Bus. Intell. J. **13**(1), 30–36 (2008)
31. Kletti, J.: Manufacturing Execution System: MES. Springer, Heidelberg (2007)
32. Koch, M., Baars, H., Lasi, H., Kemper, H.G.: Manufacturing execution systems and business intelligence for production environments. In: Proceedings of the 16th Americas Conference on Information Systems, Peru (2010)
33. Lasi, H.: Industrial intelligence—a BI-based approach to enhance manufacturing engineering in industrial companies. In: Proceedings of the 8th CIRP Conference on Intelligent Computation in Manufacturing Engineering (CIRP ICME), Italy (2012)
34. Lasi, H.: Decision support within knowledge-based engineering—a business intelligence-based concept. In: Proceedings of the 18th Americas Conference on Information Systems (AMCIS), USA (2012)
35. Lasi, H., Hollstein, P., Kemper, H.G.: Heterogeneous IT landscapes in innovation processes—an empirical analyses of integration approaches. In: Proceedings of the International Conference Information Systems (IADIS), Portugal (2010)
36. Lyytinnen, K., Yoo, Y.: Issues and challenges in ubiquitous computing. Commun. ACM **45**(12), 42–65 (2002)
37. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: the Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute (2011)
38. Marjanovic, O.: The next stage of operational business intelligence: creating new challenges for business process management. In: Proceedings of the 40th Annual Hawaii International Conference on System Sciences. IEEE Comput. Soc., New York (2007)
39. Mell, P., Grance, T.: The NIST definition of cloud computing. National Institute of Standards and Technology, Special Publication 800-145 (2011)
40. Platter, H.: A common database approach for OLTP and OLAP using an in-memory column database. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, Providence, USA (2009)
41. Plattner, H., Zeier, A.: In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg (2011)
42. Pourshahid, A., Chen, P., Amyot, D., Weiss, M., Forster, A.J.: Business process monitoring and alignment: an approach based on the user requirements notation and business intelligence tools. In: Proceedings of the 10th Workshop of Requirement Engineering, Canada, pp. 149–159 (2007)
43. Röhner, S., Breitsprecher, T., Wartzack, S.: Acquisition of design-relevant knowledge within the development of sheet-bulk metal forming. In: Proceedings of the International Conference on Engineering Design (ICED11), Denmark (2011)
44. Song, M., van der Aalst, W.M.P.: Towards comprehensive support for organizational mining. Decis. Support Syst. **46**(11), 300–317 (2008)
45. Strauch, C.: NoSQL databases (2011). Available via http://www.christof-strauch.de/nosqldbs.pdf. Accessed 25 July 2012
46. Thomson, W.J.J., van der Walt, J.S.: Business intelligence in the cloud. South African J. Inf. Manag. **12**(1), 1–5 (2010)
47. van der Aalst, W.M.P., Weijters, J.M.M.: Process mining: a research agenda. Comput. Ind. **53**(3), 231–244 (2004)
48. Weber, P.: Digital Mock-up im Maschinenbau. Shaker, Aachen (2003)
49. Weiser, M.: Ubiquitous computing (1996). Available via: http://sandbox.xerox.com/ubicomp. Accessed 25th July 2012
50. zur Mühlen, M.: Process-driven management information systems—combining data warehouses and workflow technology. In: Proceedings of the 4th International Conference on Electronic Commerce Research (ICECR-4), USA, pp. 550–566 (2001)

Chapter 3
Linking the Operational, Tactical and Strategic Levels by Means of CPM: An Example in the Construction Industry

Peter Rausch and Michael Stumpf

Abstract During the last decade, much progress has been made in the field of CPM. However, there are still some issues to master. Many companies have implemented only isolated CPM bricks, and controlling systems which supply real-time information are still missing. To exploit the whole potential of CPM, it is necessary to integrate the CPM components and to link the operational, tactical and strategic levels. In this chapter, an integration grid is used to address all interconnections between the different components systematically and to fill the gaps between the operational, tactical and strategic levels. The grid is applied to a new generation of CPM systems for the construction industry. The CPM approach presented is based on data of a satellite-supported, machine control & guidance system. This data is combined with data from other sources to enable intelligent analyses. Benefits and open issues are discussed. Finally, possibilities for further developments of the presented approach are mentioned.

3.1 Introduction

For many years, the construction industry has been in a process of change due to different influences, for example, globalisation, legal changes in many countries and technical progress. In general, two aspects are important to be successful in such an environment. On the operational level, companies have to be able to complete their projects on time and within budget. Smooth construction sequences and real-time information about work progress are crucial. In case of incidents or development of delays during a project, it is important to counteract these as fast as possible to avoid a loss of profit. Unfortunately, controlling systems which supply real-time information to the production process are missing in the construction industry [23]. On the

P. Rausch (✉) · M. Stumpf
Department of Computer Science, Georg Simon Ohm University of Applied Sciences,
Kesslerplatz 12, 90489 Nuremberg, Germany
e-mail: perausch@prof-rausch.de

M. Stumpf
e-mail: michael.stumpf@ohm-hochschule.de

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_3, © Springer-Verlag London 2013

tactical and strategic levels, a sophisticated strategy and smart analyses of project opportunities are necessary. Especially managers and controllers are interested in intelligent analyses of cost and performance data of completed projects. For example, they want to know which project types were successful or what can be done to improve the defined target figures.

In many cases the different levels are non-affiliated. There is a lack of coupling of the operational, the tactical and the strategic levels in terms of cost and performance management. The Corporate Performance Management (CPM) approach, which is explained in Chap. 1, can fill this gap. In this chapter, the benefits and issues of CPM as an instrument to couple the operational, tactical and strategic levels will be examined. It is shown how this approach can be applied to the construction industry. The findings are based on the research project EPOS (*E*fficient *P*r*O*cess design by *S*atellite-supported software in the earth moving and road construction industry). It should be pointed out that the aspects discussed in this chapter are not only relevant for the construction industry, but can be transferred to other industries as well.

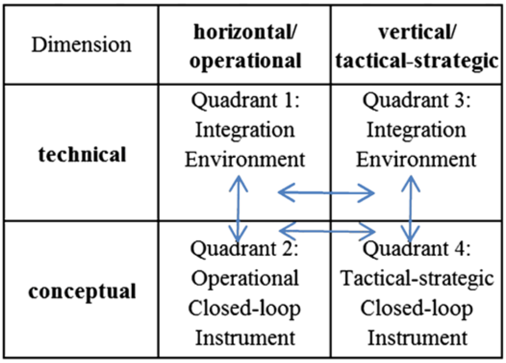
The following sections will show the details. After introducing the CPM integration grid to outline the scope of the approach presented, prior research related to possible CPM building blocks or rather technological prerequisites will be explained in the following section. In the subsequent section, details of the research project EPOS will be given. The intention of EPOS was to expand the scope of the controlling approach to a fully automated multi-layered, closed-loop system for the earth moving and road construction industry. Possible benefits, issues and organisational impacts are discussed subsequently. Afterwards, further developments of the CPM approach presented are outlined. The last section summarizes the chapter and provides final conclusions.

3.2 The CPM Integration Grid

To successfully introduce a CPM approach as described in Chap. 1, some requirements have to be fulfilled. First of all, the existence of an enterprise strategy and its communication is essential. Additionally, it is crucial to have access to a high quality data base. Hence, it can be helpful to benefit from a data warehouse. A data warehouse is defined as "a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions" [15].

Furthermore, technical solutions which integrate methods, metrics and processes enterprise-widely are needed [7]. It is important to have a multidimensional integration approach, as illustrated in Fig. 3.1.

Technical integration concerns the transfer of data, while conceptual integration covers methods and instruments to integrate data logically. The technical horizontal dimension (quadrant 1) is focused on the integration of business process data across functional borders on the operational level, for example from product demand to shipment [21, 34]. In contrast to that, the technical vertical dimension (quadrant 3) aims to integrate the various management levels of an organisation. This includes the



Dimension	**horizontal/ operational**	**vertical/ tactical-strategic**
technical	Quadrant 1: Integration Environment	Quadrant 3: Integration Environment
conceptual	Quadrant 2: Operational Closed-loop Instrument	Quadrant 4: Tactical-strategic Closed-loop Instrument

Fig. 3.1 The CPM integration grid

decision flows [34] as well as the supply of planning and controlling functions [21]. The conceptual dimension focuses on management and controlling aspects from a business perspective, by means of closed-loop information systems (quadrants 2 and 4). In addition, integration could be differentiated in terms of other aspects, like organisational units, etc. In any case, the grid can be used as a checklist to determine whether a company is well-positioned in terms of CPM or if there are still gaps which have to be filled. If all quadrants are addressed and all connections are linked to fulfil the needs of the corresponding connected quadrants, full transparency can be achieved.

The integration environment in quadrants 1 and 3 depends on the special requirements of an enterprise. It has to be established individually. For example, if an enterprise has deployed a service-oriented architecture (SOA) including an enterprise service bus (ESB), it is reasonable to use this architecture for the vertical integration as well as for the horizontal integration. This is illustrated by the double arrow between the quadrants 1 and 3. A SOA is a systems architecture concept. The basic idea is to provide functions as reusable and technically loosely coupled software services, which should be logically and technologically independent [26]. Service providers register their services in a particular registry or service directory, which is frequently called the ESB [6, 30]. Service requesters also direct their requests to the ESB, which tries to match the request to a registered service, based on meta-data that describe each service. The ESB is also responsible for establishing data transformations between interacting services at runtime. Independent of their implementation, they are accessible by their interfaces [26]. More information on service-oriented architectures can be found in [17] and [20]. Of course, other architectures and integration technologies can also be used.

The conceptual horizontal integration (quadrant 2) is achievable by using process controlling instruments. An example can be found in Sect. 3.4.1. The conceptual vertical integration (quadrant 4) can be established by a closed-loop approach, which will be explained in Sect. 3.4.2. This section includes examples for instruments. To link the vertical and the horizontal level on a conceptual basis, Key Performance Indicators (KPIs) can be used. This is depicted by the double arrows between the quadrants 2 and 4.

Also, the horizontal integration (link between quadrant 1 and 2) as well as the vertical integration (link between quadrant 3 and 4) have to be covered. Adequate software and network technologies help to implement the conceptual ideas. Examples for applications in the construction industry will be given in the following sections.

If any link between the different dimensions of the grid is missing, the potential benefits of CPM won't be fully exploited. Unfortunately, most CPM approaches consider only single aspects of the grid. In this chapter, a complete concept is outlined.

Another important issue which especially concerns the entries in quadrants 1 and 3 are governance aspects. Leadership, organisational structures, and processes should ensure that the organisation's IT sustains and extends the corporate strategies and objectives [16]. To fulfil this requirement, new technologies demand that the IT management has to be aligned to the company's strategy, which can be visualized by a strategy map (see Sect. 3.4.2.2). Additionally, new roles and an extension of the scope of the IT department are necessary. Communication and information chains have to be established. The corresponding service ownerships or responsibilities should be assigned [3, 25]. In terms of the closed-loop principle, it is recommendable to define service level agreements and to measure the performance of the integration environment.

The next sections describe how the CPM approach can be applied to the construction industry.

3.3 CPM in the Construction Industry: Prior Research

Most research work on CPM in the construction industry is still focused on horizontal integration. This might be due to a lack of a centralised high quality data source, which is an important requirement for CPM. In addition, controlling systems which supply real-time information on the production process are missing [23]. However, over the last few years some progress has been made in the field of satellite-supported, machine control & guidance systems. These systems collect data which can be used in terms of business activity monitoring, see Chaps. 2 and 15. Therefore, some examples should be mentioned. Satellite-supported systems for bucket wheel excavators in the coal mining industry have already been applied successfully. They deliver mining data and input for material planning [5, 13, 35]. A GPS-based system to control trucks is described in [10, 11]. Gut [12] reports on a system for GPS-based control of bulldozers. Additionally, data from a "virtual construction site" could be used for controlling purposes. The research cooperation "virtual construction site" (ForBAU) aimed to adapt concepts and practices of modern industrial organisations, their production technologies and logistics systems to those of the construction processes [33]. It was intended to collect data from different phases of a construction project and to transfer it to virtual landscapes. Further information about ForBAU can be found in [33]. Another research project, called Mefisto, which was completed in 2012, aimed to develop a platform for the execution of construction projects. An information system enabled visually descriptive real-time simulations, which are based on operational data on all levels of abstraction. The intention was to provide different stakeholders with information and to improve transparency [8].

In this chapter, a CPM example which includes a machine control & guidance system for excavators using GPS positioning will be introduced. In 2008, a research team worked to improve the productivity of excavators, which resulted in a research project. Now, the satellite-supported system is an important part of a complete information chain which is extended to a CPM approach by the research project EPOS (Efficient process design by satellite-supported software in the earth moving and road construction industry). To prove the system's functionality, extensive preparatory work was necessary: Members of the project team equipped a 20 ton excavator with antennas, GPS and other sensors, cables and an on-board computer [31]. The machine control & guidance system called the DTM Navigator was developed by Schreiber and Diegelmann [31]. It was originally designed for excavators, but can be also used for other heavy equipment, such as bulldozers. The system helps operators to navigate construction machines. Additionally, the machine control & guidance system provides users with valuable real-time information. It records data sets consisting of GPS-time, coordinates of an excavator's bucket, heading and angles. A so-called differential GPS (DGPS) is used to achieve the necessary accuracy. A more detailed description of the technical set-up can be found in [32].

By means of the DTM Navigator software, the excavator's performance for a certain performance period can also be computed. The procedure is described in [31]. The excavated volumes of the performances are processed and stored for further analyses. In terms of the closed-loop approach, which was explained in Chap. 1, the machine control & guidance system supports process automation, process execution and activity monitoring on the different levels. In addition to the horizontal data integration (quadrant 1, Fig. 3.1), the data could be also used for further analyses in terms of a vertical integration over multiple management levels (quadrant 3), in combination with conceptual CPM approaches (quadrants 2 and 4). The research project EPOS started in July 2009 in order to extend the scope of the system described.

3.4 Research Project EPOS

The idea of EPOS was to develop a multi-layered, closed-loop system supporting process performance analyses and further analyses on the tactical/strategic level. In the first step, a wireless data transfer for locally collected performance data of excavators to a central Production Activity Control (PAC) component was established. With the focus on excavators, a fully automated, multi-layered, closed-loop system for the earth moving and road construction industry in terms of the integration grid was achieved. As Fig. 3.2 shows, a multi-layered architecture with new components was introduced for this purpose.

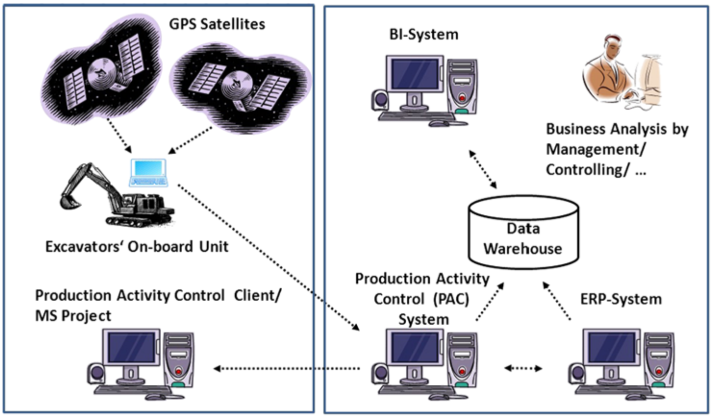

GPS Satellites
Excavators' On-board Unit
Production Activity Control Client/ MS Project
BI-System
Business Analysis by Management/ Controlling/ ...
Data Warehouse
Production Activity Control (PAC) System
ERP-System

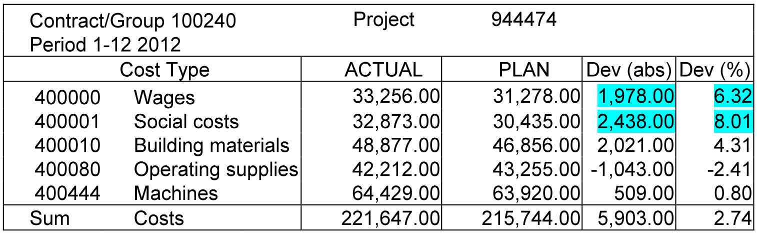
Fig. 3.2 EPOS architecture

3.4.1 The Operational Level

As already mentioned, the machines' data can be collected at the construction sites. The DTM Navigator provides this data from the operational level (quadrant 1, Fig. 3.1) to another EPOS component. The excavated volumes and their corresponding time stamps are transferred to a central PAC server by means of a wireless network. The PAC receives the data and covers the operational level of CPM by offering support for the controlling and planning activities of construction supervisors. They can access the analyses with their mobile devices, for example with a mobile phone. The web-based application is remotely accessible from all construction sites. Additionally, data which is manually entered by the construction supervisors, like lighting conditions or soil classes during the excavators' operation, can be added for further analyses. The PAC component can also deliver a graphical representation of key figures, such as operating and down times of construction machines and a comparison between planned and actual performance parameters. In case of incidents, immediate action can be taken.

It is also possible to combine data from the PAC with cost parameters of an ERP system, for example, details of machine operators' wages, costs of fuel and related operating materials, costs of the machine itself (depreciation, interest charges and average repair charge rates), costs of wear and tear, and miscellaneous costs. It can be very useful to distribute the information by means of ERP reporting components. Thus, users can get a more detailed view of the data from a business perspective. Actual cost figures, which are based on the performance data for two following periods, can be generated and compared by the ERP system [27]. Also, comparisons for actual and planned values of different cost types were implemented. As seen in Fig. 3.3, cost violations which exceed a certain threshold (here 6 %) are highlighted.

The performance data can also be used to derive revised cost calculations from comparable machine operations. Because most of these analyses don't concern the



Contract/Group 100240 Project 944474
Period 1-12 2012

Cost Type		ACTUAL	PLAN	Dev (abs)	Dev (%)
400000	Wages	33,256.00	31,278.00	1,978.00	6.32
400001	Social costs	32,873.00	30,435.00	2,438.00	8.01
400010	Building materials	48,877.00	46,856.00	2,021.00	4.31
400080	Operating supplies	42,212.00	43,255.00	-1,043.00	-2.41
400444	Machines	64,429.00	63,920.00	509.00	0.80
Sum	Costs	221,647.00	215,744.00	5,903.00	2.74

Fig. 3.3 Example of an ERP report [27]

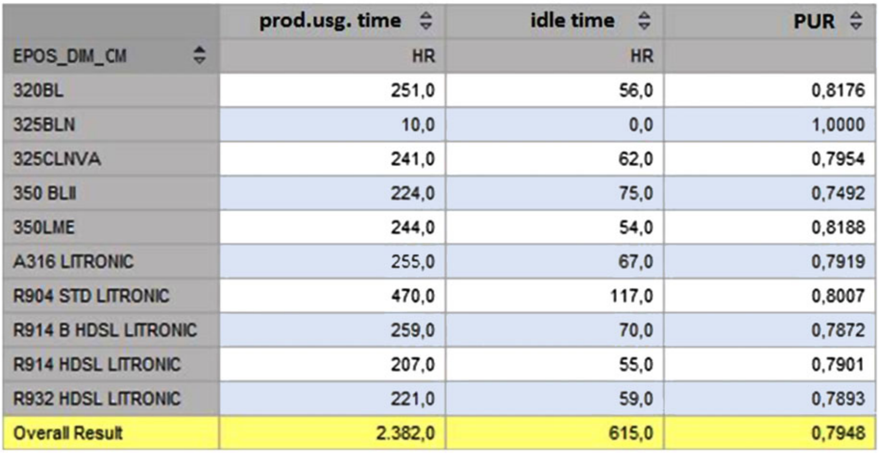
tactical and strategic levels, this module of the EPOS project can be assigned to quadrant 2 of Fig. 3.1.

3.4.2 The Tactical and the Strategic Level

On the strategic level a broad variety of instruments can be used. A widespread instrument is the Balanced SCorecard (BSC). The basic idea of the BSC is to achieve a balanced mix of figures to measure quality, finance, process efficiency, customer satisfaction as well as progress in terms of learning and growth [18]. Another instrument which will be explained in Sect. 3.4.2.2 is an enhancement of the BSC approach—the strategy map. Furthermore, other CPM instruments also include forecast, planning and budgeting components [2]. These instruments support the ability to adapt the processes and the business model. Software products and instruments are discussed in [24] and in the Chaps. 8, 9 and 11. For analysis purposes classical business intelligence tools, like reporting systems, dashboards, data mining components or OnLine Analytical Processing (OLAP) tools can be applied. These instruments can also support the tactical level. In the context of the EPOS project OLAP was used. A possible extension to the strategic level by means of strategy maps in combination with BSC can also be applied in this field and will be explained in Sect. 3.4.2.2.

3.4.2.1 Online Analytical Processing

At the top level of the EPOS architecture, a Business Intelligence (BI) system is used for business analyses of project data and other data from different sources, such as ERP systems. It addresses the tactical and strategic levels of the CPM approach and supports several stakeholders, for example managers and controllers (integration grid quadrant 4). For this purpose, the data of the ERP and PAC component can be transferred to a BI system by an Extraction, Transformation and Loading (ETL) process (quadrant 3). It is stored in a data warehouse. Though it is not necessary to use a data warehouse, it helps to ensure a high data quality for OLAP. OLAP allows



	prod.usg. time	idle time	PUR
EPOS_DIM_CM	HR	HR	
320BL	251,0	56,0	0,8176
325BLN	10,0	0,0	1,0000
325CLNVA	241,0	62,0	0,7954
350 BLII	224,0	75,0	0,7492
350LME	244,0	54,0	0,8188
A316 LITRONIC	255,0	67,0	0,7919
R904 STD LITRONIC	470,0	117,0	0,8007
R914 B HDSL LITRONIC	259,0	70,0	0,7872
R914 HDSL LITRONIC	207,0	55,0	0,7901
R932 HDSL LITRONIC	221,0	59,0	0,7893
Overall Result	2.382,0	615,0	0,7948

Fig. 3.4 Drill down of key figure "productive usage rate"

multidimensional data analysis [29]. Important figures are related to dimensions, which help stakeholders to get a better understanding of their business.

In our example, OLAP offers many valuable reports on strategic and tactical key figures, such as capacity utilisation, performance, cost figures and others. They can be analysed in terms of machine types, time and many other dimensions. For instance, it is possible to analyse information by drill down functions for tactical analyses, as Fig. 3.4 illustrates. The report shows the level of construction machines' utilisation and contains their productive usage rates (PUR_i, right column).

The PUR is the result of the productive usage time pt_i (left column) divided by the total amount of available time which is the sum of pt_i and the idle time it_i (left plus middle column):

$$\mathrm{PUR}_i = pt_i/(pt_i + it_i) \tag{3.1}$$

When the PUR of a deployed construction machine falls beneath a threshold for a certain period, this could be a hint that a performance-based maintenance is necessary or that the machine should be replaced. On an aggregated level, it could be analysed whether the production planning process should be redesigned. Further analyses of whether special machine types have more down times compared to others can support the procurement process. It is also possible to analyse the key figure "costs per cubic meter excavated material" depending on the projects' specific environment, for example soil conditions. One possible finding may be that the costs of projects which have to cope with difficult soil conditions (like rocky grounds) are generally much higher than expected when compared to other projects. Hence, it should be analysed whether the construction machine inventory is suitable for this type of projects or whether the cost schedule is still sufficient. As a conclusion, it might be advisable to purchase specialised equipment or to withdraw from projects with difficult soil conditions. This finding can also indicate an inadequate business strategy, for example revenue maximization. Thus, in this case, an adjustment of the strategy and the business objectives might be the result.

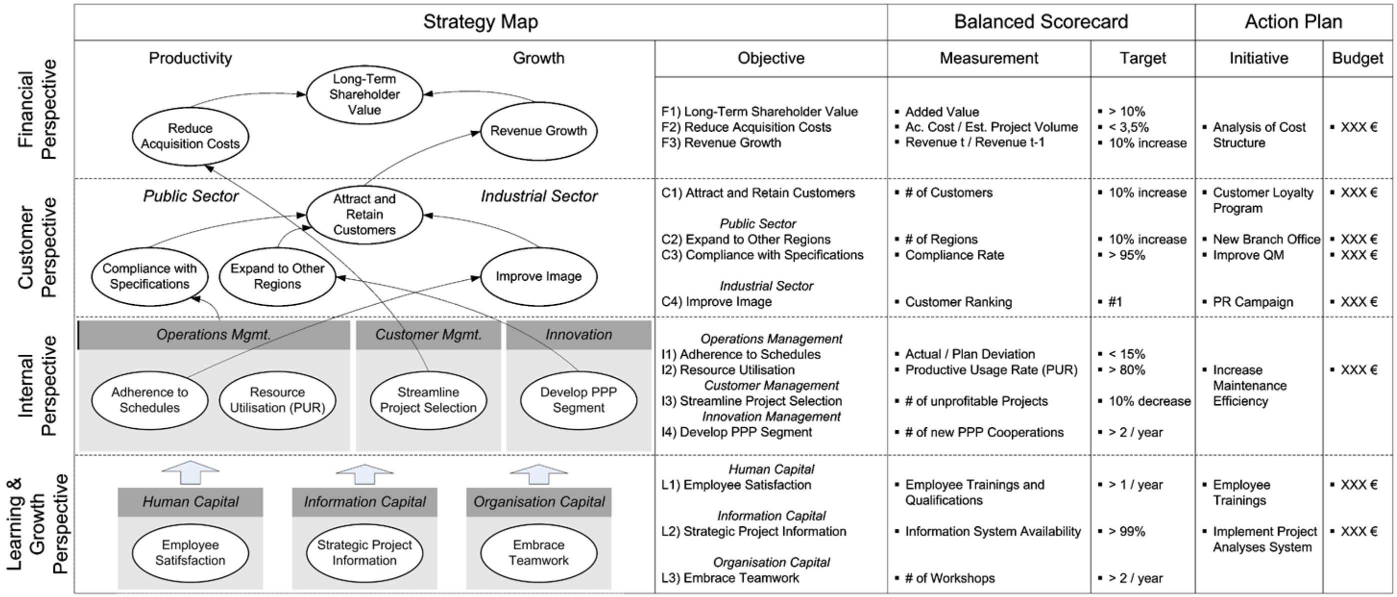
3.4.2.2 Balanced Scorecard and Strategy Maps

The BSC framework is a management instrument which can be used to control and align an organisation to its strategy. Kaplan and Norton propose four perspectives to control and evaluate the organisations' performance: finance, customer, internal business processes, and learning & growth. Additionally, there are further developments which add another perspective regarding sustainability. This approach is called Sustainable Balanced SCorecard (SBSC). Since the lifetime of products in the construction industry (e.g. buildings, bridges, and roads) can easily span many decades, it is advisable to take sustainability into account, for example, by means of avoiding waste or by saving energy [9]. The goal of strategic management is to balance these perspectives. By doing so, the focus shifts from traditional management of financial measures towards integrated approaches which also include non-financial measures [18].

The combination of a BSC in conjunction with a strategy map empowers organisations to communicate the strategy internally and to better understand the impact of certain KPIs and events. A strategy map supplements a BSC with the possibility to describe cause-and-effect relationships between strategic goals. Also, the cause-and-effect relationships between non-financial and/or financial measures are described and help to reach objectives of the customer perspective (non-financial measures), as well as objectives of the financial perspective concurrently [14].

Kaplan and Norton provide a generic strategy map which has to be customised to the needs of a specific organisation [14]. The typical procedure for creating an individual strategy map is described as a top–down method, which starts by defining strategic objectives for each perspective [14]. Initially, the objectives of the financial perspective are defined. In BSC approaches, the financial perspective is retained as the ultimate objective of profit-maximizing organisations, since financial measures indicate whether the organisation's strategy was successful on the bottom-line [19]. The customer perspective is considered as another central element. Targeted customer segments are defined, as well as the value proposition the organisation will offer to them. The former two perspectives describe the strategy and its economic consequences. A description of how the strategy will be accomplished can be found in the internal and the learning & growth perspectives [19]. The internal perspective includes objectives which are necessary to achieve the value proposition by means of process measures. The last perspective—learning & growth—defines objectives concerning intangible assets.

Once the logic of the strategy and the objectives is defined by means of a strategy map, a BSC is used to translate them into concrete KPIs and a set of targets [14, 19]. In Fig. 3.5, an example of a strategy map, including a derived BSC, is shown. Additionally, strategic investments and action plans have to be defined to ensure that targets can be achieved in the designated timeframe [19].



	Balanced Scorecard			Action Plan	
	Objective	Measurement	Target	Initiative	Budget
Financial Perspective	F1) Long-Term Shareholder Value F2) Reduce Acquisition Costs F3) Revenue Growth	▪ Added Value ▪ Ac. Cost / Est. Project Volume ▪ Revenue t / Revenue t-1	▪ > 10% ▪ < 3,5% ▪ 10% increase	▪ Analysis of Cost Structure	▪ XXX €
Customer Perspective	C1) Attract and Retain Customers *Public Sector* C2) Expand to Other Regions C3) Compliance with Specifications *Industrial Sector* C4) Improve Image	▪ # of Customers ▪ # of Regions ▪ Compliance Rate ▪ Customer Ranking	▪ 10% increase ▪ 10% increase ▪ > 95% ▪ #1	▪ Customer Loyalty Program ▪ New Branch Office ▪ Improve QM ▪ PR Campaign	▪ XXX € ▪ XXX € ▪ XXX € ▪ XXX €
Internal Perspective	*Operations Management* I1) Adherence to Schedules I2) Resource Utilisation *Customer Management* I3) Streamline Project Selection *Innovation Management* I4) Develop PPP Segment	▪ Actual / Plan Deviation ▪ Productive Usage Rate (PUR) ▪ # of unprofitable Projects ▪ # of new PPP Cooperations	▪ < 15% ▪ > 80% ▪ 10% decrease ▪ > 2 / year	▪ Increase Maintenance Efficiency	▪ XXX €
Learning & Growth Perspective	*Human Capital* L1) Employee Satisfaction *Information Capital* L2) Strategic Project Information *Organisation Capital* L3) Embrace Teamwork	▪ Employee Trainings and Qualifications ▪ Information System Availability ▪ # of Workshops	▪ > 1 / year ▪ > 99% ▪ > 2 / year	▪ Employee Trainings ▪ Implement Project Analyses System	▪ XXX € ▪ XXX €

Fig. 3.5 Strategy Map (extract) and selected objectives/measures, based on [19]

In the next step, the objectives and targets are passed down, and subsequent organisational units derive their own BSCs in accordance with higher levels. Thus, quadrant 4 of the integration grid is addressed. The BSC in turn is supplied by data from the operational level (in our case PAC and ERP data), integrating quadrants 2 and 4. The corresponding monitoring results can trigger actions. Furthermore, a validation of the cause-and-effect relationships can be advisable. Hence, strategies are dynamic and the closed-loop principle is fulfilled.

Now, the translation of a strategy map by means of a BSC will be applied to the EPOS case. A few examples of figures of the different perspectives are selected to outline the principle. At the beginning, when the figures were first defined, an important issue came up. Within the construction industry it is important to consider that different types of projects, for example, public sector and industrial projects, underlie different strategies [23]. Hence, in our case, it was reasonable to have special figures and objectives for the variety of project types. The bidding, the implementation and the post-implementation phase have to be distinguished with respect to the projects' life cycle.

In the bidding phase, the acquisition costs related to the expected project volume (see Fig. 3.5, F2) can be a useful figure to represent the financial perspective, while in the implementation phase a comparison between planned and actual cost parameters is necessary. Also, the measurements for revenue growth (F3) and for the long-term shareholder value (F1) are important key figures. Rejection and reworking costs can indicate whether the post-implementation phase was successful or not [23]. According to Kaplan and Norton, the overall financial objectives in this phase are the operating cash flow (before depreciation) and reductions in working capital requirements [18].

Since it would be suboptimal for the total result just to optimise financial key figures like F2, other perspectives have to be considered. Concerning customer aspects, customer satisfaction is very important. It is driven by compliance of the project results with customer specifications (C3) and the promised progress timelines. For example, adherence to schedules can be measured by means of deviation rates. To cover the internal perspective during the implementation phase, a comparison between planned and actual performance parameters, for example, for earth works, can be valuable. Details concerning the EPOS project can be found in [28].

The productive usage rates which were explained in Sect. 3.4.2.1 can be another example in the context of the internal perspective (I2). They represent a valuable figure to measure the efficiency of business operations and can also be used for further analyses in the context of the learning & growth perspective. PURs do not necessarily have to refer to machines. They can also be used for project teams. Based on multidimensional OLAP analyses, key figures in terms of the specific project environments can be analysed. Dependent on the results of the measurements, initiatives for employee training can be triggered to improve the performance and the employees' satisfaction (L1) in future. Further project analyses can help to distinguish profitable from unprofitable projects. This could trigger initiatives to improve the project margins [23] or an elimination of unprofitable project types.

3.5 Benefits, Issues and Impacts

The examples in Sect. 3.4 show that the fully automated transfer of locally collected cost and performance data of construction machines or other data sources to an information system improves transparency and helps to avoid costs. Additionally, the preparation of the data for several stakeholders, for example, construction supervisors, controllers or managers, is valuable. Performance assessments help to indicate whether the actual performance is according to plan. Transparency of progress is important to avoid backlogs. A permanent comparison of the actual and planned performance plays an important role in completing projects on time and is necessary to avoid contractual penalties. In particular, construction supervisors and managers benefit from cost transparency. Problems on the operating level become immediately visible. So action can be taken as quickly as possible to avoid a loss of profit. Another important advantage is the fast distribution of information. Information on finished work is important in terms of solvency, because in the construction industry it is common that invoicing depends on finished construction phases. Additionally, processed cost parameters can be very useful for post-calculation of projects and calculating subsequent bids for similar construction projects. The instruments described are also important for planning and initialisation of construction work.

Furthermore, tactical and strategic management can benefit from the ideas presented. In Sect. 3.4 it was shown how intelligent cost analysis helps to avoid future costs. The variety of analyses can increase the basis of information for decisions. Managers and controllers are enabled to identify trends which have an impact on the tactical or strategic level early, so they can take action promptly. Due to business performance monitoring and process performance analyses, continuous process improvements can be achieved. The strategic guidelines are translated into measurable results. Processes are aligned with the business strategy on the tactical as well as on the operating level, and management is supplied with consistent information. For instance, analysing the performance data of construction machines in terms of special project environments can increase efficiency, as the OLAP example shows. The coverage of all quadrants of the integration grid and their interconnection saves time and costs in processing information. Additionally, the quality of information is improved, and management is supported by efficient planning, measuring and control instruments.

In spite of all the benefits, the costs for the technical equipment have to be considered before the CPM approach presented can be introduced in a different environment. Additionally, the employees have to be trained, and operating costs for the CPM components and the technologies should be taken into account. Initially, it might be necessary to establish the links between the quadrants of the integration grid, or rather the data flows between the components which were illustrated in Fig. 3.2 and described in the sections above. The costs of horizontal and vertical integration depend on the prerequisites which companies already possess. If, for example, a data warehouse is already established, the additional costs for the introduction of CPM components will be lower than those for an environment without a centralised data source. In addition, there are some other issues to master. As already mentioned in Sect. 3.2, the introduction of a CPM system also has organisational impacts. Roles, for example in the IT department, have to be adapted or created. For example, it has to be clear who is responsible for the quality of data or what should be done in case of transfer failures. Tasks have to be reassigned or reengineered, service level agreements must be defined and responsibilities distributed [1, 22]. Another aspect is the necessity to communicate and reward improvements in performance. Hence, a performance management approach should be coupled with an incentive system.

Very often this implies the necessity of a cultural change. Especially in Tayloristic organised companies, the extension of the employees' responsibilities and the enlargement of their scope of tasks can be exhausting, due to acceptance problems and internal resistance. Besides these issues, legal aspects also have to be considered. For example, in Germany, the analysis of personal data is restricted by the Data Protection Act [4]. Performance analysis of personal data can require the anonymisation of data. In other countries, similar regulations might also need to be respected. Another challenging aspect is the definition of appropriate key figures. This task requires a deep insight into business processes and the causal relations of the value chain. Of course, the relevance of key figures depends on the company's strategy and its special environment. Workshops which are moderated by external consultants can be very useful, because external staff can introduce a neutral perspective. On the other hand, this results in higher costs.

Whether the introduction of the CPM approach described is advisable or not depends on the costs and the benefits. While costs are quite easy to predict, it is usually difficult to quantify the benefits. However, it can be assumed that the benefits will exceed costs, because solely the avoidance of contractual penalties due to missed deadlines can save hundreds of thousands of euros. Considering additional strategic competitive advantages, the summary should be positive.

3.6 Further Developments

So far, a complete information chain to deliver a detailed view of excavators' performances in the earth moving and road construction industry has been established. Benefits on the operational, tactical and the strategic levels can be exploited. However, there are still many ideas to extend the scope of the CPM application outlined. It would be very useful to combine EPOS with other machine control & guidance systems, which were designed for bulldozers, trucks, etc. The system design already considers the possibility of extending the approach to other machine types and to extend it in terms of an operational BI. It would be also very interesting to integrate subcontractors' data. Based on corresponding additional facts, the analytical components can be enhanced on all levels. This could be addressed by extending the reporting components (BI, ERP and PAC).

3.7 Summary and Conclusions

In the last decade, much progress has been made in the field of CPM. Technical as well as conceptual ideas have been realised. CPM applications can be a valuable instrument for the construction industry to cope with the challenging environment. To complete projects on time and within budget, controlling of costs and performance is crucial. Data from the operational level which is used for controlling can be also very useful in terms of tactical and strategic analysis. In order to exploit this potential, it is necessary to establish an all-embracing approach and to integrate all of the components.

In this chapter, an integration grid was introduced to address all of the important linkages between the different components systematically. As an example, a new generation of controlling systems for the earth moving and road construction industry illustrates the idea of a complete integration. The CPM approach is based on data from a satellite-supported, machine control & guidance system which delivers real-time operating data from construction sites. This data combined with data from other sources, for example from an ERP system, enables intelligent analyses. So, the gap between the operational, tactical and strategic levels is filled by offering intelligent analyses of key figures. The reports provide different stakeholders, such as construction supervisors, managers or controllers of construction companies, with valuable information. Benefits, for example, the possibility to implement countermeasures as fast as possible in case of incidents or developing delays during a project, were examined and possible issues of the approach outlined. The example of the construction industry shows that the CPM approach can have a strong impact on business strategy and on business processes. It offers a chance to improve the competitiveness of construction companies. The findings of the EPOS project encourage further developments.

Acknowledgements The authors are particularly grateful for the generous support of the following cooperating partners and sponsors: STAEDTLER-Stiftung, Nuremberg/ Germany and BMTI GmbH, Munich/ Germany.

Last but not least: Thanks a lot to the EPOS project team for their excellent work.

References

1. Baars, H.: Distribution von Business-Intelligence-Wissen. In: Chamoni, P., Gluchowski, P. (eds.) Analytische Informationssysteme, 3rd edn., pp. 409–424. Springer, Berlin (2006)
2. Barrett, R.: Planning and Budgeting for the Agile Enterprise. Elsevier, Oxford (2007)
3. Bashiri, I., Engels, C., Heinzelmann, M.: Strategic Alignment. Springer, Heidelberg (2010)
4. BfDI (ed.) Bundesdatenschutzgesetz (BDSG) (2009). Available via http://www.bfdi.bund.de/DE/GesetzeUndRechtsprechung/BDSG/BDSG_node.html. Accessed 7 Sep 2012
5. Bulowski, T., Körber, T.: Betriebliche Neuerungen von GPS im heimischen Braunkohlenrevier. World Mining – Surf. Undergr. **56**(6), 413–421 (2004)
6. Chappell, D.A.: Enterprise Service Bus. O'Reilly, Sebastopol (2004)
7. Coveney, M.: CPM: What it is and how it is different from traditional approaches?—Part One (2003). Available via http://www.businessforum.com/Comshare01.html. Accessed 7 Sep 2012

8. Dresden University of Technology: Mefisto (2012). Available via http://www.mefisto-bau.de/objective. Accessed 7 Sep 2012
9. Girmscheid, G.: Strategisches Bauunternehmensmanagement. Prozessorientiertes integriertes Management für Unternehmen in der Bauwirtschaft, 2nd edn. Springer, Berlin (2010)
10. Günthner, W.A., Kessler, S., Sanladerer, S.: EDV gestützte Fahrzeugdisposition und -abrechung im Baubereich zur Optimierung der Prozesskette. In: Marquardt, H.-G. (ed.) Tagungsbeiträge 2, pp. 17–26. WGTL-Fachkolloquium, Dresden (2006)
11. Günthner, W.A., Kessler, S., Sanladerer, S.: Transportlogistik in der Baubranche – Optimierung durch den Einsatz eines Flottenmanagementsystems. In: Jahrbuch Logistik, pp. 252–256. free beratung GmbH, Korschenbroich (2007)
12. Gut, O.: Automatische Steuerung von Baumaschinen mit GPS. Der Gartenbau **43**, 2–4 (2007)
13. Heck, V., Weber, P., Doll, M.: Satellitengestützte Baggereinsatzsteuerung im Tagebau. Autom.tech. Prax. **46**(11), 42–47 (2004)
14. Hügens, T.: Balanced Scorecard und Ursache-Wirkungsbeziehungen. Gabler, Wiesbaden (2008)
15. Inmon, W.H.: Building the Data Warehouse, 4th edn. Wiley, New York (2005)
16. IT Governance Institute: Board Briefing on IT Governance, 2nd edn. (2003). Available via http://wikimp.mp.go.gov.br/twiki/pub/EstruturaOrganica/AreaMeio/Superintendencias/SINFO/Estrategia/BibliotecaVirtual/MaterialExtra/26904_Board_Briefing_final.pdf. Accessed 7 Sep 2012
17. Josuttis, N.M.: SOA in Practice: The Art of Distributed System Design. O'Reilly, Sebastopol (2007)
18. Kaplan, R.S., Norton, D.P.: Balanced Scorecard: Translating Strategy Into Action. Harvard Business School Press, Boston (1996)
19. Kaplan, R.S., Norton, D.P.: Strategy Maps. Harvard Business School Press, Boston (2004)
20. Krafzig, D., Banke, K., Slama, D.: Enterprise SOA—Service-Oriented Architecture Best Practices. Prentice Hall, New York (2007)
21. Lehner, F., Wildner, S., Scholz, M.: Wirtschaftsinformatik. Hanser, Munich (2007)
22. Maier, R.: Knowledge Management Systems–Information and Communication Technologies for Knowledge Management, 2nd edn. (2004). Berlin
23. Nebe, L.: Kennzahlengestütztes Projekt-Controlling in Baubetrieben. Ph.D., Dortmund (2003). Available via https://eldorado.tu-dortmund.de/bitstream/2003/2887/1/Nebeunt.pdf. Accessed 7 Sep 2012
24. Oehler, K.: Unterstützung von Planung, Forecasting und Budgetierung durch IT-Systeme. In: Chamoni, P., Gluchowski, P. (eds.) Analytische Informationssysteme, 3rd edn., pp. 329–360. Springer, Berlin (2006)
25. Ott, C., Korthaus, A., Böhmann, T., Rosemann, M., Krcmar, H.: Towards a reference model for SOA Governance (extended version) (2010). Available via http://eprints.qut.edu.au/31057/1/c31057.pdf. Accessed 7 Sep 2012
26. Rausch, P., Landes, D.: New opportunities and challenges in dynamic environments through flexible process design and service-oriented architectures: the example of the German insurance business. In: Proceedings of CONQUEST, September 2007, pp. 333–342 (2007), dPunkt-Verlag
27. Rausch, P., Schreiber, F., Diegelmann, M.: Effiziente Prozessgestaltung im Erd- und Straßenbau durch den Einsatz von satellitengestützten Entscheidungsunterstützungssystemen. WIRTSCHAFTSINFORMATIK **4**, 305–313 (2008)
28. Rausch, P., Schreiber, F., Diegelmann, M.: Satellitengestütztes Projektcontrolling bei Erd- und Straßenbauprojekten. In: Küpper, A. (ed.) Proceedings of the 7th GI/KuVS-Technical Discussion, Ortsbezogene Anwendungen und Dienste, 23rd–24th September 2010. Deutsche Telekom Laboratories/TU, Berlin (2010)
29. Rob, P., Coronel, C.: Database Systems: Design, Implementation, and Management, 8th edn. Course Technology. Thomson Learning, Boston (2009)
30. Schmidt, M.-T., Hutchinson, B., Lambros, P., Phippen, R.: The enterprise service-bus: making service-oriented architecture real. IBM Syst. J. **44**(4), 781–797 (2005)

31. Schreiber, F., Diegelmann, M.: Entwicklung eines DGM-basierten Maschinenführungssystems für Bagger. In: Chesi, G., Weinold, T. (eds.) 14. Internationale geodätische Woche Obergurgl, pp. 83–93. Wichmann, Heidelberg (2007)
32. Schreiber, F., Diegelmann, M., Rausch, P.: Use of a machine control and guidance system, determination of excavator performance, cost calculation and protection against damaging of pipes and cables. In: Ingensand, H., Stempfhuber, W. (eds.) Proceedings of the 1st International Conference on Machine Control and Guidance, ETH Zurich, Switzerland, June 24th–26th 2008, pp. 21–30 (2008)
33. TUM: ForBAU (2010). Available via http://www.fml.mw.tum.de/forbau/. Accessed 7 Sep 2012
34. Vernadat, F.: Enterprise Modeling and Integration: Principles and Applications. Chapman and Hall, London (1996)
35. Weber, P., Cerfontaine, P.A.: SABAS – Satellitengestützte Baggereinsatzsteuerung. In: Virtual Reality Center Aachen (2006). RWTH Aachen (ed.): Jahresbericht 2005/2006. Available via http://www.vrca.rwth-aachen.de/jabe/2006/pdf/VRCA-Jahresbericht_2006.pdf. Accessed 7 Sep 2012

Chapter 4
Adaptive Business Intelligence: The Integration of Data Mining and Systems Engineering into an Advanced Decision Support as an Integral Part of the Business Strategy

Zafer-Korcan Görgülü and Stefan Pickl

Abstract IT-based decision support is in the heart of business intelligence. It should be based on a successful integration of data analysis techniques and certain system engineering (like system dynamics) concepts. This contribution introduces in the large realm of IT-based decision support and its meaning for a modern business strategy. Central is the relationship to Business Intelligence with its own characteristics and requirements. The relevant data mining techniques are summarized and characterized by its special role within traditional business intelligence approaches.

As an holistic approach this chapter tends to combine a classical data-centric approach with a modern system-engineering concept ("system of systems"-thinking). As a result, this new approach leads to an advanced concept of Adaptive Business Intelligence. It will be characterized and described by several successful examples.

4.1 Introduction: IT-Based Decision Support

4.1.1 IT-Based Decision Support

The notion *IT-based decision support (IDS)* used in this chapter describes a computer-based information system that assists in decision making regarding business in all its three main levels such as designing, processing and managing.

Another purpose of an *IDS* with information state $\mathcal{I}$ is to offer, once the problem P of the decision maker is formulated and delivered to the *IDS*, an (assessable) set $\mathcal{C}$ of beneficial decisions/alternatives from which the decision maker can choose the one he considers the best in the current situation $\mathcal{I}$. This implies that the *IDS* is capable of solving and listing a compilation of selected solutions of formulated decision problems that are explicitly addressed to the *IDS*.

Further, an *IDS* has a cooperative quality which means that the decision maker can simulate all consequent and related scenarios based on the suggested decisions tendered by the *IDS*.

Z.-K. Görgülü (✉) · S. Pickl
COMTESSA, Chair for Operations Research, Universität der Bundeswehr Müchen,
85577 Neubiberg-München, Germany
e-mail: korcan.goerguelue@unibw.de

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_4, © Springer-Verlag London 2013

The core of an *IDS* in *BI* is the expertise of specialized problem solving techniques for a wide range of business related problems deposited as actual data or equivalent assets, i.e., procedures, algorithms, set of rules, available tools, library of solvers, etc., in short: as information.

The inventory of actual and available information or equivalents like exploitable assets may be strongly fluctuating even over small periods of time, i.e., $\mathcal{I}=\mathcal{I}(t)$. Thus, through the possibly high volatility of the available data and assets itself the potential advantageousness of a decision becomes highly uncertain and, moreover, fluctuating as well. In particular, the inventory of data and equivalent assets are interacting with the decision maker.

Generally speaking, an *IDS* consists of an inventory of information that encompasses raw data as well as equivalent assets, an environment for formulating, editing and processing decision context, and finally an user interface to connect the decision maker along with his personal knowledge and expertise with the *IDS* for interaction. However, the interface also serves as editor, worksheet and display where *IDS* and decision maker meet.

As the suggested decisions of the *IDS* are replies of engineered scenarios played through the *IDS*, the decision itself is engineered. In this chapter we will focus on building an *IDS* in the distinct domain of *BI* including management: a *IT-based Decision Support (System) for Business Intelligence (BIDS).*

Further details concerning the *BI* basics can be found in the introduction chapter of this book.

4.1.2 Business Intelligence

BI signifies the transformation of an organization's capabilities into useful knowledge with the potential to gain competitive advantage in the market. Obviously, *BI* is an enterprise-centered view on the market including participation in it. Whereby, the *IDS* is the core of every *BI*-approach that is devised to provide solutions or beneficial alternatives to a specific problem that the decision maker embedded into an accurate scenario. Stringing together a set of separate potential decisions enables the decision maker to test, modify and improve the business strategy he has in mind.

"While the business world is rapidly changing and the business processes are becoming more and more complex making it more difficult for managers to have comprehensive understanding of business environment. The factors of globalization, deregulation, mergers and acquisitions, competition and technological innovation, have forced companies to re-think their business strategies and many large companies have resorted to Business Intelligence techniques to help them understand and control business processes to gain competitive advantage. *BI* is primarily used to improve the timeliness and quality of information, and enable managers better understand the position of their firm as in comparison to competitors." [17]

On that note, *BI* utilizations and methods enable companies to assess changing trends in market share, alteration in customer attitude, spending behavior, company capabilities, and, even, the condition of the market itself.

BI qualifies the analyst/manager to ascertain which strategic or operational adjustments are the best to answer to changing trends in an overall beneficial way.

In this chapter we understand *BI* as an area of mostly IT-based Decision Support, i.e., an information system that is purposed to support complex decision making, including solving complex, semi-structured, or even ill-structured problems [6, 27, 31]. "The first reference to *BI* was made by [19], which has replaced other terms such as Executive Information Systems and Management Information Systems [26, 35, 36]." [17]

Being resident in the *IDS* discipline, *BI* attracts large interest from both the industry and researchers [3, 10, 14, 15, 26, 29, 30]. *BI* appears as an architecture/system that collects and stores data, analyzes it using designated analytical tools, and delivers information, including, intrinsic relationships: that ultimately enables organizations to improve their process of decision making and finding [10, 18, 24–26, 28, 34, 36].

4.2 Data Mining and Its Role in Business Intelligence

4.2.1 Data Mining

Data Mining in a *BIDS* is the attempt to detect (prevalent and hidden) patterns or functional (inter)dependencies in an existing inventory of information $\mathcal{I}$. Before starting the process of Data Mining the inventory $\mathcal{I}$ has to be preprocessed by suitable filtering $\mathfrak{f}$. The engineering of a suitable filter strongly depends on what the decision maker declares as purpose. Preprocessing of data is a vital part in the data mining process, not least because data collection is loosely implemented and controlled. The collected data may contain, sporadically, values that are out of range, missing, or plain impossible (e.g., pupil of age 40 at local primary school). Interpreting the collected set of data, without examining the current data for further usability, may easily bring about misleading results and possibly wrong conclusions. The final good of data preprocessing is the training set, meaning, a set of data used to discover potentially predictive relationships.

The general tasks of Data Mining are [4]:

- *Clustering* which means creating a set of categories as containers for information of special characteristics,
- *Classification* which deals with compiled information of the past and present and deciding whether an information fits in a given category or not,
- *Prediction* which deals with forecasting/estimating information to occur on the basis of the current inventory,
- *Association*, i.e., detecting information that occur often at the same time or a certain order or with delay, in other words: detecting patterns,

- *Text Analysis* which deals with finding key terms and phrases in a text.

Information may posses the following property such as, to be [13]:

- *qualitative*, meaning without immanent ordering, e.g., Boolean values or a product line etc.,
- *quantitative*, meaning that the information is in $\mathbb{R}^m$ for some $m \in \mathbb{N}$,
- *set valued*, meaning it has more than one attribute,
- *ordinal*, meaning categorical but with obvious ordering, e.g., screen diagonal.

In this day and age businesses, indeed, are more capable of addressing and accessing their target customers. Data mining is a catalyzer for their success, it uses real data collected in real business cases and helps providing models and modifying evolving business strategies. Today, data mining in business intelligence implies a variety of business-oriented applications and has become a indispensable tool for detecting dependencies between business variables or gaining conception of causal relations.

4.2.2 *Utilization of Data Mining in Business Intelligence*

In fact, businesses are faced with an explosive growth of data that is, mainly, caused by automated data collection tools or database systems. In addition, businesses have access to abundant data that can be obtained from various sources. Of course, overwhelmed with partly disjoint and unrelated data the businesses wish a automated analysis of these massive data sets. That is exactly what data mining achieves. Data mining is detection of patterns from a huge amount of data, it can be used with a both predictive and descriptive purpose. Exemplary for data mining with predictive purpose, we cite classification, regression, time series analysis and prediction. Further, examples for data mining with descriptive purpose are clustering, association rules, summarization/clustering and sequence discovery. Data mining is successfully utilized in market analysis and management (such as target marketing, customer relationship management, market basket analysis, cross selling and market segmentation) as well as risk analysis and management, including forecasting, customer retention, improved underwriting, quality control and competitive analysis.

We pick up on a few common techniques in data mining that can be used to classify, predict, cluster and/or simply associate present business information, e.g., namely:

• *Decision Trees*: from a business perspective decision trees denote a segmentation of the original inventory $\mathcal{I}$ or parts of it where each segment/class is one leaf of the tree. In fact, segmentation is applicable in classifying, e.g., clients, market trends, behavioral responses to enterprise policies, products, branch offices, their sales districts and service coverage. The classification is essentially a expedient for intended (further) prediction. Information gathered in each leaf of the decision tree is pooled there due to bearing significant resemblance to information being predicted. Whereas, algorithms for generating decision trees may be quite complex, the use of decision trees remains popular, precisely because once presented it can be grasped almost immediately. If not else, decision trees are preferred because of their property to easily build rules, which, moreover, is an implication of the tree structure itself. E.g., let us assume the prediction of a highly probable demand slump on the part of a certain client population concerning a certain class of products. In order to decide, whether these clients are indispensable and thus the enterprise should conduct an expensive marketing intervening, or, a cheaper intervening is acceptable, the decision maker has to test diverse cost models on the current decision tree. Since, otherwise the expenses would exceed the revenues. Further sample applications of decision trees to business issues, besides investigation/exploration and estimation/prediction in general, are concretely the prediction of loan default, the crude oil price, or the exchange rate of currencies [8].

- *Rule-based Methods/Rule Induction*: this method aims at finding rules of interest and value for an enterprise's business, to put it another way, this approach might extract yet undetected dependencies, properties or correlations in parts of an information inventory. E.g., the inventory might provide the information, that if a television is purchased then a DVD player is purchased in 7 out of 10 cases. Rule induction exposes all possible/predictive patterns that can be expressed on the basis of the current information inventory. Consequently, it is inevitable that the presentation of all correlations found by this method exceeds the decision makers capacity to penetrate its response. Rather, the response of rule induction provides an overwhelming collection of isolated alternatives that, furthermore, may even be based on different questions/problems and, in the end, rather appear more to be a collection of mainly disjointed opinions than problem-specific solution suggestions. Basically, rules expose what functional operation holds between information (or sets of it such as classes/segments) captured in the information inventory. The accuracy of a rule is a criterion for its reliability meaning how often the rule turns out to be valid, whereas, coverage is a criterion for how often that rule applies to the current inventory.

- *Nearest Neighbor Classification*: this method exploits the concept of a metric in order to decide whether an information is nearer to a class than to another one. E.g., from a business perspective, a notebook is nearer to a desktop than to a cell phone, because they are alike concerning their resources and capabilities. Often, the decision maker can draw on a variety of conceivable metrics that all may make sense in one way or the other. Given an information i (or a set of it) and given classes $\mathfrak{C}_1, \ldots, \mathfrak{C}_h$ ($h \in \mathbb{N}$) the nearest neighbor method determines whether i should be considered belonging to class $\mathfrak{C}_1, \ldots, \mathfrak{C}_{h-1}$ or $\mathfrak{C}_h$. In the case of predicting stock prices where the enterprise essentially deals with time series, prediction means forecasting the next value of the stock price. Especially, the nearest neighbor approach can be used in early prediction of time series [39]. Further, it should be mentioned that the results obtained by a nearest neighbor approach depend very much on the selected distance measure.

- *Bayesian Classification*: the Bayesian classification method is a knowledge-based graphical representation that shows a set of information and their probabilistic relationships with each other. A less formal representation of that kind would be, e.g., a representation using system dynamics. System dynamics describes a method for depicting the mode of operation of a complex system over time.

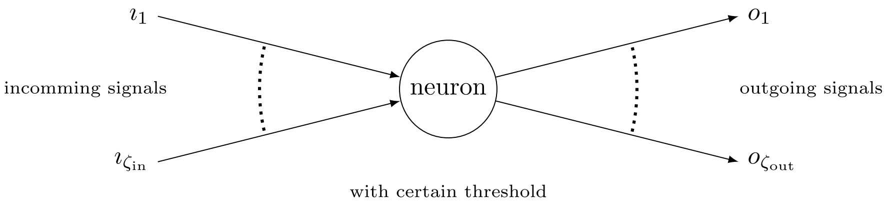


Fig. 4.1 An isolated artificial neuron

This method, especially, highlights how the single elements of the complex system effect each other. In other words, this approach highlights the relationships of the components of a dynamical system, meaning that, e.g., such relations might be circular, interlocking or even time-delayed. More precisely, system dynamics uses, inter alia, so-called internal feedback loops, stocks and flows in order to express the dynamical behavior of the entire (complex) system.

Certainly, the Bayesian classification method is based on conditional probabilities, i.e., the probability of an event given the occurrence of another event. Although, the Bayesian method is computationally nasty, nevertheless, can tender a potential benchmark for algorithms that may apply in classification. Bayesian classification provides, mainly, probabilistic prediction, i.e., it has the potential to predict multiple hypotheses that are weighted by their probabilities.

• *Support Vector Machines (SVM)*: a *SVM* classifies a set of information in such a way in categories/classes that around the boundaries of these categories/classes one obtains a strip as wide as possible that is free of current information, meaning, a strip which contains no data points. By this means, a *SVM* is, essentially, a large margin classifier. *SVM* apply to classification as well as regression, and, indeed, it is a mathematical technique for pattern recognition. E.g., two classes with their corresponding information/ data, at best, can be separated by a hyperplane, i.e., in the 2-dimensional case this reduces to a straight line. Albeit, at worst, there might not exist a separating hyperplane with the result that two classes can be separated properly. In this particular case, other curved surfaces should be taken into account in order to achieve the desired separation. This can be done by using what is called the kernel trick which means implicitly mapping their inputs into high(er)-dimensional feature spaces. In the case of linear classification *SVM* can be considered as special cases of Tikhonov regularization, the most commonly used method of regularization of ill-posed problems, as well. *SVM* apply in web services like spam categorization in email accounts [12].

• *Neural Networks*: inspired by the biological neural network, e.g., in the human brain, a (artificial) neural network denotes a network of (artificial) neurons intending to mathematically emulate, although in a simplified manner, the mechanics of a biological neuron. In the biological case of a neuron one has to keep in mind that the neuron sends out a signal/information only if a default threshold is exceeded by incoming signals(s)/information(s) (Fig. 4.1).

Indeed, several groups of neurons (ordered in so-called layers) in interconnection to each other build up a neural network. In general, the neural networks expresses an

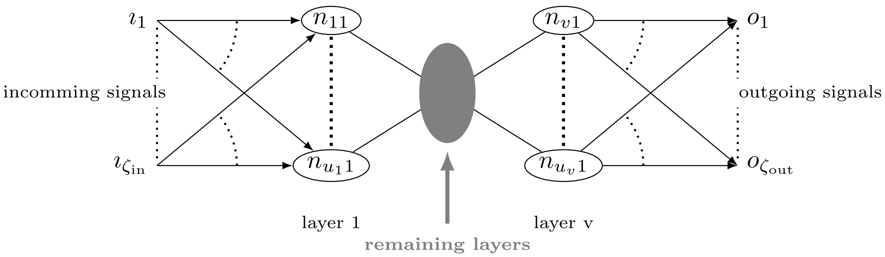


Fig. 4.2 An exemplary neural network

adaptive network that is capable of renewing and adjusting its connectivity, hence, transforming its very structure according to the information flow through the entire network. A neural network is a tool predestined to model functional relations that are inherent in the entirety of input and output, which, explicitly, includes, among others, pattern recognition, prediction as well as association.

Neural networks (Fig. 4.2) apply in a variety of additional business disciplines such as providing *Web Services/Information Retrieval* like intelligent search engines that are based on the clients's preferences, satisfying *Security* needs like making voice recognition systems available for authentication to client access, helping doctors to diagnose/analyse on the basis of afflictions and additional information like X-ray images, hereby, offering *Medical Decision Support*, assisting in *Portfolio Management* and *Optimization* as well as in predicting, e.g., future trends forming or the development of selected stocks or securities, assigning a clients credit rating as application in the field of *Finance & Banking*, right up to, *target recognition* or recognition of critical infrastructure in complex networks and energy grids as actual application in *Military Red Teaming*.

4.3 System of Systems: Challenges and Limitations within Business Intelligence

4.3.1 System of Systems

Whereas systems engineering focuses on designing an appropriate system, *System of System Engineering (SSE)* focuses on choosing the most appropriate system out of all available existing systems in order to satisfy the requirements. Mainly, an *SSE* aims at creating new capabilities through joining systems (may they be separate or sub systems) to a bigger structure, called *System of Systems (SoS)*. In doing so, the *SoS* becomes more than the sum of all its components, for, the *SoS* enables to master tasks none of the *SoS*-components (or even subsets of it) could cope with alone. An *SoS* should be regarded as meta system or, more abstract, as meta model.

Multidimensional data spaces have to be integrated in order to enable complex business analytics processes. Acknowledging modern *BI*-approaches in terms of an

SoS view supports the structured integration of existing (legacy) and additional systems as well as their data spaces within the respective business context.

We mention Maier's crucial characteristics [21] to distinguish a *System of Systems (SoS)* from other big and complex, but, however, monolithic systems:

1. *Operational Independence of Elements:* Separate components disjoint from the system itself can operate independent from each other, for, there is no such join between component and *SoS* that may restrict the functionality of the component or sets of it.
2. *Managerial Independence of Elements:* Separate components or sets of it can be acquired or developed independently.
3. *Evolutionary Development:* An *SoS* is a ever-changing structure whose capabilities, i.e., components are added, removed or modified over time.
4. *Emergent Behavior:* Through interaction of all components of an *SoS* new capabilities form and gain shape that can not be gained by the separate components independently without join to the encompassing *SoS*.
5. *Geographical Distribution of Elements:* The components of an *SoS* are geographically distributed so that the *SoS* covers sufficiently the sphere of its influence.

With reference to Sect. 4.1.1 the transition from inventory $\mathcal{I}$ to the filtered inventory ${}^{f}\mathcal{I}$ describes the compilation of a set of components ${}^{f}\mathcal{I}$ on the basis of the starting inventory $\mathcal{I}$ that is exactly the *SoS* at time $t = 0$.

Considering the development of integrated *BI* systems [16] present an incremental process model as a possible approach that includes a macro and a micro level view. The macro level determines the conceptional frame that includes decisions that are closely connected to the strategic view of the management [16]. The derived framework will have to be continuously audited and adapted to the dynamically changing business context. The micro level comprises the development and reengineering processes with respect to the single *BI* application systems within the integrated adaptive *BI* system. These processes are closely synchronized to the framework developed at the macro level [16].

When it comes to major unsolved problems in the IT industry, we have to mention the problem of managing semi-structured and unstructured data. Because of the difficulty of assessing unstructured/semi-structured data, companies usually do not incorporate these vast reservoirs of information into the decision making process. This ultimately leads to uninformed decision making.

4.4 Adaptive Business Intelligence as Integral Part of a Business Strategy

A *BIDS* as introduced above is a highly adaptive meta model, so that it is justified to speak of a IT-based decision support (system) in the realm of *Adaptive Business Intelligence (ABI)*, in short: *ABIDS*. An *ABIDS* not just encompasses all the advantages and abilities highlighted yet, it also includes optimization that is a very important feature we left unmentioned so far. Optimization in a business context can mean, e.g., maximizing the market penetration of a specific product line of a company, minimizing costs in terms of production, warehousing and/or transport etc.

An *BIDS* lacking in adaptivity implying flexibility in response to an eventually rapidly changing information inventory/business reality would be rather a burden than a helpful decision support tool. The adaptivity of a *BIDS* is a vital part of the applicability of the *BIDS* in real(-time) business.

An *integrated* adaptive *BI*-approach is needed in order to provide the basis for effective management decisions in the context of a dynamic and rapidly developing market environment. Traditional single business systems are often not able to deal with such a level of complexity. A modern *BI*-approach is needed that is able to efficiently integrate the different information spaces and that adapts to the dynamic decision environment.

4.5 Examples

4.5.1 Clinical Decision Support

"*Clinical decision support (CDS)* systems provide clinicians, staff, patients, and other individuals with knowledge and person-specific information, intelligently filtered and presented at appropriate times, to enhance health and health care." [7]

CDS enables providers and their patients to make the best decision based on the respective circumstances. By comparing a patient's electronic records/information with fixed clinical guidelines, an IT-based *CDS* system can, for instance, remind a provider to ensure that a patient receives, e.g., recommended immunizations, track a diabetic patient's blood sugar level over time or notify a provider that the medication that is about to be prescribed may lead to an undesirable side effect. Obviously, the ultimate goal of *CDS* is to supply the right information, to the right person, in the right format, through the right channel, at the right point in the clinical workflow to improve health and health care decisions and outcomes [38]. Using reminder systems, front office staff can be alerted to make sure that important lab work is done prior to the visit. Documentation of key elements of a patient's exam can be obtained before the provider even sees the patient. *CDS* can support disease management by tracking long-term issues that a given patient may need to have addressed for optimal health outcomes. Also, by using *CDS* with electronic prescribing, the selected drug can be checked against the patient's allergy list, against other drugs for possible interaction, for contraindication based on the patient's problem list [38].

"The ultimate objective of clinical decision support parallels the objective of providers themselves: Provide the best possible care for every patient. Health information technology holds a vast potential to help providers and their patients manage their overall health in the context of daily life. Modern quality improvement theory suggests that sustainable improvement happens when individuals or groups make a series of small, manageable changes over time. This is a logical approach to implementing health information technology and clinical decision support. One practice may choose to use electronic prescribing software as preparation for the ultimate leap to electronic health records. Another practice may use the information available in its practice management system to begin issuing preventive care reminders. However it happens, the important thing is that each practice acknowledges the ongoing need for improvement and takes action as a result." [38]

4.5.2 Decision Support in Airlines: Business Intelligence in Aviation Management

Before discussing a case study in the field of aviation management, one should understand the latest problems airlines constantly face. Generally speaking, airlines constantly have to deal with operational disruptions such as delays, cancellations and diversions, resulting in considerable inconvenience to passengers and costs to the airlines. For that specific reason the so-called airline operations control centers need decision-making processes to mitigate the effects of these disruptions. This crucial fact strengthens the necessity of business intelligence in the aviation management sector [9]. One example which can be found happened in the former Continental Airlines company, which did a company fusion with United Airlines in 2010 [37].

Supported by a newly designed data warehouse, Continental Airlines changed their business model. In details, the case study shows that Continental Airlines invested approximately 30 million US Dollar into real-time warehousing technique [2]. These investments were the core part of the business intelligence initiative. With the help of latest hardware, software and personnel training, the company was able to increase their revenues and costs savings by the factor ten. Real-time business intelligence started to become a significant business benefit. Especially the powerful real-time warehouse enabled Continental Airlines after the year 2000 to develop and deploy different application in the area of revenue management, customer relations, flight and ground operations, fraud detection and security, and others. Internal publicity helped a lot to preserve the excitement around the warehouse use and in addition encouraged business users to support warehouse expansion efforts. Finally it can be said that this example shows clearly the benefits of real-time data warehousing and business intelligence. Especially the quick data access seems to be a key for the support of current decision making and business processes, which can directly affect the company's actual situation [2].

4.5.3 IT-Based Complex Decision Support with System Dynamics: Strategic Management

Generally speaking, the types of decisions based on a project can be differentiated in three different types: strategic, tactical and operational. The actual use of system dynamics usually affects the strategic/tactical area of a project. In the context of business intelligence strategic project management can be described as follows: It includes the decision support in the project development phase, the support in making decisions concerning the project schedule with a long-term focus on the realization of these decisions [20].

In particular, the complexity involved in a project in most cases exceeds the human imagination and therefore requires a computer-aided modeling method, such as System Dynamics. One of the strengths of system dynamics is the representation of the interdependencies within a project and the subsequent tracking of changes in the model. It can be said that System Dynamics consists of one of the most developed plans for action, the optimal representation, analysis and detailed explanation of dynamics in complex technical systems as well as in entrepreneurial systems [32].

Especially large, long-term projects are now among the most important, while the least organized activities in the modern society. Large-scale projects are for example the construction of civilian equipment and infrastructure or military projects of all types (e.g.: construction of aircraft, development of weapons systems). Projects of all types typically experience additional costs, delays and quality problems also. Over several years Cooper and Mullen analyzed some major projects in different industries [11]. They reported that commercial projects are more expensive by about 140 % than planned and lasted longer about 190 % as originally scheduled. For military projects, his analysis reported that there were even 310 % additional costs and 460 % delay. Generally speaking, time delays, extra costs and quality problems, especially in connection with advanced technology are significant problems in the long-term planning and the design of a long-term project. Associated with these project-related problems, especially the regional economic situation and their ability to defend are affected dramatically.

In [33] Sterman provided a very good example of the successful use of system dynamics in a large defense project of the US Navy. In this particular case the firm of Ingalls Shipbuilding Pascagoula should build 30 newly developed destroyers for the US Navy in 1969. Mainly due to the fact that the US Navy caused Ingalls various problems, especially with the added cost of the project, they started with the help of system dynamics, to analyze the effects of time delays and the additional costs and to argue correctly about the extra costs to the US Navy. Especially the requirements of the US Navy rebuilding even more modern and effective weapons systems on the destroyers caused extreme costs for Ingalls.

Pugh-Roberts Associates of Cambridge developed a particular model, which included all phases of the project, simulated from the contract creation until the delivery of the ships, even with a forecast of five years. The main result was a system dynamics model with thousands of functions, which needed the latest computer technology of that time to complete. But it began as a much smaller model, which included the most important feedback effects of project delays and cost overruns. Of particular note at this point is the fact that at that time, the modeling team worked closely with all decision makers from various levels of management of the firm Ingalls together and created the first modeling designs [33].

Finding mistakes in the project and especially the extra work for the correction of errors, in both civil and military projects, can cause delays up to nine months. At this point it may be mentioned that changes or adjustments in the project by the customer do not have the same effect as the just mentioned detection and correction of errors in the project.

In general it can be said that the true value lies in the proactive use of system dynamics models. Additional costs and delays can be detected earlier. System Dynamics should be regarded as an additional method for decision support in project management to the existing, traditional project management methods. Especially when handling *complex project dynamics*, based on causal relationships, feedback loops, time delays and non-linearity System Dynamics can regarded as a potential method [32].

4.5.4 Portfolio Management – Example Energy Portfolio in Germany: Risk Management and Performance Optimization

When talking about energy portfolio, the term energy portfolio can have different meanings. On the one hand energy portfolio can stand for a combination of energy sources used for electricity generation and on the other hand energy portfolio can also mean a combination of either private or state energy investment assets. For this chapter only the first explanation of portfolio is relevant. Nevertheless, one of the core advantages of an energy portfolio concept is the following fact. It gives an opportunity to evaluate energy sources and technologies not separately but as a combination or collection of diversified asset. Originally a financial instrument, energy portfolio concept deals with risks and costs of energy supply. In the authors' point of view energy portfolio includes mainly the investment decision problem. This problem has to be further evaluated to manage risk and to maximize the performance of the energy portfolio concept.

There exist few basic approaches that aim to optimize energy portfolio of a certain country. Before speaking about energy portfolio concepts, the one of the main approaches has to be explained in details which is based on H. Markowitz's Modern Portfolio Theory [22]. One of the core issues of Markovitz's modern portfolio theory is the way to calculate cost and risk by diversifying them for achieving an efficient portfolio. Markovitz general idea for defining the optimal portfolio concepts is not only to include the possible profit, he also includes possible risk. Nevertheless the Mean-Variance Portfolio based approach has therefore been criticized for being concentrated on production costs of electricity-generation technologies. Although not production costs but rather expected risks and returns usually serve as a basis for private investment decisions [23]. Especially in the modern energy economy in

Germany this crucial fact can change the amount of energy investment assets and furthermore every portfolio concept. Nevertheless this approach gives an opportunity to evaluate different portfolio concepts of energy sources used for electricity generation in a certain country (see, e.g., [5]). In many cases risk in energy portfolios is mostly associated with the volatility of fossil fuels prices. Therefore the already mentioned diversification is achieved by adding increased share of renewable energy. Generally speaking, Germany will increase the share of renewable energy sources in the energy portfolio dramatically in the next years. Nevertheless Germany has to invest besides renewable energy sources strongly in all kinds of fossil energy sources to stabilize the energy system while phasing out the nuclear power production. Which means that not only Germany's energy portfolio will be more diversified in the year 2025. Although some energy scenarios conclude that photovoltaic will remain uncompetitive until 2030, photovoltaic can be still regarded as a relevant policy option for Germany's energy portfolio in the year 2025. Especially the rapid photovoltaic market growth generates cost reductions in the near future. In addition, photovoltaic can help to limit sudden energy price shocks. Furthermore, photovoltaic and wind energy reduce risk from fossil-fuel dependence (see, e.g., [1]). Similar to Germany, China has also started to adapt their actual energy portfolio concept and is willing to diversify the energy portfolio for a more efficient energy future [40].

4.6 Outlook and Perspectives: IT-Based Decision Support and Critical Infrastructures

Future society depends decisively on the availability of infrastructures such as energy, telecommunication, transportation, banking and finance, health care and governmental and public administration. Even selective disruption of one of these infrastructures may result in disruptions of governmental, industrial or public functions. Vulnerability of infrastructures therefore offers spectacular leverage for natural disasters as well as criminal actions. Threats and risks are part of the technological, economical, and societal development. Increasing complexity of our critical infrastructures exacerbates consequences of natural and/or man-made disasters.

Not only primary effects but also cascading effects as a result of increasing dependencies and interdependencies of our technological and societal systems demand intelligent simulation and optimization techniques in the area of IT-based decision support and computer-based information systems for a comprehensive safety and security management. At the end of this contribution we might mention that business intelligence has to consider also this new *external* dimension.

Therefore, one key element to estimate, analyze and simulate these special aspects within complex supply networks is *a new kind of business intelligence*: Innovative methods like computational intelligence, evolutionary algorithms, system dynamics and data farming should be combined within modern heuristics to master such complex networks via modern soft computing approaches.

This contribution summarizes some of these actual and future decision support approaches in the area of general IT-based decision support to design, process and manage complex systems.

4.7 IRIS Intelligent Reachback Information System – Smart Control Towers

In order to analyze such complex adaptive systems in the future agent-based modeling and simulation might be appropriate. Agent-based modeling and simulation as part of innovative business intelligence methods are an approach for modeling real world systems that are of complex and adaptive nature, such as an adaptive supply chain network. They enable to design each single actor of a supply chain network individually, based on their own decision rules. Moreover, it allows simulating the aggregate behavior of these heterogeneous organizations. At this way, emergent, non-linear behavior can be captured. To adapt as flexible as possible to *unexpected changes*, information along a supply chain network have to be visible. Yet, achieving visibility in a supply chain network still remains a problem. A new approach to overcome this difficulty represents the concept of a supply chain control tower which should be embedded in the global reachback concept as part of a comprehensive business intelligent approach in the future via a special service-orientated approach. This is the center of the research project IRIS (Intelligent Reachback Information System) at COMTESSA where business intelligence is understood as flexible IT-based decision support which is based on intelligent services concerning design, processing and management of an holistic business strategy.

References

1. Albrecht, J.: The future role of photovoltaics: a learning curve versus portfolio perspective. Energy Policy **35**, 2296–2304 (2007)
2. Anderson-Lehman, R., Watson, H., Wixom, B., Hoffer, J.: Continental airlines flies high with real-time business intelligence. MIS Q. Exec. **3**(4), 163–176 (2004)
3. Arnott, D., Pervan, G.: Eight key issues for the decision support systems discipline. Decis. Support Syst. **44**(3), 657–672 (2008). doi:10.1016/j.dss.2007.09.003
4. Arnth-Jensen, N.: Applied Data Mining for Business Intelligence. Kongens Lyngby (2006)
5. Awerbuch, S., Jansen, J., Beurskens, L.: Portfolio-Based Electricity Generation Planning: The Role of Renewables in Enhancing Energy Diversity and Security in Tunisia. United Nations Environment Programme (2005)
6. Azevedo, A., Santos, M.: Business intelligence: state of the art, trends, and open issues. In: Proceedings of the First International Conference on Knowledge Management and Information Sharing, KMIS 2009, pp. 296–300 (2009)
7. Berner, E.: Clinical Decision Support Systems: State of the Art. AHRQ Publication No. 09-0069-EF. Agency for Healthcare Research and Quality, Rockville (2006)
8. Berson, A., Smith, S., Thearling, K.: Building Data Mining Applications for CRM. McGraw-Hill, New York (1999)

9. Bruce, P.: Decision-making in airline operations: the importance of identifying decision considerations. Int. J. Aviation Manag. **1**(1,2), 89–104 (2011)
10. Clark, T., Jones, M., Armstrong, C.: The dynamic structure of management support systems: theory development, research, focus, and direction. Manag. Inf. Syst. Q. **31**(3), 579–615 (2007)
11. Cooper, K., Mullen, T.: Swords and plowshares: the rework cycles of defense and commercial software development projects. Am. Program. **6**(5), 41–51 (1993)
12. Drucker, H., Wu, D., Vapnik, V.: Support vector machines for spam categorization. IEEE Trans. Neural Netw. **10**(5), 1048–1054 (1999)
13. Grabmeier, J., Rudolph, A.: Techniques of cluster algorithms in data mining. Data Min. Knowl. Discov. **6**, 303–360 (2002)
14. Hannula, M., Pirttimäki, V.: Business intelligence empirical study on the top 50 finish companies. J. Am. Acad. Bus. **2**(2), 593–599 (2003)
15. Hoffman, T.: 9 hottest skills for ’09. Comput. World **1**(1), 26–27 (2009)
16. Kemper, H.G., Mehanna, W., Unger, C.: Business Intelligence – Grundlagen und praktische Anwendungen, 2. Aufl. Vieweg, Wiesbaden (2006)
17. Khan, R., Quadri, S.: Business intelligence: an integrated approach. Bus. Intell. J. **5**(1), 64–70 (2012)
18. Kudyba, S., Hoptroff, R.: Data Mining and Business Intelligence: A Guide to Productivity. Idea Group Publishing, Hershey (2001)
19. Lunh, H.: A business intelligence system. IBM J. Res. Dev. **2**(4), 314–319 (1958). doi:10.1147/rd.24.0314
20. Lyneis, J., Cooper, K., Els, S.: Strategic management of complex projects: a case study using system dynamics. Syst. Dyn. Rev. **17**, 237–260 (2001)
21. Maier, M.: Architecting principles for system of systems. Syst. Eng. **1**(4), 267–284 (1998)
22. Markowitz, H.: Portfolio selection. J. Finance **7**(1), 77–91 (1952)
23. Markowitz, H.: Portfolio Selection: Efficient Diversification of Investments. Wiley, New York (1959)
24. Michalewicz, Z., Schmidt, M., Michalewicz, M., Chiriac, C.: Adaptive Business Intelligence. Springer, Berlin (2007)
25. Moss, L., Shaku, A.: Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications. Pearson Education, Upper Saddle River (2003)
26. Negash, S.: Business intelligence. Commun. Assoc. Inf. Syst. **13**(1), 177–195 (2004)
27. Nemati, H., Steiger, D., Iyer, L., Herschel, R.: Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. Decis. Support Syst. **33**(1), 143–161 (2002). doi:10.1016/S0167-9236(01)00141-5
28. Raisinghani, M.: Business Intelligence in the Digital Economy: Opportunities, Limitations and Risks. Idea Group Publishing, Hershey (2004)
29. Richardson, J., Schlegel, K., Hostmann, B.: Magic quadrant for business intelligence platforms. Core research note: G00163529, Gartner (2009)
30. Richardson, J., Schlegel, K., Hostmann, B., McMurchy, N.: Magic quadrant for business intelligence platforms. Core research note: G00154227, Gartner (2008)
31. Shim, J., Warkentin, M., Courtney, J., Power, D., Sharda, R., Carlsson, C.: Past, present, and future of decision support technology. Decis. Support Syst. **32**(1), 111–126 (2002). doi:10.1016/S0167-9236(01)00139-7
32. Sterman, J.: System dynamics modeling for project management (1992). http://web.mit.edu/jsterman/www/SDG/project.pdf, visited 11.08.2011
33. Sterman, J.: Business Dynamics – Systems Thinking and Modeling for a Complex World. McGraw-Hill, New York (2000)
34. Thierauf, R.: Effective Business Intelligence Systems. Quorum Books, West Port (2001)
35. Thomsen, E.: BI’s promised land. Intell. Enterprise **6**(4), 21–25 (2003)
36. Turban, E., Sharda, R., Aroson, J., King, D.: Business Intelligence: A Managerial Approach. Pearson, Upper Sadle River (2008)

37. www.airliners.de (2012). Continental und United unter einem Dach. http://www.airliners.de/management/strategie/continental-und-united-unter-einem-dach/22288, visited September 2012
38. www.pcpcc.net (2010). Clinical decision support in the medical home – an overview. http://www.pcpcc.net/files/clinical-decision.pdf, visited September 2012
39. Xing, Z., Pei, J., Yu, P.S.: Early prediction on time series: a nearest neighbor approach. In: Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09), pp. 1297–1302 (2009)
40. Zhu, L., Fan, Y.: Optimization of china's generating portfolio and policy implications based on portfolio theory. Energy **35**, 1391–1402 (2010)

Chapter 5
How to Introduce KPIs and Scorecards in IT Management

Martin Kütz

Abstract IT management is based on continuously updated facts and uses scorecards and KPIs. However, the real life in IT organizations is different. On one hand data assets are wasted in traditional reports which are ignored by IT management, on the other hand scorecard introductions fail despite of powerful technologies. This chapter starts with a description of how management and scorecards are linked up against the background of control cycles and the principal agent model. It continues with the discussion of processes to build, use and continually improve scorecards. Finally it shows that the establishment of a scorecard based IT management will be successful only if appropriate processes are implemented, management behaviour is changed towards a culture of measurement and the usage of KPIs is driven by top management representatives.

5.1 Introduction

5.1.1 Preliminary Considerations

Transparency depends on measurement and measurability. It is generally accepted, that KPI's (Key Performance Indicators) and scorecards can help IT managers to better plan and control their area of responsibility. A lot of consulting is done in performance management and a variety of tools can be used (management cockpits, dashboards, etc.). But the results are all too often disappointing as the author has experienced in many consulting projects and discussions with IT managers and IT performance analysis experts. Many IT scorecards are technically installed but neither accepted nor used by the responsible managers. It has turned out, that just providing a scorecard technically or building a KPI warehouse and providing a manager friendly user interface is not enough.

The foundation of a successful scorecard based IT management is an appropriate management culture. If the scorecard and its KPIs are not integrated into the

M. Kütz (✉)
Fachbereich Informatik und Sprachen, Hochschule Anhalt, Lohmannstr. 23, 06366 Köthen, Germany
e-mail: martin.kuetz@inf.hs-anhalt.de

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*, Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_5, © Springer-Verlag London 2013

management process and not considered by the managers to be their most important management tool, there will be costs, but no benefits. Finally, it is the manager's person, who makes the success or failure.

5.1.2 Approach

This chapter starts with a definition of KPIs and scorecards and how they are connected to management control cycles. It then describes the processes, which must be established if an IT organisation wants to base its performance management onto scorecards. It also presents an approach, which has been repeatedly successful in supporting IT scorecard introductions and has helped to successfully integrate IT scorecards into managerial work. Finally the major success factors for introducing scorecards are pointed out.

5.2 Basics and Definitions

5.2.1 Control Objects

Each manager has to ensure that the system which has been resigned to him/her can work and realize its objectives. Because in IT a system is predominantly considered to be a technical system consisting of hardware and software components subsequently in this chapter the term "control object" will be used. Typical control objects in IT are services, projects, processes, functions, technical systems and subsystems, organisational units, etc. With any category of management there is always a related control object.

5.2.2 Indicators

An indicator is a metric which quantifies a specific feature of a control object or its surroundings. Sometimes it can be taken from measurement data directly, but in most cases it is built from several measurement data, which are combined by a well defined calculation rule. The measurement data are taken from the control object or its surroundings at a specific point of time or for a specific period of time. Thus the (numerical) values of an indicator may change over time and each value refers to a specific point or period of time. If one writes down the indicator values in timely order one gets a time series.

However, there is not only one single value of an indicator for each point of time. Really an indicator is represented by a set of values [4], namely:

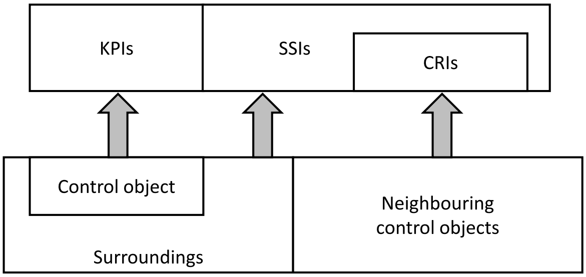


Fig. 5.1 Categories of indicators

- *An actual value*
  The actual value is generated from the actually measured data by the pregiven calculation rule.
- *A planned or expected value*
  The planned or expected value is given and is the result of the corresponding planning process for the specific control object. Within a planning period there will be several checkpoints and the planning process has to deliver the planned values for the total group of checkpoints. Thus the planned or expected indicator value may change over time. Ideally the actual value and the planned value for a specific checkpoint are identical.
- *Prescribed boundaries*
  The prescribed boundaries are lower and upper values for the considered indicator. A control object may have several prescribed boundaries, which show different levels of deviation from the planned or expected values. If the actual value of an indicator is below a lower boundary or exceeds an upper boundary then the deviation from the planned or expected value has reached a specific level of significance and management has to react accordingly. The prescribed boundaries may be time dependent.
- *A forecast value*
  A forecast value is generated from the time series of actual values of the indicator. It estimates the actual value for a future point of time, normally for the next checkpoint or the end of the planning period. Compared to planned or expected values the forecast values are not based on objectives but on actual values.

There are three categories of indicators [4] (see Fig. 5.1):

- *Key performance indicators (KPI)*
  KPIs are metrics of the control object corresponding to the most important features of the control object. Management aims at reaching the planned values of the KPIs.
- *Surroundings status indicators (SSI)*
  SSIs deliver information about the surroundings of the control object. They help management to act and react correctly. This information is pregiven for the management and cannot specifically be influenced. If there are deviations between actual and expected values management knows, that the surrounding of the control object or the control object itself has changed.

- *Cross reference indicators (CRI)*
  CRIs are a subset of SSIs and deliver information about other control objects in the neighborhood of the control object to be considered.

As an example of those categories of indicators a service desk can be considered. The first-line resolution rate is a KPI because the defined level is part of the service desk manager's target agreement. The number of incoming incidents or service requests is a SSI, because the service desk manager cannot actively increase or decrease this number. And the number of changes or incoming change requests is a typical CRI, because it indicates subsequent incidents or service requests.

5.2.3 Scorecards

A scorecard is a structured set of indicators. It must always contain a set of KPIs. It will normally contain a set of SSIs and it may contain one or more CRIs. It is a model of the control object and documents all information which is relevant for the management of that control object. It is a control panel for the responsible manager.

Each scorecard has a specific structure. It may have different segments with respect to different aspects or feature groups of the control object as it is typical for a Balanced Scorecard [2]. Usually those segments are more or less based on stakeholder groups:

- *Customers*
  These are those external persons or organisations or other internal control objects, who take over the output of the considered control object.
- *Owners*
  The owners provide funds and resources to enable that the control object can work. They expect a specific Return on Investment (ROI). The financial segment of the scorecard is related to the interests of the owners.
- *Managers*
  Managers are responsible for creating the ROI which is expected by the owners. They have to care for the operations of the control object and this is considered in the process segment of the scorecard.
- *Employees*
  Creativity and innovation comes from human beings. They are the mould to make an organisation succeed. This, of course, includes the managers. As far as that goes each manager has two roles: On the one hand he/she manages (a part of) the organisation, and on the other hand he/she contributes to the organisation. The corresponding scorecard segment is often denominated with "Learning and development".

There are also segmentations oriented towards critical success factors or organisational subunits. Speaking abstractly a scorecard is segmented by different control areas. A typical structure of a scorecard for an IT organisation might have these segments:

- Financial Management
- Customers & Services
- Process management
- Project management
- Supplier management
- Personnel management
- Innovation Management
- GRC management (GRC = Governance, Risk and Compliance)

Financial management, customers & services and process management refer directly to the usual BSC labels. Project, supplier and GRC management represent IT specific views. And the learning and innovation area has been split up into personnel and innovation management.

Scorecards may also have different segments according to various periodicities among the selected KPIs.

5.2.4 Management Control Cycles

Management is done in control cycles. Basis for the subsequent considerations is the so-called PDCA or Deming cycle, which was introduced in the early days of modern quality management. The acronym PDCA stands for Plan-Do-Check-Act. It represents a 4-phase model [3].

In the planning phase (P = Plan) the objectives of the control object are fixed and the conversion strategy is worked out. In the conversion phase (D = Do) the organisation works on reaching the planned objectives. In the deviation analysis phase (C = Check) it is investigated, how good or bad the performance has been in the considered interval of time. In the final adjustment phase (A = Act) corrective action is initiated and taken to eliminate or reduce detected deviations between planned or expected and actual status.

In complex organisations each manager reports to a superior manager. The superior manager appoints inferior managers to do specific jobs for him, which he cannot do by himself for different reasons. In this relation between different management levels the superior manager is called the principal and the inferior manager is called the agent. The principal is not sure that the agent really does what he should do. The principal agent model [1] shows what are the problems and challenges in this relationship and what the superior manager can do to master the agent's egoism. On that background the management control cycle should be extended by two phases, namely a contract phase (C = Contract) and a communication phase (C = Communication), leading to a new acronym CPDCCA (see Fig. 5.2):

- *Contract phase*
  The contract phase opens the extended management control cycle. Here the objectives are agreed between principal and agent or simply pregiven by the principal. The agent is the manager being responsible for the control object. The principal is an upper manager or owner to whom the agent reports.

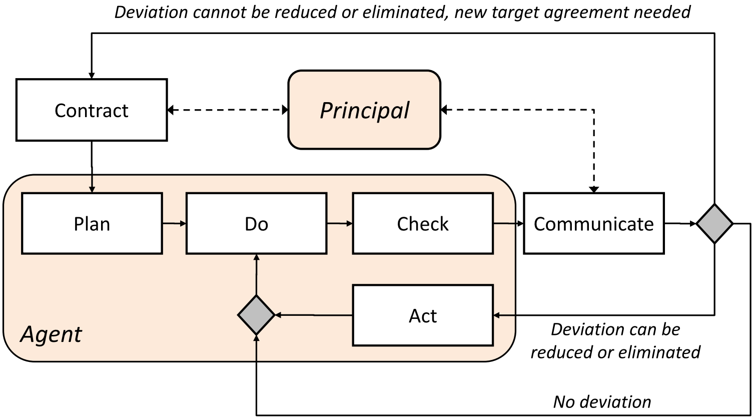

Deviation cannot be reduced or eliminated, new target agreement needed
Contract
Principal
Plan
Do
Check
Communicate
Act
Agent
Deviation can be reduced or eliminated
No deviation

Fig. 5.2 Extended management control cycle (CPDCCA)

- *Communication phase*
  When the deviation analysis has been closed and the necessary actions are defined, then the agent has to inform the principal. The principal has to agree to the defined actions. It may happen, that the formerly defined objectives cannot be reached or not be reached within the remaining period of time. In that case the process must switch to the contract phase and principal and agent have to fix new objective (or stop the activity).

5.2.5 Control Cycles and Scorecards

It is assumed, that a scorecard describes a specific control object exhaustively. Then it is clear immediately, how a scorecard can support the management process. With the vector of actual KPI values the actual position or status of the control object is identified; the actual SSI and CRI values show the status of the relevant surroundings and adjacent control objects. With the vector of planned KPI values the planned position of the control object is described for the considered point of time; similarly the expected SSI values and planned CRI Values describe the "ideal" status of the surroundings. The difference of both vectors shows the extent and the direction of the deviation.

The scorecard attends the whole management process (see Fig. 5.2):

- In the contract phase ("Contract") objectives are agreed between principal and agent. The final values of the defined KPIs as well as the expected status or the surroundings are discussed an agreed.
- In the planning phase ("Plan") the planned values of the KPIs for the periodic checkpoints are derived according to the conversion strategy (the chosen way to achieve the agreed targets).

- In the conversion phase (“Do”) the necessary monitoring and measurement is conducted to get the actual indicator values for each checkpoint.
- In the deviation analysis phase (“Check”) the vectors of actual indicator values and planned respectively expected indicator values are compared and the differences are analyzed.
- In the subsequent correction phase (“Act”) which ends in the conversion phase monitoring and measurement are conducted again.
- In the communication phase (“Communicate”) all figures are reported to the principal combined with results of the analysis and the corresponding recommendation for the further operation.

Thus the scorecard is the vehicle for quantifying objectives, planning and deviation analysis. A scorecard increases management maturity, because it forces all involved persons to a fact based management. People discuss about facts which are based on measurements. They are not longer talking about sentiments.

However there is one big risk when controlling with scorecards; it is outlined by the subsequent questions:

- Is the scorecard appropriate for the specific management duty?
- Is the model of the control object a good model?
- Do the selected indicators consider all aspects which are relevant?
- Do all involved parties understand and agree to the defined metrics?

The answers to these questions have to be found out by the involved people at working with the scorecard.

5.2.6 Scorecard Networks

IT departments are complex organisations and are composed of a variety of control objects: departments, services, processes, functions, systems, projects, etc. All control objects build a network with many relationships and dependencies between those control objects. Each single control object, each subset of control objects and of course the total IT organisation can be controlled with the help of a scorecard. As there are interrelationships between the control objects there are interrelationships between the scorecards, too.

Traditionally a pyramid of KPIs was built (see the Du Pont system of financial control, [1]) and the dependencies between different KPIs were modelled by mathematic formulas. However this mechanical approach has not been successful for different reasons. The main reasons of the failure of that approach are, that on one hand it is nearly impossible to transfer it to the approach of multidimensional control as it is represented by Balanced Scorecard and on the other hand a network of mathematical formulas is somehow tautologic. This leads to an oversimplification of control.

As each scorecard itself represents a strongly simplified and stand-alone model of the specific control object the set of all scorecards will cover the total IT organisation only like a patchwork blanket with a lot of gaps. A method is needed to tie

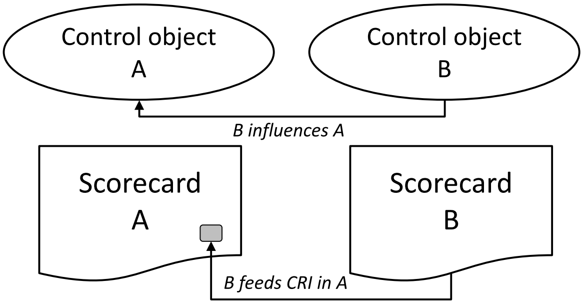


Fig. 5.3 Usage of cross reference indicators (example)

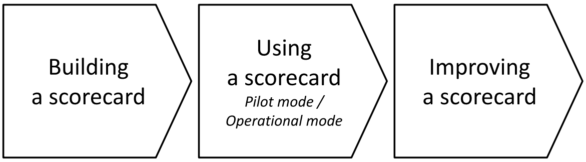


Fig. 5.4 The scorecard process

these scorecards together. This can successfully be done by loose coupling and this coupling is realized with the CRIs.

If there are two control objects A and B, where A is influenced by B, then the manager of A has to know if B deviates from its plan. To inform the manager of A about such a deviation one or more CRIs are added to the scorecard of A which are fed from the deviations occurring at B (see Fig. 5.3). The advantage of that approach is that each scorecard can be modelled specifically for the corresponding control object. The CRIs can be added step by step and later. Thus the scorecard network can be built and improved iteratively as managers identify the interrelationships between the different control objects.

5.3 Scorecard Processes

5.3.1 Preliminary Considerations

There is a lot of experience in introducing Balanced Scorecards. This can be used for any scorecard. The real challenge is not the design of the scorecard, but its integration into the management processes. A three staged process has been helpful in several scorecard introduction projects (see Fig. 5.4).

5.3.2 Building a Scorecard

To set up an IT scorecard as a management tool the following steps are proposed:

- Identify the control object.
- Identify the responsible persons and stakeholders.
- Define and document the mission for the control object and agree upon it. (What is the reason for its existence?)
- Identify the surroundings and the constraints and agree upon it.
- Define the objectives and agree upon it.
- Identify control areas for the scorecard segmentation and corresponding main features of the control object.
- Identify corresponding indicators, describe the calculation rules and define the measurements.
- Define the supporting processes, which are necessary to generate and provide the actual indicator values for the scorecard.
- Identify the management processes, which are affected by the scorecard and have to be adopted to the new control system:
  - Extend the planning processes to let them deliver the planned and expected values of the indicators.
  - Extend and change the management communication so that the scorecard becomes the basic vehicle of that communication.
  - Change the reporting systems accordingly.
  - Modify the management meeting culture and organisation to adopt the scorecards.
- Define or change the affected operational processes. (Change or include measurements of necessary data.)
- Define the indicator database.

Ideally a central data warehouse should be used to store time series of measurement data and indicators. It is often more efficient not to look for the best new solution but to take the already implemented second-best solution. There is a big advantage if existing platforms can be used. Those systems are already up and running. There is no time needed to select and implement such a system. Also the operation and administration of the systems from access management to archiving is already established. The organisation has people who understand the technology and know how to work with the system. The installed system can be used tomorrow and one is not forced to wait weeks or months until the technical infrastructure is available.

There are some mature organisations which have decided, that all management information must be delivered from one central source. Only data from that source are authorized to be presented to top management. The reason behind such a strategy is quality. Data must be quality assured and if there is a central platform for all management information, a specific and common level of quality can be assured.

- Identify systems or tools.
  Here are the same arguments valid as for the database. It is strongly recommended to use already implemented and used systems or tools. IT management should not have the ambition to find the ideal solution. No system or tool will be perfect. But the more parts of an organisation use it the higher will be the benefits from the system or tool.

- Define roles and responsibilities:
  - Scorecard owner:
    This is the manager (agent) who is responsible for managing a specific control object. This person is the indicator owner, too, for all indicators being element of the scorecard. He/she has to ensure that scorecard is fit for purpose.
  - KPI manager(s):
    These are persons in the manager's (agent's) organisation who are responsible for the operational management of the KPI. That means that those persons are the primary addressees for the scorecard owners questions. They have to watch the KPI, analyse the actual KPI value respectively its change or development and prepare first recommendations for further action.

    This role seems to be artificial. But it has been shown for several times that it is an appropriate vehicle to get the whole management team involved and stimulate communication and cooperation.

    If those persons report to the scorecards owner, then they have a principal agent relationship with the scorecard owner and there will be a scorecard by which they manage a control object on the next lower hierarchy level. There may be a strong relationship between the "managed" KPI in the upper level scorecard and the "owned" KPIs in the lower level scorecard, but this is not a Must.
  - Scorecard addressee:
    This is the principal to whom the scorecard owner is reporting and who is the person finally defining the objectives which have to be reached. The discussion between principal and agent must be completely based on the scorecard. Otherwise all people around will conclude that the scorecard is unimportant and not "the" management tool.

5.3.3 Using a Scorecard

A scorecard will be used in two modes, first and for a limited time span the piloting mode and afterwards and permanently the operational mode. In the piloting mode changes can be implemented fast and easily. Learnings and experience can be transformed into improvements of the scorecard. Changes are normal. In the following operational mode the scorecard is an element of continuity and steadiness. Changes are exceptions.

5.3.4 Piloting a Scorecard

A scorecard should become mature through its use. Thus the concept phase should be as short as possible and then the management team should start to actively use the scorecard in daily operation. Most and best ideas for completing and improving the scorecard as well as its related processes and activities will raise in day-to-day's work. In contrast to reducing the concept phase as much as possible the pilot phase should have plenty of time. This time is needed because management has to learn to work with scorecards.

Scorecards will change managerial work seriously. Actual values of KPIs are documented facts. What people say must be consistent to those facts. Deviations from targets will be detected early and the discussions within the management team will become strongly focussed on the achievement of the measurable targets. Reasonable actions will be defined and responsibilities can be clearly defined. There will be a higher level of commitment in the management team.

The pilot phase should start even if not all indicators are well defined or not all measurement data can be provided. According to experience the practical work with the scorecard will generate ideas to get better data or get the data easier or faster. In many organisations managers are sure to a high degree, which indicators might be helpful for them. But they have to find out, which indicators really help them to improve their management performance. At the beginning they also do not have an idea what are the "right" planned or expected indicator values. They will find it out by working seriously with the scorecard.

It is important that even in the pilot phase management has to work with the scorecards as if it were fully in operation. Otherwise the pilot phase will be a fake and will not lead to any progress. It will rather let all stakeholders ignore the scorecard and thus finish with the result that the scorecard is not useful at all.

5.3.5 Regular Operation with the Scorecard

To run a scorecard in an operational mode the following chain of subprocesses is recommended:

- *Conduct measurements*

  Indicators are built from measurement data. Each measurement has to be described technically and organisationally. Data must be taken periodically at defined points of time. The measurement processes should run automatically. However, from an economic point of view it sometimes makes sense to do measurements manually, e.g. measure the process duration for processes with as low number of realisations.

  The quality of measured data must be ensured. Measurements must be integrated into regular monitoring and event management. Control objects must be changed if necessary to enable measurement. One cannot measure process duration if at the beginning and the end of a process time stamps are not taken.
- *Run ETL process*

  Data from different measurements must be taken together (E = Extraction) and prepared for storage in the database (T = Transformation). Before being loaded (L = Loading) into the database there must be an appropriate clearing and preparation of data. A careful quality assurance must be conducted. The ETL process should have a high degree of automization but may stay a manual process if this is the most economic approach.

- *Generate Indicators*

  When the database is up-to-date then the defined indicators have to be calculated from the actual time series. This step should be totally automated to avoid errors. The indicator values themselves should also be stored in time series. This leads to some kind of redundancy but eases reporting and data analysis.

  Usually additional values are calculated which refer to the underlying indicators:
  - moving averages
  - difference of actual and planned value
  - difference of actual value and boundary if boundary is exceeded
- *Provide scorecard*

  After the indicator generation has been finished, the results can be grouped into scorecards and then presented and forwarded to the addressees. Traditionally this was done by paper reports. Nowadays those paper reports are replaced by electronic media:
  - management cockpits
  - management dashboards
  - digitised paper reports

  The addresses are performance analysis experts (members of controlling service unit) and KPI managers.
- *Enrich scorecard with first analysis*

  Experts and KPI managers have to analyse data and document results and first recommendations. Experts from controlling service units must coordinate results and complete scorecards accordingly.
- *Communicate scorecard with added comments to addressees*

  When this first analysis is done the scorecard is provided to the scorecard owners.
- *Discuss scorecard and deduce actions and recommendations*

  Establish management meetings or use already existing management meetings to work out the final interpretation of the KPIs. The results of those meetings have to be minuted. This shall be done based on the 3W-method: Who is doing what until when?
- *Document and communicate results*

  Finally the manager (agent) will communicate the results to his principal. This might lead to new or changed results and actions and in extreme cases lead to new target agreements.

5.3.6 Improving a Scorecard

Business is permanently changing, often smoothly and continuously and sometimes extremely and singularly. According to that control objects will change and subsequently the scorecards must be adapted. Also experience shows which indicators are successful and helpful and which indicators are obsolete or not fit for purpose. On that background each scorecard should regularly be reviewed and improved, if possible. An annual scorecard review should be conducted. This can be integrated into the annual planning cycle, where all planned or expected values of indicators have to be actualised.

5.4 Success Factors

Scorecards are one of the most powerful management tools we have. The introduction often leads to a cultural revolution in (IT) management, because they make things transparent and force all stakeholders to discuss about facts and figures, not about opinions or sentiments. But not all concerned people will like them. The success will not come automatically.

There are some success factors which help to introduce a scorecard based IT management:

- All management people must be involved actively from the beginning. Managers who are not involved will not be committed.
- There must be a clear sponsorship and promotion from the top IT management (If he or she is not clearly committed then there will be no sustainable success). Ideally promotion comes from the principal level. If the principal demands it then there will be enough power to get it.
- There must be clear responsibilities. Everybody in the organisation must know, who the scorecard owner is, who the KPI managers are and who is responsible for collecting the data and providing the indicators.
- The actual scorecard must be actively discussed in the management team, not only registered in a passive manner. There must be a dispute on the figures. This discussion must be a regular action and be a standard topic for the agenda of the periodic management meetings in the IT organisation internally as well as in the communication between the IT manager with his or her boss.
- All discussions must refer to the scorecard and its indicators. Any change in the control object should have an effect in the scorecard. If not, then there might be a need for improvement of the scorecard.
- The metrics must allow easy and quick generation of the scorecard. As a rule of thumb indicators should be available 2–3 days after closing the measurements.
- The quality of data must be assured from the beginning. All people being involved should be able to trust the data. If stakeholders do not trust the data they will not work with the scorecard.
- The value of the KPIs increases if it can be seen how the figures change over time: Is there a trend? Are there specific significant changes?
- The development of a scorecard should be done quickly. It should not last longer than six months, normally less. If the objectives are clear with regards to the content then the conceptual phase should take not more than 2–3 months.

- It is essential that working with the scorecards starts as soon as possible. All managers in an IT organisation must learn, how to work with a scorecards, what a scorecard can deliver. Thus one should start as early as possible even if only a few KPIs can really be generated.
- The pilot phase should have enough time so that all people involved know, how the scorecard helps them in their managerial work. If a scorecard is established for the management of a complete IT organisation the pilot phase will take one complete planning period, after all 12 months. This has been turned out to be necessary in several scorecard projects. This time span coincides with the duration of 2 years for BSC introductions, which is communicated by many people having done such projects. If the control object is smaller, then the pilot phase could be shortened [4].

5.5 Summary and Conclusions

Scorecards are an essential element of modern IT management. A successful implementation first of all needs participation of the complete management team and clear sponsorship of the highest involved management level. This initial sponsorship must be perpetuated until working with scorecard has become self evident in the organisation. The control objects of IT have to be prepared for scorecard management. Working with scorecards needs the establishment of specific processes for collecting and processing data and providing the scorecard. Scorecards will not only improve transparency in IT management, but also lead to a better understanding of IT objectives and higher planning quality.

The major success factors for successfully introducing a scorecard based IT management is really not a system or tool. It is on the one hand the manager who works with his/her KPIs and on the other hand a set of processes to process the data and ensure management communication and discussion.

However there are two additional success factors as everywhere in performance management:

- Start working with scorecards!
- And keep up with it!

References

1. Gladen, W.: Performance Measurement, 4th edn. Gabler, Wiesbaden (2008)
2. Kaplan, R.S., Norton, D.P.: Balanced Scorecard. Harvard Business Review Press, Boston (1996)
3. Kütz, M.: IT-Controlling für die Praxis. dpunkt, Heidelberg (2005)
4. Kütz, M.: Kennzahlen in der IT. dpunkt, Heidelberg (2011) (4th, revised and extended edition)

Part III
BI/PM Applications to Business Development

Chapter 6
Identifying Suspicious Activities in Company Networks Through Data Mining and Visualization

Dieter Landes, Florian Otto, Sven Schumann, and Frank Schlottke

Abstract Company data are a precious asset which need to be truly authentic and must not be disclosed to unauthorized parties. In this contribution, we report on ongoing work that aims at supporting human IT security experts by pinpointing significant alerts that really need closer inspection. We developed an experimental tool environment to support the analysis of IT infrastructure data with data mining methods. In particular, various clustering algorithms are used to differentiate normal behavior from activities that call for intervention through IT security experts. Before being subjected to clustering, data can be pre-processed in various ways. In particular, categorical values can be cleverly mapped to numerical values while preserving the semantics of the data as far as possible. Resulting clusters can be subjected to visual inspection using techniques such as parallel coordinates or pixel-based techniques, e.g. circle segments or recursive patterns.

Preliminary results indicate that clustering is well suited to structure monitoring data appropriately. Also, fairly large data volumes can be clustered effectively and efficiently. Currently, the main focus is on more elaborate visualization and classification techniques.

D. Landes (✉) · F. Otto
Coburg University of Applied Sciences and Arts, Coburg, Germany
e-mail: landes@hs-coburg.de

F. Otto
e-mail: florian.otto@hs-coburg.de

S. Schumann
HUK COBURG, Coburg, Germany
e-mail: Sven.Schumann@huk-coburg.de

F. Schlottke
Applied Security, Stockstadt, Germany
e-mail: frank.schlottke@apsec.de

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_6, © Springer-Verlag London 2013

6.1 Introduction

Nowadays information security is an essential factor for the long-term business success of any company. Simultaneously the complexity of IT infrastructures as well as the quantity and quality of attacks on these structures are growing. As a reaction, companies generally introduce new and more sophisticated security mechanisms, which gradually and significantly increase the complexity of the IT security architecture. Furthermore, the amount of data generated by the security mechanisms grew enormously. In contrast, human resources available for their interpretation and analysis remained essentially unchanged. As a result, the management of IT security architectures and in particular the timely detection and defence against attacks are rendered even more difficult.

Controlled allocation of access rights, shielding systems by firewalls, or intrusion protection systems all contribute to protecting data and IT systems. Yet supplementary mechanisms are required to detect potentially harmful activities. Such security-related activities can be initiated both from inside and from outside the company's network.

Thus, it makes sense for companies to optimise the continuous monitoring of the level of IT security by introducing uniform analysis procedures across different areas and systems. It is necessary to consolidate and analyse all security-related incident reports in a central place. It is highly important to process the information quickly and automatically. This is the only way that allows detecting, evaluating, and escalating attacks and situations in the IT environment which pose a threat to security quickly and with consistent quality.

In order to address this problem, a large number of usage data need to be logged and subjected to rule-based filtering in order to detect potential security problems. Apart from a relatively small number of incidents which are indeed significant, most of the identified events are factually permissible and supposedly harmless. Due to the large number of events found, a purely manual analysis by the company's security experts is unthinkable—it is necessary to support these experts through automatic pre-selection of events to allow them to focus on the incidents which are really significant and to adopt suitable countermeasures. A pre-selection based exclusively on static criteria is insufficient since the characteristics of attacks on IT infrastructures are continuously refined. Both new and modified patterns of attack need to be identified through a suitable anomaly detection system. Therefore, an approach for supporting security experts is required which is flexible enough to highlight incidents, which are yet unknown but may be relevant and pose a threat to security.

Each potential solution for these problems must deal with the conflicting areas of technology versus human resources. Based on the goal of achieving cost savings and the limited availability of human resources, a primarily technologically oriented approach to these problems should be selected.

Evidently the solution to this complex of problems cannot be found in deploying additional human resources. This would merely provide temporary relief and postpone the necessity of solving the problem.

The SecMine project aims at enabling the responsible persons in the companies to analyse data generated by various security mechanisms jointly and efficiently. To that end, an innovative decision-support system is needed for assessing current security threats in complex IT infrastructures. This system needs to exhibit intelligent behaviour, be capable of learning, and must support the identification of new types of attack through interaction with the experts.

Data mining seems to be a promising approach. Using largely automated processes, data mining tries to identify and make use of patterns in large data sets which would be impossible or very hard to discover with the naked eye. One category of data mining techniques consists of so-called clustering methods which arrange data in groups of similar data records.

A combination of data mining methods appears to be appropriate for the problem in question. Promising methods include techniques for offline clustering of hybrid, high-dimensional data, for visualising high-dimensional data with the aim of decision support and validation, and for classifying new data using the clusters identified.

The employed data mining and data visualisation methods are embedded in an experimental tool environment called Cluster Utility Framework (CUF) which allows the flexible combination and easy addition of data mining processes.

In our particular case, data are generated by a security information and event management (SIEM) system that supervises various sources such as, e.g., firewalls, intrusion protection systems, or login procedures in the network of an insurance company. Potentially security-related incidents are identified using heuristic rules and recorded as so-called events in a common format, irrespective of their original source. In this joint format, properties such as source and target IP addresses and port numbers of an event are recorded together with some 60 other aspects. Since the heuristic rules are fairly weak, thus giving rise to many false-positives, the SIEM system potentially generates several millions events per day that may be fed into CUF for further analysis.

In the next section, we will outline clustering methods in general, before we highlight some features of CUF. In particular, we discuss the general architecture, clustering techniques that we currently use in CUF, and data pre- and post-processing to facilitate analysis. Finally, we present the current status of our research in the SecMine project and further steps to take.

6.2 Data Clustering

Clustering, or cluster analysis, aims at discovering patterns in a potentially large amount of data. Each data object is characterized by a fixed set of attributes, or features. Attributes can be of various types, ranging from numeric through ordinal to categorical [14]. Numeric attributes are quantitative, measured in integer or real values, thus allowing basic calculations. Ordinal attributes can take values that can be ordered in a meaningful way, yet the magnitude between successive values cannot be determined. In categorical data, there is not even such an ordering of values, let alone meaningful numeric operations on the values.

Clustering splits a set of data objects into subsets, so-called clusters. This is done such that data objects within the same cluster are more similar to each other than to data objects in other clusters [14]. Consequently, clustering requires some form of similarity measure for the data objects at hand. In numeric data, similarity can easily be defined as a function of the distance between data objects: the closer two data objects are in feature space, the more similar they are. Determining similarity between data objects containing categorical attributes is much harder. In general, this requires either a meaningful mapping from categorical to numeric values or domain-specific heuristics to establish some form of ordering.

A key characteristic of clustering lies in the fact that normally the resulting clusters are not known in advance. Often, it is even unknown how many meaningful clusters there are in the data. Cluster algorithms partition data objects in an automated fashion, without human intervention or feedback. Therefore, clustering is a form of unsupervised learning in contrast to classification where class information is known and used for feedback to optimize the classifier's performance. However, there is a connection between clustering and classification as clustering can be used to identify meaningful clusters from a particular set of data. These clusters can then be used as classes to which a classifier can assign new data objects.

Unfortunately, there is no single best cluster algorithm. Thus, several distinct approaches to clustering and a multitude of cluster algorithms following these approaches exist. Approaches to clustering can be categorized into partitioning, hierarchical, density-based, and grid-based ones [14, 18, 19]. Most clustering algorithms generate non-overlapping or disjoint clusters, i.e. each data object is assigned to exactly one cluster or rejected as being an outlier. Some algorithms, however, establish fuzzy clusters, i.e. assign data objects to different clusters simultaneously, although with varying degree of membership.

6.2.1 Partitioning Clustering Methods

Partitioning clustering methods directly decompose the set of data objects into disjoint clusters based according to a goodness criterion.

A well-known partitioning cluster algorithm is k-means [24, 25]. In essence, k-means tries to minimize intra-cluster variance. To that end, k-means arbitrarily seeds k clusters and assigns additional data objects to the nearest cluster. The nearest cluster is the one with the smallest distance between the cluster centroid and the new data object. The cluster centroid is the mean value of all data objects in a particular cluster. Once new data points have been assigned to clusters, cluster centroids are updated. This process is repeated, including reassignment of data points, until no further relocations are required.

Clearly, computing mean values works well for numerical data, but cannot be done easily for categorical attributes. Therefore, a number of variants of k-means have been proposed, notably k-modes and k-prototypes [15], to cope with categorical or mixed types of attributes.

Expectation Maximization (EM), a second influential partitioning cluster algorithm, follows a different approach. Clusters can be viewed as regions in feature space with a probability density that differs from its surroundings as well as from other clusters. Thus, a set of clusters is generated by a mixture of distinct probability distributions each of which represents one cluster. Detecting clusters is then equivalent to reasonably estimate the probability distributions. EM [7] consists of two steps, namely an expectation and a maximization step. The expectation step estimates probabilities that a new data object is "generated" by a particular probability density function. The maximization step adapts the parameters of the density functions by maximizing a goodness function. Thus, EM iteratively establishes maximum-likelihood estimates for the parameters of k probability density functions, corresponding to k clusters in the underlying data. Since data objects that are closer to each other have a higher probability of being members of the same cluster, EM can be viewed as a generalization of k-means.

EM forms fuzzy clusters, allowing each data object to be a member of multiple clusters with different degrees of membership. In contrast, k-means and variants assume that each data object is assigned to exactly one cluster.

6.2.2 Hierarchical Clustering Methods

Hierarchical clustering methods generate hierarchies of clusters by either splitting clusters into smaller ones or merging two clusters into a larger one. The former strategy is called divisive, the latter agglomerative.

Linkage algorithms are well-known agglomerative hierarchical clustering algorithms. Initially, each data object is a cluster in its own right. Then, pairs of clusters are iteratively merged whenever the representatives of the two respective clusters are closer to each other than to a representative of any other cluster. Single linkage, for instance, chooses the two data objects that are closest to each other as cluster representatives.

ROCK [12] follows a different approach in merging the pair of clusters with the maximum number of links. A link indicates that two data objects are similar to each other and, thus, likely to end up in the same cluster. A link is established if two data objects have at least one common neighbor. ROCK is claimed to be particularly suitable to handle categorical data.

6.2.3 Density-Based Clustering Methods

The core concept of density-based clustering methods is the number of data objects in distinct regions of the feature space, i.e. the density of data. Clusters are regions with high density surrounded by regions of considerably lower density.

DBSCAN [10], a widely known density-based algorithm, identifies dense regions by finding core data objects with sufficiently many neighbors within a neighborhood of a given radius. Dense regions are extended by including all data objects in the neighborhood of core objects, eventually turning them into core objects as well. This process is repeated until no additional core objects can be found, i.e. density falls below a specified threshold. DBSCAN is capable of finding clusters of arbitrary shape.

6.2.4 Grid-Based Clustering Methods

Grid-based clustering methods impose a grid structure on the feature space, i.e. the range of each attribute is quantized into intervals of equal length. Clusters are identified using the cells of the grid rather than individual data objects. Often, grid-based clustering algorithms also rely on density.

The latter is also the case for CLIQUE [1] which views a grid cell as dense if the number of data objects in that cell exceeds a certain threshold. Dense cells are identified iteratively by first examining one-dimensional subspaces of the feature space, i.e. individual attributes. Afterwards, k-dimensional cells are examined only if each of its $(k - 1)$-dimensional projections is already known to be dense. Adjacent dense cells in the same subspace are merged to form clusters. CLIQUE is capable to identify clusters in subspaces of the full feature space, i.e. only a subset of attributes may be relevant in some clusters. Thus, CLIQUE, like other so-called subspace algorithms, automatically solves the feature selection problem, i.e. the task of identifying those attributes that exhibit certain patterns in a cluster.

6.3 Cluster Utility Framework

The objective of the SecMine project is to develop a methodology, which allows analysing large amounts of data generated by a security information and event management system. To this end, various data mining methods, notably clustering, classification, and visualization, need to be explored in combination. However, there are many different approaches in data mining and a large number of possible combinations. The Cross-Industry Standard Process for Data-Mining (CRISP) [26] distinguishes different phases including data preparation, modelling, and evaluation. After preparation, e.g. cleaning or converting data, the modelling phase defines a specific configuration of different approaches. Different configurations are then compared with respect to their quality in the evaluation phase. An appropriate workflow can only be found by experimenting with different combinations on real data and evaluating the results to figure out which configuration fits best. Hence, a highly flexible experimental environment is required which allows to easily set up and combine different methods and provides a range of tools for visualization and analysis. In SecMine we developed the Cluster Utility Framework (CUF) based on a pipes-and-filters architecture to fulfil these requirements.

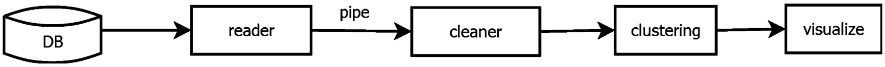


Fig. 6.1 Data mining workflow

6.3.1 Pipes-and-Filters Architecture

A pipes-and-filters architecture [3] is a pattern which is often used in software applications that handle and process data streams. Filters constitute independent working steps which manipulate incoming data. Pipes connect pairs of filters and pass data on to other filters.

Filters in CUF are instances of services with a common interface.

Figure 6.1 shows a sample workflow where a database is read, the data are cleaned (e.g. inconsistent data are filtered out), and then clustered and visualized. Since each of these steps is implemented as an individual filter in CUF, different workflows can easily be arranged and executed. If different clustering methods need to be compared, only the filter for clustering is exchanged and everything else left untouched. As an enhancement to the pure pipes-and-filters pattern, CUF also allows to split the workflow and establish parallel paths. Thus, different clustering filters or multiple instances of one filter with different parameter settings can be combined to process data in a single run and compare their results directly. Therefore, experiments can easily be set up and changed in order to evaluate different approaches under equal circumstances.

In addition, a pipes-and-filters pattern provides an easy way of integrating new filters, e.g., new clustering algorithms or filters for visualization.

Since filters work independently, each of them can be executed in its own thread. Thus, CUF can exploit multicore architectures to reduce execution time.

6.3.2 Flexible Handling of Heterogeneous Data

Data normally consist of different types like numbers, text, timestamps etc. This is also true in SecMine where we have IP addresses (as strings), port numbers, event occurrences as timestamps and many more. To ensure that the framework can cope with heterogeneous data types, all attribute values are stored as strings. CUF provides a data structure based on hash maps that can carry metadata for each entry where the original attribute type can be stored. The data structure also provides methods to safely convert attributes to their original type. The conversion increases flexibility with respect to data input at the expense of execution time.

6.3.3 Cluster Algorithms in CUF

In clustering, there is no a priori knowledge on what patterns may be expected in the data. The first goal is to get a better understanding of the data before useful information can be extracted. Appropriate clustering techniques only be determined experimentally due to the lack of information on which patterns might be found at all. In order to not restrict the range of techniques, CUF integrates clustering algorithms from all categories mentioned in Sect. 6.2. Possibly, only a combination of techniques will provide acceptable results or particular techniques are only suitable in special cases. Due to the flexible architecture, additional ones can be added easily.

6.3.3.1 k-Means and k-Modes

As representatives of partitioning clustering algorithms, k-means and k-modes [4] are available in CUF. k-Means and k-modes perform well on a large amount of data. Yet, they expect the number of clusters in the data as an input parameter. This parameter is hard to determine because clustering aims to find out if there are patterns or clusters at all. Therefore, the number of clusters should rather be the result than an input of a cluster algorithm.

6.3.3.2 CLIQUE

CLIQUE is a grid-based clustering method that is able to find clusters in arbitrary subspaces and can handle large amounts of data.

Figure 6.2 shows a dataset with attributes “age” and “salary”. The range of both attributes is divided into equal-sized intervals, so-called units. At the bottom the projection of all data objects onto the dimension “age” is shown with dense units ranging from 20 to 30, 40 to 50, and 50 to 60. The unit ranging from 20 to 30 and the interval from 40 to 60 constitute two clusters in the subspace “age”. Likewise, there is a cluster from 2000 to 4000 in the subspace “salary”. Combining these two dimensions leads to the 2-dimensional subspace “age $\times$ salary” where the borders of the intervals form the “grid”. In this subspace there are two clusters: one with an age between 20 and 30 and a salary between 2000 and 3000, the other with an age between 40 to 60 and a salary between 3000 and 4000.

In addition to CLIQUE, CUF provides three clustering algorithms based on MAFIA [11], an optimized variant of CLIQUE. In contrast to CLIQUE, MAFIA uses adaptive intervals which are determined from the data distribution in each dimension. This leads to more accurate cluster descriptions. Yet, the main advantage of MAFIA-based algorithms is that they are much faster, i.e. they can handle several millions of data points within a few hours runtime. Due to the large data volume in SecMine, this is an important feature.

CLIQUE and descendants rely on the density of regions. Hence outliers—single points which are isolated or sparsely surrounded by other points—will be neglected

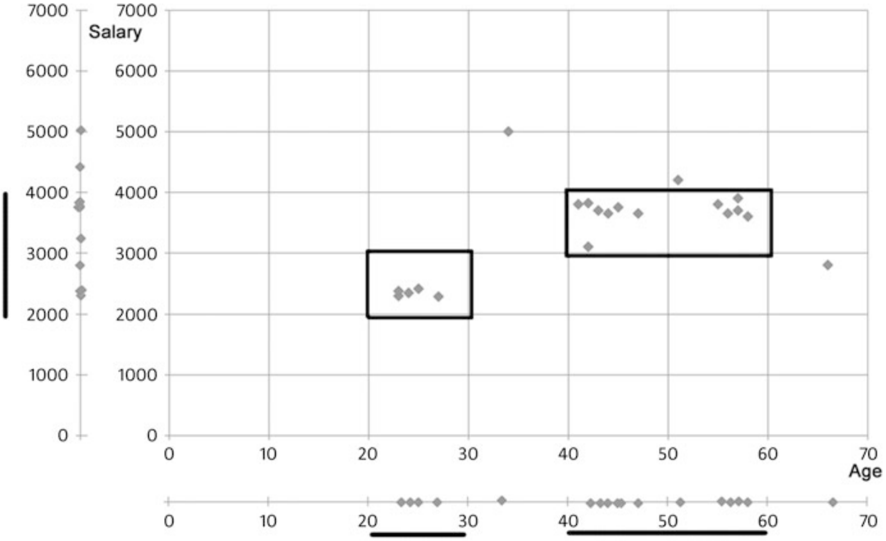


Fig. 6.2 Clusters in CLIQUE

during the procedure. These points, however, might be very interesting for detecting anomalies. To avoid this problem, the original data set can be compared with the clustered results and points which are no members of clusters can be marked. This allows recovering outliers but lacks further information on relationships between these data points.

Any CLIQUE-based algorithm can handle categorical attributes by defining an interval for each single value and inserting empty intervals between them. This workaround ensures independence of the order of values in categorical attributes, but it is still an open issue if this leads to meaningful results.

6.3.3.3 ROCK

ROCK is an agglomerative hierarchical algorithm. The pair of clusters to be merged is determined by a so-called goodness measure which takes into account the number of links between the pair and the current size of each cluster. The latter avoids large clusters which might dominate the analysis due to their sheer size.

Links are established whenever two data objects share a common neighbor. ROCK defines neighbors as points with a sufficiently high similarity. If there are clusters in a data space then a pair of data objects can be expected to have many common neighbors in such a cluster. Possibly objects from different clusters may also be quite similar, but the number of common neighbors should be considerably lower. Figure 6.3 depicts data objects as nodes; edges indicate the neighborhood between two points. It shows quite well that the number of common neighbors or "links" might work well as a similarity measure to find and separate two clusters.

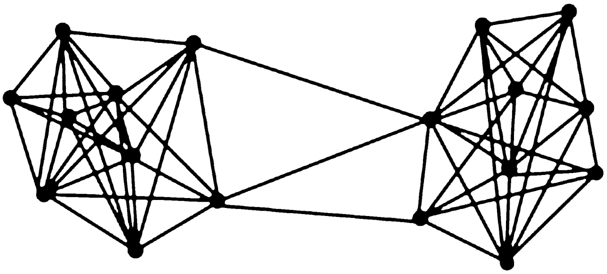


Fig. 6.3 Data space with neighbourhoods

Like k-means, Rock expects the number of clusters as an input parameter, which is problematic. To alleviate this, we introduced an additional criterion, which terminates clustering when the goodness measure falls below a limit, i.e. minimal similarity. There is no need to guess how many clusters there are.

ROCK is extremely resource-consuming since the goodness measure is calculated and stored for every pair of data points. This leads to quadratic storage complexity and even higher execution time complexity. Thus, the use of ROCK is limited to samples of less than 100,000 data points.

QROCK [9] is an optimization of ROCK without goodness measure. QROCK merges any pair of clusters if there is at least one link between them. This reduces the needed resources to a fraction of those of ROCK. Unfortunately, QROCK tends to merge almost all data objects into a single cluster since the link count between pairs of clusters does not matter anymore. Since ROCK's ability to handle categorical data is based on links and their count, QROCK also lacks the capability to find meaningful clusters in categorical data. Nevertheless, QROCK is well suited for detecting outliers in the data.

6.3.4 Preprocessing Categorical Attributes

A major challenge in data mining in general and SecMine in particular is the treatment of categorical data. The absence of an inherent ordering impairs the definition of meaningful similarity measures as well as visualization. Also, choosing appropriate representatives or descriptions of clusters is hard in the presence of categorical attributes. Numerical attributes allow for the description of clusters by means of their borders, e.g. "The cluster extends from 3 to 10". For a categorical dimension, e.g. colour, there is no point in saying "There is a cluster between blue and red" unless the attribute dimension is ordered in some way.

In SecMine, an "artificial" ordering can be based on either the data itself or domain-specific heuristics. CUF contains filters that transform specific categorical attributes to numerical values. In addition, CUF provides similarity measures, which rely on additional information.

6.3.4.1 Data-Driven Transformation

Several approaches directly extract information from the data to transform categorical values into numerical ones on the one hand, and to derive meaningful similarity measures on the other hand. Often they are based on probabilities of individual values of an attribute [2].

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{6.1}$$

Formula (6.1) Jaccard index.

A common similarity measure is the Jaccard index [18] which computes the ratio of the size of the intersection and the size of the union of the attributes of two data vectors. This measure can be refined by weighting a pair of attributes in the intersection, which share the same value. In particular, the probability that the values are in fact equal can be used as a weighting factor. For instance, if the attribute "color" has the value "red" in both data objects while the probability for "red" is 99 %, this might be less conclusive as if the probability for "red" were 1 %. As an alternative, categorical values can be transformed into numerical ones by simply replacing them with their probability.

Drawbacks of data-driven approaches lie in the high effort for preprocessing and difficult interpretation of transformed data. Also, the dependency of numerical values on the underlying data might lead to improper results. For instance, if data change and so should the measure, chances are that the change goes unnoticed.

6.3.4.2 Knowledge-Driven Transformation

Another source of useful information is domain-specific knowledge. A trivial transformation might assign a numeric value to each categorical one manually or define a similarity matrix with entries for every pair of data objects. This works well if the attribute's cardinality is sufficiently small and almost constant.

If the cardinality is large, as it is in most cases, a heuristic approach is needed which can determine the values automatically. CUF contains filters, which either transform the values or establish a similarity measure based on an implicit hierarchy in the values of a categorical attribute. Such a hierarchy can mirror real world phenomena, e.g. the structure of a network for IP addresses. If such a hierarchy exists, it can be represented as a tree from which similarity measures or appropriate mappings to numerical values can be derived.

Let a company be organized into several departments that can have different subnets. This leads to a tree similar to the one in Fig. 6.4. Similarity between two components can now be measured as the path length between the nodes representing these components: The shorter the path, the higher the similarity and vice versa. In that way, components a and b will be assessed to be more similar than, e.g., components a and d, which is also intuitively correct.

Knowledge-driven transformations avoid direct dependencies among data. Transformed data allow for simple interpretation as they reflect real domain facts.

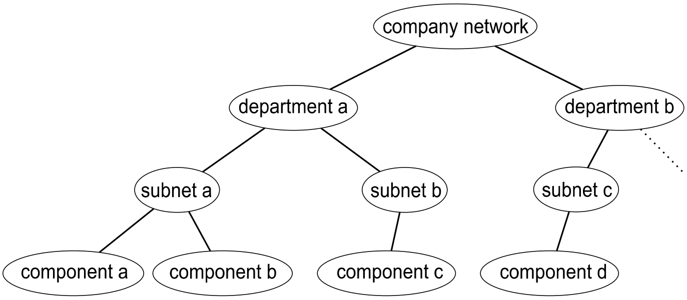


Fig. 6.4 Network hierarchy

6.4 Semantic Validation of Clusters

Cluster algorithms perform unsupervised learning. Consequently, they identify clusters on a purely syntactical basis, given a set of data objects. Chances are that the resulting clusters largely represent artefacts of the algorithm rather than semantically meaningful patterns. Therefore, clusters need to be validated semantically by domain experts. To that end, we currently use two techniques in SecMine, namely visualization and cluster validation measures.

6.4.1 Cluster Visualization

Visualization is a powerful tool since it exploits the human capability to quickly understand pictorial information and intuitively find patterns by simply looking at some displayed information. “Data analysis without data visualization is no data analysis” [21]—although this may not be true in any data analysis task, it is in clustering. In the context of cluster validation, visualization techniques need to fulfil three requirements:

- Visualization needs to be capable of displaying a large number of data objects, before and after they have been subjected to a cluster algorithm.
- Visualization must be able to deal with high-dimensional data since data objects are generally characterized by ten or more relevant attributes.
- Visualization should be capable of handling numerical and categorical attributes equally well since categorical attributes are frequent in real-world data.

In CUF, we use parallel coordinates for visualization and recently began to use pixel-based techniques, in particular recursive pattern, as a supplement.

Parallel coordinates explicitly aim at visualizing high-dimensional data [16, 17]. Parallel coordinates represent each attribute dimension independently as a vertical axis in a two-dimensional plane. Thus, all attributes are displayed on parallel axes and any data object is represented as a polygonal line that intersects each axis in the position of the respective attribute value (Fig. 6.5).

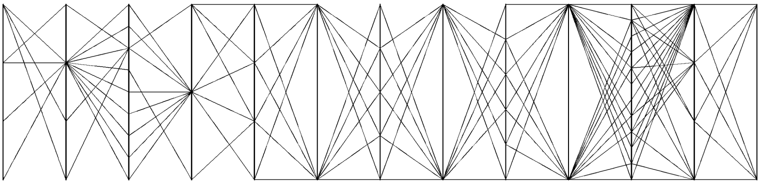


Fig. 6.5 Parallel coordinates

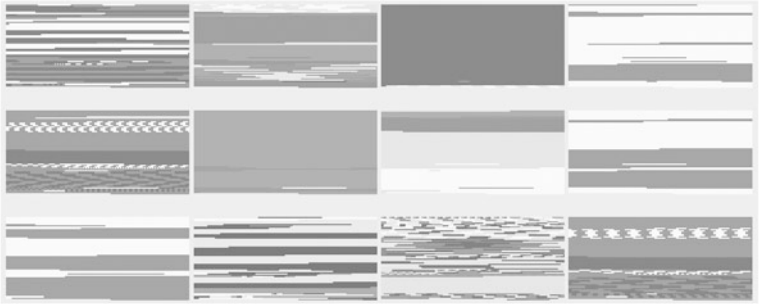


Fig. 6.6 Pixel-based visualization

Parallel coordinates do not pay special attention to categorical attributes. In particular, they do not prescribe any specific order of categorical values on a vertical axis. Thus, suitable pre-processing or transformation of categorical attributes is required for a meaningful display. Parallel coordinates do not in themselves account for a large number of data objects to be displayed but this can be supported by features such as, e.g., zooming and brushing.

Recursive pattern [20] particularly targets visualizing large quantities of multidimensional data. For that reason, recursive pattern is currently evaluated in CUF as a supplement to parallel coordinates. Recursive pattern generally represents each attribute value of a data object in a single pixel of a specific color (Fig. 6.6). The data values of each attribute dimension are presented in separate frame. In each of these frames, data objects and their corresponding pixel representations are arranged according to the same scheme e.g. left to right and back. Thus, patterns, in particular dependencies between attributes, can be identified quite easily, even for large amounts of data objects.

6.4.2 Cluster Validation Measures

Normally, any cluster algorithm is associated with a set of parameters. Examples of such parameters are density thresholds, distance metrics, expected cluster numbers etc. Yet, optimal parameter settings can only be determined experimentally, giving rise to the need to compare the quality of multiple runs of cluster algorithms with different settings. For this purpose, cluster validation measures can be useful even though they do not take the semantics of the underlying data into account. Over the years, several such measures have been developed [13, 23]. In CUF, we currently have four of these measures at hand, namely Dunn's index [8], Davies-Bouldin index [6], CS measure [5], and CDbw index [23]. Unfortunately, these measures neither work well on clusters generated by subspace algorithms such as CLIQUE, nor handle categorical attributes well. Therefore, we intend to experiment with CPCQ as an additional measure that is claimed to be better suited for categorical attributes [22].

6.5 Classification

Once we arrive at a set of clusters which have been firmly validated to be semantically meaningful, these clusters will be used to classify incoming data in order to detect anomalies, preferably in real-time. First experiments with three different types of classifiers, namely decision trees generated by C4.5, naïve Bayes classifier, and feed-forward neural networks [14], have been conducted. As a preliminary result, it can be stated that decision trees and neural networks work reasonably well, although training of neural networks needs considerable effort. A naïve Bayes classifier exhibited relatively poor results. This might be due to the fact that attributes in our data are not statistically independent and thus violate a core assumption of naïve Bayes classifiers.

6.6 Conclusion and Outlook

The complexity of IT infrastructures as well as the quantity and quality of attacks on these structures are growing. In order to maintain the integrity of IT infrastructures, novel integrated approaches are needed.

In the SecMine project, we aim at supporting human IT security experts by pinpointing significant alerts that really need closer inspection. To that end, we are developing an experimental tool environment called Cluster Utility Framework (CUF) that provides flexible data mining capabilities to analyze data generated by a security information and event management system. In particular, we use clustering algorithms to differentiate normal behavior from activities that call for intervention through IT security experts, such as hacker attacks from outside or unauthorized data access through insiders. This is accomplished in offline mode using a variety of alternative clustering algorithms with different strengths and weaknesses. In particular, we developed improved versions of subspace clustering algorithms like CLIQUE/MAFIA, agglomerative hierarchical clustering algorithms like ROCK/QROCK, but also "classical" partitioning clustering algorithms such as k-means or k-modes.

Before being subjected to clustering, data can be pre-processed in various ways. This is essential for categorical values in the data, such as, e.g., IP addresses or port numbers. Categorical data lack an underlying ordering relation and pose specific problems to clustering algorithms like k-means that expect purely numerical values, but also to visualization. In order to cope with categorical attributes, we devised heuristics to map categorical values cleverly to numerical values such that the semantics of the data are preserved as far as possible, while making them more easily accessible for clustering and visualization.

Resulting clusters can be subjected to visual inspection in CUF using parallel coordinates, but also pixel-based techniques such as recursive pattern. Thus, clusters can be categorized as reflecting normal or un-normal behavior.

Relevant clusters are then used as training data for real-time classification of the incoming data stream to identify security-related events.

Preliminary results using real network data of one of our industrial partners, an insurance group, indicate that clustering is well suited to structure monitoring data appropriately. Also, fairly large volumes of data (several millions of data records) can be clustered effectively and efficiently, in particular using parallel versions of MAFIA. Since network traffic and security-related events should be similar across industries, these results should be transferable also to other domains.

Current work focusses on the semantic validation of clusters. As it seems, this might give rise to a need for more elaborate visualization and cluster validation measures. In parallel, we examine how validated clusters can be used to classify streams of incoming filtered network data in real-time.

The main contribution of our research lies in an integrated approach to solving a real-world problem, namely identifying significant events from a huge data quantity in order to subject them to closer inspection by human experts. To that end, we rely on fairly general data mining and visualization techniques and integrate them in a flexible tool environment called CUF. Due to the generality and flexibility of the environment, CUF could be useful in other domains where relevant unknown pattern need to be identified in large quantities of multi-dimensional data, containing a significant share of categorical values.

Acknowledgements The SecMine project is supported under grant no. 17049X10 by Bundesministerium für Bildung und Forschung (BMBF). We thank Christian Bergmann, Toni Böhnlein, Sebastian Detsch, Thomas Geus, Steffen Hammer, Johannes Henninger, Matthias Herrmann, Sebastian Jakob, Daniel Klett, Adrian Köhlein, Evelyn Krüger, Benjamin Krull, Andreas Kühntopf, Hannes Müller, Marc Pieruschek, Markus Pütz, Markus Ring, Martin Rosenbaum, Manuel Schnapp, Tobias Schmidtlein, Christopher Schramm, Elena Tereshko, Melanie Westendorf, Thomas Worch, and Bernhard Sick for their contributions.

References

1. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high-dimensional data for data mining applications. In: Proc. 25th Int. Conference on Management of Data (SIGMOD'98), pp. 94–105 (1998)
2. Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: a comparative evaluation. In: Proc. SIAM Int. Conference on Data Mining, pp. 243–254 (2008)

3. Buschmann, F., Meunier, R., Rohnert, H., Sommerlad, P., Stal, M.: Pattern-Oriented Software Architecture—A System of Patterns. Wiley, Chichester (1996)
4. Chaturvedi, A.D., Green, P.E., Carroll, J.D.: k-Means, k-medians, and k-modes: special cases of partitioning multiway data. In: Classification Society of North America Meeting, Houston (1994)
5. Chou, C.-H., Su, M.-C., Lai, E.: A new cluster validity measure and its application to image compression. PAA Pattern Anal. Appl. **7**(2), 205–220 (2004)
6. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **1**(2), 224–227 (1979)
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **39**(1), 1–38 (1977)
8. Dunn, J.C.: Well separated clusters and optimal fuzzy partitions. J. Cybern. **4**, 95–104 (1974)
9. Dutta, M., Kakoti Mahanta, A., Pujari, A.K.: QROCK: A quick version of the ROCK algorithm for clustering of categorical data. Pattern Recognit. Lett. **26**, 2364–2373 (2005)
10. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. 2nd Int. Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226–231 (1996)
11. Goil, S., Nagesh, H., Choudhary, A.: MAFIA: Efficient and scalable subspace clustering for very large data sets. Technical report CPDC-TR-9906-010, Northwestern University, Evanston (1999)
12. Guha, S., Rastogi, R., Shim, K.: ROCK; a robust clustering algorithm for categorical attributes. In: Proc. 15th Int. Conference on Data Engineering (ICDE'99), pp. 512–521 (1999)
13. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. J. Intell. Syst. **17**(2/3), 107–145 (2001)
14. Han, J., Kamber, M., Pei, J.: Data Mining—Concepts and Techniques, 3rd edn. Morgan Kaufmann, Waltham (2012)
15. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. In: Data Mining and Knowledge Discovery, vol. 2, pp. 283–302 (1998)
16. Inselberg, A.: The plane with parallel coordinates. Vis. Comput. **1**, 69–91 (1985)
17. Inselberg, A., Dimsdale, B.: Parallel coordinates: a tool for visualizing multidimensional geometry. In: Proc. 1st IEEE Conference on Visualization (Visualization'90), pp. 361–378 (1990)
18. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs (1988)
19. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. **31**(3), 264–323 (1999)
20. Keim, D., Kriegel, H.-P., Ankerst, M.: Recursive pattern: a technique for visualizing very large amounts of data. In: Proc. 6th IEEE Conference on Visualization (Visualization'95), pp. 279–286 (1995)
21. Kozak, M.: Watch out for superman: first visualize, then analyze. IEEE Comput. Graphics Appl. **32**(3), 6–9 (2012)
22. Liu, Q., Dong, G.: CPCQ—contrast pattern based clustering quality index for categorical data. Pattern Recognit. **45**, 1739–1748 (2012)
23. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of internal clustering validation measures. In: Proc. 10th Int. Conference on Data Mining (ICDM 2010), pp. 911–916 (2010)
24. Lloyd, S.P.: Least squares optimization in PCM. Technical report, Bell Labs (1957). Also IEEE Trans. Inf. Theory **28**(2), 129–137 (1982)
25. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967)
26. Wirth, R., Hipp, J.: CRISP-DM: towards a standard process model for data mining. In: Proc. 4th Int. Conference on the Practical Application of Knowledge Discovery and Data Mining, pp. 29–39 (2000)

Chapter 7
Exploring the Differences Between the Cross Industry Process for Data Mining and the National Intelligence Model Using a Self Organising Map Case study

Richard Adderley

Abstract All Police Analysts in the UK, and many Forces in Europe and the USA, use the National Intelligence Model as a means to provide relevant, timely and actionable intelligence. In order to produce the required documentation analysts have to mine a variety of in-house data systems but do not receive any formal data mining training. The Cross Industry Standard Process for Data Mining is a database agnostic data mining methodology which is logical and easy to follow. By using a self-organising map to suggest offenders who may be responsible for sets of house burglary, this study explores the difference between both processes and suggests that they could be used to complement each other in real Police work.

7.1 Introduction

In our society today both private and public sector organisations are reducing costs and shedding staff in order to remain profitable. The Police are no exception to this rule. In the last three years the policing budget has fallen on average 3.86 % annually [1–3], the number of Police Officers has been reduced by 4.24 % and support staff reduced by 11.24 % [4]. Since 1996 there has been a downward trend in the number of offences that have been recorded in the UK [5] however, although not statistically significant, the year 2010/11 showed a rise in reported offences over the previous year. As reported in the Mail Online [6], approximately 850 more crimes per day were recorded 2012 than the previous year. Police analysts are at the forefront of providing actionable intelligence to front line staff to prevent and detect crime but they also have been part of the staffing cuts. With cost cutting and rising crime it is important that the analysts' work is focussed according to the Force's priorities. This study will discuss how a policing intelligence framework, the National Intelligence Model, can be integrated with a data mining process, the Cross Industry Process for Data Mining to assist the analyst in the provision of actionable intelligence.

R. Adderley (✉)
A E Solutions (BI) Ltd, Badsey, Evesham, Worcestershire, WR11 7AA, UK
e-mail: rickadderley@a-esolutions.com

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_7, © Springer-Verlag London 2013

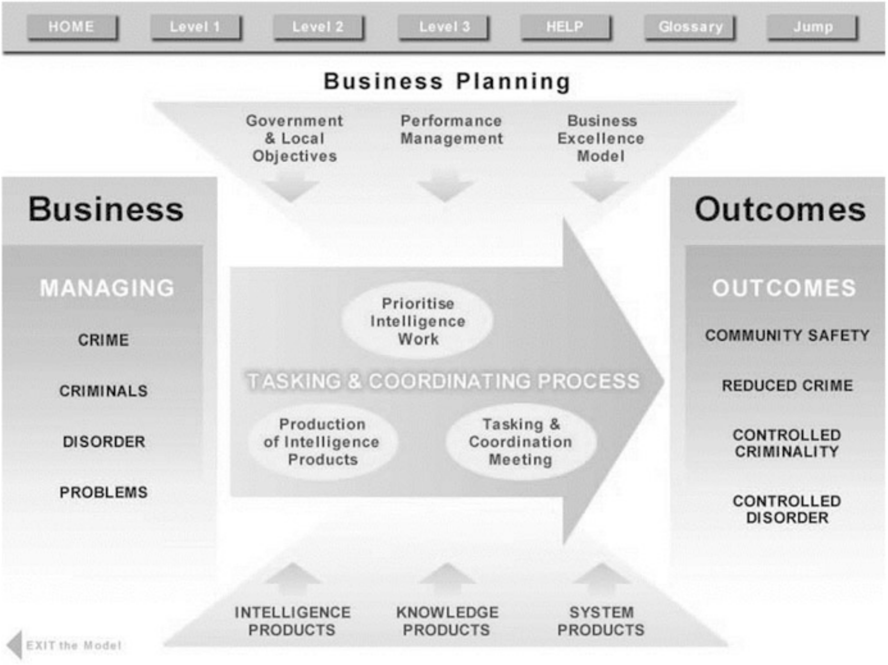


Fig. 7.1 National Intelligence Model

7.1.1 National Intelligence Model

All Police Forces within the UK, many Forces in Europe, the USA and beyond, use the National Intelligence Model (NIM) [7] as a framework for the provision of intelligence. It was developed by the UK National Criminal Intelligence Service on behalf of the Crime Committee of the Association of Chief Police Officers and provides a set of common standards and discipline facilitating the access of large amounts of data to make inferences about current crime.

The model has been designed to work at three levels:

1. Local issues—these are crimes, criminals and other problems affecting an area within a large Force such as a Basic Command Unit (BCU) or the entire area of a small force. A BCU is a geographic area with a Force usually led by a Chief Superintendent and has a common demographic breakdown and crime profile.
2. Cross border issues—frequently, criminals do not restrict their criminality to small areas within a Force, they cross internal boundaries and/or Force boundaries; these are the cross border issues.
3. Serious and organised crime—such crimes that operate nationally or internationally such as drug importation and people trafficking.

The model as illustrated in Fig. 7.1, focuses on the tasking and coordinating process (T&C) to manage the business of crime, criminals, disorder and other problems to provide actionable outcomes that reduce crime, enhance community safety and control both criminal behaviour and disorder. It takes into account objectives set locally and nationally, performance management and business excellence. The intelligence products provide the main input into the T&C process comprising strategic and tactical assessments and target and problem profiles, but they are also supported by the knowledge and system products. Knowledge products are the formal documents, guidelines and protocols that provide the necessary in which the analysts operate; for example Data Protection Act, Human Right Act etc. There are three types of system products;

1. During the research process there is a requirement for access to data storage, retrieval and comparison
2. The provision of access to systems in order to acquire new information
3. The provision of operational security systems

The T&C meetings comprise strategic and tactical sessions. The former establishes or amends the control strategy for the area ensuring that appropriate resources are committed at this level. The latter is the assessment that drives the day-to-day policing by targeting offenders, managing crime hot spots, identifying crime series and applying crime prevention measures. At levels two and three of the NIM the T&C process will typically involve multi-agency personnel and, particularly at level three, agencies that cover the entire UK such as the Serious and Organised Crime Agency (SOCA).

At the tactical level the assessment comprises a number of target and problem profiles. A target profile is an in-depth analysis of an offender and associates (criminal networks/gangs) describing in detail spatial and temporal activities and current intelligence. A problem profile is concerned about the geography of crime and disorder, identifying hot spots and crime series. As criminals work within geographic areas there will be some overlap within these profiles.

The knowledge products mentioned above, in practice, are the policing and open source data systems into which the analysts will interrogate to provide the target and problem profiles.

The T&C process aims at providing timely and actionable intelligence. Personnel are physically tasked to deliver measurable results and held accountable for that delivery. This means that each task is evaluated on its outcomes; for example, has offender X been arrested, has the burglary crime in the area been reduced, has more intelligence been gathered about the late night dance club.

7.1.2 Cross Industry Standard Process for Data Mining

The Police store a large amount of data on a daily basis which is held in disparate date sets depending on the type of incident or crime that is reported. For example;

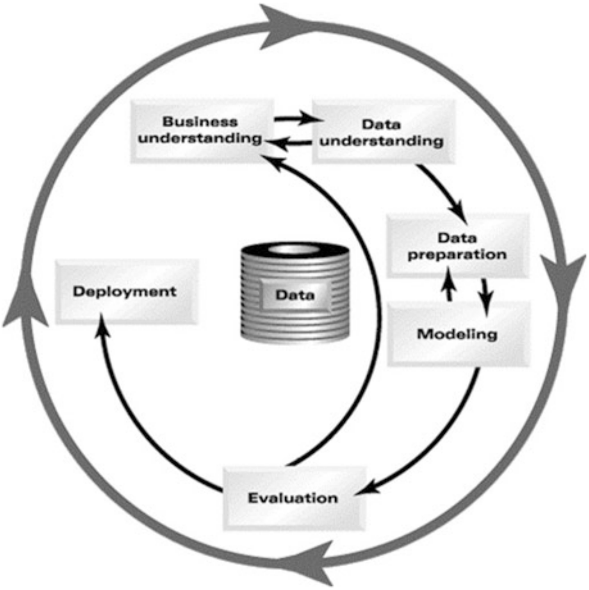


Fig. 7.2 CRISP-DM

when a crime occurs the victim will often report this fact via a Command and Control system (another term for Computer Aided Dispatch), an Officer will attend and take the report which is then entered onto a crime system. During the investigation a variety of contacts with suspected offenders occur and their details are entered into systems that include Stop & Search, Intelligence, Custody (when arrested) and Nominals (when the crime has been solved). There is no commonality or unique reference number to connect many of these systems, the analyst has to mine the data to provide operationally effective intelligence. The NIM does not provide guidance on data mining and the majority of Police analysts have not received training in data mining nor the few specific tools to provide technical assistance. It would be useful to place a data mining framework within the NIM.

There are a number of data mining frameworks that are available many of which are linked to specific database systems. The Cross Industry Standard Process for Data Mining (CRISP-DM) is database agnostic. It was developed within a European Project by four companies, SPSS, Teradata, Daimler and OHRA and was released as a process in 1999 [8].

The process, as illustrated in Fig. 7.2, is cyclic and has six sections:

1. *Business understanding* In this section the business objectives and the data mining goal are determined together with the success criteria. A situational assessment is undertaken to cover the resource requirements, constraints, risks and contingencies and cost benefits. A project plan is produced in this stage.
2. *Data understanding* The initial data set(s) are collected, described and documented. The data is preliminarily explored and its quality assessed.

3. *Data preparation* referring to the 80/20 rule, 80 % of the project's time is consumed in this section. The data is cleansed, merged, reformatted, new attributes are derived and refined ready for the next stage.
4. *Modelling* There are a number of modelling techniques that can be used many of which can be used in combination. Chapter 4 of the book, Data Mining Practical Machine Learning Tools and Techniques [9] and the paper Top 10 algorithms in data mining [10], both provide a very good overview of a large number of modelling techniques. In order to model effectively the data will need to be prepared into training set(s), testing set(s) and holdback set(s) and assessed by using cross-validation techniques [11].
5. *Evaluation* This section embellishes on the model assessment to evaluate the results of the data mining process according to the success criteria in Sect. 7.1, Business Understanding, above. The results of this analysis will provide a list of possible actions and decisions.
6. *Deployment* A deployment plan will be produced to bring this new data mining process into the organisation.

This chapter will discuss how the NIM and CRISP-DM can be used in a complimentary way to solve a real world policing problem.

7.2 Integrating CRISP-DM with the NIM

Each area within a Force may only have one or two analysts to provide a wide spectrum of reports and actionable analysis for the T&C process. Data is constantly arriving in the policing systems as crimes and incidents occur 24 hours a day, 7 days a week. The analysts need to keep abreast of this incoming data to maintain their target and problem profiles and also to become aware of emerging crime series or patterns in crime. A crime series is a set of crimes that have strong similarities in one or more of geographical/temporal/modus operandi features. Currently, the analyst will need to manually read each report and make (mental) notes. If the current report appears to have some similarities with previous reports, a SQL like search can be undertaken to retrieve such data which itself has to be manually read. This process is time consuming, prone to error and inefficient. To extract specific data, complete preliminary analysis and prepare a short report can take 1 to 2 hours and is typically 5 % accurate. In this instance "accuracy" means the crimes are so similar to the offender's MO that they would be selected. If there are two analysts in the same office, the same process is repeated twice. By utilising neural network modelling techniques there is a good probability that analysts' time can be saved and the accuracy of intelligence improved reducing the amount of time to minutes and improve the accuracy to 85 % [12].

Table 7.1 Number of Crimes Committed at Each Dwelling Type

Dwelling Type	Number of Crimes
Bungalow	16
Detached	14
Flat	87
Maisonette	9
Semi Detached	227
Terraced	107
Town House	10

7.2.1 Business Understanding

The example to be used in this analysis, which is also the business understanding project goal, is to automatically identify a house burglary crime series and suggest which offenders may be responsible for those crimes. The success criteria will be the reduction of crime (house burglaries) and the number of offenders arrested for those crimes.

7.2.2 Data Understanding

The analyst will extract data from the crime system and begin to explore and understand the nature of the problem. At this stage a problem profile will be started and enhanced as more of the data is understood. After removing outliers, in this illustrative but accurate data set there are 470 domestic burglaries distributed as illustrated in Table 7.1.

Utilising such a small data set from a single area within a Force, the records would have been input by no more than two clerks so it could be assumed that the data quality would be of a good standard. However this is not the case. When an Officer records the report of a crime it is hand written onto a paper form and submitted to a clerk for logging into the crime system. The clerk has to read handwriting which is often very poor and then subsequently transcribe it into check boxes and free text fields. Until current Police legacy systems are replaced with modern technology such as mobile data terminals, this will continue to be a problem with which analysts will have to contend.

7.2.3 Data Preparation

The modus operandi (MO) free text field contains the data needed to be parsed in order to extract the relevant fields necessary for modelling (Table 7.2). The data mining workbench tool, Authority Miner [13], was used to create the parsing process

Table 7.2 Free text modus operandi data

Num	Modus Operandi
1	OFFENDER BELIEVED APPROACHED REAR DOOR AND WITH BODILY FORCE FORCED SAME CAUSING YALE LOCK TO DISCONNECT AND ALLOW ACCESS ONCE INSIDE SEARCHED GROUND FLOOR STOLE PROPERTY EXIT VIA OPENING LOUNGE WINDOW ALLOWING ACCESS TO REAR SIDE GARDEN AREA
2	OFFENDERS UNKNOWN WENT TO REAR GROUND FLOOR WINDOW USING UNKNOWN IMPLEMENT FORCED OPEN SAME CAUSING DAMAGE ENTRY GAINED UNTIDY SEARCH OF ALL ROOMS PROPERTY STOLEN EX AS EN MADE GOOD ESCAPE
3	OFFENDERS UNKNOWN USING UNKNOWN INSTRUMENT GAINED ENTRY THROUGH FRONT KITCHEN WINDOW AND STOLE COMPUTER AND ASSOCIATED ITEMS MADE GOOD ESCAPE VIA KITCHEN DOOR LEAVING VALUABLE ITEMS BEHIND SUCH AS CD PLAYER TELEVISION AND VCR
4	PERSON ENTERED FIRST FLOOR FLAT BY FORCING LOCK ON FRONT DOOR ENTERED PREMISES STOLE PROPERTY EX AS ENTRY
5	OFFENDER FRIEND OF IP ENTERED VIA INSECURE FRONT DOOR SEARCHED FLAT STOLE PROPERTY EXEN
6	U/K OFFENDERS SMASHED KITCHEN WINDOW AT REAR OF PREMISES AND ENTERED THROUGH SAME TIDY SEARCH OF DOWNSTAIRS REMOVED TELEVISION AND MADE EXIT THROUGH REAR KITCHEN DOOR
7	OFFENDER APPROACHED FRONT DOOR SMASHED GLASS OF SAME INSERTED HAND AND OPENED DOOR ENTERED MADE UNTIDY SEARCH STOLE PROPERTY DISTRACTED BY OCCUPIER WHO RETURNED HOME SHOUTED UPSTAIRS PUNCHED BY OCCUPIER TO SIDE OF FACE OFFENDER SLAPPED OCCUPIER AROUND

Table 7.3 Parsed modus operandi fields

Num	ENTRY	FEATURE	FEATURE TYPE	METHOD	ROOMS SEARCHED	SEARCH TYPE
1	Rear	Door	Fixed	Forced	DownStairs	Tidy
2	Rear	Window	Casement	Forced	All	UnTidy
3	Front	Window	Casement	NOT KNOWN	NOT KNOWN	NOT KNOWN
4	Rear	Window	Casement	Forced	NOT KNOWN	UnTidy
5	Rear	Door	Casement	Insecure	NOT KNOWN	NOT KNOWN
6	Rear	Window	Fixed	Smash	DownStairs	Tidy
7	Front	Door	Casement	Smash	UpStairs	UnTidy

which resulted in Table 7.3. A number of data mining workbench tools could have been used such as IBM SPSS Modeler [14], WEKA [15] and SAS predictive analytics and data mining [16], however Authority Miner is the author's own software which has specific algorithms for policing.

For example number 1 in Table 7.2, once parsed, results in number 1 in Table 7.3. Where attributes cannot be assigned the value "NOT KNOWN" is entered.

Six fields are created from this process;

ENTRY: The point of entry which contains the values Above, Below, Front, Rear and Side.

FEATURE: The feature at the point of entry which contains the values of Door and Window.

FEATURE_TYPE: The type of feature at the point of entry which contains the values Casement, Fixed, Louvre, Patio, Sash and Transom.

METHOD: This is the method that the offender used to actually gain entry to the feature and contains the values Climbed, Cut, Duplicate (key), Forced, Insecure, Rammed, RemGlass, (Removed glass) and Smash.

ROOMS_SEARCHED: Once the offender has gained entry to the building this field identifies the rooms entered and contains the values All, Down (downstairs), Many, One and UpStairs.

SEARCH_TYPE: This field contains the values Tidy and UnTidy determining the type of search that the offender conducted.

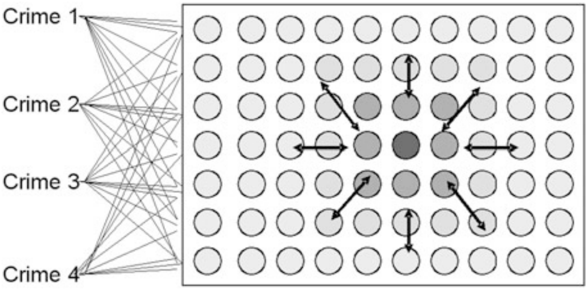
These six fields are common among many of the UK Police crime recording systems. Some have only check/list boxes where the inputer enters this information and others have the free text as well as the check/list boxes. On many occasions there is additional information in the free text which can also be parsed to add further information to the modelling process.

There are methods that can be employed to semi automate the extraction of entities from free text by employing text mining software. However these are not often used in policing as they are very expensive and time consuming to configure with domain relevant dictionaries and concepts. Even when using such software it often does not take into account misspelling due to poor typing input as in entry number 2 (UNKNOWN). Several UK Forces have trialed such software with little success.

7.2.4 Modelling

An unsupervised learning algorithm, the Self Organising Map (SOM) [17] was used to cluster the crimes into groups of similar MOs. A SOM is a neural network based algorithm which finds similarities in data and forms clusters based on those similarities. It is an unsupervised algorithm that learns from the data. This means that the algorithm is informing the analyst of similarities in the data without the person actually intervening. Each crime is connected to every neuron as illustrated in Fig. 7.3; the connection weights are randomly set between zero and one. An arbitrary neuron is selected as the start point and those crimes that immediately surround this start point and are similar are drawn towards it into the same cluster. Those that are

Fig. 7.3 Self organising map



dissimilar are pushed away into different clusters. The connection weights are adjusted each time. This process occurs for a number of cycles, in this analysis—150 cycles, and then the neighbourhood widens to two neurons around the start point. This again occurs for a number of cycles, in this analysis 200 cycles. Depending upon the actual SOM algorithm used, a number of parameters can be set; the grid size and, as stated above, the number of cycles in each neighbourhood. In this study a 10 × 10 grid forming 100 clusters was used. Empirical work has established that clusters containing between four and ten crimes after the modelling has concluded are the most accurate. As stated in Sect. 7.2.2 above there were 470 crimes in the data set and if evenly distributed over every cluster there would be between four and five in every cluster. That is a good rule of thumb. Some algorithms will automatically allocate the cycle numbers based on the change in connection weights. When the weights do not change for a number of cycles the neighbourhood changes or the algorithm completes.

7.2.5 Evaluation

Evaluation within the NIM has a different meaning to that within CRISP-DM. The former evaluates the outcomes of the actionable T&C documents. For example; has crime in area A been reduced? Is offender X still offending after being targeted? The latter evaluates the model prior to being deployed within the organisation. It is primarily concerned with the accuracy of the model within the business domain. By combining both techniques, the model can be evaluated for accuracy within the domain and then on its outcomes when the live data is turned into actionable intelligence.

The model in this study, as stated above, utilised a 10 × 10 grid using the MO features. Empirically those clusters that contain four to ten crimes are the most accurate, accuracy being determined by the MO similarities between each of the crimes in each of the clusters. Of the maximum 100, 80 clusters were populated by crimes totalling between one and 25. 44 clusters remained after removing those that contained less than four and more than ten crimes totalling 216 remaining crimes.

Table 7.4 provides an example of similarities within a cluster. In cluster K6-6 there are a total of five crimes where entry has been gained by smashing or forcing a

Table 7.4 Cluster similarities

ENTRY	FEATURE	FEATURE TYPE	METHOD	ROOMS SEARCHED	SEARCH TYPE	Crime Num	SOM Key
Front	Door	Fixed	Smash	All	UnTidy	607/05	K6-6
Front	Door	Fixed	Forced	Many	UnTidy	1497/05	K6-6
Front	Door	Fixed	Forced	One	UnTidy	3755/05	K6-6
Front	Door	Fixed	Smash	All	UnTidy	3772/05	K6-6
Front	Door	Fixed	Forced	Many	UnTidy	4413/05	K6-6
Front	Window	Casement	Insecure	Down	Tidy	3415/05	K5-1
Front	Window	Casement	Forced	Down	Tidy	3626/05	K5-1
Front	Window	Casement	Insecure	Down	Tidy	4301/05	K5-1

Table 7.5 Adjacent clusters

ENTRY	FEATURE	FEATURE TYPE	METHOD	ROOMS SEARCHED	SEARCH TYPE	Crime Num	SOM Key
Rear	Window	Fixed	Climbed	Down	Tidy	3239/05	K5-3
Rear	Window	Fixed	Climbed	UpStairs	Tidy	4549/05	K5-3
Rear	Window	Fixed	Forced	Down	Tidy	3083/05	K5-3
Rear	Window	Fixed	Insecure	Down	Tidy	2894/05	K5-3
Rear	Window	Fixed	RemGlass	Down	Tidy	2358/05	K5-3
Front	Window	Fixed	Climbed	One	Tidy	1864/05	K5-4
Front	Window	Fixed	Forced	Down	Tidy	2958/05	K5-4
Front	Window	Fixed	Insecure	All	Tidy	1430/05	K5-4
Front	Window	Fixed	RemGlass	Down	Tidy	2028/05	K5-4
Front	Window	NOT KNOWN	Forced	Down	Tidy	1414/05	K5-4
Front	Window	Patio	Forced	Down	Tidy	193/05	K5-4

fixed front door and an untidy search was conducted in a variety of room categories. In cluster K5-1 entry was gained through a front casement window that was either insecure or smashed, a tidy search was conducted in down stairs rooms.

Crimes that are in clusters which are next to each other (adjacent) are similar in structure, and they have similar themes. For example; Table 7.5 below illustrates the clusters K5-3 and K5-4 where entry has been gained through a fixed window either at the front (K5-3) or rear (K5-4) of the property where a tidy search has been conducted down stairs. These themes also continue with clusters above and below. When there are one or more empty clusters between, then the themes have a greater variability.

Although the SOM results in a grid structure and, as stated above, clusters that are adjacent have strong similarities. There is no wrapping at the grid edges. For example; there are three clusters in Table 7.6; K0-0 would wrap and be adjacent to

Table 7.6 Grid wrapping

ENTRY	FEATURE	FEATURE TYPE	METHOD	ROOMS SEARCHED	SEARCH TYPE	Crime Num	SOM Key
NOT KNOWN	NOT KNOWN	NOT KNOWN	NOT KNOWN	NOT KNOWN	NOT KNOWN	169/05	K0-0
NOT KNOWN	NOT KNOWN	NOT KNOWN	NOT KNOWN	NOT KNOWN	NOT KNOWN	155/05	K0-0
NOT KNOWN	NOT KNOWN	NOT KNOWN	NOT KNOWN	NOT KNOWN	NOT KNOWN	262/05	K0-0
Rear	Door	Patio	Insecure	Down	Tidy	402/05	K9-9
Rear	Door	Patio	Insecure	One	Tidy	1484/05	K9-9
Rear	Door	Patio	Insecure	One	Tidy	2157/05	K9-9
Front	Door	NOT KNOWN	Insecure	NOT KNOWN	NOT KNOWN	2670/05	K0-9
Front	Door	NOT KNOWN	Insecure	NOT KNOWN	NOT KNOWN	4091/05	K0-9
Front	Door	NOT KNOWN	NOT KNOWN	NOT KNOWN	NOT KNOWN	1573/05	K0-9

K0-9 and K0-9 would wrap to K9-9. It is clear to see that there are few similarities at these edges.

On examination of the relevant clusters the analyst would have a greater understanding of the current crime situation. By adding spatial (crime location) and temporal (crime date and time) information to these clusters it will be possible to identify emerging crime series' which can be used to enhance the T&C problem profile report. This will also enhance business understanding, looping back in the CRISP-DM process as in Fig. 7.2.

7.2.6 Assigning Offenders

Having identified clusters of crimes it is now necessary to suggest which offenders may be responsible for those crimes. The target profiles within the T&C process will contain information about currently active criminals within a Force or an area but to match this information to the crime clusters would be a manual and time consuming process.

Utilising the CRISP-DM process, similar steps are undertaken as described above.

Data Understanding: Offenders that have been arrested and either charged or cautioned for an offence are determined to have committed that offence [18] and are labelled as "detected".

Data Preparation: An offender set of data is extracted from the data table. Each record has a crime reference number which can be matched to the crime data set in order to obtain a list of crimes associated with offenders. The MO free text field (Table 7.2) is parsed to obtain a list of new fields as described above (Table 7.3).

Modelling: The crime SOM model that was previously created was used to cluster the offender set of crimes, the results of which were overlaid onto the crime clusters. This provided a set of clusters that had both detected and undetected crimes and that have strong similarities.

Evaluation: There were 23 unique offenders in the data set, 21 were represented in single clusters, one in two clusters and the remaining offender in five clusters. Six clusters had multiple offenders represented within, as illustrated by Table 7.7. There are only a finite number of ways in which domestic property can be burgled so it is not surprising that more than one offender has a similar MO.

There are seven crimes in cluster K9-5 (Table 7.7), three of which are associated with offenders. By examining the individual features it is possible to suggest that one or more of these offenders are responsible for the undetected crimes.

The offender Vijay Khan is represented in five different clusters (Table 7.8). Many offenders have a varied repertoire of MOs due to them learning from and teaching others [19].

Deployment: Please see Sect. 7.2.7 below.

7.2.7 Deployment

This is the last stage in the CRISP-DM process. Within the data mining cycle, deployment typically means deploying the model into the business arena. Within the NIM, deployment means the utilisation of the data within the T&C process where Officers are tasked to take action on the reported intelligence. The results of Sect. 7.2.6 can actually be deployed in this instance as the spreadsheets form actionable intelligence.

In this study there is little difference in the two meanings of deployment. The model was used on live data to provide the relevant information to enable the analyst to deliver actionable intelligence thereby enabling operational personnel to be tasked with solving the problem.

7.3 Conclusion

Analysts receive specific training to work within the NIM producing the relevant documentation and actionable intelligence. In order to accomplish this they are required to interrogate a multitude of in-house data sets but have no formal training in data mining. There are numerous data mining methodologies but to provide a framework to work within the NIM, this study has chosen CRISP-DM as it is database agnostic and logical to follow.

This study has demonstrated that both processes can be complimentary. With the requirement to produce and maintain T&C documentation, there is a necessity to

Table 7.7 Multiple offenders

ENTRY	FEATURE	FEATURE TYPE	METHOD	ROOMS SEARCHED	SEARCH TYPE	CrimeNum	SOM Key	LAST NAME	FIRST NAME
Rear	Door	Casement	Forced	All	UnTidy	ZZ/2846/05	K9-5		
Rear	Door	Patio	Forced	All	UnTidy	ZZ/3269/05	K9-5		
Rear	Door	Patio	RemGlass	All	UnTidy	ZZ/3813/05	K9-5		
Rear	Door	Casement	Forced	All	UnTidy	ZZ/3921/05	K9-5		
Rear	Door	Casement	Forced	All	UnTidy	ZZ/4143/05	K9-5	HUSSAIN	MAZAR
Rear	Door	Casement	Forced	Many	UnTidy	ZZ/507/05	K9-5	MOORE	SHAUN
Rear	Door	Patio	Forced	All	UnTidy	ZZ/3997/05	K9-5	SANDS	DAVID

Table 7.8 Multiple clusters for a single offender

ENTRY	FEATURE	FEATURE TYPE	METHOD	ROOMS SEARCH	SEARCHED TYPE	Crime Num	SOM Key	LAST NAME	FIRST NAME
Rear	Window	Fixed	Smash	Down	Tidy	ZZ/1201/05	K5-3	KHAN	VIJAY
Rear	Window	Fixed	Forced	All	Tidy	ZZ/1681/05	K6-3	KHAN	VIJAY
Front	Window	Patio	Climbed	Down	UnTidy	ZZ/3425/05	K6-4	KHAN	VIJAY
Side	Door	Fixed	Forced	All	UnTidy	ZZ/2074/05	K7-5	KHAN	VIJAY
Rear	Door	Patio	Insecure	Down	Tidy	ZZ/4367/05	K9-9	KHAN	VIJAY

continuously monitor the policing systems by mining the data. Using the CRISP-DM structured methodology to provide the framework for mining, the subsequent results can readily assist in T&C document maintenance.

This type of modelling using a SOM is eminently transferrable to other business types. Once the data has been prepared this data mining algorithm will automatically find similarities in data sets. For example; fraudulent insurance claims where a certain type of claim is always reported by the same working group of individuals; targeted marketing where a range of goods are purchased from a demographically constant group of people indicating future marketing opportunities etc.

References

1. Police Grant (England and Wales) 2012/13, The Stationary Office, London, HC 1797
2. Police Grant (England and Wales) 2011/12, The Stationary Office, London, HC 1771
3. Police Grant (England and Wales) 2010/11, The Stationary Office, London, HC 278
4. Dhani, A.: Police Service Strength England and Wales, 30 September 2011, HOSB 3/12, London (2012)
5. Chaplin, R., Flatley, J., Smith, K.: Crime in England and Wales 2010/11 findings from the British Crime survey and police recorded crime (2nd edn.), HOSB: 10/11 (2011)
6. Crime shows biggest rise for a decade (2012), Mail online. http://www.dailymail.co.uk/news/article-123421/Crime-shows-biggest-rise-decade.html. Accessed 30th August 2012
7. The National Intelligence Model (2000), National Criminal Intelligence Service, http://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=1&cad=rja&ved=0CCAQFjAA&url=http%3A%2F%2Fwww.intelligenceanalysis.net%2FNational%2520Intelligence%2520Model.pdf&ei=E9pqUJyjDoPB0QXBw4GACQ&usg=AFQjCNFQosPKkpMvj9RRKNhW9WJpeuzMLQ. Accessed on 2nd October 2012
8. Shearer, C.: The CRISP-DM model: the new blueprint for data mining. J. Data Warehous. **5**, 13–22 (2000)
9. Witten, I.H., Frank, E., Hall, M.A.: Data Mining Practical Machine Learning Tools and Techniques. Morgan Kaufman, San Mateo (2011)
10. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., Steinberg, D.: Top 10 algorithms in data mining. Knowl. Inf. Syst. **14**(1), 1–37 (2008)
11. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2, vol. 12, pp. 1137–1143. Morgan Kaufmann, San Mateo (1995)
12. Adderley, R.: The use of data mining techniques in active crime fighting. In: International Conference on Computer, Communication and Control Technologies and the 9th International Conference on Information Systems Analysis and Synthesis, Orlando, 31 July–3 August 2003, pp. 356–361 (2003)
13. A E Solutions (BI) Ltd (2012) Authority miner, http://a-esolutions.com/index.php?id=authority-miner. Accessed 4th September 2012
14. IBM SPSS Modeler (2012), http://www-01.ibm.com/software/analytics/spss/products/modeler. Accessed 2nd October 2012
15. WEKA (2012), http://www.cs.waikato.ac.nz/ml/weka. Accessed 2nd October 2012

16. SAS (2012), http://www.sas.com/technologies/analytics/datamining/index.html. Accessed 2nd October 2012
17. Kohonen, T.: Self organising formation of topologically correct feature maps. Biol. Cybern. **43**(1), 59–69 (1982)
18. Home Office Counting Rules For Recorded Crime, HMSO 2012
19. Brantingham, P., Brantingham, P.: Crime pattern theory. In: Wortley, R., Mazerolle, L. (eds.) Enviromental Criminology and Crime Analysis, pp. 78–93 (2008)

Chapter 8
Business Planning and Support by IT-Systems

Klaus Freyburger

Abstract Business planning is one of the basic tasks of corporate management. Although a detailed presentation of all different business aspects is beyond the scope of this chapter some important characteristics of business planning will be presented. The main focus will be on support by IT-systems, starting by identifying different areas like modeling and manual planning. Then different system categories used for planning purposes will be compared. Important examples are spreadsheets and OLAP based systems. Last but not least some fundamental concepts within planning systems will be discussed, like handling of hierarchies within dimensions and modeled calculations. For this purpose, some implementations will be outlined using software of SAP, Microsoft and the open source solution Palo.

8.1 Basics of Business Planning

Business planning as mental anticipation of future action is one of the basic tasks of corporate management and has been subject of numerous publications by researchers and practitioners in the past. Although a detailed presentation of all different business aspects and trends[1] is beyond the scope of this chapter some important characteristics of business planning will be presented.

The concept of planning is characterized by a particular complexity with multiple dimensions like matter, organization, elements, characteristics and time [2].

Partial plans are widely used because most of the leading companies primarily are planning in the different functional areas, such as sales, manufacturing, procurement, human resources and/or finance.

The interaction of different planning areas can be illustrated by an example based on [3] which is shown in Fig. 8.1. In this example of a manufacturing company, starting point of planning activities could be a sales plan, whereas the goal is to obtain

[1] As a recent important trend the beyond budgeting initiative can be mentioned, cf. [1].

K. Freyburger (✉)
Hochschule Ludwigshafen am Rhein, Fachbereich III, Ernst-Boehe-Str. 4, 67059 Ludwigshafen, Germany
e-mail: klaus.freyburger@hs-lu.de

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_8, © Springer-Verlag London 2013

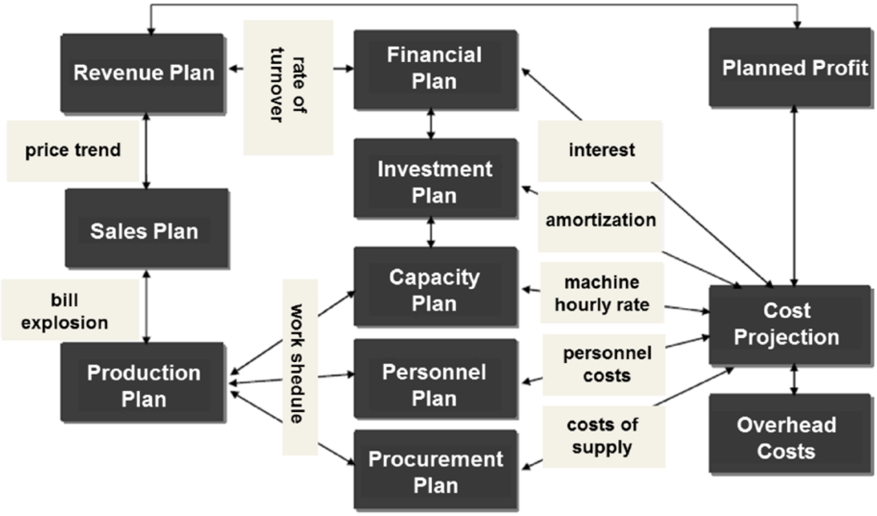
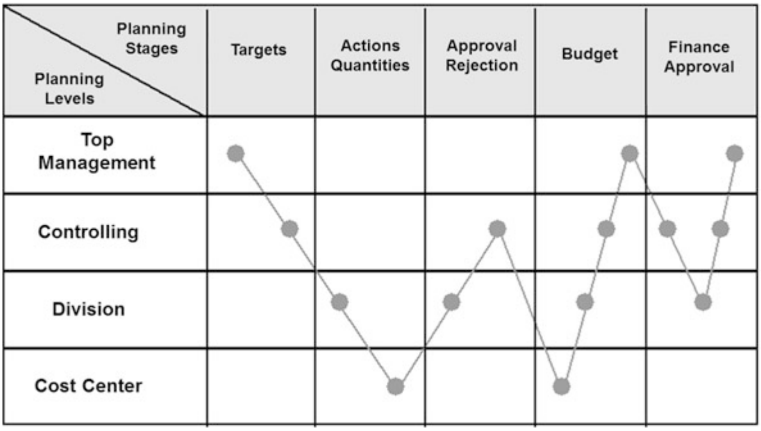


Fig. 8.1 Example of different planning areas interacting

planned profit. The edges in the graph represent the rules to connect different planning areas. E.g., the revenue plan can be derived from the sales plan by multiplying with sales prices and taking sales deductions into account.

Besides, content and organization planning is determined by the planning process. There are different main methods distinguished, namely top-down approach, bottom-up method and a combination of both. An example of the mixed approach is shown in Fig. 8.2, which distinguishes different levels within the organization as well as different stages [2].



Planning Stages / Planning Levels	Targets	Actions Quantities	Approval Rejection	Budget	Finance Approval
Top Management	●			●	●
Controlling	●		●	●	● ●
Division		●	●	●	●
Cost Center		●		●	

Fig. 8.2 Top–down/bottom–up planning

However it should be noted that these are only examples. So, in order to adapt arbitrary planning models it is necessary not to have fixed structures but a flexible framework [4].

8.2 Areas of Software Support

In the following, rather than focusing on specific functional planning areas, like sales or manufacturing, some general generic issues will be discussed which occur in many planning areas.

One approach is to divide the requirements concerning flexible planning software into the following groups [5]:

- Modeling,
- Manual planning/analysis,
- Automatic planning functions and
- Process control.

These areas will now be discussed in more detail.

8.2.1 Modeling

The planning model has to reflect the structure of corporate/organizational model.

Of course, normally, a data model exists which reflects the operational business processes. However, in many cases this data model is not appropriate for planning purposes for several reasons.

The first reason is granularity. E.g., let's look at an IT consulting company which is divided in different business units each of those might reflect a consulting topic. In the revenue planning process it is common to plan the revenue for the most important customers by customer and business unit, however all other customers should be planned on an aggregated level.

Second, the planning data has to anticipate structural changes which occur in the future. A simple example could be a new product that should be planned the following period. A more complicated example might be a reorganization of customers to sales regions or profit centers to business lines. This shows a possible conflict of goals. On the one hand planning software should use the existing organizational models. However, any disturbance of operational systems has to be avoided.

Third, planning often has a simulative aspect. Because of this, it is a good practice to add a planning version to the data model which can reflect different scenarios like base case, optimistic etc. Last but not least the planning model needs flexibility for rapid adjustment in case of model changes.

In real life, the setup of a planning model is quite a comprehensive task. An example of an adequate approach can be found in [6].

8.2.2 Manual Planning/Analysis

Business planning is a very creative process, and planning software should support the user in this sense as much as possible. Here are some examples how software can support in this process.

A sales manager has to achieve a yearly revenue target by planning sales volume where sales prices are planned by a product manager in a preceding planning step. Then the planning system should calculate and accumulate revenue after entering sales volume.

A sales manager has to plan special discounts which he wants to enter as percentage share of revenue whereas in a database normally the storage of absolute values is appropriate. So again, after storing the entered values a calculation rule has to be executed.

Planning has to happen on a monthly basis and the system should support in distributing yearly to monthly figures by appropriate algorithms, like seasonal factors. However, the planning user should have the possibility to override system proposals.

8.2.3 Automatic Planning Functions

In most cases, planning does not start from scratch but is based on past figures or external data. Very often actual figures of the previous year (sometimes combined with a very simple calculation, like adding a fixed percentage) serve as a starting point of the planning process. In some cases, statistical forecast methods are applied to predict measures, like sales volume. Another example is the support of disaggregation of measures, e.g. yearly values should be allocated to months by certain rules in order to allow a plan/actual comparison in the preceding year.

Since the calculation methods to be applied are specific by company and planning area the planning software has to provide a framework to define arbitrary calculations within the planning model. An example of this will be shown below.

8.2.4 Process Control

In many companies there are quite a lot of participants who are contributing to a planning process, cf. the process example which was shown in Fig. 8.2 at the beginning of this chapter. So, there has to be a planning coordinator (in most cases a controller) who is monitoring the planning process in general and the progress of all tasks to be done by the planning users. Software support at this point should include collaboration possibilities as well as options to represent an approval/rejection workflow.

8.3 System Categories

In this section we will look at different approaches to support business planning including spreadsheets, planning on the basis of ERP Systems, special planning applications and flexible OLAP-based planning systems.

8.3.1 Spreadsheets

Spreadsheets, like Microsoft Excel, are widely used in this context. The usage concept is very intuitive and well-known by most users. By means of the combination of measures and formulas, calculation models can be set up very fast. This technique can be used to implement what-if scenarios, since any change enforces a recalculation of the entire model. However, when models get large spreadsheets can be confusing since formulas are hidden in the background.

Additionally, data and model logic is combined in one local document which creates big problems in collaborative scenarios. Therefore, in most cases to ensure parallel access and for security reasons multiple documents are used, one per planning user. This makes model changes very difficult and requires a more or less complicated consolidation process. Although, most spreadsheet software includes some data connection techniques it is quite a challenge to combine data of different sources in real life. A more detailed discussion on this topic can be found in [7] and [8].

8.3.2 ERP Systems

ERP systems are widely used to support all business processes in a company and, therefore, incorporate a very detailed data and process model, which potentially can be used for planning purposes. However, planning on a very detailed level is not appropriate in many cases. In addition, there may be additional restrictions in ERP systems which limit some planning options.

To illustrate the pros and cons an example from cost center planning will be considered in the following [9]. As a prerequisite for standard cost accounting, it is necessary to plan cost, activities and cost center allocation on a very detailed level in order to be able to determine transfer prices (rates) even before the actual costs are known. In this case, it is useful to align all sender/receiver rules with rules used in actual data flow. On the other hand looking at SAP ERP, the world wide leading ERP system, it is difficult to implement a rolling cost center planning, because the fiscal year within the planning system has an essential role. E.g., it is not possible to plan the next 12 month beginning June of the current year.

Planning within SAP ERP is characterized by different planning solutions, which are optimized for specific planning stages. The integration of these solutions to an overall plan is a sequential process that consists of several additional steps. Therefore, to establish an integrated, enterprise-wide total plan is very complex and only of limited use for simulation purposes.

In a talk at a SAP user conference Prof. Sven Piechota expressed this as follows [10]: “I only know of three controllers, who claim to create their business plan using SAP ERP. In two cases only the final plans figures are entered in SAP ERP, the actual planning is carried out in advance using other systems. The third one actually has done the whole planning using SAP ERP. However, he got sick. He will never do his planning this way again.”

8.3.3 OLAP Systems

OLAP systems are widely used in the management support context for multidimensional analysis. Although optimized at first for data retrieval, some software vendors established techniques for data input in OLAP cubes. When enriched with additional functions to implement calculation logic, this is a very promising approach to implement a company-wide modeled planning. Some aspects on this will be discussed later.

8.3.4 Special Planning Applications

Special planning and budgeting software is available, and vendors promise ready-to use planning models with intuitive usage concepts and quick implementation times. This mostly works for standardized planning areas, like balance sheet planning. However, the planning process in a lot of companies and planning areas is very special, so that the usefulness of standard software with pre-configured models is limited.

8.3.5 Guideline

Comparing the different software categories it is also very important not only to look at initial implementation but application life cycle as well. E.g., spreadsheets are very flexible at configuration time but very difficult to maintain in the operating stage. A comparison of the different software categories in terms of flexibility during life cycle can be found in an illustration of Oehler [7], cf. Fig. 8.3. Although being purely qualitative and maybe being subject to discussion in some details this graph shows very concise and compact some fundamental trends. Since the detailed discussion of the various aspects is beyond the scope of this chapter, it is referred to the original source at this point.

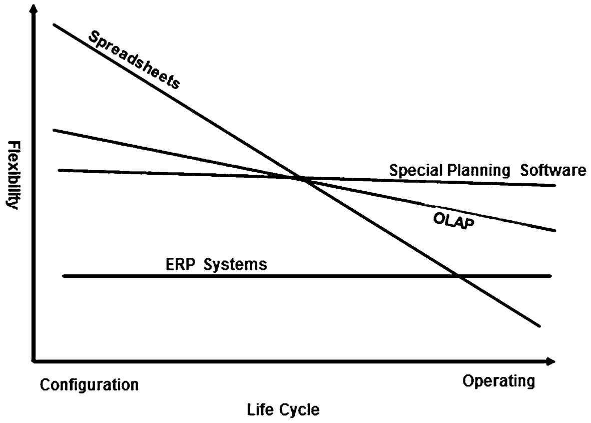


Fig. 8.3 Flexibility of planning software in life cycle

As summary, it can be stated as a kind of general guideline. Planning based on spreadsheets is suitable for individual, singular planning problems which require a high degree of flexibility. ERP Systems are advantageous if very detailed planning analogous to the methods and billing structures of the actual data is required. Special planning software allows rapid deployment of pre-configured planning models. OLAP is the first choice if company-wide modeled planning is needed. Therefore, the following paragraphs focus on OLAP-based planning systems.

8.4 Important Concepts and Sample Implementations

In this section, some important concepts within OLAP-based planning systems will be discussed. One key issue before starting the modeling task is to understand how a planning system handles hierarchies within dimensions, e.g. the entry of a planning value on an aggregated level. This will be demonstrated by using software of SAP and Microsoft. Moreover, an example of an end-user interface will be shown. For this, an open source solution, called "Palo", is used. Finally, a technique to set up arbitrary calculations will be introduced.

8.4.1 Hierarchies in OLAP Based Planning Systems

8.4.1.1 SAP BW Integrated Planning

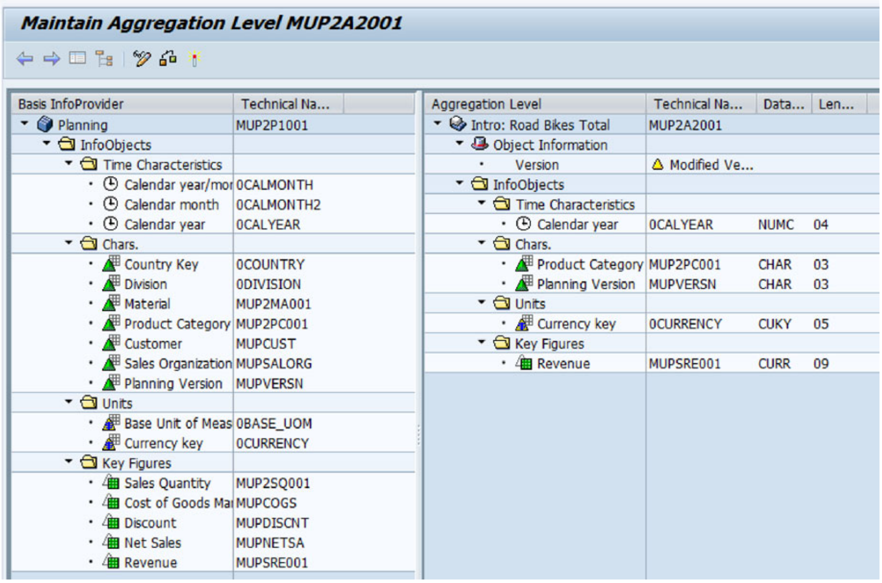
SAP BW Integrated Planning is fully integrated into SAPs business intelligence solution SAP BW and provides business experts with an infrastructure for realizing and operating planning scenarios. Planning covers a wide range of topics from simple data entry to complex planning scenarios. A detailed description of the various concepts has been subject to some other publications (e.g. see [11]) and is beyond the scope of this section. In the following a simplified business example will be presented to demonstrate the handling of hierarchies. In this example, a sales manager of a bike selling company plans revenue for most important bicycles per product. All other bikes will be planned as a total on the product category level.

In order to model this scenario, two so called aggregation levels, which determine the sections of an OLAP cube used in planning, have to be defined. The following Fig. 8.4 shows the definition of the aggregation level for planning product categories with the underlying OLAP cube on the left hand side and the attributes for the aggregation level product category on the right hand side. To keep it simple, only yearly values for a single measure revenue are considered.

The planning proceeds as follows. In a first step, the sales manager plans revenue for the three most important products cf. (1) in Fig. 8.5. Then the sales manager looks at product category level where the system adds up the planned revenue figures of 300, 100 and 200 to 600 (not seen in figure). Next the sales manager enters the estimated total revenue for race bikes as 1,000 (2). The difference of 400 is not distributed to the existing products planned in step (1) but remains on a special system generated product Not Assigned (3). However, if required the 400 can be distributed by appropriate rules (4) (result not shown in figure).

It can be noted that this behavior is kind of special to SAP BW Integrated Planning. In most other planning systems (like the ones discussed later in this chapter) values entered on a node level are distributed to existing leaf members.

OLAP based planning with SAP BW Integrated Planning normally begins with a modeling task, where appropriate OLAP cubes and aggregation levels have to be identified. The next step is to define appropriate rules to connect these items to

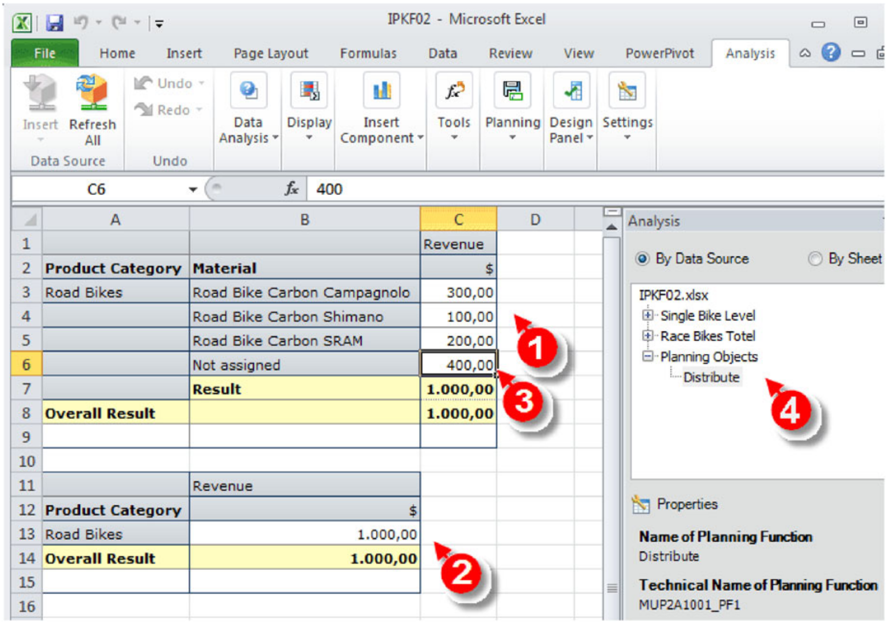


Maintain Aggregation Level MUP2A2001

Basis InfoProvider	Technical Na...
Planning	MUP2P1001
InfoObjects	
Time Characteristics	
Calendar year/mor	0CALMONTH
Calendar month	0CALMONTH2
Calendar year	0CALYEAR
Chars.	
Country Key	0COUNTRY
Division	0DIVISION
Material	MUP2MA001
Product Category	MUP2PC001
Customer	MUPCUST
Sales Organization	MUPSALORG
Planning Version	MUPVERSN
Units	
Base Unit of Meas	0BASE_UOM
Currency key	0CURRENCY
Key Figures	
Sales Quantity	MUP2SQ001
Cost of Goods Mai	MUPCOGS
Discount	MUPDISCNT
Net Sales	MUPNETSA
Revenue	MUPSRE001

Aggregation Level	Technical Na...	Data...	Len...
Intro: Road Bikes Total	MUP2A2001		
Object Information			
Version	△ Modified Ve...		
InfoObjects			
Time Characteristics			
Calendar year	0CALYEAR	NUMC	04
Chars.			
Product Category	MUP2PC001	CHAR	03
Planning Version	MUPVERSN	CHAR	03
Units			
Currency key	0CURRENCY	CUKY	05
Key Figures			
Revenue	MUPSRE001	CURR	09

Fig. 8.4 Definition of aggregation level product category



C6 400

	A	B	C	D
1			Revenue	
2	**Product Category**	**Material**	$	
3	Road Bikes	Road Bike Carbon Campagnolo	300,00	
4		Road Bike Carbon Shimano	100,00	
5		Road Bike Carbon SRAM	200,00	
6		Not assigned	400,00	
7		**Result**	**1.000,00**	
8	**Overall Result**		**1.000,00**	
9				
10				
11		Revenue		
12	**Product Category**	$		
13	Road Bikes	1.000,00		
14	**Overall Result**	**1.000,00**		
15				
16				

Name of Planning Function
Distribute

Technical Name of Planning Function
MUP2A1001_PF1

Fig. 8.5 Planning steps in SAP BW Integrated Planning

end up with a company-wide modeled plan. At the end of this chapter, it will be discussed, how this can be achieved.

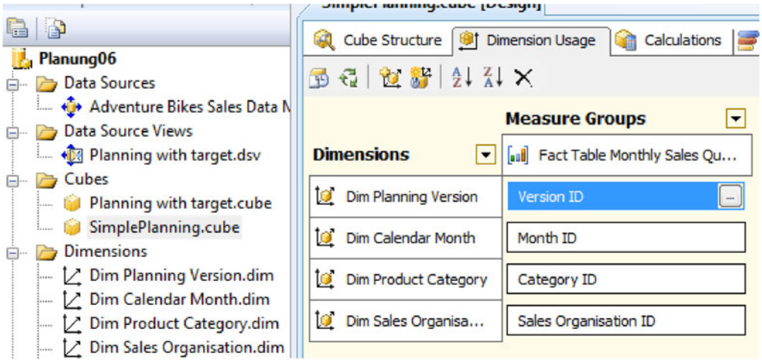
8.4.1.2 Microsoft SQL Server Analysis Services

Although Microsoft does not claim to incorporate a planning solution in its product portfolio, an example based on [12] which contains a planning process supported by using Microsoft SQL Server 2008 R2 in conjunction with Excel 2010 Pivot tables will be shown.

Again, a simple example is presented which is similar to the example above. However, the lowest level of detail of the time dimension is a month and concerning the product dimension only the product category is considered. The modeling task starts by defining an ordinary OLAP cube with dimensions and measures in Microsoft SQL Server Analysis Services, as shown in Fig. 8.6. For this, SQL Server Business Intelligence Studio is used. It is an extension of visual studio, which is widely used in the Microsoft development context.

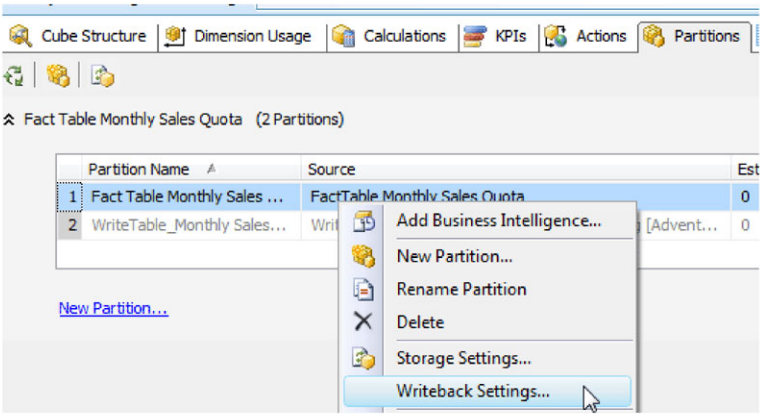
The next step is to enable this cube for data input. For this, a so called write back partition has to be created as is shown in Fig. 8.7.

Since Excel version 2010, a pivot table based on this cube can be defined which can be enabled for data input and write back to the OLAP cube. An example will be shown below.



Dimensions	Fact Table Monthly Sales Qu...
Dim Planning Version	Version ID
Dim Calendar Month	Month ID
Dim Product Category	Category ID
Dim Sales Organisa...	Sales Organisation ID

Fig. 8.6 OLAP cube in Microsoft SQL Server Analysis Services



	Partition Name	Source	Est
1	Fact Table Monthly Sales ...	FactTable Monthly Sales Quota	0
2	WriteTable_Monthly Sales...	Wri [Advent...	0

Fig. 8.7 Definition of write back partition

Whenever a user of a planning software changes a measure on a leaf level of the cube the system determines the difference between old and new value and saves this delta in an SQL Server database. This very simple scenario already might facilitate the planning process in many organizations. Currently, many users of planning software enter the planning data in local excel sheets which they have obtained from the planning administrator. Then a complicated collection and consolidation process of all excel workbooks is necessary. Using the new write back technique, data directly can be submitted to an OLAP cube by users, so that consolidation is done automatically by the OLAP system.

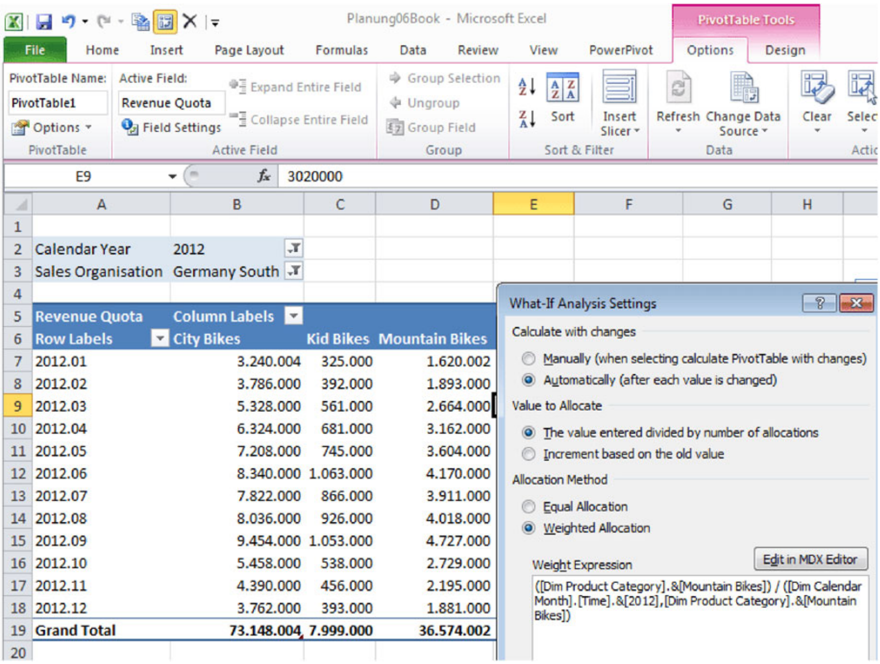
The interesting question is how the system behaves on data entry on a node level. In contrary to SAP BW Integrated Planning, the MS SQL Server always distributes data to the leaf level of the OLAP cube, even if the leafs are not shown in the pivot table. In real life this is critical for the system setup, since there might be quite a large number of leafs which can cause long system run times and memory shortage.

So, normally, special OLAP cubes have to be designed which are used for planning purposes.

By default data distribution to the leaf level is done equally. So, e.g., when distributing data from total years to the monthly level, every month has the same value. Since this is not appropriate in all cases, there is a possibility to define special distribution rules using Multidimensional Expressions (MDX), which is a language for querying OLAP databases, quite similar to SQL which is a query language for relational databases. Using the MDX rule, shown in Fig. 8.8, any total yearly revenue value entered leads to same monthly distribution as for the mountain bikes. In this example the total yearly value of city bikes was entered, and therefore any monthly value for city bikes doubles the corresponding value for mountain bikes.

Although planning users might not be familiar with MDX, it is good practice to have an MDX expert as part of every Microsoft BI project who can assist setting up those rules before distributing the spreadsheet to the planning user [13].

As described above, in most cases, planning does not start from scratch but is based on past figures which often have higher granularity. In Microsoft's BI solution a technique can be used which combines different fact tables (in Microsoft terminology: Measure Groups) with dimensions. In Fig. 8.9 an example is shown where the actual data is on the product level (Fact Table Monthly Sales) whereas the plan data is on a product category level (Fact Table Sales Target). Whenever appropriate this principle can be applied to other dimensions as well.



PivotTable Name: PivotTable1

Active Field: Revenue Quota

E9 3020000

	A	B	C	D
1				
2	Calendar Year	2012		
3	Sales Organisation	Germany South		
4				
5	Revenue Quota	Column Labels		
6	Row Labels	City Bikes	Kid Bikes	Mountain Bikes
7	2012.01	3.240.004	325.000	1.620.002
8	2012.02	3.786.000	392.000	1.893.000
9	2012.03	5.328.000	561.000	2.664.000
10	2012.04	6.324.000	681.000	3.162.000
11	2012.05	7.208.000	745.000	3.604.000
12	2012.06	8.340.000	1.063.000	4.170.000
13	2012.07	7.822.000	866.000	3.911.000
14	2012.08	8.036.000	926.000	4.018.000
15	2012.09	9.454.000	1.053.000	4.727.000
16	2012.10	5.458.000	538.000	2.729.000
17	2012.11	4.390.000	456.000	2.195.000
18	2012.12	3.762.000	393.000	1.881.000
19	Grand Total	73.148.004	7.999.000	36.574.002
20				

What-If Analysis Settings

Calculate with changes

○ Manually (when selecting calculate PivotTable with changes)
● Automatically (after each value is changed)

Value to Allocate

● The value entered divided by number of allocations
○ Increment based on the old value

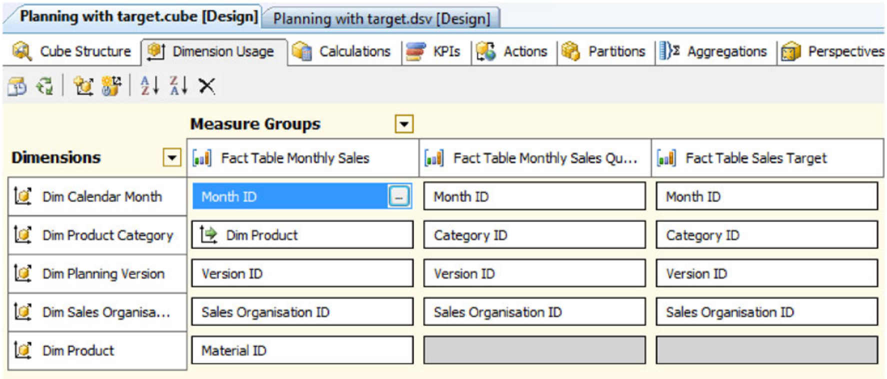
Allocation Method

○ Equal Allocation
● Weighted Allocation

Weight Expression [Edit in MDX Editor]

```
([Dim Product Category].&[Mountain Bikes]) / ([Dim Calendar Month].[Time].&[2012],[Dim Product Category].&[Mountain Bikes])
```

Fig. 8.8 Allocation by rule using MDX

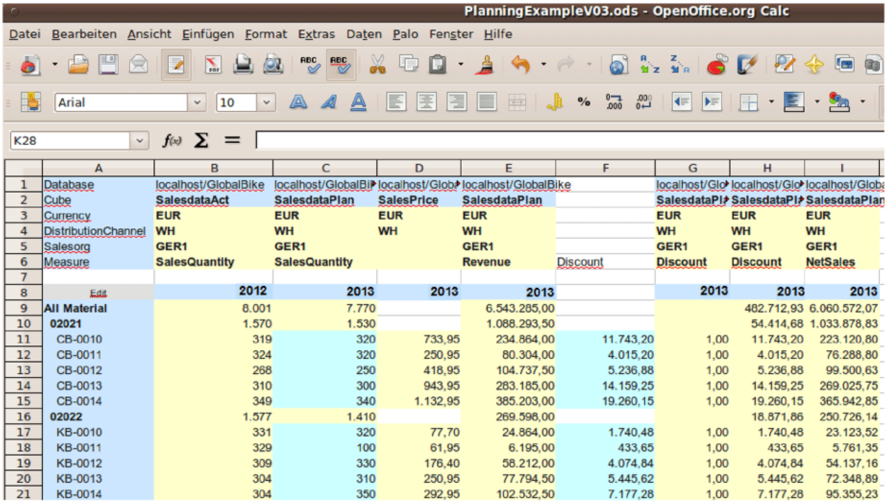


	Measure Groups		
Dimensions	Fact Table Monthly Sales	Fact Table Monthly Sales Qu…	Fact Table Sales Target
Dim Calendar Month	Month ID	Month ID	Month ID
Dim Product Category	Dim Product	Category ID	Category ID
Dim Planning Version	Version ID	Version ID	Version ID
Dim Sales Organisa…	Sales Organisation ID	Sales Organisation ID	Sales Organisation ID
Dim Product	Material ID		

Fig. 8.9 Fact tables with different granularity

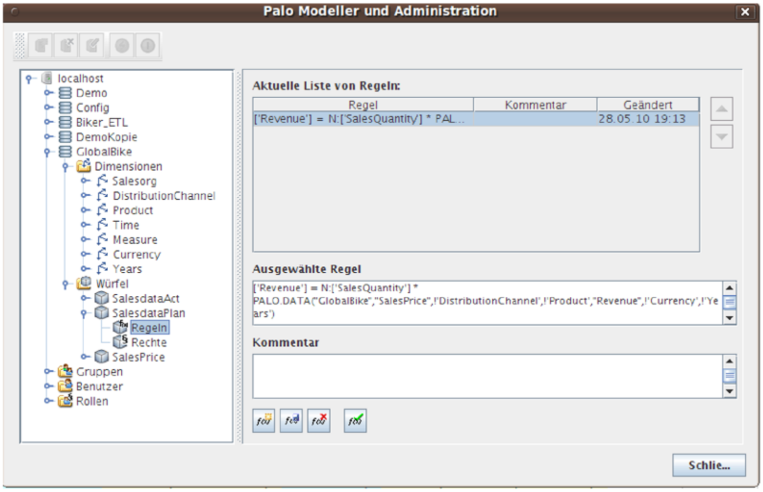
8.4.2 User Interface for Manual Planning

In the following an example of a bit more elaborated user interface for manual planning will be considered [14]. One interesting aspect of this example will be the interaction of a centrally provided planning model with local formulas created by the planning user. For this the open source OLAP server Palo [15] will be used, that offers write back functionality similar to the Microsoft offering sketched above for quite some time. Palo has add-ins for Microsoft excel as well as open source spreadsheet LibreOffice Calc, which is shown in Fig. 8.10.


PlanningExampleV03.ods - OpenOffice.org Calc
Datei Bearbeiten Ansicht Einfügen Format Extras Daten Palo Fenster Hilfe
Arial
10
K28
f(x) Σ =

	A	B	C	D	E	F	G	H	I
1	Database	localhost/GlobalBike	localhost/GlobalBi	localhost/Glob	localhost/GlobalBike		localhost/Glo	localhost/Glo	localhost/Globa
2	Cube	SalesdataAct	SalesdataPlan	SalesPrice	SalesdataPlan		SalesdataPl	SalesdataPl	SalesdataPlan
3	Currency	EUR	EUR	EUR	EUR		EUR	EUR	EUR
4	DistributionChannel	WH	WH	WH	WH		WH	WH	WH
5	Salesorg	GER1	GER1		GER1		GER1	GER1	GER1
6	Measure	SalesQuantity	SalesQuantity		Revenue	Discount	Discount	Discount	NetSales
7									
8	Edit	2012	2013	2013	2013		2013	2013	2013
9	All Material	8.001	7.770		6.543.285,00			482.712,93	6.060.572,07
10	02021	1.570	1.530		1.088.293,50			54.414,68	1.033.878,83
11	CB-0010	319	320	733,95	234.864,00	11.743,20	1,00	11.743,20	223.120,80
12	CB-0011	324	320	250,95	80.304,00	4.015,20	1,00	4.015,20	76.288,80
13	CB-0012	268	250	418,95	104.737,50	5.236,88	1,00	5.236,88	99.500,63
14	CB-0013	310	300	943,95	283.185,00	14.159,25	1,00	14.159,25	269.025,75
15	CB-0014	349	340	1.132,95	385.203,00	19.260,15	1,00	19.260,15	365.942,85
16	02022	1.577	1.410		269.598,00			18.871,86	250.726,14
17	KB-0010	331	320	77,70	24.864,00	1.740,48	1,00	1.740,48	23.123,52
18	KB-0011	329	100	61,95	6.195,00	433,65	1,00	433,65	5.761,35
19	KB-0012	309	330	176,40	58.212,00	4.074,84	1,00	4.074,84	54.137,16
20	KB-0013	304	310	250,95	77.794,50	5.445,62	1,00	5.445,62	72.348,89
21	KB-0014	304	350	292,95	102.532,50	7.177,28	1,00	7.177,28	95.355,23

Fig. 8.10 Sales planning using Palo and LibreOffice Calc



Aktuelle Liste von Regeln:

Regel	Kommentar	Geändert
['Revenue'] = N:['SalesQuantity'] * PAL...		28.05.10 19:13

Ausgewählte Regel

```
['Revenue'] = N:['SalesQuantity'] *
PALO.DATA("GlobalBike","SalesPrice",!'DistributionChannel',!'Product',"Revenue",!'Currency',!'Ye
ars')
```

Kommentar

Fig. 8.11 Datamodel and central rule in Palo OLAP server

The business scenario assumes that a manger of a sales organization wants to plan sales quantities (column C) and discounts (column F) on a product level. To keep it simple, only annual values are considered.

The data in this example is combined from different cube sources. While the plan data is stored in a cube SalesdataPlan, there is also a cube SalesdataAct involved, which has a higher granularity in time dimension, and serves as a source for the actual data of the last year in column B. Column D shows sales prices, which were planned by a central department and a stored in a cube SalesPrice. Column E shows the revenue which is calculated by a central rule. The corresponding formula processes data of the cubes SalesdataPlan (quantity) and SalesPrice (cf. Fig. 8.11 right hand side).

Column F is used by sales manager to plan discount values. Here he might want to use local formulas involving revenues. Column G is used for write back discounts to OLAP server. In real life situations, this column will be hidden in most cases. Finally, columns H and I show discount and centrally calculated net sales. An overview of the data model with dimensions and cubes is shown on the left hand side of Fig. 8.11.

By means of the combination of local formulas with centrally provided cubes and rules, a sales manager can do the planning using known environment and tools avoiding the main disadvantages of spreadsheet based planning.

8.4.3 OLAP Based Calculations

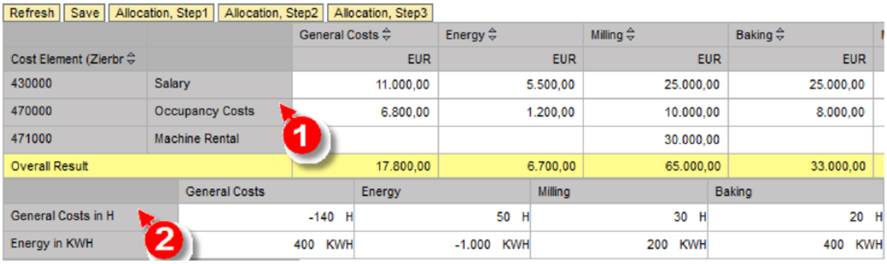
As described before, one of the strengths of OLAP based planning systems is the possibility to model dependencies between different planning areas to obtain an overall company-wide planning model which can be used for simulations. In this section an example will be shown how this can be achieved using SAP BW Integrated Planning. This example is based on a textbook which original intention is to teach the controlling module of SAP ERP [16]. It is intended to plan the profit of a fictitious company which is manufacturing garden fountains by means of techniques of multidimensional contribution accounting.

At the beginning, a planning user can enter primary costs (1) and activities (2) for different cost centers as shown in Fig. 8.12.

Other important modeled domains are the sales plan for the three different fountains including sales quantity and sales price, the bill of materials, and the work schedule (not shown in figure).

The result of the planning task is a contribution margin schema (cf. Fig. 8.13) and a cost center report as shown in Fig. 8.14.

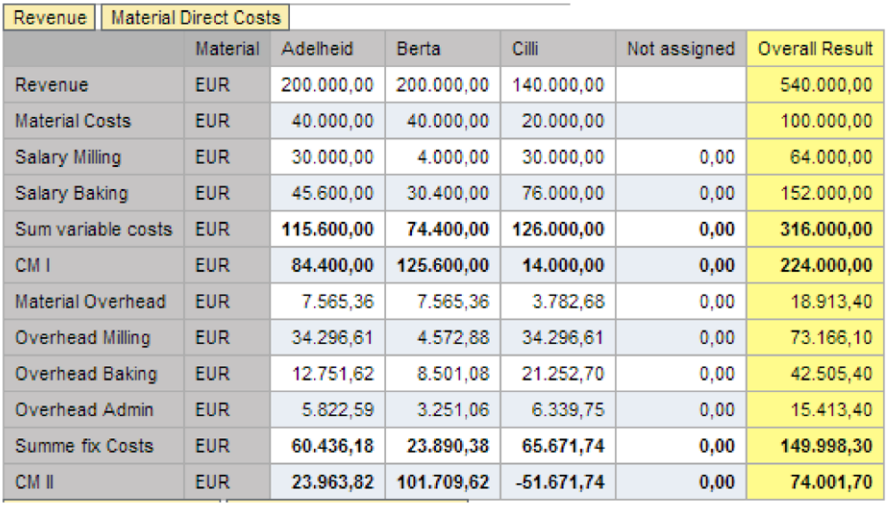
Now, the interesting part is the connection of the different areas by means of formulas. SAP BW Integrated Planning allows any desired user-defined calculation



Cost Element (Zierbr		General Costs EUR	Energy EUR	Milling EUR	Baking EUR
430000	Salary	11.000,00	5.500,00	25.000,00	25.000,00
470000	Occupancy Costs	6.800,00	1.200,00	10.000,00	8.000,00
471000	Machine Rental			30.000,00	
Overall Result		17.800,00	6.700,00	65.000,00	33.000,00

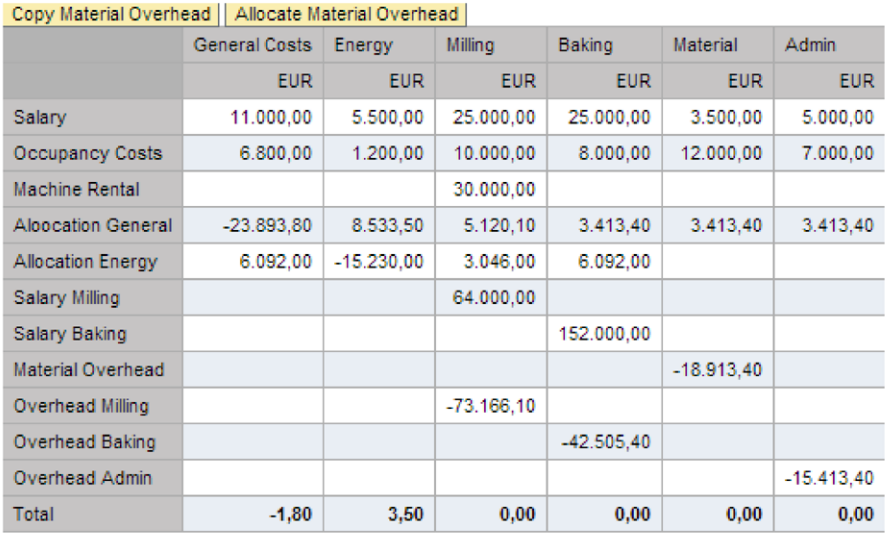
	General Costs	Energy	Milling	Baking
General Costs in H	-140 H	50 H	30 H	20 H
Energy in KWH	400 KWH	-1.000 KWH	200 KWH	400 KWH

Fig. 8.12 User interface for multidimensional contribution planning, part 1



	Material	Adelheid	Berta	Cilli	Not assigned	Overall Result
Revenue	EUR	200.000,00	200.000,00	140.000,00		540.000,00
Material Costs	EUR	40.000,00	40.000,00	20.000,00		100.000,00
Salary Milling	EUR	30.000,00	4.000,00	30.000,00	0,00	64.000,00
Salary Baking	EUR	45.600,00	30.400,00	76.000,00	0,00	152.000,00
Sum variable costs	EUR	**115.600,00**	**74.400,00**	**126.000,00**	**0,00**	**316.000,00**
CM I	EUR	**84.400,00**	**125.600,00**	**14.000,00**	**0,00**	**224.000,00**
Material Overhead	EUR	7.565,36	7.565,36	3.782,68	0,00	18.913,40
Overhead Milling	EUR	34.296,61	4.572,88	34.296,61	0,00	73.166,10
Overhead Baking	EUR	12.751,62	8.501,08	21.252,70	0,00	42.505,40
Overhead Admin	EUR	5.822,59	3.251,06	6.339,75	0,00	15.413,40
Summe fix Costs	EUR	**60.436,18**	**23.890,38**	**65.671,74**	**0,00**	**149.998,30**
CM II	EUR	**23.963,82**	**101.709,62**	**-51.671,74**	**0,00**	**74.001,70**

Fig. 8.13 User interface for multidimensional contribution planning, part 2

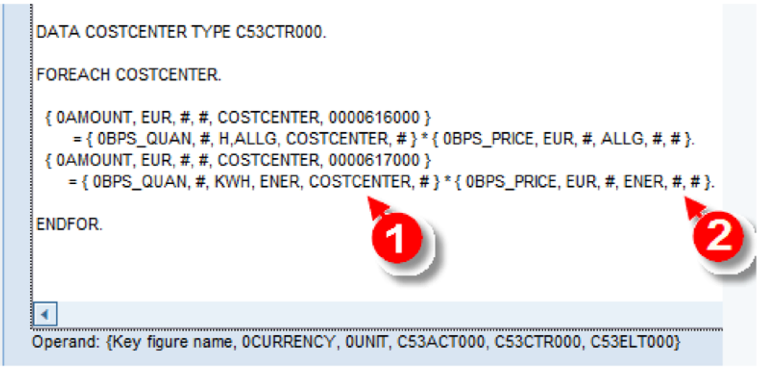


	General Costs	Energy	Milling	Baking	Material	Admin
	EUR	EUR	EUR	EUR	EUR	EUR
Salary	11.000,00	5.500,00	25.000,00	25.000,00	3.500,00	5.000,00
Occupancy Costs	6.800,00	1.200,00	10.000,00	8.000,00	12.000,00	7.000,00
Machine Rental			30.000,00			
Aloocation General	-23.893,80	8.533,50	5.120,10	3.413,40	3.413,40	3.413,40
Allocation Energy	6.092,00	-15.230,00	3.046,00	6.092,00		
Salary Milling			64.000,00			
Salary Baking				152.000,00		
Material Overhead					-18.913,40	
Overhead Milling			-73.166,10			
Overhead Baking				-42.505,40		
Overhead Admin						-15.413,40
Total	**-1,80**	**3,50**	**0,00**	**0,00**	**0,00**	**0,00**

Fig. 8.14 User interface for multidimensional contribution planning, part 3

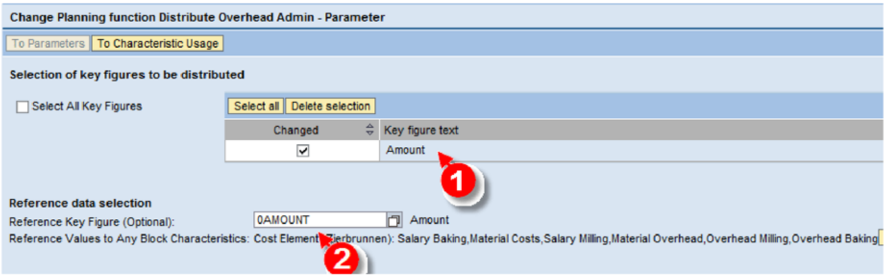
within an OLAP cube. For each formula it can be configured which dimensions should serve as operands.

When allocating costs within cost center, accounting cost centers are debited according to their activity consumption. On the bottom line of Fig. 8.15 it is shown how the operands are composed in this case, namely starting with the technical name of key figure followed by Currency, Unit of Measure, Activity, Cost center and Cost element. Now, there are two calculations that are executed for all cost centers ("FOREACH COSTCENTER"). The lower formula calculates cost element 61,700 (energy cost) as consumption quantity ("0BPS_QUAN") of activity "ENER" multiplied by consumption price ("0BPS_PRICE") of energy. Note that the con-


```
DATA COSTCENTER TYPE C53CTR000.

FOREACH COSTCENTER.

 { 0AMOUNT, EUR, #, #, COSTCENTER, 0000616000 }
      = { 0BPS_QUAN, #, H,ALLG, COSTCENTER, # } * { 0BPS_PRICE, EUR, #, ALLG, #, # }.
 { 0AMOUNT, EUR, #, #, COSTCENTER, 0000617000 }
     = { 0BPS_QUAN, #, KWH, ENER, COSTCENTER, # } * { 0BPS_PRICE, EUR, #, ENER, #, # }.

ENDFOR.
```

Operand: {Key figure name, 0CURRENCY, 0UNIT, C53ACT000, C53CTR000, C53ELT000}

Fig. 8.15 Cube formula calculation



Change Planning function Distribute Overhead Admin - Parameter

Selection of key figures to be distributed

□ Select All Key Figures

Changed	Key figure text
☑	Amount

Reference data selection

Reference Key Figure (Optional): 0AMOUNT Amount

Reference Values to Any Block Characteristics: Cost Element (Zierbrunnen): Salary Baking,Material Costs,Salary Milling,Material Overhead,Overhead Milling,Overhead Baking

Fig. 8.16 Allocation by reference data

sumption quantity is specific to the corresponding cost center ("COSTCENTER") (1), whereas the consumption price is independent on cost center ("#" which means not assigned) (2).

Another example of an OLAP cube based calculation is the distribution of overhead administration costs according to reference data, in this case direct costs. Here instead of a formula a predefined allocation function is used as shown in Fig. 8.16. This allows distributing arbitrary values from one aggregation level (cf. Sect. 8.4.1 "*Hierarchies in OLAP Based Planning Systems*") to a lower level.

8.5 Summary

The software support of business planning will be an area of growing interest by researchers and professionals in the near future. Currently, spreadsheets, like Microsoft Excel, are widely used in this context, but there is a lot of potential for improvement. Most promising approaches for company-wide modeled planning are solutions based on OLAP technology which ease the adaption of arbitrary planning models. As a key issue, it was discussed (using SAP and Microsoft technology) how planning systems handle the entry of planning values on aggregated levels. Some requirements on an end-user interface were shown using the open source solution Palo. Finally, it was demonstrated how to set up OLAP based calculations in order to model dependencies between different planning areas. This can lead to a company-wide planning model for simulation purposes.

References

1. Hope, J., Fraser, R.: Beyond Budgeting. Harvard Business Review Press, Boston (2003)
2. Freyburger, K., Lehmann, P., Seufert, A., Zirn, W., Grasse, S., Suhl, C.: Unternehmensplanung mit SAP BW (2005). Steinbeis edition, Stuttgart
3. Gluchowski, P., Gabriel, R., Dittmar, C.: Management Support Systeme und Business Intelligence. Springer, Heidelberg (2008)
4. Oehler, C.: Beyond Budgeting, was steckt dahinter und was kann Software dazu beitragen? Kostenrechnungspraxis **46**(3), 151–160 (2002)

5. Freyburger, K., et al.: BI1 & BI2: SAP NetWeaver business Warehouse and SAP business explorer (2008). https://cw.sdn.sap.com/cw/docs/DOC-7223. Accessed 06 July 2012
6. Fischer, R.: Business Planning with SAP, SEM. Galileo Press, Fort Lee (2004)
7. Oehler, C.: Unterstützung von Planung, Forecasting und Budgetierung durch IT-Systeme. In: Chamoni, P., Gluchowski, P. (eds.) Analytische Informationssysteme, 3rd edn. Springer, Berlin (2006)
8. Rasmussen, N., Eichorn, C.J.: Budgeting. Wiley, New York (2000)
9. Seufert, A., Lehmann, P., Freyburger, K.: Zukunftsorientierte Unternehmenssteuerung auf der Basis von Business Intelligence – Herausforderungen und Potenziale für das Controlling. Controller-Leitfaden. Weka Verlag, Zürich (2006)
10. Brück, U.: (1999) Käsehersteller plant sein Geschäft mit R/3. In: Computerwoche Nr. 5/1999. http://www.computerwoche.de/1079849. Accessed 06 July 2012
11. Srinivasan, K., Srinivasan, S.: SAP NetWeaver BI Integrated Planning for Finance. Galileo Press, Boston (2007)
12. Business intelligence competence network. http://www.bicn.info/BICN-EN/. Accessed 06 July 2012
13. Mundy, J., Thornthwaite, W.: The Microsoft Data Warehouse Toolkit. Wiley, Indianapolis (2011)
14. Freyburger, K.: Anwendungsszenarien. In: Haneke, U., et al. (eds.) Open Source Business Intelligence (OSBI): Möglichkeiten, Chancen und Risiken quelloffener BI-Lösungen. Carl Hanser Verlag, München (2010)
15. Palo Open Source business intelligence. www.palo.net. Accessed 06 July 2012
16. Friedl, G., Hilz, C., Pedell, B.: Controlling mit SAP R/3. Vieweg, Braunschweig (2002)

Chapter 9
Planning Purchase Decisions with Advanced Neural Networks

Hans Georg Zimmermann, Ralph Grothmann, and Hans-Jörg von Mettenheim

Abstract In this chapter we investigate a typical situation of a corporate treasurer: on an ongoing basis some kind of transaction is performed. This may be a regular monthly investment in equities for a pension plan, or a fixed income placement. It might be a foreign exchange transaction to pay monthly costs in another currency. Or it could be the monthly supply of some commodity, like fuel or metal.

All these cases have in common that the treasurer has to choose an appropriate time for the transaction. This is the day on which the price is the most favorable. Ideally, we want to buy at the *lowest* price within the month, and we also want to invest our money at the *highest* available interest rate.

This problem is complex, because the underlying financial time series are not moving independently. Rather, they are interconnected. In order to truly understand our time series of choice, we have to model other influences as well: equities, currencies, interest rates, commodities, and so on. To achieve this we present a novel recurrent neural network approach: Historically Consistent Neural Networks (HCNN). HCNNs allow to model dynamics of entire markets using a state space equation: $s_{t+1} = \tanh(W \cdot s_t)$. Here, W represents a weight matrix and s_t the state of our dynamic system at time t. This iterative formulation easily produces multi step forecasts for several time points into the future.

We analyze monthly purchasing decisions for a market of 25 financial time series. This market approximates a world market: it includes various asset classes from Europe, the US, and Asia. Our benchmar, an averaging strategy, shows that using HCNNs to forecast an entry point for ongoing investments results in better prices for every time series in the sample.

H.G. Zimmermann · R. Grothmann
Siemens AG Munich, Corporate Technology, Munich, Germany

H.G. Zimmermann
e-mail: hans_georg.zimmermann@siemens.com

R. Grothmann
e-mail: ralph.grothmann@siemens.com

H.-J. von Mettenheim (✉)
Institut für Wirtschaftsinformatik, Leibniz Universität Hannover, Hannover, Germany
e-mail: mettenheim@iwi.uni-hannover.de

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_9, © Springer-Verlag London 2013

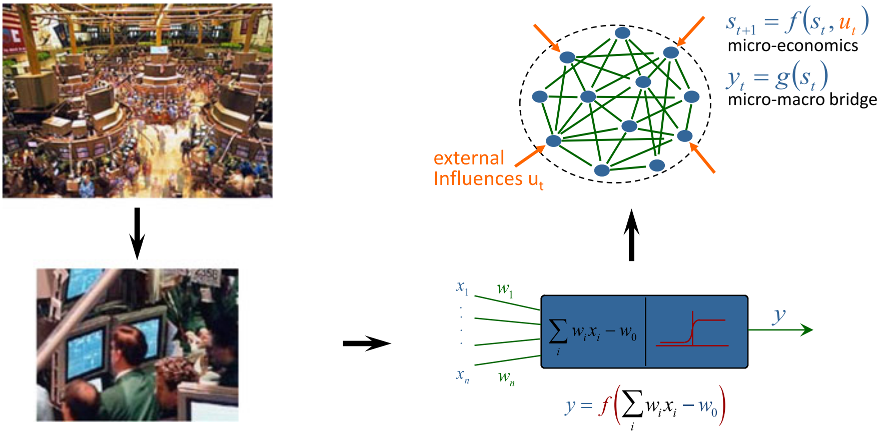


Fig. 9.1 Schematic representation of the analogy between markets, decision makers, single neurons and a neural network

9.1 Introduction and Motivation

It is an important task in today's BI to mine huge amounts of data and extract knowledge. This task is even more important when considering the spread of data warehouses. At the end the generated knowledge should be integrated into a decision support system (DSS) that helps the company's decision makers. In the present chapter we focus on a specific task and on specific data. The task at hand is to build a DSS for a corporate treasurer with regular purchases of a traded asset. The data we are considering are time series data from financial markets. A typical regular purchase concerns raw material in the case of industrial production. Another possibility is the regular purchase of a financial asset, for example a stock index fund when the treasurer invests money for a pension plan regularly. In both cases we cannot assume the treasurer to be particularly familiar with market timing techniques. The DSS we propose helps the treasurer in finding an appropriate entry time.

For this analysis we will use a novel tool, the Historically Consistent Neural Network (HCNN). Before we present this tool in detail we want to illustrate our motivation for this complex task of system identification, see Fig. 9.1. The system we want to identify is composed of economic markets, symbolized by a trading floor. Many agents are acting there. They may influence each other and also get some common influences from observable time series. As this system is *very* complex, we try to break it down to individual agents, symbolized by a single trader. Now, imagine what this trader is doing. Information systems give him millions of time series to choose from. Then, there are fellow traders, whose behavior he may observe. His first action is to narrow his focus. That is, he selects a few (perhaps a dozen) series he wants to observe. Observing alone, does not make him any profits. He aggregates his observable series into a decision model. This is symbolized by the formula in the lower right of Fig. 9.1. This formula is a weighted sum of his observables x_i minus a bias value w_0. This sum then passes through a squashing function f which

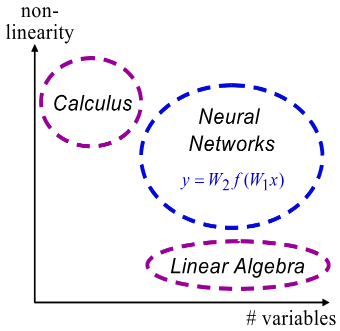


Fig. 9.2 Neural networks as powerful tool to model complex systems

is used to produce some kind of confidence. For example: will the price of gold rise or fall? Or: should I buy now or rather later? However, computing a confidence still does not make any money. So, the last step is to act, that is to output his decision to the world. Other agents will again pick up this decision. This is symbolized by the graphical network representation in the upper right part of the figure. Although the presented model is quite general it is an accurate (yet simplified) description of a neural network. The basic building blocks of neural networks are a concatenation of linear algebra and a non-linearty, symbolized in the figure by the function f.

The figure also introduces the notion of state, s_t, which is used to compute the following state s_{t+1}. We may also have external influences u_t and can think of state s as something that is not directly visible to us. Rather, it's via the micro-macro bridge $g(s_t)$ that we get an observable reaction y_t. The micro-macro bridge is a function g that translates and aggregates individual agents' behavior ("micro" economics) into an observable behavior of the world ("macro" economics). This illustrates the correspondence between economical dynamical systems and neural networks.

However, there is also a more technical argument, why we propose a neural network to model a complex system, see also Fig. 9.2. When we consider a complex dynamical system a system that

- includes many variables and
- moves along a state space trajectory according to unknown rules

we can see, why a neural network is a good fit to model it. On the one hand, we can model highly complex systems with few variables using calculus. Mostly one uses Taylor expansions (or some related method) to achieve this. A series expansion tries to model a function as a (typically infinite) power series, see for example [5], p. 671. However, these methods rapidly become cumbersome in high dimensions. On the other hand, we are also very good at modeling linear systems with a high amount of variables. But then, think again: do you believe that the economy out there moves linearly? A common argument is that the system should behave linearly as a first approximation, because the first derivative is used in the Taylor expansion. However, one often forgets to add that this linear behavior is only valid around the expansion point of our model. The question remains: do we believe that our economy is in equilibrium and that its state simply oscillates around our expansion point? If we neither believe in linear behavior nor in an equilibrium economy we should look at neural networks. By using simple building blocks (neurons, illustrated in the lower right part of Fig. 9.1) they allow to model the behavior of complex systems with many variables. By the simple concatenation of linear algebra and a (non-linear) squashing function like the hyperbolica tangent neural networks are universal approximators that are able to approximate any differentiable function on a compact domain, see [4]. A neural network is able to deal with high-dimensionality and nonlinearity at the same time. For that reason it is positioned in the upper right part of the figure.

Our DSS uses an advanced neural network. However, there are other interesting alternatives for dealing with complex systems which we won't discuss here. These include, for example, evolutionary methods like Genetic Programming.

After this brief motivation we will shortly introduce the mathematics of HCNN in Sect. 9.2. Section 9.3 presents the dataset we use to model the corporate treasurer's decision problem and discusses practical considerations of data preprocessing for model building. Then, Sect. 9.4 outlines results of applying HCNN to our dataset to forecast the best moment within the next twenty days to buy an asset. Finally, Sect. 9.5 outlines limitations of our work and concludes.

9.2 Historically Consistent Neural Networks

To understand the concept behind HCNN we still have to introduce another aspect of neural networks: recurrency. The most commonly known neural network, the three layer perceptron is not recurrent. That means, that output values do not influence further computations. The three layer perceptron is a kind of *feedforward* neural network. However, when modeling time series, recurrency naturally arises: recurrency describes the concept that network output may influence the computation at succeeding time steps. This feature is particularly useful because it mostly frees the modeler from the obligation of determining input time lags to the model. Indeed, there are various statistical tools for determining lagged inputs to a time series model. Often it is difficult to give a convincing explanation (based in the economy and not on statistical test values) why the model includes such and such lag of a time series. With recurrency the training process determines the amount of lag that is necessary to best approximate the dynamical system.

HCNN were first introduced by [11]. In the following several studies analyze HCNN in more detail, see [1, 7–9, 12] for a selection of follow-up literature. Essentially, a HCNN is a state space based model that follows the simple equation

$$s_{t+1} = W \tanh(s_t) \tag{9.1}$$

where s_t is the state vector at time t and W is a weight matrix of weights that we want to optimize (that is learn or train). Figure 9.3 illustrates several iterations of Eq. (9.1). State vector s is typically very high-dimensional, for example $\dim(s) = 500$ in the following application. The reason for this is that it fulfills a

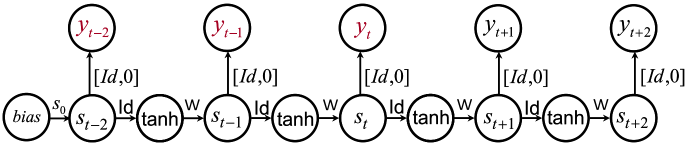


Fig. 9.3 The basic HCNN architecture

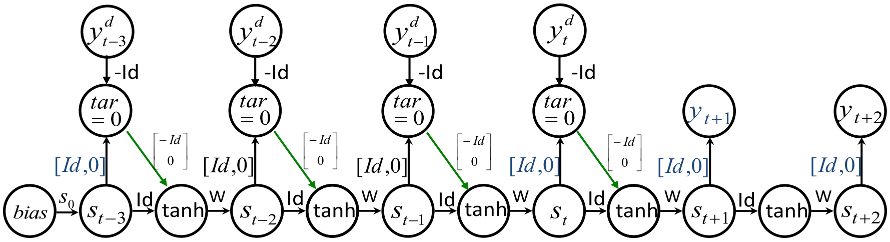


Fig. 9.4 The extended HCNN architecture with teacher forcing converges to the basic HCNN architecture

double role: on the one hand it contains our observable time series in its upper part. We can extract this upper part (y_t in Fig. 9.3) by applying an appropriate identity matrix and a zero matrix to state vector s. We can only train on y_t, obviously, as we don't know how the other members of s_t should look. The other members are hidden variables. Functionally, we can think of hidden variables as representing other unknown time series which we would need to model but can't observe. Technically, hidden variables have the purpose to act as a memory.

We have to consider an important aspect of weight matrix W. To achieve proper learning behavior this matrix has to be sparsely populated. In our application we use a sparsity of 12.5 percent. That means than only one eights of the weights in W are initialized to a random non zero value. All other weights are set and kept at zero during the entire learning process. Sparsity is necessary because otherwise a single large number would propagate to all other states within a few iterations and cause a numerical overflow. While a sparsity of 12.5 percent is generally a good starting value, [11, 12] explain in more detail how to choose an appropriate level of sparsity.

The basic architecture in Fig. 9.3 is nicely suited for illustrative purposes. However, it does not converge well. To remedy this we use the extended architecture of Fig. 9.4 which converges towards the basic architecture when learning terminates. The key is to set the HCNN back on track at every iteration by feeding it the observed values (if we know them) instead of burdening this task entirely on the learning algorithm. We call this procedure teacher forcing. It allows for faster convergence (or may facilitate convergence at all) because teacher forcing keeps the state space on the right track. All formulae related to learning can be found in [7].

So now: how do we generate a forecast once we have found a suitable weight matrix W by learning? We simply iterate further, as illustrated by the states $s_{t+1}, s_{t+2}, \ldots$ in Figs. 9.3 and 9.4. In this chapter's application we use twenty iteration steps to forecast the next twenty days of the dynamics of our observables *without* updating the network in between. So this is a true 20 day ahead forecast. We have successfully used up to 60 forecast iterations.

As we train HCNN until a very low error, we cannot differentiate among models based on training error. Every model that converges looks good a priori. Without hindsight we cannot decide which is the *true* model. For this reason we do not use a *single* model but rather an *ensemble* of HCNN. For the following application we train 200 HCNN (with different random sparse initializations) and take the average as forecast value. The ensemble width is a measure of uncertainty: it represents forecast *and* model uncertainty. A more thorough discussion on uncertainty can also be found in [12].

9.3 Dataset

Our dataset approximates the observables that a treasurer of a multi-national corporation faces, see Table 9.1. By design, it is a *world model*. That means that we acknowledge that it is probably impossible to establish a clear causal relationship between the observables found in the table. For this reason we include the most important stock indices, relevant long and short term interest rates, the most liquid foreign exchange rates and a few world-wide traded commodities. All observables in our dataset are likely to be needed at some time or another by the treasurer. For example, a pension plan may require regular purchases of stock index exchange traded funds. Correctly assessing interest rate movements is important for credit and investment decisions. A good timing on the exchange rates may be important to decide when to repatriate gains or pay foreign contractors. And, finally, forecasting, e.g., freight rates as per the BDI (Baltic Exchange Dry Index) is important when purchasing shipping capacity. In any case, we may be more confident with a model that suitably forecasts *all* the observables, rather than with separate models which are specialized on a *single* observable but do not perform well on the others. The goal of the treasurer is to buy (or sell, but we will focus on *buy* for now) an asset at the best (lowest) price within the next, say, 20 days. In doing this, the treasurer has no constraints. Also, we will not consider interest credited on free funds.

Although the reason for inclusion of most observables may be clear because of their worldwide following and importance, we will still motivate some of them. We include the Kospi Index of South Korea because it is considered a leading-indicator. The economy of South Korea is significantly biased towards electronics and even more importantly ship-building. A slow-down in worldwide economic activity—so goes the argument—should in principle impact South Korea's economy first, because insiders know that less goods will ship. Relatedly, we include the BDI for a similar reason. This index represents freight rates for dry shipments. A slow-down in economic activity tends to be accompanied by softening freight rates.

When we model market dynamics it is important that we avoid (or at least are aware of) a look-ahead bias in our data. As all values in our dataset are close values

Table 9.1 The time series used in the following application, ordered by instrument type. The given region indicates not necessarily where the instrument is quoted but rather that it's important and widely followed in the given area. Refer to the text for further explanation and especially time zone considerations. Note that we condense long and short term interest rates to a yield curve. We quote the Datastream mnemonic to allow readers to follow along with their own investigations

Name	Instrument	Region	Datastream
FTSE 100 Index	Equities	United Kingdom	FTSE100
DAX 30 Index	Equities	Germany	DAXINDX
CAC 40 Index	Equities	France	FRCAC40
FTSE MIB	Equities	Italy	FTSEMIB
Dow Jones Euro Stoxx 50	Equities	Europe	DJES50I
S&P 500 Index	Equities	United States	S&PCOMP
NASDAQ 100 Index	Equities	United States	NASA100
Nikkei 225 Index	Equities	Japan	JAPDOWA
Kospi Index	Equities	South Korea	KORCOMP
3 months LIBOR	Interest rate	United Kingdom	ECUK£3M
12 months LIBOR	Interest rate	United Kingdom	BBGBP12
Germany 3 months	Interest rate	Germany	ECWGM3M
France 3 months	Interest rate	France	ECFFR3M
Italy 3 months	Interest rate	Italy	ECITL3M
EURIBOR 3 months	Interest rate	Euro area	ECEUR3M
Eurodollars 3 months	Interest rate	United States (Europe)	ECUS$3M
Benchmark Bond 3 months	Interest rate	Japan	ECJAP3M
Benchmark Bond 10 years	Interest rate	United Kingdom	UKMBRYD
Bund Future 10 years	Interest rate	Germany	BDBRYLD
Benchmark Bond 10 years	Interest rate	France	FRBRYLD
Benchmark Bond 10 years	Interest rate	Italy	IBRYLD
US Treasuries 10 years	Interest rate	United States	USBD10Y
Benchmark Bond 10 years	Interest rate	Japan	JPBRYLD
US Dollar to Great British Pound	Exchange rate	United Kingdom	USDOLLR
US Dollar to Swiss Franc	Exchange rate	Switzerland	SWISFUS
US Dollar to Euro	Exchange rate	Euro area	USEURSP
Yen to US Dollar	Exchange rate	Japan	JAPAYE$
Gold Bullion	Commodity	United Kingdom (world)	GOLDBLN
Brent Crude Oil	Commodity	Europe (world)	OILBREN
CRB Index	Commodities	United States (world)	NYFECRB
Baltic Exchange Dry Index	Commodities	world	BALTICF

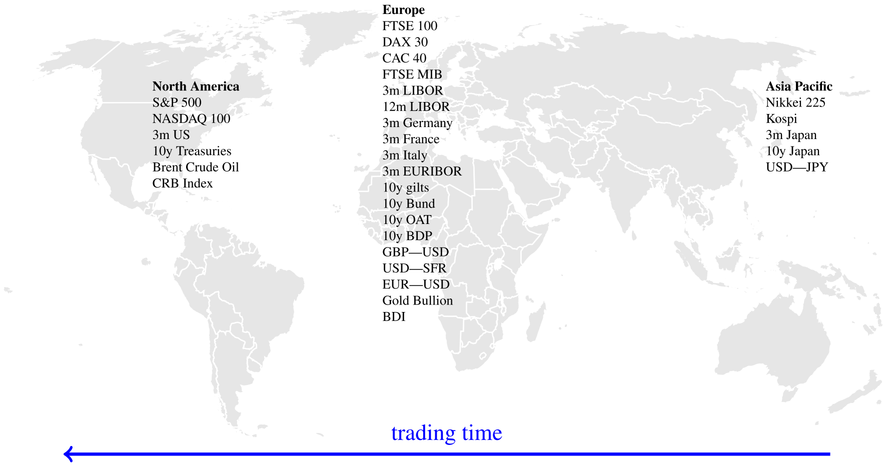


Fig. 9.5 Schematic representation of data used for subsequent analysis, ordered roughly geographically by *source*. Note that you have to be careful when considering which lags to use: the trading day begins in Asia Pacific in the East, goes to Europe and ends in North America. At the end of trading in North America Asia Pacific again takes over with a *new* trading day. Consequently, when using time series for forecasting only data which comes from a market place *East* of the series to forecast should be taken into account for the same day. As a rule of thumb Asia Pacific data is available at 11.00 GMT, European data at 17.00 GMT and North America data at 22.00 GMT

they do not occur at the same time of day. The Asian close happens first, then the European, and finally the American. Figure 9.5 illustrates this on a world map. That means that we are not allowed, strictly speaking, to use American or European data to forecast the next movement for Asian data. In our case this bias does not unduly influence the results, because we are forecasting over a time period of twenty days. The bias would only be relevant for the very next forecasting step. To avoid a look-ahead bias we could lag, for example, European and American data by one time step when forecasting Asian data. We tried the three possible combinations and could not find any meaningful difference in the results. Therefore we may safely assume that data from the same day are sufficiently close together.

Before we can apply any system identification algorithm to a dataset we generally need some kind of preprocessing of the data. The intention of preprocessing is to highlight the salient features of the dataset and facilitate the following system identification task. You may have already noticed that our dataset is heavily biased towards interest rates. This bias disappears in the final dataset because we build a coarse yield curve by computing the difference of ten year interest rates and corresponding three months interest rates. We only keep 12 months LIBOR and 3 months EURIBOR as independent (non yield-curve) interest rates because these are very widely followed. This operation reduces the total number of observables to 25 for the final model.

Finally, we have to avoid time series with shifting means. A shifting mean leads to distorted learning—not only with neural networks but with most other statistical methods, too. The learning algorithm will simply learn a trend in the data but will

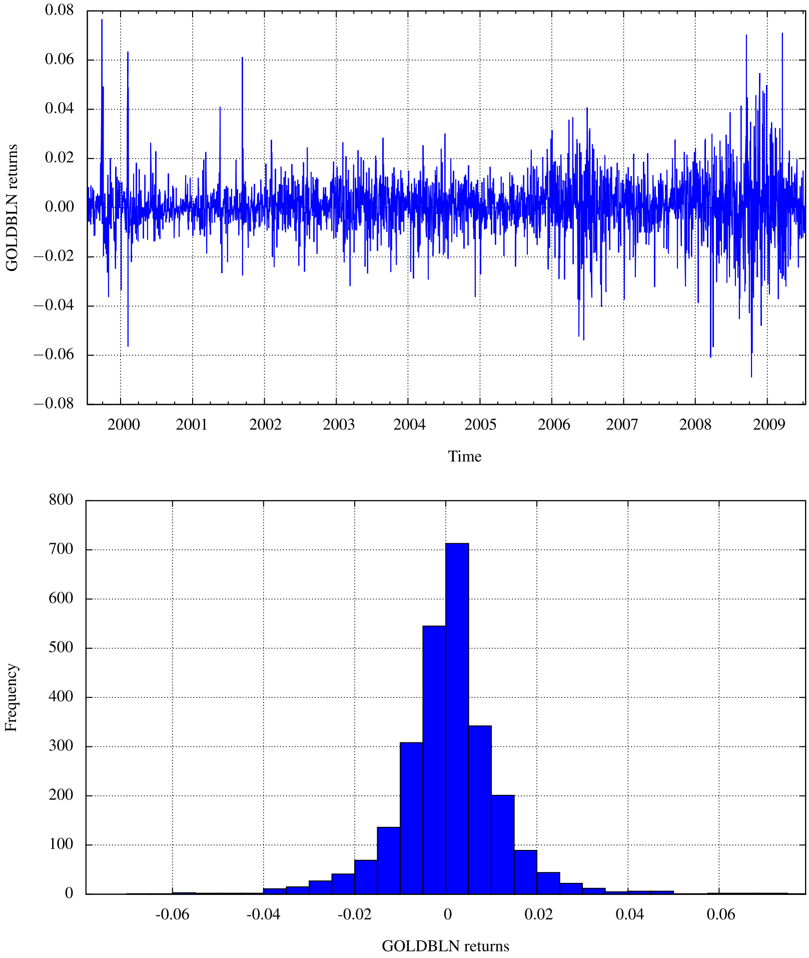


Fig. 9.6 The Gold Bullion daily returns and distribution

generally be unable to react correctly when the trend breaks. To avoid a shifting mean (that is non-stationarity) we make our observable time series stationary by computing simple returns or differences (in the case of interest related time series):

$$r_t := \frac{x_t}{x_{t-1}} - 1 \tag{9.2}$$

where x_t is the level series. This gives us returns centered around 0. The result of applying this transformation to an observable time series can be seen exemplarily on Fig. 9.6 for the Gold Bullion. As a general note keep in mind that a histogram is often more informative than a simple plot of values. In this case we see that the histogram is unimodal and centered at zero. This characteristic is good for the ensuing learning process.

Model building always involves the selection of the time window that will be used to learn a model. In the present case we want to be able to use a model approximately six months. We choose an appropriately long preceding learning period of approximately two years or 440 trading days. Every six months we train a new HCNN ensemble and discard the previous one. Our dataset spans ten years from the beginning of July 1999 till the end of June 2009.

9.4 Results

Recall our application: a corporate treasurer has to purchase (or sell) some kind of asset regularly (monthly in our example). The treasurer has to choose an appropriate time for the transaction. This is, of course, the day on which the price is the most favorable. For example, when we buy, we want to achieve the *lowest* price within the given time frame (one month or twenty days). And we also want to get credit at the lowest possible interest rate. Conversely, we could also look at *highest* prices, but for the following application we will stick to realizing the lowest price, hence taking the position of a buyer. Here, a multi step forecast proves useful because it gives us an idea of the probable price fluctuations.

In the following we look at a 20 day ahead multi step forecast for different assets. The benchmark against which we will evaluate the quality of our market timing is the *realized potential*, *RP*. *RP* is a number between 0 and 1, where 0 indicates that our transaction takes place at the *worst* possible moment. $RP = 1$ indicates a perfect fit, i.e., we get the *best* possible price. We define the twenty day ahead realized potential at time t of a transaction as

$$RP_t(20) := 1 - \frac{p_t^{\text{realized}} - p_t^{\min}(20)}{p_t^{\max}(20) - p_t^{\min}(20)}. \tag{9.3}$$

$p_t^{\max}(20)$ and $p_t^{\min}(20)$ represent the maximum and minimum prices in the twenty day ahead window starting at time t. p_t^{realized} is the price realized when following the forecast at time t. Please note, that p_t^{realized} is *not* the forecast price but the actual price at which it would have been possible to trade. Indeed, there is no obligation for the price to follow our forecast…

We compare *RP* from the twenty day ahead forecast with *RP* resulting from buying on a fixed day in the month, i.e., buying always on the 1st, or on the 15th, and so on. We then use the HCNN ensemble to make a twenty day ahead forecast and evaluate *RP*. Then we update the network with new data from the day (teacher forcing) and move one time step forward, forecast, evaluate *RP* and so on. This is a forecast with a twenty day rolling window. Note, that we only *retrain* the ensemble every six months. The daily updates generally occur with the same weights.

To illustrate how the forecast works we will have a look at two typical examples. First, have a look at Fig. 9.7 which shows a forecast for the Baltic Exchange Dry Index. The forecast starts at day 441, which is the first day on which the HCNN

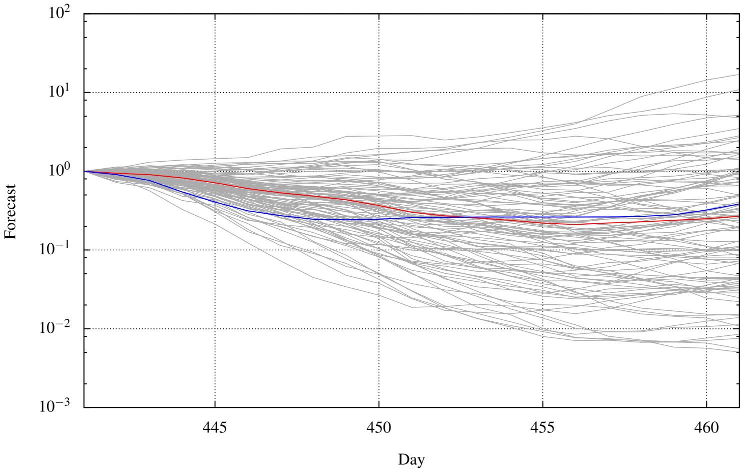


Fig. 9.7 An almost ideal example of a twenty day forecast for the Baltic Dry Index. Realized values in *blue*, forecast in *red*. Additionally the *gray lines* show forecasts of all different networks. Note, how the networks gets the general tendency *right*: first down, then flat, then rising again a little bit. It also suggests a buy at a very sensible low point although the actual value of the low is not quite hit. Keep in mind, that this is a genuine twenty day ahead forecast: the network runs freely for 20 time steps *without* input of realized data. To allow comparisons with other forecasts the values are rebased at 1. That means that the value for the first day of the forecast is set to 1. This also applies for the following figure

ensemble has not been trained. Then, for the next twenty days until day 461, we get level forecasts. These level forecasts have been computed by a back transformation of the return forecasts. All forecasts are rebased at one to facilitate comparisons between different forecasts. Our first observation for the BDI is that the networks follow the general tendency of the realized values. I.e., they first indicate prices to go down, then stay flat for several days and finally go up again slightly.

This has been forecast *without* ever seeing the values between $t = 442$ and $t = 461$. If we wanted to buy dry bulk shipping capacity within the next twenty days: which would be the best day? Clearly, with hindsight, it is $t = 448$ where the BDI hits its low at 1423 points. The HCNN ensemble suggests a buy at $t = 456$ or 1428 points. Note the following:

- The HCNN ensemble avoids the high values at the beginning of the period under consideration.
- The suggested buy point is not very time sensitive: it is a good timing whether you buy one day earlier or later.
- The HCNN ensemble avoids the price rise at the end of the investigated period. For market timing this is very important.

Using concrete values for the BDI we have a high, $p_{441}^{\max}(20) = 1505$ points, right at the beginning of the period. The low occurs at $t = 448$ with $p_{441}^{\min}(20) = 1423$ points. We buy at $p_{441}^{\text{realized}} = 1428$ points. Our realized potential for the BDI using

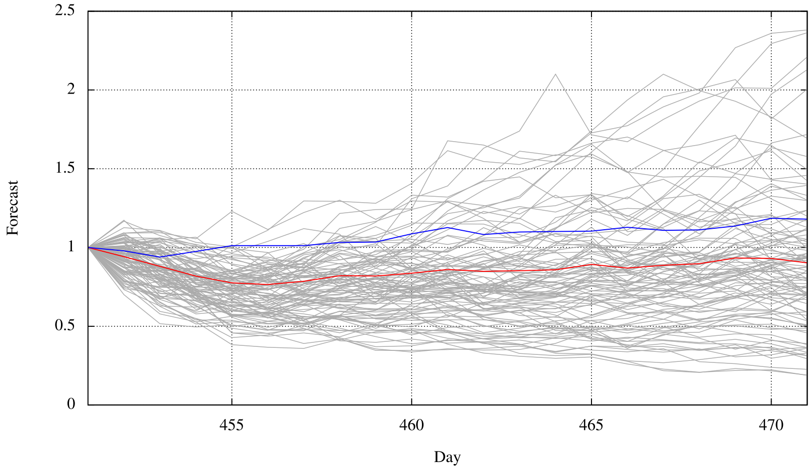


Fig. 9.8 Forecast for the Gold Bullion showing a typical case. The network misses out on the absolute minimum but the suggested buying point on day 456 is not bad at all, considering that the Gold price continues to *rise* afterwards. Again, we note that the network appropriately gets the general tendency *right*, but an overshooting in the first four days causes the values to be skewed. Realized values in *blue*, forecast in *red*. The *gray lines* show forecasts of all different networks

Eq. (9.3) is therefore

$$RP_{441}(20) := 1 - \frac{p_{441}^{\text{realized}} - p_{441}^{\min}(20)}{p_{441}^{\max}(20) - p_{441}^{\min}(20)} = 1 - \frac{1428 - 1423}{1505 - 1423} = 0.94.$$

This means that we achieved 94 percent of the best possible price.

Figure 9.8 shows another forecast. We start at $t = 451$ and forecast the price of the Gold Bullion for the next twenty days, i.e., until $t = 471$. In this case the HCNN suggests a buy at $t = 456$ which is actually not the absolute low but still close to it. We see what happens, when the HCNNs overshoot *at the beginning* of the forecast, here in the direction of low values: as the forecasts are based on *returns* rather than *levels*, small errors at the beginning tend to skew subsequent values.

This is a phenomenon only observable in true multi step forecasts. In our case the HCNN tends to exaggerate the low and only gets the trend right again in the following. And this is substantial: a corporate treasurer basing a decision on the ensemble would still get a feel for where the gold price is headed, even without having an *exact* level forecast.

The treasurer would see that it makes sense to buy in three to six day's time, because the gold price is expected to decrease. She would see that the exact decision is not very time sensitive, because the gold price should stay flat for some time—and it does. But she would also note, that the gold price will rise at the end of the period—and this is indeed what happens. We get *RP* as

$$RP_{451}(20) := 1 - \frac{p_{451}^{\text{realized}} - p_{451}^{\min}(20)}{p_{451}^{\max}(20) - p_{451}^{\min}(20)} = 1 - \frac{258.30 - 257.05}{264.85 - 257.05} = 0.84.$$

We realize 84 percent of the high-low span over twenty days.

The above examples make for nice illustrations. However, two questions remain:

- Do the forecasts outperform the benchmark? That means, do the forecasts beat the typical strategy of a corporate treasurer whose primary goal is *not* to predict the financial markets. Or would a simple strategy of buying always at the same day in a four week cycle perform better? The latter strategy is currently often implemented.
- Are the results consistent over time and over all assets? Or does the model age and looses forecasting power? Do certain assets or certain type of asset classes perform better?

To answer the first question we look at a near term forecast over the next 110 days and calculate the excess realized potential. I.e., we compare the realized potential of the neural networks with the strategy of buying always at some fixed day in a four weeks—or twenty days—cycle. Formally we define the excess realized potential as

$$ERP = RP_{\text{neural}} - RP_{\text{fixed day}} \tag{9.4}$$

with *RP* from the two respective strategies. *ERP* is a value in the range $[-1 \dots 1]$. -1 signals that the simple strategy was perfect but the neural network suggested to buy at the high—the worst possible case. 1 signals the inverse: the simple strategy performed worst and the neural network hit the low exactly. Clearly, we expect $ERP > 0$ for our forecasts to offer any added value. If we consistently had $ERP < 0$ we would be better off not using the forecasts. To get a performance measure for different time spans we calculate the cumulated excess realized potential as

$$cERP = \sum_{t=t_{\min}}^{t_{\max}} ERP_t \tag{9.5}$$

where $t_{\min}$, $t_{\max}$ represent the time span of interest.

The results are shown in Fig. 9.9. To interpret this figure you should first consider that in every case *cERP* is positive. That means, for near term forecasts our model *always* performs better than any fixed day strategy. This is true for every asset. Here, we omit a detailed tabular analysis of the results due to space constraints. The results are available from the authors.

We also note that there are fixed days, mostly in the range 12–18, which show only relatively low *cERP*. We may concede that an *optimized* fixed day strategy can be a hard benchmark to beat with the benefit of hindsight. This is already recognized by [3]. Because of, e.g., futures expiry certain days of a week and of a month will consistently exhibit certain return patterns. Conveniently optimized, a strategy might

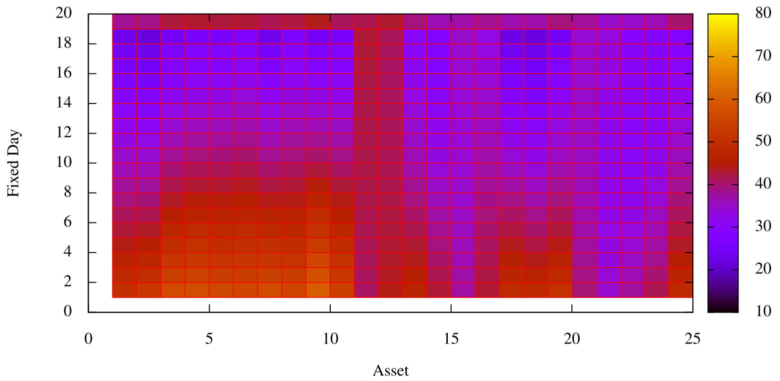


Fig. 9.9 Cumulated yearly excess realized potential for a typical time span of 110 days and different fixed day strategies. Note, that for *every* asset and *every* day of the month the excess realized potential is positive. I.e., the forecasts add value consistently. See also Fig. 9.10 for a comparison to an 8 year forecast

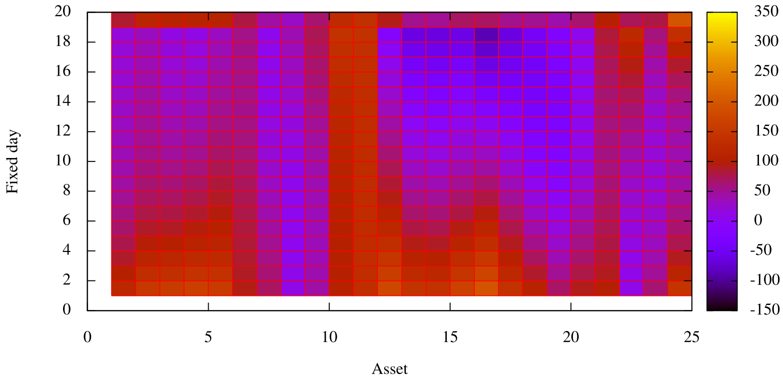


Fig. 9.10 Cumulated excess realized potential for the entire time span of 8 years and different fixed day strategies. What you should take from this figure is that the model is remarkably robust. Several assets still show positive excess realized potential, although in this case the model has only been trained and validated on the first 2 years. Note that, although for some assets and for some days negative excess potential is shown, this is only with the benefit of *hindsight*. We couldn't have possibly known *before*, which day of the month would lead to the best results

simply exploit this although we can doubt if the pattern will *really* be persistent. Obvious patterns are generally rapidly exploited by arbitrageurs.

We now address the second question of model consistency over time. For this we look at *cERP* over a time span of eight years. That means that we train a HCNN ensemble over two years and then use this ensemble for the entire time span *without retraining*. Figure 9.10 presents these results. Generally we note from the figure that often inferior performance occurs clustered at the end of the month. This includes the FTSE 100, Dow Jones Euro Stoxx, Nikkei, German, French, Italian, United

States and Japanese yield curve, GBP|USD, USD|SFR, USD|JPY. A few assets show inferior performance at the beginning of the month: Kospi, EUR|USD and—only to some extent—Brent oil. Several assets are very good performers without any negative values: DAX 30, CAC 40, FTSE MIB, S&P500, NASDAQ 100, 12 months LIBOR and 3 months EURIBOR, UK yield curve, Gold Bullion, CRB Index and Baltic Exchange Dry Index.

We note that from the Western equity indices only FTSE 100 performs inferiorly. On day 19 we also have a small negative value for the Dow Jones Euro Stoxx. The other well known indices all show overwhelmingly good performance. The two indices from Asia Pacific, however, are not so convincing. Still, the best performing days are at least twice as good as the wort performing, and the underperforming days are a minority.

Considering interest rates we note that interest rate differences are easier to forecast than yield curve shifts. Still, in every case the good forecasts outperform the bad. Further analysis if market interaction of other governments is different compared to the United Kingdom could lead to interesting results. For now, its very satisfying to see such good performance for short term interest rates: 3 and 12 months.

We note that currencies are difficult to predict in the long term. Foreign exchange is notorious for having very little exploitable inefficiencies, see [10]. This work confirms the findings. Especially EUR–USD underperforms in a majority of cases. Very interestingly, USD–JPY performs surprisingly well. We only have four negative days and the worst performance of -20.21 is dwarfed by the best performance of 228.87 *cERP*. According to [10], USD–JPY is the second most actively traded currency pair.

Commodities perform very satisfactorily. We only have a small underperformance for oil on five days. All other values are positive. We especially note consistently high values for the CRB index and the Baltic Exchange Dry Index. As commodities *always* fulfill some real economic purpose inefficiencies are not uncommon. Especially the less traded commodities are prone to these, see, e.g., [2, 6]. This is even more true for the BDI: prices are always *real* in the sense that they are always tied to existing or soon to exist capacity. The Baltic Exchange is a market where real capacities are traded by those who need them and those who have them. As demand and supply is inelastic and likely to stay so we are not surprised to see that our network can exploit the inefficiencies of this market. Additionally, an electronically tradeable version of the BDI future is only available since June 2008 through Imarex. It remains to be seen if this market attracts enough quantitative strategies to void returns.

In summary, this section shows the application of an HCNN ensemble to a real world dataset of practical relevance to corporate treasurers. It is intended as a decision support system for the purchasing department and could provide a useful BI tool. Especially, when additional data comes from company data warehouses the non-linear HCNN ensemble of publicly and only privately available information together could yield new insights.

9.5 Conclusion

This chapter investigates how to combine several observable time series into a DSS for purchasing at the right moment. The tool used is an ensemble of HCNN, an advanced neural network. We motivate using a neural network by the analogy of a complex market that breaks down into individual actors. In a neural network simple elements (neurons) combine to build a powerful entity (the neural network). The prevalent characteristic of HCNN ensembles is that they are robust *by design*: each network approximates all observable series equally well. Furthermore, the ensemble width allows to gauge the forecast uncertainty, see Sect. 9.2.

The method also has its limitations which you should be aware of before applying it to your problem domain. The computational complexity currently limits the number of usable observables to approximately forty. In our experience this number leads (together with a state space dimension of 800) to a training time of three hours for an ensemble of 200 members on a modern server with eight cores. As this problem parallelizes well, adding more cores helps almost linearly. Equally, ongoing hardware development plays in favor of including more observables. Note, that only training time is long. Model evaluation is almost instantaneous in less than one second.

Another aspect to be aware of is the selection of observable time series. We did not discuss this in detail and instead motivated series selection by the wish to build a world model. However, there may be situations where you want to be more specific and select series that fit *your* problem particularly well. We describe methods to do this in the referenced literature.

Finally, all means of modeling time series using some kind of historical training set may suffer from structural breaks between training and out-of-sample data. Although HCNNs are quite robust in absorbing external shocks they cannot forecast them. Also, suitable training data may not always be available.

To summarize we can state that in the context of data mining in data warehouses HCNN ensembles offer a solution that is able to perform a multivariate analysis with an adequately complex model without being intractable. HCNN ensembles could therefore offer a valuable addition in the BI toolbox.

References

1. Breitner, M.H., Luedtke, C., von Mettenheim, H.J., Rösch, D., Sibbertsen, P., Tymchenko, G.: Modeling portfolio value at risk with statistical and neural network approaches. In: Dunis, C., Dempster, M., Breitner, M.H., Rösch, D., von Mettenheim, H.J. (eds.) Proceedings of the 17th International Conference on Forecasting Financial Markets, Hannover, 26–28 May 2010. Advances for Exchange Rates, Interest Rates and Asset Management (2010)
2. Dunis, C.L., Laws, J., Evans, B.: Modelling and trading the soybean-oil crush spread with recurrent and higher order networks: a comparative analysis. Neural Netw. World **13**(3/6), 193–213 (2006)
3. Gibbons, M.R., Hess, P.: Day of the week effects and asset returns. J. Bus. **54**(4), 579–596 (1981)

4. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**, 359–366 (1989)
5. Kreyszig, E.: Advanced Engineering Mathematics. Wiley, New York (2011)
6. Lindemann, A., Dunis, C.L., Lisboa, P.: Probability distribution architectures for trading silver. Neural Netw. World **15**(5), 437–470 (2005)
7. von Mettenheim, H.J.: Advanced neural networks: finance, forecast, and other applications. Ph.D. thesis, Faculty of Economics, Leibniz Universität Hannover (December 2009)
8. von Mettenheim, H.J., Breitner, M.H.: Robust forecasts with shared layer perceptrons. In: Dunis, C., Dempster, M., Breitner, M.H., Rösch, D., von Mettenheim, H.J. (eds.) Proceedings of the 17th International Conference on Forecasting Financial Markets, Hannover, 26–28 May 2010. Advances for Exchange Rates, Interest Rates and Asset Management (2010)
9. von Mettenheim, H.J., Breitner, M.H.: Neural network model building: a practical approach. In: Dunis, C., Dempster, M., Girardin, E., Péguin-Feissolle, A. (eds.) Proceedings of the 18th International Conference on Forecasting Financial Markets, Marseille, 25–27 May 2011. Advances for Exchange Rates, Interest Rates and Asset Management (2011)
10. Weithers, T.: Foreign Exchange: a Practical Guide to the FX Markets. Wiley, Hoboken (2006)
11. Zimmermann, H.G.: Forecasting the Dow Jones with historical consistent neural networks. In: Dunis, C., Dempster, M., Terraza, V. (eds.) Proceedings of the 16th International Conference on Forecasting Financial Markets, Luxembourg, 27–29 May 2009. Advances for Exchange Rates, Interest Rates and Asset Management (2009)
12. Zimmermann, H.G.: Advanced forecasting with neural networks. In: Dunis, C., Dempster, M., Breitner, M.H., Rösch, D., von Mettenheim, H.J. (eds.) Proceedings of the 17th International Conference on Forecasting Financial Markets, Hannover, 26–28 May 2010. Advances for Exchange Rates, Interest Rates and Asset Management (2010)

Part IV
Methodologies

Chapter 10
Financial Time Series Processing: A Roadmap of Online and Offline Methods

Daniela Pohl and Abdelhamid Bouchachia

Abstract Because financial information is a vital asset for financial and economic organizations, it requires careful management so that those organizations can enhance and facilitate the decision making process. The financial information is usually gathered over time providing a temporal and historical trace of the financial evolution in the form of time series. The organizations can then rely on such histories to understand, uncover, learn and most importantly make appropriate decisions. The present chapter tries to overview the analysis steps of financial time series and the approaches applied therein. Particular focus is given to the classification of such approaches in terms of the processing mode (i.e., online vs. offline).

10.1 Introduction

Due to its important value, financial information is usually captured in the form of time series (TS) and stored for analysis/evaluation purposes. A time series is an ordered sequence of values of a variable (univariate) or many variables (multivariate) at equally spaced time intervals. Financial TS (FTS) pertain in particular to *stock market analysis*, *budgetary analysis* and *economic forecasting*. The analysis relies on a number of computational techniques with the overall goal of understanding the evolution of time series and finding ways to predict future movements. Usually the analysis follows different activities, like trend detection, recognizing/extracting patterns, finding correlation between TS or similar TS.

Standard analysis methods stem from statistics and probabilities [1]. However, many other computational intelligence methods, like neural networks, evolutionary algorithms, fuzzy logic and chaos theory have been successfully applied [2–5] as well. Generally speaking these techniques and in particular neural networks have

D. Pohl
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
e-mail: daniela.pohl@aau.at

A. Bouchachia (✉)
Bournemouth University, Bournemouth, UK
e-mail: abouchachia@bournemouth.ac.uk

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*, Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_10, © Springer-Verlag London 2013

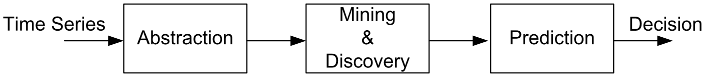


Fig. 10.1 Stages/processing steps of time series analysis

been applied in order to estimate general models without specifying an exact functional form.

Processing TS and especially FTS consists of a certain number of consecutive processing steps: *Abstraction*, *Mining & Discovery* and *Prediction*. Abstraction is about finding a suitable representation of the time series for computational analysis. Mining is an important task aiming at uncovering and extracting/mining knowledge elements such as associations, patterns etc. Discovery aims at finding predefined patterns in TS. Prediction is almost the ultimate goal of TS processing and aims at predicting the future evolution of TS.

Additionally, the algorithms in each processing step can be categorized into *offline* and *online* methods. Considering the former mode, the processing steps are executed in a batch mode, that is, once the data are gathered in the form of historical streams, the analysis is conducted in a one-shot experiment. The knowledge obtained from such a type of analysis is then considered as valid for the future data. In contrast, in the latter type, data are processed sequentially online as they arrive over time. The online processing allows incrementally updating the knowledge extracted from the data. Particularly online processing is desired as financial data changes over time. Thus, it has the potential of coping with non-stationarity. Moreover, online analysis seems appealing because of the sequential nature of the data flow. It is currently emerging quickly giving birth to various adaptations of the offline analysis methods to work online.

The present chapter intends to cover the analysis steps and methods pertaining to offline processing as well as online processing. The steps mentioned earlier will be discussed from both processing perspectives (offline and online) putting focus on the recent research developments in this area.

The chapter starts with the major processing stages without distinguishing between online and offline modes (Sect. 10.2). Section 10.3 gives an overview of computational methods for FTS. The section distinguishes the different algorithms based on the processing mode (offline vs. online) and the processing step (abstraction, mining & discovery, prediction). Sections 10.4 and 10.5 give an outlook and conclude the work.

10.2 Generic Processing Steps

As mentioned in Sect.10.1, the financial time series processing can be subdivided into three major processing steps: *Abstraction, Mining & Discovery* and *Prediction* which are sequentially carried out as illustrated in Fig. 10.1.

Traditionally, the techniques and methods dedicated to realize these stages operate in an offline mode. That is, they assume that the data can be collected before initiating any kind of processing. However, there exist some methods that operate online especially in relation with the last two stages as will be explained later. The relevance of online processing stems form the power of dealing with streaming data incrementally, thus alleviating storage constraints and handling of non-stationary data.

10.2.1 Abstraction

Abstraction aims at summarizing the time series by capturing the important semantical contents of the time series resulting in a reduced compact time series.

Given a continuous data stream of the form $\{(t, x)\}$ where t indicates the time stamp and x the value of the asset (e.g., stock, share, etc.) at time t. The problem of this representation is the huge amount of data to handle. Therefore, a compact structure that preserves the information content must be found. This step facilitates the rest of the processing steps like mining, discovery and prediction.

There exit several techniques that deal with time series summarization:

- *Linguistic summarization*: The idea of linguistic methods is to summarize TS in the form of linguistic description [3, 6, 7]. A summary of a time series may be expressed as a set of linguistic statements of the form: *Most trends with a low variability are constant* or something like: *The title dropped between the 1rd and the 6th. Then it rose between the 6th and the 12th…* etc.
- *Trend-based summarization*: The idea here is to segment TS into small and closely-related sections that describe the up/down trends of TS [8, 9].
- *Transformation-based summarization*: The time series could be transformed into a mathematical description like with the Discrete Fourier Transformation (DFT), the Discrete Wavelet Transform (DWT) and the Bezier-Curve used for capturing similar shaped curves [10].

Other representation alternatives have been as well investigated. For instance, the original time series can be converted into a picture representation (e.g., bitmap, jpeg, etc.) to be used in similarity analysis, in a way that patterns, commonalities, or similarity between the time series can be found by comparing the corresponding pictures. For such a purpose typical image processing algorithms are applied as in [11–14].

It is important to note that the analysis techniques used heavily depend on the representation/abstraction (enabling offline or online processing) of the time series conducted in the first stage of the processing cycle.

10.2.2 Mining and Discovery

After the preprocessing and transformation steps, the analysis steps/tasks can be effectively conducted. Typically the *Mining & Discovery* task relies on the resulting representation and can be discussed from the two perspectives: offline and online.

10.2.2.1 Approaches

The analysis of financial time series has been the center of investigation from many different communities: statistics, data mining, soft computing, pattern recognition and computational intelligence.

From the statistical perspective, prominent methods focus on the correlation analysis to understand the evolution of individual and collective time series and relation between the time series. Relying on the correlation between time series with different time shifts, prediction can be performed. Moreover, regression techniques have been extensively used to understand the behavior of the time series and particularly to capture the trends occurring within a time series to enable prediction.

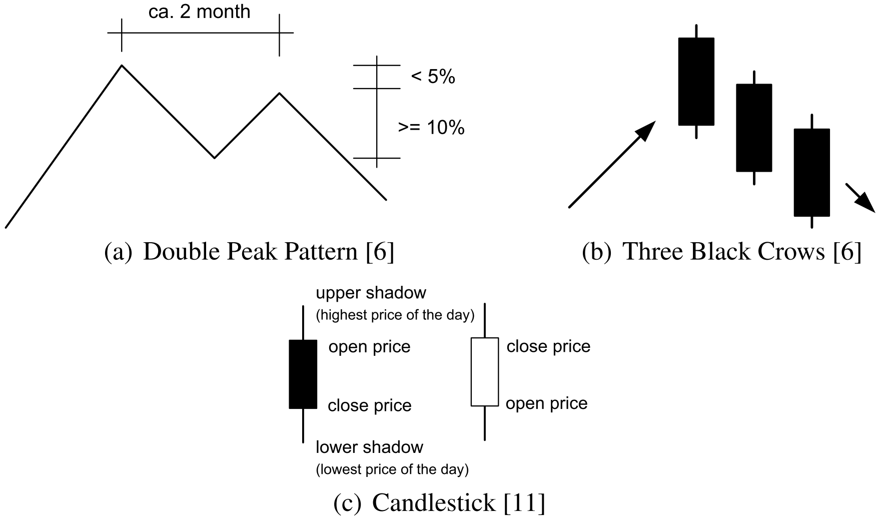
Data mining techniques also focus on event, change and temporal aspects, while text mining can be applied when data have a textual form (such as news) [15, 16]. Techniques like clustering and classification [17–19] are usually applied during the recognition or discovery processes, where the aim is to recognize the different and possibly already known patterns in the literature (e.g. Double Peak [20]). For more information on predefined financial patterns see Sect. 10.2.2.2.

There exist different techniques for recognizing predefined patterns [21] and [22], e.g. neural network based methods [23], fuzzy rule-based systems [10, 24]. Patterns are a very important instrument for FTS analysis.

10.2.2.2 Financial Pattern

Financial patterns are of paramount importance in these steps. They are formed based on the observation of price movements of different stock prices over many years and are now known as best-practices. They describe reoccurring situations of the market—exactly of one stock—that impacts the future price movement. Important patterns can be found in [20]. In particular, patterns are collected based on either close-price charts which show the price of a stock at the end of the day, or on candlestick charts (see Fig. 10.2(b) for a chart snapshot) which describe the price movement from different perspectives of a day (opening, close, low and highest price of the day). Predefined patterns rely on these two kind of charts.

An example of a chart pattern is the so called Double Top or Double Peak [20] (see Fig. 10.2(a)). This pattern is characterized by two separated peaks where the distance between these two peaks is only about two months and the price varies between 0 % and 5 %. The price tendency after the pattern occurs takes the form of an increasing movement. The double peak pattern is specified by a time scale of two



(a) Double Peak Pattern [6]

(b) Three Black Crows [6]

(c) Candlestick [11]

Fig. 10.2 Examples of financial patterns including a visualization of the structure of a candlestick

months [20]. This does not have to be true for other chart patterns as patterns can spread over different time scales. One pattern is spread over weeks and the next time over some months—but it is specified/identified as the same pattern. These different time scales make the handling of those pattern within the recognition process not trivial.

A candlestick consists of the body, the upper shadow and the lower shadow (see Fig. 10.2(c)). The body is colored green/white, if the close price is higher than the open price of the stock exchange, otherwise it is red/black. The upper shadow shows the highest state of the stock price in the trading day; the lower shadow specifies the lowest state of the stock price. Note that candlestick charts contain chart patterns as well.

One example of a pattern, called the Three Black Crows, is illustrated in Fig. 10.2(b). This candlestick pattern predicts the reversal of the current uptrend. The pattern contains three black colored candlesticks where the shadows are more or less not existing and the close price is lower than that of the previous day.

Patterns are used to categorize down-trends or up-trends. Further information can be found in the *Encyclopedia of Chart Patterns* [20]. Based on the analysis results (e.g., identified patterns), prediction, buy and sell decisions can be made.

10.2.3 Prediction

As explained earlier, prediction relies on patterns and uses computational algorithms to construct models during the phase of *Mining & Discovery*. FTS analysis is performed in order to predict future values and understand the market factors that influence the evolution course (e.g. global crisis).

Various amounts of properties can be predicted, like future trends, future stock price values or the volatility. Furthermore, it is possible to make predictions in the long term (e.g., some days) or short term (e.g., within some hours or minutes).

A very simple example can be drawn from seasonal charts which describe the fluctuation of the price within a specific time span (mainly one year). Due to seasonal occasions/events, the price can be influenced. For instance, before the Valentine's Day the floristics companies have a higher turnover than on other days and so the movement of the corresponding time series is influenced. Seasonal charts can be easily calculated through the available time series information of the last years. The resulting compact curve is another model.

Released news are also a very important source for performing better prediction and can therefore be integrated into the different analysis steps as an additional source of information to enhance the quality of the prediction.

10.3 Computational Methods for FTS

Time series analysis is not easy to perform manually, because of the fact that there are many factors influencing the evolution of the time series. Automatic analysis approaches are therefore recommended in order to cope with the huge amount of data that arrives over time.

Several approaches come from different research areas, like statistics, soft computing, machine learning, linguistics and data/text mining. These approaches fit into the analysis of single time series (intra-relation) and into the analysis of multiple time series (inter-relation) [16]. They can be categorized into offline and online methods. Offline methods explore stationary data sets, whereas online methods analyze streaming data.

The next section introduces the details of different computational approaches. The offline and online analysis techniques for analyzing FTS are described as well. Table 10.2 categorizes the introduced methods based on the processing mode (offline and online).

10.3.1 Abstraction

It is important to reduce the amount of data to enable an efficient processing. The abstraction process aims at eliminating unnecessary information or to shorten the information thus allowing therefore speeding up the analysis process. A detailed overview of different abstraction methods can be found in [25].

10.3.1.1 Time-Series Segmentation

Time series segmentation has been extensively discussed in the literature. The idea is to partition the time series into a set of segments on which the analysis is performed [26]. For instance in [9], an offline segmentation approach is presented. It consists of generating segments of multi-dimensional time series, where the segments are based on a polygonal approximation.

For a higher reduction rate in more than one dimension, time series segmentation is applied. This technique transforms the time series into trends which represent closely related points with the same characteristics. The slope of the trend line uncovers positive or negative tendency within the time span of the trend. The algorithms for segmenting the time series are separated into three categories [8]: Top–Down, Bottom–Up, and Sliding Window. The bottom–up and top–down approaches are offline methods, while the sliding window is online.

The top–down algorithm divides the original time series into segments [8]. It starts with the complete time series and splits them continuously at particular locations. This is done by approximating the time series with each possible partitioning. The segmentation is performed recursively until a stopping criterion (e.g., a predefined amount of trends found) is reached.

With the bottom–up algorithm, the time series is subdivided into to highest possible amount of trends. This means that one point in time represents exactly one trend. Within each iteration, trends which are closely related are grouped together. This results in new combined trends. The algorithm stops until there are no trends to combine.

With the sliding window [8], a specific amount of data is previously visible to the algorithm. The sliding window is an online approach. The current identified segment grows until a threshold is reached.

Beside the traditional (top–down, bottom–up and sliding window) algorithms, Keogh et al. [8] describe a new approach using a hybrid algorithm, combining the advantage of the sliding window and bottom–up technique. The SWAB (Sliding Window and Bottom–up) algorithm is an online algorithm and reduces the shortcomings of both algorithms. A buffer is used to provide the possibility to perform a lookahead of the next data.

10.3.1.2 Symbolic Representation

A binary transformation of a time series could be seen as a symbolic time series. Symbolic representations have different advantages such as efficiency, low measurement noise and low complexity [7].

Singh and Stuart [27] describe the online creation of a binary representation. The proposed algorithm replaces the value of the time series. That is, a positive change (increasing) in TS is marked with 1, if the previous value is less than the current value. A 0 is marked, if the opposite (decreasing) is observed. This representation reduces the range of possible values for later analysis process. Another example for a binary representation is clipped series described in [25]. There the value of the time series is replaced by 1, if it is above a specific value (e.g. the average) and 0 otherwise. Note that if the reference point for the mapping is the average value for the time series, only an offline mode is possible.

The problem of most symbolic representations, including the binary one is that the symbols are meaningless to the user. The work of Lee et al. [7] (online) describes a symbolic representation over fuzzy linguistic variables which are understandable by the user. The symbolic transformation is based on candlestick charts. Each candlestick is described by seven parts in the algorithm: sequence, open style, close style, upper shadow, body, body color, and lower shadow [7]. To describe the characteristic of a candlestick, linguistic phrases are used.

10.3.1.3 Linguistic Descriptions

Detailed linguistic descriptions turn out to be a useful abstraction method. For extracting a human-readable form out of time series, linguistic summarization is applied. For instance [6, 28, 29] generate duration-based summaries via an online method. The time series is examined in terms of how fast the values change, the variability of the values and the persistence of the trend. The theory of fuzzy information granulation for the computation with words from Yager and Zadeh is applied [30]. It specifies the so called portoforms (= templates for linguistically quantified propositions). An example of such a portoform is *Among All Segments, Q are P*, where Q is a temporal classifier and P is the characteristics (like fast decrease) [29]. A more detailed explanation of the approach can be found in [6]. Summarization concerns the number and the characteristics of trends. An example of such a summarization in terms of trend number is: *Most of the trends are of large variability* [6].

A full description of all possible abstraction techniques can not be covered in this chapter due to space limitation. Besides the introduced techniques, there are other methods like, discrete Fourier transformation, cosine transformation, singular value decomposition which are abstraction techniques. An illustrative application of these techniques, such as Discrete Fourier Transformation, within the data mining can be found in the next section. Moreover the indexing approaches are presented within this section as it is an essential part of FTS similarity analysis. For further examples it is referred to Bagnall et al. [25] providing an overview of the abstraction methods.

10.3.2 Mining and Discovery

Stock price prediction is not easy to perform because of the number of factors influencing the price. Nevertheless, it is important to know the optimal time interval to buy and sell titles. To cope with the problem of prediction accuracy, technical financial analysis can be applied. Different approaches have been developed to deal with the extraction and uncovering of information in historical data. Such approaches can be categorized in mining and discovery approaches. The former ones which focus on events and changes within the FTS have their origin in data mining, while the latter ones which focus on pattern recognition stem from statistics and soft computing.

The approaches can also be categorized into offline and online. In both modes the goal is to understand the future stock price movement and to enable prediction and decision taking (see Table 10.2 for summarization).

10.3.3 Mining

Chiang et al. [31] propose an online approach that mines data in the natural order of the time series relying on the work by Yager and Zadeh [30]. To find useful knowledge, the authors summarize data using fuzzy sets by searching unique characteristics within a group of objects. Initially, the user pre-classifies objects into groups based on the time series attributes like for instance, the *busiest time of a CPU* in a time series modeling the CPU usage. Once the groups are identified, the similarity/membership degree of an object to a group is measured using the fuzzy relation. A linguistic summary is utilized to describe textually the group according to a set of attributes possibly valid for that group.

News (texts) are a very important source for predicting the stock market. Studies revealed that utilizing news leads to 'more accurate predictions' [15]. Text mining approaches are therefore appealing for financial analysis. One of the most important activities are news labeling. A news is labeled with a predefined class, such as UP or DOWN depending on the positive or negative influence of the news on the stock price. String matching techniques can then be applied to uncover patterns. After training of a classifier to identify the impact of released news on the stock, the approach could perform in an online mode.

Fung et al. [16] propose an online approach relying on news to analyze multiple time series, departing from the idea that single time series are not usually useful and sufficient to make accurate prediction of the stock price movement due to the multiplicity of factors impacting the price. First, trends are extracted from the time series. Then the news texts are aligned with the time series assuming that the news have an immediate influence on the market and therefore on the time series. In parallel, important words and their weights are extracted from the news texts. The next step tries to identify possible relation between the stocks, meaning that relations of the form: *time series A triggers time series B*.

Another approach handling time series mining and prediction combining time series data and news information is developed by Schumaker and Chen [32]. It describes a system called AZFinText (Arizona Financial Text System). Financial news texts are represented as the set of nouns occurring in the texts. A binary representation of the texts is adopted: 1 indicates that a particular noun appears in the text and 0 indicates its absence. A support vector regression model is then developed to estimate the stock price in the future. After building the model, the approach could be used in an online manner for new arriving articles.

Qin and Shi [33] describe an algorithm for extracting association rules from time series. The approach finds inter-transactional association rules, where the time difference of the occurrence of an event is considered aiming at expressing rules of the form: '*If the value of time series A goes up the first day and B goes up the second day then C goes up the third day.*' [33]. The algorithm works on static data and is therefore an offline algorithm.

Jiang and Gruenwald [34] describes an online algorithm for mining rules from data streams. They discuss the main aspects that need to be considered in different research applications such as medical application and stock ticker. The most important issue that must be considered is the memory management, as most algorithms perform full scans over the database. Efficient data structure must also be devised to have a compact form. Additionally, the generation of frequent item sets used to derive the association rules must be optimal using exact and approximate algorithms. The most important issue is the frequent updates of the created association rules over time when new data points arrive.

An example for the creation of frequent itemsets from data streams (online algorithm), which are needed for the creation of association rules can be found in [35]. There data points of the recent past are higher weighted than other data points. This means historic data has not so much influence on the current values or future predictions.

The most common data mining problem in the area of FTS is similarity search. The goal is to find the most similar time series based on a query describing the reference point. A popular similarity measure is the Euclidean distance. Indexing structures help in supporting the search for similar time series. Indices do not scale well within a very high dimensional space and loose their advantage [36]. Therefore, dimensionality reduction is needed (e.g. via Fourier Transformation, Wavelet Transformation, Singular Value Decomposition or Piecewise Aggregate Approximation).

An example that uses Piecewise Aggregate Approximation for bivariate and multivariate correlation is introduced in [37]. It describes segments with the mean-value and the length of the individual segment. Due to the preprocessing and the time span that must be considered within the similarity analysis, the index approaches are considered as offline methods. Among the popular similarity search approaches in data mining is the Generic Multimedia Indexing (GEMINI) Framework [26]. This framework applies dimensionality reduction to overcome the difficulty (= high dimensionality) of indexing time series.

In [19] time series correlation is discussed. It uses the Support Vector Clustering as offline algorithm. Clustering is a data mining technique that enables finding similarities between time series. Much of the research work focuses on the creation of new and optimized similarity measures [26]. These can mainly be categorized into: raw-data-based, feature-based, and model-based [38]. Raw-data means that there is no preprocessing, like feature reduction. This could only be used for short time series. Feature-based approaches transform the FTS into a feature (vector) representation. The model-based techniques try to describe the clusters using models like the Gaussian Mixture Model. Time series clustering is very common since the earliest nineties. The approaches use partitional clustering like fuzzy c-mean or k-mean and agglomerative hierarchical clustering. In addition, it is possible to perform clustering of subsequences or the whole time series.

Subsequence clustering is used in many other algorithms as a building block [39]. Moreover, it is possible to describe the occurrence of patterns within a series only (e.g., pattern A after pattern B). Rules of the form *if A occurs, then B occurs within time T* [39] can be derived.

10.3.4 Discovery

The discovery process aims at extracting/uncovering information out of the FTS. Knowledge can be either transparent (for instance in the form of rules) or encoded (e.g. neural networks) depending on the prediction model used to infer such knowledge. In the following some of the representative approaches are presented.

Guo et al. [23] applied a three-layered feedforward neural network to recognize predefined chart patterns as described in Sect. 10.2.2.2. The output nodes correspond to the patterns, while the input nodes correspond to segments of FTS. For the segmentation an bottom–up approach (offline algorithm) is used.

Anand et al. [21] described patterns with templates which can be easily extended by the user. This offline approach gives the possibility to describe patterns over a specific language called *Chart-Pattern Language*. The definition language is described by means of the Haskell programming language. Using the templates, it is possible to define both simple as well as composed or more complex patterns as combinations of simple patterns. This approach performs the pattern recognition processes in white box fashion, in contrast to neural networks.

Suh et al. [10] applied expert systems using the Backward Screening Pattern Recognition Algorithm (BSPRA). It is developed as a semi-realtime approach. The rule-based expert system is responsible for identifying patterns in time series. The time series is smoothed by applying the moving average. The BSPRA Algorithm is executed to identify a specific pattern [20].

Leigh et al. [40] used also a template-based approach for online pattern identification. The template which specifies the pattern is stored as a 10 × 10 matrix. The template is then compared to a pictographic image of the time series.

In [41] an approach based on collecting the most important points of the time series is proposed. These points are called Perceptually Important Points (PIPs) (e.g. peaks). Once a set of PIPs are identified, a similarity measure is used to perform the pattern identification. The identified patterns can be used to predict future movements. The pattern matching algorithm is applied to a ‘real-time series’, hence it could be applied for online detection.

In addition to neural networks, template and rule-based approaches, fuzzy sets have been applied for pattern discovery. For instance in [7], candlestick patterns are described by fuzzy linguistic variables with membership functions of the form: long, short, middle, etc. Fuzzy modifiers such as somewhat, not, very, extremely, etc. can be used as well.

Fuzzy rule-based systems (considered as realtime system) have been used in [24] relying on regression and clustering techniques. The analysis of the stocks is based on indices which are the input parameters for the algorithm. An index describes the collective movement of stocks which are summarized due to equal characteristics (e.g. handled on the same stock exchange).

In contrast to the previously mentioned approaches, Kasabov [42] proposed Dynamic Evolving Neural-Fuzzy Inference System (DENFIS) that can operate both online and offline to solve the problem of time series prediction. It consists of a neuro-fuzzy system. Once the neuro-fuzzy system has been trained, the prediction on unseen data can be done. Similar approach has been proposed in [43] for time series forecasting.

In [44] an online approach based on Fibonacci sequences is proposed. The idea comes from Elliot waves that describe psychological strength behind the movement of prices. Many of the standard chart patterns are based on the Elliot waves. The time series is partitioned into segments. The time series is transformed into a matrix representation used for future prediction.

10.3.5 Prediction

Prediction is drawn based on the approaches applied in the *Mining & Discovery* phase. The results of the mining and of course of the discovery steps are models, that can be applied to perform prediction and help in decision making. Examples of approaches encountered in the literature are shown in Table 10.1.

Generally speaking when patterns are newly discovered, a deep analysis must be performed to validate the discovery using, for instance, historical data. The impact of such patterns must be observed and identified. Pattern recognition methods are very useful but sometime require expert knowledge. The prediction process can be successful only if the models built during mining and discovery step are reliable.

10.4 Discussion and Perspectives

Time series analysis explores past and present data to understand the behavior of the market and predict the future developments. The analysis aims at finding reoccurring or predefined patterns. The interrelationship among FTS's are usually extracted using correlation mechanisms that uncover dependencies having (mutual) influence on each other.

Recent developments show that the analysis procedure can be subdivided into three steps (*Abstraction, Mining & Discovery* and *Prediction*). It is possible to use offline and online methods. Table 10.2 portrays the different methods described in this chapter providing an overall picture of offline and online methods within the three steps. Which kind of approach should be applied, depends on the source of data, the environment and on the goal of the financial time series analysis.

Table 10.1 Overview of some prediction models

Technique	Summary	Pros	Cons
Rule-Based Sys.	Rules are extracted from FTS	Centralized knowledge extendable by users	Continuously rule updates, concept drift
Fuzzy Rule Sys.	Rules extraction	Centralized knowledge linguistic variables	Rule updates, def. of membership functions
Data Warehouse	Mining of huge data bulks	Predef. operation, performance, recognition	Update of the data, def. of the data structure
Neural Networks	Applied in Pattern Recognition	Adaptive, complex pattern	Huge number of training examples, concept drift
Template-based	Defining templates for	Performance	User must learn
System	Pattern recognition	Extendable by user	A new notion, or only machine readable
Classifier	Create classifier like	Adaptive learning	Update needed due
(-Neural Net.)	Support Vector Machines		To concept shift
Fibonacci Sequence	Pattern described as Elliot waves	Detailed description possible	Results not directly understandable by users
Correlation	Correlation of different FTS	Predef. correlation coefficient	Results valid for a small time span, performance
Simulation	Monte Carlo Simulation and game theory [32]	Mathematical models already proofed	Set-up of the simulation not easy to understand
Clustering	Performing Clustering Algorithm	No labeled training data	Concept shift, ongoing recalculation
Association Rules	Market Basket Analysis	Transactional based	Memory, and time consuming

The past research shows the efficiency of offline methods to analyze time series. Due to the non-stationary environment of the financial market, online methods are becoming increasingly investigated. They are tailored to streaming data (financial ticks) on the fly with the overall goal of uncovering patterns and regular events. Sustainable effort is still required to let the offline analysis methods be adapted to operate online. Currently rule-based and template-based methods can already operate online and be triggered for each arriving tick.

The future will focus on the inclusion of additional sources of information, e.g. news. In recent years, also social media turns out to be an important source [45, 46]. Through the inspection of those information the influence on the stock price can be estimated. Sources which contain different information about companies, advertisement etc. can be very valuable for tracing the evolution of the titles. However, deep insight into the relevance of such sources have to be considered. Because of the di-

Table 10.2 Overview of discussed offline and online methods

	Offline	*Online*
Abstraction		
Segmentation	Debled-Rennesson et al. [9], Keogh et al. [8] (top–down, bottom–up)	Keogh et al. [8] (SWAB, sliding window)
Linguistic/Symbolic Representation (Rep.)	Bagnall [25]	Lee et al. [7], Kacprzyk et al. [3, 6, 28, 29], Singh et al. [27], Bagnall [25]
Picture Rep.	Kumar et al. [11]	Weekley et al. [13]
Classification (without Neural Net.)	Kasetty et al. [12]	Weekley et al. [14]
Mining		
Fuzzy Rules		Chiang et al. [31]
Data Mining & Association Rules &	Qin & Shi [33]	Nikfarjam & Emadzadeh [15], Fung & Xu-Yu [16], Jiang & Gruenwald [34], Chang & Lee [35], Schumaker & Chen [32]
Streaming Analysis		Mueen & Keogh [22]
Indexing	Nguyen & Shiri [37], Keogh et al. [36]	
Clustering	Liao [38], Das et al. [39]	
Discovery		
Classification (Neural Networks)	Guo et al. [23]	
Rule-based Systems	Kasabov [42], Kablan [43]	Chang & Liu [24] Fu et al. [41], Kasabov [42]
Statistic Methods		Chen et al. [44], Suh et al. [10]
Symbolic Rep./Templates	Anand et al. [21]	Lee et al. [7], Fu et al. [41], Leigh et al. [40]

versity of information sources, information must be checked according to different dimensions, namely source, quality, frequency and scope of influence.

For example, the information about a company within the yellow press is less important than from financial newspapers. This has also an influence about the quality of the information. Frequency considers the opinion of the readership. The more people have the same opinion (or confirm the same information), the more relevant the information is.

Systems or approaches that react very fast on the changing market and the influencing environment will have the most success. Therefore, they must react on concept drifts so that prediction can be kept at a reasonable level. They also must be easy to understand and have a white-box model to understand and follow the results/conclusions drawn from these systems. This increases the confidence of the users which are mainly finance experts.

10.5 Conclusion

In this chapter an overview of different methods applied for time series analysis is presented. Financial time series analysis has emerged from the great interest of the finance segment of the economy. The huge amount of data, the different influencing factors and the global market make it difficult to perform analysis manually. Therefore, a pre-selection of informative and important effects must be performed automatically to support the financial experts.

As discussed the analysis task can be divided into three different phases. The first phase, called *Abstraction*, can be seen as a preprocessing step where the time series is transformed into another (shorter) representation.

The *Mining and Discovery* phase analyzes the input data based on its characteristics. Mining aims at uncovering new (not-already known) patterns whereas discovery aims at detecting predefined financial patterns. The analysis approaches applied in this step are diverse (e.g., statistical methods, soft computing methods, data mining methods, machine learning methods, etc.). Such methods can be categorized into offline and online methods. A summary of offline and online methods discussed in this chapter is presented.

The third phase deals with the *Prediction*. Depending on the method and the granularity of the input data (hourly, daily or monthly) short or long term prediction is performed. Based on the goal of the model, rates or trend movements can be estimated as well.

Ongoing research shows the trend to include additional data sources from the Internet, especially news texts and RSS feeds. As different factors influence the stock price, information sources must be filtered before including them during the analysis and the development of forecasting models (for instance news or public statements are key information). Additionally, the created systems must react on the rapid changing environment of the market and must be therefore self-adaptive.

References

1. Tsay, R.: Analysis of Financial Time Series, 2nd edn. Wiley, New York (2005)
2. Flores, P., Anaya, C., Ramirez, H.M., Morales, L.B.: Automated linear modeling of time series with self adaptive genetic algorithms. In: Proceedings of the International Joint Conference on Neural Networks, pp. 1389–1396 (2007)
3. Kacprzyk, J., Wilbik, A.: Linguistic summaries of time series using a degree of appropriateness as a measure of interestingness. In: International Conference on Intelligent Systems Design and Applications, pp. 385–390. IEEE Comput. Soc., New York (2009)
4. McNelis, P.D.: Neural Networks in Finance: Gaining Predictive Edge in the Market. Elsevier, Amsterdam (2005)

5. Sapankevych, N., Sankar, R.: Time series prediction using support vector machines: a survey. Comput. Intell. Mag. **4**(2), 24–38 (2009)
6. Kacprzyk, J., Wilbik, A., Zadrozny, S.: Linguistic summarization of time series using a fuzzy quantifier driven aggregation. In: Fuzzy Sets and Systems, vol. 159, pp. 1485–1499. Elsevier, Amsterdam (2008)
7. Lee, C.-H.L., Liu, A., Chen, W.-S.: Pattern discovery of fuzzy time series for financial prediction. In: IEEE Transaction on Knowledge and Data Engineering, vol. 18, pp. 613–625 (2006), IEEE Educational Activities Department
8. Keogh, E.J., Chu, S., Hart, D., Pazzani, M.J.: An online algorithm for segmenting time series. In: Proceedings of the IEEE International Conference on Data Mining, ICDM 2001, Washington DC, USA, pp. 289–296. IEEE Comput. Soc., New York (2001)
9. Debled-Rennesson, I., Tabbone, S., Wendling, L.: Fast polygonal approximation of digital curves pattern recognition. In: International Conference on Pattern Recognition, vol. 1, pp. 465–468. IEEE Comput. Soc., New York (2004)
10. Suh, S.C., Li, D., Gao, J.: A novel chart pattern recognition approach: a case study on cup with handle. In: Artificial Neural Network in Engineering Conference (2004)
11. Kumar, N., Lolla, N., Keogh, E., Lonardi, S., Ratanamahatana, C.A.: Time-series bitmaps: a practical visualization tool for working with large time series databases. In: Data Mining Conference, pp. 531–535. SIAM, Philadelphia (2005)
12. Kasetty, S., Stafford, C., Walker, G.P., Wang, X., Keogh, E.: Real-time classification of streaming sensor data. In: IEE International Conference on Tools with Artificial Intelligence, pp. 149–156 (2008)
13. Weekley, R.A., Goodrich, R.K., Cornman, L.B.: An algorithm for classification and outlier detection of time-series data. J. Atmos. Oceanic Technol. **27**(1), 94–105 (2010)
14. Weekley, R.A., Goodrich, R.K., Cornman, L.B.: Fuzzy image processing applied to time series analysis. In: 3rd Conference on Artificial Intelligence Applications to the Environmental Science, California, USA (2003)
15. Nikfarjam, A., Emadzadeh, E.M.S.: Text mining approaches for stock market prediction. In: The 2nd International Conference on Computer and Automation Engineering (ICCAE), pp. 256–260. IEEE, New York (2010)
16. Pui Cheong Fung, G., Xu-Yu, J., Wai Lam: Stock Prediction: integrating text mining approach using real-time news. In: Proceedings of IEEE International Conference on Computational Intelligence for Financial Engineering, 2003, pp. 395–402. IEEE, New York (2003)
17. Lee, J.W.: Stock price prediction using reinforcement learning. In: Proceedings of IEEE International Symposium on Industrial Electronics, 2001, ISIE, pp. 690–695. IEEE, New York (2001)
18. Kasetty, S., Stafford, C., Walker, G.P., Wang, X., Keogh, E.: Real-time classification of streaming sensor data. In: Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial Intelligence, ICTAI '08, pp. 149–156. IEEE Comput. Soc., New York (2008)
19. Yankov, D., Keogh, E., Kan, K.F.: Locally constrained support vector clustering. In: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, ICDM '07, pp. 715–720. IEEE Comput. Soc., New York (2007)
20. Bulkowski, T.N., Bulkowski, T.N. (eds.): Encyclopedia of Chart Patterns. Wiley, New York (2005)
21. Anand, S., Chin, W.-N., Khoo, S.-C.: Charting patterns on price history. In: Proceedings of the Sixth ACM SIGPLAN International Conference on Functional Programming, ICFP '01, pp. 134–145. ACM, New York (2001)
22. Mueen, A., Keogh, E.: Online discovery and maintenance of time series motifs. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, pp. 1089–1098. ACM, New York (2010)
23. Guo, X., Liang, X., Li, X.: A stock pattern recognition algorithm based on neural networks. In: Proceedings of the Third International Conference on Natural Computation, ICNC '07, pp. 518–522. IEEE Comput. Soc., New York (2007)

24. Chang, P.-C., Liu, C.-H.: A TSK Type Fuzzy Rule Based System for Stock Price Prediction. Expert System Application, vol. 34, pp. 135–144. Pergamon, Elmsford (2008)
25. Bagnall, A., Ratanamahatana, C., Keogh, E., Lonardi, S., Janacek, G.: A Bit Level Representation for Time Series Data Mining with Shape Based Similarity Data Mining and Knowledge Discovery vol. 13, pp. 11–40. Springer, Netherlands (2006)
26. Keogh, E., Kasetty, S.: On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. Data Mining and Knowledge Discovery, vol. 7, pp. 349–371. Springer, Netherlands (2003)
27. Singh, S., Stuart, E.: A pattern matching tool for time-series forecasting. In: Proceedings of the 14th International Conference on Pattern Recognition, ICPR '98, p. 103. IEEE Comput. Soc., New York (1998)
28. Kacprzyk, J., Wilbik, A., Zadrozny, S.: Analysis of time series via their linguistic summarization: the use of the sugeno integral. In: Proceedings of the Seventh International Conference on Intelligent Systems Design and Applications, ISDA '07, pp. 262–270. IEEE Comput. Soc., New York (2007)
29. Kacprzyk, J., Wilbik, A.: Using fuzzy linguistic summaries for the comparison of time series: an application to the analysis of investment fund quotations. In: ISA-EUSFLAT (2009)
30. Zadeh, L.A.: Toward a theory of fuzzy information granulation and its centrality. In: Human Reasoning and Fuzzy Logic. Fuzzy Sets and Systems, vol. 90, pp. 111–127 (1997)
31. Chiang, D.-A., Chow, L.R., Wang, Y.-F.: Mining Time Series Data by a Fuzzy Linguistic Summary System. Fuzzy Sets System, pp. 419–432. Elsevier, Amsterdam (2000)
32. Schumaker, R.P., Chen, H.A.: Discrete Stock Price Prediction Engine Based on Financial News Computer, vol. 43, pp. 51–56. IEEE Comput. Soc., New York (2010)
33. Qin, L.-X., Shi, Z.-Z.: Efficiently mining association rules from time series. Int. J. Inf. Technol. **12**, 30–38 (2006)
34. Jiang, N., Gruenwald, L.: Research Issues in Data Stream Association Rule Mining. SIGMOD Rec., vol. 35, pp. 14–19. ACM, New York (2006)
35. Chang, J.H., Lee, W.S.: Finding recent frequent itemsets adaptively over online data streams. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 487–492. ACM, New York (2003)
36. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Locally adaptive dimensionality reduction for indexing large time series databases. In: Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, pp. 151–162. ACM, New York (2001)
37. Nguyen, P., Shiri, N.: Fast correlation analysis on time series datasets. In: Proceeding of the 17th ACM Conference on Information and Knowledge Management, pp. 787–796. ACM, New York (2008)
38. Liao, T.W.: Clustering of Time Series Data—A Survey. Pattern Recognition, vol. 38, pp. 1857–1874 (2005)
39. Das, G., Lin, K.-I., Mannila, H., Renganathan, G., Smyth, P.: Rule discovery from time series. In: Proc. of the 4th Intl. Conference on Knowledge Discovery and Data Mining KDD, pp. 16–22. AAAI Press, Menlo Park (1998)
40. Leigh, W., Modani, N., Hightower, R.: A Computational Implementation of Stock Charting: Abrupt Volume Increase as Signal for Movement in New York Stock Exchange Composite Index Decis, vol. 37, pp. 515–530. Elsevier, Amsterdam (2004), Support Syst.
41. Fu, T.-c., Chung, F.-l., Luk, R., Ng, C.-M.: Stock time series pattern matching: template-based vs. rule-based approaches. Eng. Appl. Artif. Intell. **20**, 347–364 (2007). Pergamon, Elmsford
42. Kasabov, N.: DENFIS Dynamic Evolving Neural-Fuzzy Inference System and Its Application for Time Series Prediction. IEEE Transactions on Fuzzy Systems, vol. 10, pp. 144–154 (2002)
43. Kablan, A.: Adaptive neuro fuzzy inference systems for high frequency financial trading and forecasting. In: International Conference on Advanced Engineering Computing and Applications in Sciences, pp. 105–110. IEEE Comput. Soc., New York (2009)

44. Chen, T.-L., Cheng, C.-H., Teoh, H.J.: Fuzzy Time-Series Based on Fibonacci Sequence for Stock Price Forecasting. Statistical Mechanics and Its Applications, vol. 380, pp. 377–390 (2007)
45. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. J. Comput. Sci. **2**, 1–8 (2011)
46. Jin, X., Gallagher, A., Cao, L., Luo, J., Han, J.: The wisdom of social multimedia: using Flickr for prediction and forecast. In: Proceedings of the International Conference on Multimedia, pp. 1235–1244. ACM, New York (2010)

Chapter 11
Data Supply for Planning and Budgeting Processes under Uncertainty by Means of Regression Analyses

Peter Rausch and Birgit Jehle

Abstract Planning and Budgeting (P&B) is an important part of Performance Management (PM). The corresponding processes for medium-sized and large organisations are usually very resource-intensive, time consuming and costly. These issues are mainly caused by uncertainty, which is a big challenge for companies. It is shown that available software and tools do not address this challenge in an appropriate way. Before possible issues and solutions are analysed in detail, an overview of different types of uncertainty is given. Afterwards important steps of the P&B process which suffer from uncertainties are outlined. Quite often it is not really clear which parameters have an impact on the planning object and how strong the planning object is influenced by certain parameters. Additionally, forecasts of the most important parameters which anticipate uncertainties are needed at an early stage of the P&B process. To resolve these issues, the application of different types of regression analyses will be explored. Also, ideas for further processing of fuzzy data in the following P&B steps are given. Furthermore, organisational and cultural prerequisites for the successful application of the outlined approaches will be indicated.

11.1 Introduction

Planning and Budgeting (P&B) is an important part of Performance Management (PM) approaches. As we learnt in Chap. 8, business planning is applied in many different areas. However, despite all effort in the last decades, P&B is still a challenging task. Especially for medium-sized and large organisations, the budgeting process is usually very resource-intensive, time consuming and costly [1]. Due to the fact that P&B processes have to anticipate the future, many assumptions about

P. Rausch (✉)
Department of Computer Science, Georg Simon Ohm University of Applied Sciences,
Kesslerplatz 12, 90489 Nuremberg, Germany
e-mail: perausch@prof-rausch.de

B. Jehle
Noris Treuhand Unternehmensberatung GmbH, Virchowstr. 31, 90409 Nuremberg, Germany
e-mail: b.jehle@noris-treuhand.com

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_11, © Springer-Verlag London 2013

uncertain conditions have to be made. Hence, P&B processes should be able to cope with uncertainties. Efficiency is another important requirement [26]. On the one hand, a high level of automation is desirable. On the other hand, it has to be considered that according to the Artificial Intelligence Applications Institute, “black box or fully automated solutions are not acceptable in many situations” [32]. Another requirement is that methods should be easy to handle.

In this chapter, the aspect of data supply for P&B processes and the related issues will be explored. At first, the subject of this chapter will be aligned in terms of Chap. 1. Afterwards, practical issues and challenges of P&B processes will be outlined in general. As we will see, the ability to manage uncertainty in a professional way could contribute a lot to business development. Before possible solutions for the P&B data supply problem will be discussed, an overview of different types of uncertainty is given. It is necessary to know that they require different approaches to be anticipated in an appropriate way. To identify and forecast planning parameters and to address the aspect of uncertainty, the application of different types of regression analyses will be explored. In the following section, benefits and issues of the approaches are discussed. Additionally, important organisational and cultural prerequisites will be mentioned. The chapter concludes with a summary and ideas for further extensions in terms of PM.

11.2 Planning and Budgeting—Issues and Challenges

According to Chap. 1, P&B can be classified in the field of BI and PM. Data for P&B purposes is usually supplied by internal as well as by external data sources. To improve the quality of P&B and to get the data from a “single point of truth”, it makes sense to load this data into a data warehouse and to obtain the P&B input from it as data source. As we will learn in the following sections, data mining methodologies can be used to analyse the data for P&B purposes. Within the scope of PM, P&B is an important task. Goals on the operational and on the strategic level should be achieved by measures which are related to plans and budgets. They are controlled by closed-loop approaches, which were discussed in Chap. 1.

Despite the importance of P&B in the field of PM, the traditional P&B process in most companies is rather old and associated with issues. The process was first defined at the beginning of the 20th century [34]. Most of the companies still use this method or variants of it [22]. The company’s vision is the starting point. In the first step of the P&B process, targets and strategic guidelines are determined, as shown in Fig. 11.1. Normally this is a top down process. The next step is the specification of strategic goals to get the operational objectives for the planning period. The responsible divisions have to specify the different operational objectives, for instance, sales, costs, staff or investment targets for the business units. Then, a bottom-up planning follows. Once the budget is adopted, the periodical comparison of budget figures with actual figures follows. Deviations have to be analysed and necessary actions have to be taken if the defined targets are threatened [14]. The permanent

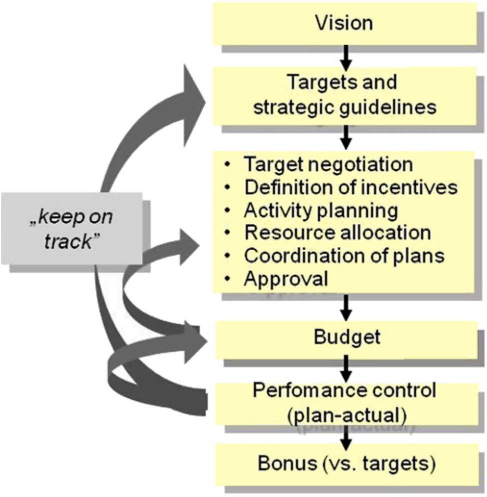


Fig. 11.1 Process of P&B in large companies (adapted from [5])

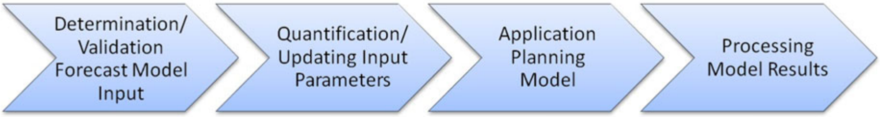

Determination/ Validation Forecast Model Input
Quantification/ Updating Input Parameters
Application Planning Model
Processing Model Results

Fig. 11.2 Determination of the planning output

control whether targets can be achieved at the end of the planning period is one of the main tasks of the controlling department.

For the bottom-up planning, some preliminary work is necessary (see Fig. 11.2). First, a forecast model has to be determined and validated. Relevant parameters and interdependencies have to be identified and, if necessary, the complete model has to be updated, for instance, if manufacturing processes have been reengineered. The next step is the quantification or updating of the input parameters, for instance, material costs, labour costs, capital expenditures, etc. Afterwards, the application of the planning model and the processing of the model results follow. In practice, steps 3 and 4 often have to be repeated several times to get results which are consistent with the targets.

Due to several issues, the top-down/bottom-up approach, which is described in Fig. 11.1, is not appropriate anymore. Especially for medium-sized and large companies, the budgeting process is usually very resource-intensive, time consuming and costly. It can take up to 30 % or more of management's time [7]. However, in many cases the results produced do not justify the time and money invested. The fixed budgets are inflexible, there is no incentive to review or even undercut the budget, and adaptations are very complex. Also, linguistic misunderstandings have to be mentioned. For instance, if managers assume that the turnover will increase "slightly", what do they think of? Is it of a range of 5 % to 10 %, or rather 1 % to 3 %? In traditional approaches, this issue is not anticipated. Deterministic values are

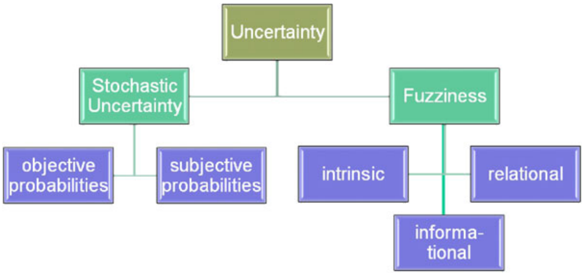


Fig. 11.3 Types of uncertainty (adapted from [24])

used, and variations in these ranges are neglected [29]. Additionally, we have to deal with external issues. In times of volatile economic conditions, plans which include crisp parameters might become obsolete as soon as they are adopted. Furthermore, because of markets' dynamics, assumptions which affect P&B might change often. Thus, uncertainty is a major concern of CEOs [9].

According to Oehler [21], nearly every company uses IT to address the issues mentioned above. A classification of different types of software can be found in Chap. 8 and in [21]. As shown in [25], tasks such as providing input data from different data sources, transferring data and information, or the automation of process chains can be covered quite well by means of available software. Other IT support includes functions which help to access and analyse P&B data. Furthermore, most software packages provide model and method support to formally describe dependencies between system states or parameters [21]. Software is also used frequently to assist managers in analyses of changes of budget premises and the resulting effects. This is an important support of P&B processes, but, unfortunately, the issue of managing uncertainty in an efficient way is not really addressed [25]. Most of the available tools only try to find the "best" deterministic solution for a certain step of the P&B process. Hence, the quality of the planning results is poor and a lot of effort is wasted. Before solutions for this issue are examined, different types of uncertainty have to be distinguished.

11.3 Types of Uncertainty

So far, we used the term "uncertainty" in a very general way. Many different models to describe uncertainties are discussed, for instance, density functions, probability intervals, possibility distributions, fuzzy sets, belief/plausibility functions and others. Hence, it is necessary to get a broader understanding what is meant by the term uncertainty and to focus on certain instruments. In this chapter, we focus on aspects of the probability and the fuzzy set theory, which are the most widely used concepts. Details concerning these theories, which can't be mentioned due to the limited space of this chapter, can be found in [2, 27]. Figure 11.3 shows the different types of uncertainty at a glance.

At the first level, stochastic uncertainty and fuzziness can be distinguished. The former is based on random experiments and can be modelled by means of probability theory. It is assumed that probabilities for events can be determined. In case the probabilities are verifiable inter-subjectively, statements for repetitive events which are independent of individual estimates are called “objective probabilities” [6]. In contrast to that, “subjective probabilities” are based on individual estimates [35]. In that case, an inter-subjective verification is not possible.

As shown in Fig. 11.3, fuzziness is another type of uncertainty. One possible manifestation is intrinsic uncertainty [27]. It is based on human perceptions, such as “high costs” or “reasonable profits”, and arises because adjectives, such as “high” and “reasonable”, are not clearly defined, and the respective terms can be interpreted from person to person in a different way. Another type of fuzziness covers “informational uncertainty”. In this case, uncertainty could, at least theoretically, be eliminated. Informational uncertainty occurs, for instance, if information acquisition is too expensive or if the collection of information would require too much effort. Additionally, relationships can be vague [37], for instance, if one production line is “not much more” profitable than the other.

In general, all of the types of fuzziness mentioned above can be found in the environment of complex systems. This is due to the fact that human capabilities to make precise statements decline with increasing complexity [36]. Zahdeh’s fuzzy set theory covers this type of uncertainty, while stochastic uncertainties are addressed by probability theory. In contrast to fuzzy set theory, probability theory is derived deductively from axioms and, therefore, its structure must fulfil tough requirements [15]. Due to the fact that both approaches address different types of uncertainty, it can be conceivable to combine them. In the following sections, we will explain a selection of instruments and outline their contribution to the P&B data supply problem.

11.4 Instruments: A Case Study

To analyse possible solutions for the challenges outlined in the last sections, we introduce a case study from the food and beverage industry. The company is medium-sized and its turnover is €8,500,000 per year. Most of the 100 employees work in manufacturing. A few are in charge of P&B. Due to many uncertainties, for instance, weather influences or volatile prices of input factors, P&B is a complex task. Once a year, the P&B department has to provide management with a budget plan, which has to be revised at the end of each quarter. For the initial plan, a certain amount of money has to be distributed to different production lines and departments. Additionally, it is necessary to control expenses during the planning period. If deviations become more and more likely, action should be taken. Reallocation of certain resources is allowed during the planning period, as long as the complete budget is not exceeded. At the beginning of a planning period, the following questions have to be answered or verified: Which are the “right” planning parameters for the planning period? Do the actual planning parameters have the greatest impact on current planning? Additionally, the parameters have to be quantified and processed to get a valid

Table 11.1 Input data for regression analysis (fiscal year 1–3)

Period	Av. sun mins./day (x_i)	Av. temp./day (z_i)	Sales (€, y_i)
1 (Jan, FY 1)	79.19	−4.15	123.26
2 (Feb, FY 1)	116.43	0.25	13,749.60
3 (Mar, FY 1)	276.58	5.75	33,905.87
4 (Apr, FY 1)	431.57	10.41	62,508.44
5 (May, FY 1)	234.77	11.99	53,657.41
6 (Jun, FY 1)	622.63	18.89	66,899.04
...	...	...	...
36 (Dec, FY 3)	86.77	1.90	355.63

plan. The following sections will outline different types of instruments which support the steps 1 and 2 of Fig. 11.2. Due to the dynamic environments of companies and the issue of uncertainty, these steps are very challenging and usually require a lot effort. Efficient instruments could save a lot of time and money.

11.4.1 Conventional Linear Regression Analysis

As shown in Fig. 11.2, our company reviews its planning model and checks whether the influences on planning parameters have changed or whether new relevant planning parameters have arisen, before starting to produce the periodic P&B results. For that purpose, assumptions concerning new parameters are collected and so-called regression analyses are performed. With this type of analysis, the nature of the relationship between parameters x_i, z_i and y_i can be examined. Different variants of this approach can be distinguished. At first, we focus on a simple variant, the univariate linear regression. It is supposed that single variables x and y are not on the same hierarchy level, which means that y is considered as dependent on a function of x. Additionally, it is assumed that x can be accurately measured and only the y-values are spread. x and y are metrically scaled.

First, an explanatory model which describes the relationship between y and x has to be determined. In case it exists already, the model has to be updated. For example, x represents the average number of minutes of sunshine per day, and y denotes the sales of ice cream. We now ask whether there is any relationship between the two variables, and if yes, how y can be derived as a function of x. From an external database, the company's data warehouse is supplied with weather data. Also, sales data, which is originally supplied by the enterprise resources planning system, is available. For all of the given values $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ the value pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_n)$ are obtained, see Table 11.1. (The z-values will be used later.)

Now, we look for a linear function which best describes the dependency of the values y_i from the x_i values. This variant of the regression analysis is called "linear

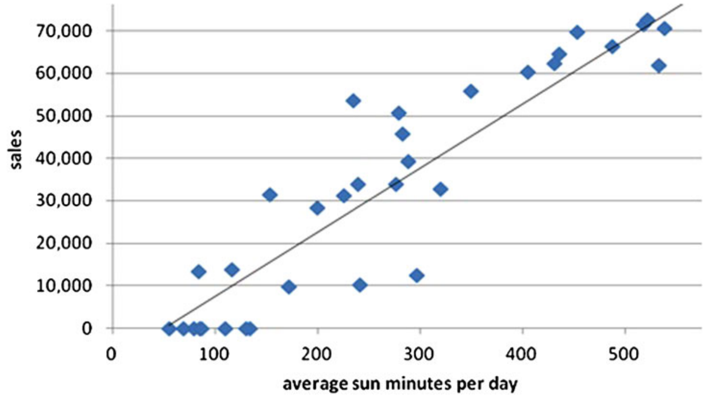


Fig. 11.4 Linear regression analysis results

regression". To determine the function, (see Eq. (11.1)) it is common to apply an optimisation approach which minimises the sum of squared vertical distances between the observed y_i-values in the dataset and the predicted values by linear approximation. This methodology is called "ordinary least squares" or "linear least squares". Thus, we calculate a line that passes through the cloud of points in such a way that the average squared distances to the points are as small as possible.

$$y_i = f(x) = \beta_0 + \beta_1 x_i \tag{11.1}$$

β_1 denotes the gradient, and β_0 is the intercept of the regression line. Other variants of the regression analysis, which can't be explained in this chapter due to the space limitations, can be found in [33]. In our case, Eq. (11.2) and Fig. 11.4 illustrate the result.

$$y_i = -7817.7 + 151.27x_i \tag{11.2}$$

$$R^2 = 0.8568 \tag{11.3}$$

R^2 in Eq. (11.3) represents the correlation coefficient which is in the interval $[-1, 1]$. The closer R^2 comes to the edges of the interval, the stronger the positive or negative relationship between the analysed variables is. In our case, it can be confirmed that there is a strong relationship between the average number of minutes of sunshine and sales of ice cream. To judge the function determined, additional measures or error rates can be used. Common software packages usually provide the most important measures like the "Mean Absolute Error" (MAE), "Root Mean Square Error" (RMSE), "Relative Absolute Error" (RAE) and "Root Relative-Squared Error" (RRSE). Low measures indicate low error rates of a function. More details concerning error measures can be found in [19].

While discussing the results, the idea came up to include the effects of the average temperature per day z_i. Thus, the impact of both parameters on sales was analysed in a so-called "multivariate regression analysis". The basic concept is very similar to the univariate regression analysis. In case of two parameters, a function of type (11.4) is provided.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i \tag{11.4}$$

In our case, we get Eq. (11.5) and the correlation coefficient (11.6) as results.

$$y_i = -7533.3027 + 121.54 x_i + 778.28 z_i \tag{11.5}$$

$$R^2 = 0.9199 \tag{11.6}$$

Considering just the correlation coefficient, our model (11.5) explains the sales figure very well and indicates that the multivariate linear approach is appropriate. Hence, the function (11.5) could be used as a predictive model. It would also be possible to extend the multivariate regression analysis by time and other parameters. To determine the input for our forecast function (11.5), management could use time series analyses to get ideas about how the average number of sun minutes per day and the average temperature will develop. Also, seasonal patterns can be anticipated. The linear regression approach is very easy to implement. Many widespread data mining tools, such as WEKA or even spreadsheet software, include a regression analysis component. The data can be provided by a data warehouse.

On the other hand, regression analyses are based on historical data. Hence, the significance of the results in the future is not reliable. Furthermore, it is important to realise that the approaches, which are discussed here, are based on the assumption that potential relationships are linear. If a linear relationship can't be assumed, it would be also possible to approximate this case by means of linear instruments. Other approaches to handle non-linear relationships, for instance, by means of non-linear regression analyses or neural networks, can be found in [28] and in Chap. 9. It is also advisable to check whether other approaches, such as time series analyses, can be useful apart from the supply of input for the forecast function. Details concerning time series analyses can be found in Chap. 10. The choice of the "right" method depends on the purpose of the analysis and the type of data. In general, it is recommended to compare the results of different approaches.

11.4.2 Advanced Ideas

Let's assume our company would use the regression functions (11.2) or (11.5) for the determination of planning parameters. As a result of our models, we get a point forecast. A point forecast is a single guess for y_t that in some way represents the whole distribution of possible values well [13]. In our case, a point forecast is not appropriate, because as we can conclude from the regression coefficients (11.3) and (11.6) there's no clear causal relationship between the x-/z- and y-values. The challenges, which were mentioned at the beginning of this chapter, wouldn't be solved, because it is quite unlikely that all of the parameters are forecasted correctly. In order to address uncertainty explicitly, we have to extend the approaches presented in

Sect. 11.4.1. A widely used instrument to address this issue is probability theory. The next section will give us a brief idea how confidence intervals can be applied to determine prediction intervals.

11.4.2.1 Quantile Regression

With quantile regressions, any desired quantile $q \in [0, 1]$ of y depending on x can be estimated [11]. Quantiles are points taken at regular intervals from the cumulative distribution function of a random variable [10]. The idea is to determine a one-sided confidence interval which includes the forecasted parameter with a certain probability p [18], for instance:

$$p(y_t < A) = 0.95 \tag{11.7}$$

Granger reports on an extension of this approach [13]. The approach determines so-called prediction intervals. They are based on confidence intervals for forecast results. Thus, a confidence interval has to be determined which includes the forecasted parameter with a certain probability p [13]:

$$p(B < y_t < A) = 0.95 \tag{11.8}$$

The "conventional" regression approaches, which were presented in Sect. 11.4.1, can be interpreted as special cases of the quantile regression. The latter technique requires more computational effort in comparison to the approaches in Sect. 11.4.1. Furthermore, quantile regression is not available in most of the widespread software suites. Thus, it would require software development effort to apply this instrument. This might be a reason for the rarity of examples for applications in the field. On the other hand, quantile regression is not very sensitive in cases with outliers or extreme values. In conventional approaches, it is necessary to filter outliers in advance to avoid contortions of the results [11]. Nevertheless, it also might make sense to analyse outliers when quantile regression is used. It should be checked whether or not structural interruptions occur.

Unfortunately, quantile regression does not cover cases when the y-values or the regression coefficients can't be determined exactly and when it is just possible to describe them in more detail by means of fuzzy sets. In these cases, fuzzy regression approaches can be appropriate.

11.4.2.2 Fuzzy Parameters and Fuzzy Regression Approaches

Because not all readers may be familiar with fuzzy set theory, a brief domain-specific introduction of the basic concept is given, before fuzzy regression approaches are introduced. Fuzzy sets can help to manage uncertainty in the field of P&B and to create more realistic P&B models. By means of fuzzy sets, it is possible to model parameters in a realistic way. To describe a fuzzy set very accurately, it

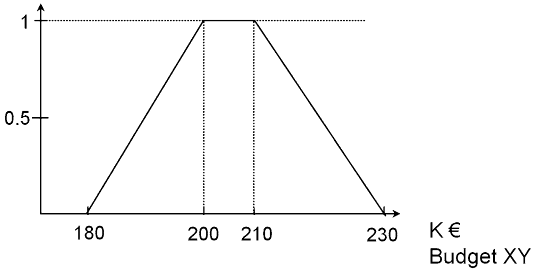


Fig. 11.5 Planning and budgeting with fuzzy parameters (reprinted from [25] with kind permission of Springer Science+Business Media)

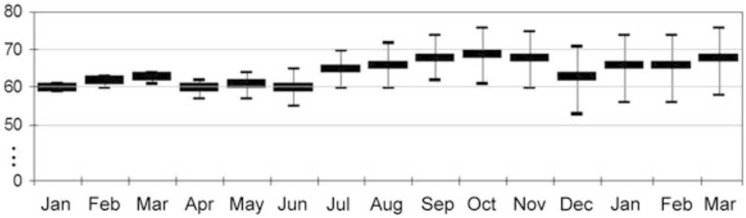


Fig. 11.6 Rolling forecasts of sales with fuzzy parameters (reprinted from [25] with kind permission of Springer Science+Business Media)

is recommended to apply a membership function $\mu_A : X \to [0, 1]$ which assigns a membership value to each element of a set [27]. To minimize the modelling effort, Rommelfanger recommends piecewise linear membership functions [27]. Their accuracy is sufficient for many purposes, such as the one described here. Figure 11.5 shows an example of a fuzzy parameter which represents an appropriate budget for the sum of expenses of a certain production line from a planner's point of view. However, different types of membership functions which would cause more modelling effort could also be used. The membership function shown in Fig. 11.5 gives an account of the budget planner's imagination that a sum between €200,000 and €210,000 is absolutely acceptable. But at minimum, the sum has to be greater than or equal to €180,000. A larger sum is desirable to ensure production, but a maximal a sum of €230,000 is acceptable.

Many planning parameters are not certain during the planning period. For instance, in our case it is not clear how the weather will be during the planning period, how much money is needed for the ingredients of ice cream and how much revenue can be earned. In most cases, the longer the forecast horizon is, the more difficult it gets to quantify parameters exactly. Thus, as Fig. 11.6 shows, the user should be allowed to work with fuzzy parameters. The upper and lower bounds represent elements with membership values which indicate a very low degree of membership to the set considered.

In our case, the head of the P&B department has the idea to use rolling forecasts with fuzzy data to support the quantification of planning parameters (step 2 of Fig. 11.2) and the generation of plans. By means of a fuzzy regression approach, a tendency combined with spreads can be determined. In contrast to that, conventional regression regards deviations between the observed and the estimated values

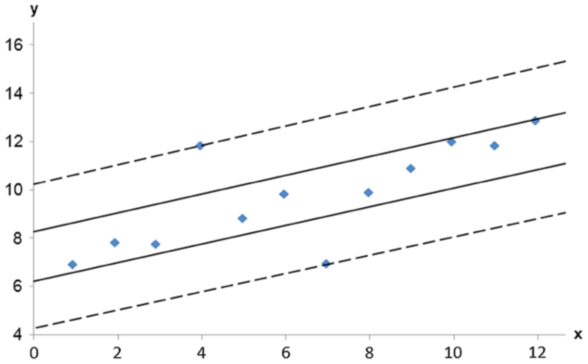


Fig. 11.7 Results of a fuzzy regression approach (adapted from [30])

as observation errors. In fuzzy regression, these deviations are considered as part of the parameters' possibility distribution [30]. Fuzzy regression analyses can be applied in cases which lack information, when it might be possible to describe the y-values in more detail by means of fuzzy sets.

In the early 1980's, the first fuzzy regression approaches were introduced. Tanaka et al. presented an approach in [31], which attracted a lot of interest and was further developed over the following years. Tanaka et al.'s first approach is based on Linear Programming (LP). It is assumed that we deal with non-fuzzy input and output data for the observations, which are used to create the forecast model. According to [30], a corresponding fuzzy regression model for prediction can be written as

$$Y = A_0 + A_1x_1 + \cdots + A_nx_n = \boldsymbol{Ax} \tag{11.9}$$

where $\boldsymbol{x} = (1, x_1, \ldots, x_n)$ is an input vector, and $\boldsymbol{A} = (A_0, \ldots, A_n)$ represents a fuzzy coefficient vector. Y denotes the corresponding estimated fuzzy output. It is assumed that A_i is a symmetric fuzzy number and can be defined by its membership function. The estimated output also becomes fuzzy due to the fuzzy coefficients. In [30], the authors show that by means of the given non-fuzzy input and output data, the optimal fuzzy coefficients A_i can be found by solving an LP problem.

Tanaka and Lee refined the approach by means of combining the least squares method with a so-called possibility approach based on quadratic programming [30]. They propose an objective function for the optimisation problem which aims to indentify a central tendency and minimises the spreads of the estimated fuzzy outputs. By means of a model parameter, the goal of minimising the spreads to discard outliers can be weighted. Figure 11.7 shows the results of an example, which is presented in [30]. The inner solid lines represent the estimated interval lines. The outer dotted lines define the estimated interval lines including an error value e [30].

Other fuzzy regression methodologies are based on different assumptions, for instance, concerning the fuzziness of input data and parameters. An overview can be found in [17, 20]. Thus, fuzzy regression offers a powerful set of instruments to address the issue of getting a range for the output of step 2 of the P&B process, Fig. 11.2. As we have learned so far in Sect. 11.3, these instruments are appropriate if the effort of gathering detailed information would be very high or if exact

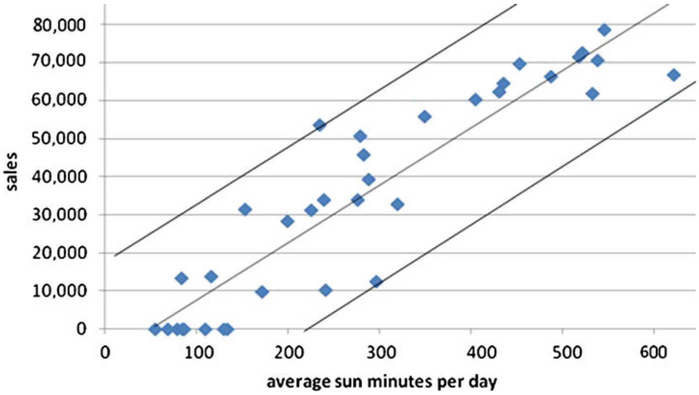


Fig. 11.8 Regression channel

information is not available. However, as in the case of quantile regression, fuzzy regression approaches require the development of software or at least the adaption of standard software. Hence, in the next section we will introduce another approach which also provides ranges and can be applied easily.

11.4.2.3 Regression Channels

So-called regression channels are an option to determine the upper and lower bounds and to extend the approach, which was presented in Sect. 11.4.1. The idea of regression channels became popular at the beginning of the 1990's when Gilbert Raff introduced the idea of the Raff regression channels to analyse stock markets [23]. Regression channels consist of equidistant parallel lines above and below a linear regression line (see Fig. 11.8). Depending on the regression channel variant, the distance between the linear regression trend line and the upper and lower channel lines is defined in a different way [8].

Raff regression channels are very popular in the field. To determine a Raff regression channel, the distance between the channel lines and the regression line is the greatest distance to a high or a low value and the regression line. The upper line of the channel acts as the resistance level and the lower line as support. It is assumed that the values fluctuate between the upper and the lower lines for a certain period [23]. A trend ends when the values break above or below the channel extensions. The upper and lower bounds of the regression channels can also be interpreted as spreads of the estimated fuzzy outputs, as shown in Fig. 11.6. Elements representing the highest degree of membership to the set considered can be derived from the classical regression model. The thick black bars in Fig. 11.6 denote these values.

On the one hand, this method is very simple. On the other hand, the determination of a start and an end for a trend can be challenging. Furthermore, there are no scientifically proven, generally valid instructions on what kind of action has to be taken when values touch the upper or lower channel lines. In such a situation, it remains unclear whether the values at the band limit will cross it and a reversal of the trend will start [12]. Even a short term crossing of the channel line doesn't necessarily mean the trend is broken. Also, other approaches which are used for chart analysis, such as Bollinger bands [4], suffer from this issue. Since historical data is used, structural interruptions may not be recognised in time. Especially in volatile markets, this is a big issue. Hence, it makes sense to analyse further information and to be attentive for news that might indicate the breaking of a trend. For instance, in the field of stock markets and commodities, the increase or decrease in volumes (trading activities) can foreshadow the end of a trend. Furthermore, it can be advisable to filter outliers when a channel is determined. Additionally, more sophisticated methods are needed if the parameters are subject to seasonal fluctuations.

Despite of all of these issues, regression channels can help to reduce the effort to determine the limits of fuzzy parameters, if the time horizon is not too long. Later, during the planning period, data can be updated, and more precise information can be gathered for the crucial parameters.

11.5 Benefits, Issues and Further Processing

As discussed in Sect. 11.4, all of the regression approaches are related with issues. Their ability to forecast parameters is limited and not absolutely reliable. However, in comparison to guesswork and the determination of crisp parameters, there are benefits. For instance, in our case, management wasn't really sure about the relationship between the average number of minutes of sun per day and the sales of ice cream. The regression approach was very useful to confirm that there is a strong correlation. Supported by ETL processes and the at least partially automated determination of uncertain parameters by means of regression channels, the effort for the P&B data supply could be substantially reduced. Costs were saved, and the data quality was improved, due to a more realistic forecast. In the near future, it is intended to compare the output with the results of fuzzy approaches. Additionally, by means of time series analysis (see Chap. 10), it was found that due to climatic changes in the sales area the average amount of sun minutes per day and the temperatures tend to rise, which will have an impact on sales.

In steps 3 and 4 of the planning process ("application of the planning model" and "processing of the model results", see Fig. 11.2), a fuzzy linear optimisation approach was applied to get a valid and realistic plan without any artificial accuracy. A description of the approach and examples for the application can be found in Chap. 12 and in [25]. By means of this method, it is possible to address fuzziness of coefficients and objectives as well as fuzziness of constraint borders. Hence, it was not necessary to investigate all of the parameters in detail or to use unrealistically crisp parameters. Decision makers could decide whether additional effort should be invested to reduce fuzziness or not and could efficiently set up a realistic model of their P&B process. Based on the model results, flexible budgets for resources were approved, and the departments were allowed to spend more if they can achieve better results. Because of the positive climatic conditions in the following period and the flexible budgets, action was taken in time, and more ice cream could be sold.

11.6 Organisational and Cultural Prerequisites

To benefit from the approaches and ideas, which were presented in this chapter, it is necessary to establish some prerequisites in the environment of the companies. Quite often, traditional management models in the field are based on the assumption that human beings must be kept on a tight leash [3]. One consequence of this policy is that traditional P&B processes lead to fixed budgets, which are controlled strictly. The rigidity of these budgets prevents rapid adaptations and demoralises the staff [7]. Hence, if budget ranges are to be introduced, the impacts on all of the management levels have to be considered. Thus, the aspect of corporate culture and management should be reviewed before innovative instruments for data supply and further P&B steps are introduced. Experience with flexible plans and budgets are documented in the field of "beyond budgeting". "Beyond budgeting" can be interpreted as a philosophy which intends to make the P&B process and its results more flexible. This includes budgets, planning cycles and the management of human capital. Details can be found in [16, 22]. Successful implementations of this philosophy are usually accompanied by a paradigm change for the companies.

11.7 Summary and Conclusion

In this chapter, challenges and possible solutions in the field of P&B data supply were outlined. Without BI tools, the related P&B processes are usually performed manually, and because of an artificial accuracy of planning parameters, the quality of the P&B results suffers. To manage these issues, we have presented a set of regression analysis approaches. They allow the identification of essential parameters for business success. Moreover, some of the instruments presented can provide the P&B process with non-deterministic parameters. Also, the reduction of manual effort was discussed, and references for further processing steps in the P&B process were given.

The avoidance of working with pseudo-exactness has a lot of advantages. The results of the P&B process are more realistic and costs of information acquisition can be reduced. Revisions of plans and budgets will become rarer. On the other hand, it has to be considered that some organisational and cultural prerequisites should be fulfilled to establish such a flexible P&B approach. Furthermore, it has to be clear that the results of the related processes are also subject to some degree of uncertainty, since the proposed P&B data supply methods use more or less historical data input. Besides, it is still impossible to predict unexpected events which have an impact on planning parameters and planning objects. Despite these issues, companies could save a lot of money by automating data supply and improving data quality. Moreover, because of flexible budgeting, the motivation of staff can be improved, and a better result can be achieved.

Apart from the ideas suggested in this chapter, there is still potential for improvements. For instance, it would be interesting to combine the instruments presented here with performance management controlling tools. These tools could include components of early warning systems or risk management. They would be another important component to complete a closed-loop approach. Thus, benefits of the flexible P&B approaches presented could be better exploited and companies would be able to respond efficiently to the dynamics of markets.

References

1. Barrett, R.: Planning and Budgeting for the Agile Enterprise. Elsevier, Oxford (2007)
2. Beer, M.: Fuzzy probability theory. In: Meyers, R. (ed.) Encyclopedia of Complexity and System Science, vol. 6, pp. 4047–4059. Springer, New York (2009)
3. Bogsnes, B.: The World has changed—isn't it time to change the way we lead and manage? In: Balanced Scorecard Report, May–June, vol. 12, No. 3. Harvard Business Publishing, Harvard (2010)
4. Bollinger, J.: Bollinger on Bollinger Bands. McGraw Hill, New York (2002)
5. Borck, G., Pflaeging, N., Zeuch, A.: Making performance management work. BetaCodex Network Associates (2009). Available via http://www.betacodex.org/sites/default/files/paper/3/BetaCodex-PerformanceManagement.pdf. Accessed 30 October 2012
6. Brose, P., Corsten, H.: Bedeutung und Bestimmungsfaktoren subjektiver Wahrscheinlichkeiten. WiSt. Wirtschaftswiss. Stud. **12**(7), 329–335 (1983)
7. Caulkin, S.: An end to the numbers game. In: The Observer, 13.04.2003, p. 17. Available via http://www.bbrt.org/bb-briefing/files/Observer_030413.pdf. Accessed 30 October 2012
8. CMC: Markets technical indicators (v2.1 10th June 2009). Available via http://www2.cmcmarkets.com.au/repository/docs/help/Charting_glossary.pdf. Accessed 30 October 2012
9. Colvin, G.: Business's real problem: uncertainty, uncertainty, uncertainty. In: CNN Money (2012). Available via http://management.fortune.cnn.com/2012/08/08/business-economic-uncertainty. Accessed 30 October 2012
10. Elsner, J.B., Jagger, T.H.: On the increasing intensity of the strongest Atlantic hurricanes. In: Elsner, J.B., Hodges, R.E., Malmstadt, J.C., Scheitlin, K.N. (eds.) Hurricanes and Climate Change, vol. 2, pp. 175–190. Springer, Dordrecht (2010)
11. Fahrmeir, L., Kneib, T., Lang, S., Lineare Regressionsmodelle. In: Fahrmeir, L., Kneib, T., Lang, S. (eds.) Regression: Modelle, Methoden und Anwendungen, pp. 59–188. Springer, Berlin (2007)
12. Gabriel, T.J.: The Gabriel linear regression angle indicator: a new indicator for intermediate-term trading. J. Tech. Anal., Summer 1997. Available via http://www.mta.org/eweb/dynamicpage.aspx?webcode=journal-technical-analysis-1997-summer. Accessed 30 October 2012
13. Granger, C.W.J.: Forecasting in Business and Economics. Academic Press, New York (1980)
14. Hansen, D.R., Mowen, M.M., Guan, L.: Cost Management, Accounting & Control. South Western Cengage Learning, Mason (2009)
15. Hanuscheck, R.: Investitionsplanung auf der Grundlage vager Daten. Schulz-Kirchner Verlag, Idstein (1986)
16. Hope, J., Fraser, R.: Beyond Budgeting: How Managers Can Break Free from the Annual Performance Trap. Harvard Business School Press, Boston (2003)

17. Jinn, J.H., Song, C., Chao, J.C.: A study of fuzzy linear regression. In: InterStat, (6), July (2008). Available via http://interstat.statjournals.net/YEAR/2008/articles/0807006.pdf. Accessed 30 October 2012
18. Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**, 33–50 (1978)
19. Krzystanek, M., Lasota, T., Trawiński, B.: Comparative analysis of evolutionary fuzzy models for premises valuation using KEEL. In: Proceedings of the 1st International Conference on Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems, Proceeding ICCCI '09, Wrocław, Poland, pp. 838–849. Springer, Berlin (2009)
20. Neubauer, D.: Fuzzy-Regression bei Fehlern in den Daten: Modellierung und Analysepotentiale (2010). Available via http://digital-b.ub.uni-frankfurt.de/files/7917/10_0303_dneubauer_fuzzy_regression.pdf. Accessed 30 October 2012
21. Oehler, K.: Unterstützung von Planung, Prognose und Budgetierung durch Informationssysteme. In: Chamoni, P., Gluchowski, P. (eds.) Analytische Informationssysteme, pp. 359–394. Springer, Berlin (2010)
22. Pfläging, N.: Beyond Budgeting, Better Budgeting. Haufe Mediengruppe, Freiburg (2003)
23. Raff, G.: Trading the regression channel. In: Stocks & Commodities, vols. 9–10, pp. 403–408 (1991)
24. Rausch, P.: HIPROFIT—Ein Konzept zur Unterstützung der hierarchischen Produktionsplanung mittels Fuzzy-Clusteranalysen und unscharfer LP-Tools. Peter Lang Verlag, Frankfurt (1999)
25. Rausch, P., Rommelfanger, H.J., Stumpf, M., Jehle, B.: Managing uncertainties in the field of planning and budgeting—an interactive fuzzy approach. In: Research and Development in Intelligent Systems, vol. XXIX. Springer, London (2012)
26. Remenyi, D.: Stop IT Project Failures Through Risk Management. Butterworth-Heinemann, Oxford (1999)
27. Rommelfanger, H.J.: Fuzzy Decision Support-Systeme—Entscheiden bei Unschärfe, 2nd edn. Springer, Berlin (1994)
28. Seber, G.A.F., Wild, C.J.: Nonlinear Regression. Wiley, New York (2003)
29. Strobel, S.: Unternehmensplanung im Spannungsfeld von Ratingnote, Liquidität und Steuerbelastung. Verlag Dr. Kovac, Hamburg (2012)
30. Tanaka, H., Lee, H.: Fuzzy linear regression combining central tendency and possibilistic properties. In: Proceedings of the Sixth IEEE International Conference on Fuzzy Systems, Barcelona, Spain, pp. 63–68. IEEE Press, New York (1997)
31. Tanaka, H., Uejima, S., Asai, K.: Linear regression analysis with fuzzy model. IEEE Trans. Syst. Man Cybern. **12**, 903–907 (1982)
32. Tate, A.: Intelligible AI planning—generating plans represented as a set of constraints. Artificial Intelligence Applications Institute (2000). Available via http://www.aiai.ed.ac.uk/oplan/documents/2000/00-sges.pdf. Accessed 30 October 2012
33. Wang, G.C.S., Jain, C.L.: Regression Analysis: Modeling & Forecasting. Graceway Publishing, Flushing (2003)
34. Weber, J., Linder, S.: Neugestaltung der Budgetierung mit Better und Beyond Budgeting? Wiley-VCH, Weinheim (2008)
35. Wittmann, W.: Entscheiden unter Ungewissheit, Sitzungsbericht der Wissenschaftlichen Gesellschaft an der Johann Wolfgang Goethe-Universität Frankfurt a.M., Vol. 13, No. 3, Frankfurt (1975)
36. Zadeh, L.A.: A fuzzy-algorithmic approach to the definition of complex or imprecise concepts. Int. J. Man-Mach. Stud. **8**, 249–291 (1976)
37. Zimmermann, H.-J.: Fuzzy Sets in Operations Research – Eine Einführung in Theorie und Anwendung. In: Operations Research Proceedings, Berlin, pp. 594–608 (1985)

Chapter 12
Minimizing the Total Cost in Production and Transportation Planning—A Fuzzy Approach

Heinrich J. Rommelfanger

Abstract In this chapter, we deal with the production and transportation planning of a household appliances manufacturer that has production facilities and central stores for resellers in several sites in Europe. Each store can receive products from all production plants and it is not necessary that all products are produced in all production units. The transport between any two bases is done by trucks. For simplicity we assume, that each truck has the same capacity of M EURO-pallets, and for each product the unit is EURO-pallet. The target of this chapter is to determine a combined production and transport plan that minimize the total sum of the production cost and the transportation cost. For working in a realistic environment we assume that the production capacities in the different plants and the demand in the sales bases are not known exactly but the management can describe the data in form of fuzzy numbers. By using an inter-active algorithm for solving the fuzzy linear programming system we achieve a stable production and a satisfactory supply of the products. Moreover, we demonstrate that this integer programming problem can adequately be solved without using computation-intensive integer programming algorithms. Additionally, in the course of the inter-active solution process the production bottlenecks get clearly visible. A numerical example illustrates the efficiency of the proposed procedure.

12.1 Introduction

Planning and budgeting in firms or non-profit organizations is a sophisticated mission because the planners are usually confronted with different forms of uncertainty. For getting realistic and consistent plans not only probabilities should be noted but also inaccuracies, which are based on missing information and human deficiencies. In this chapter, we demonstrate that planning processes can be modelled in a realistic way by means of fuzzy systems. Moreover, the proposed interactive procedures are consistent with human thinking and lead to convincing results.

H.J. Rommelfanger (✉)
Faculty of Economics and Business Administration, Goethe University Frankfurt am Main,
Niebergallweg 16, 65824 Schwalbach am Taunus, Germany
e-mail: Rommelfanger@wiwi.uni-frankfurt.de

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_12, © Springer-Verlag London 2013

Due to globalisation and the involved international expansion of companies numerous firms in Europe produce their products in several European and Overseas countries. On the one hand they take advantage of the different production cost and of higher lot sizes; on the other hand they endeavour to minimize the total cost of manufacturing and transportation.

In literature transportation planning is often considered together with facility location decisions in order to minimize the total cost [2, 4]. But total planning is only appropriate if the construction of new production plants or sales bases is considered. Mostly the production facilities and the sales outlets are given and the problem is to organize the production in the factories and the transportation between the different plants and the stores with minimal cost.

Both the facility location problem and the production and transportation problem are usually solved by means of an integer programming system [3]. But this procedure is very computationally intensive. Moreover, it is neglected that real production capacities in the different plants and the demand for the products are not known exactly. In this chapter we look on the more realistic case that the management can describe the production capacities in the different plants and the demand in the sales bases in form of fuzzy numbers.

Fuzzy sets allow a realistic modeling of these parameters. As shown in [6], it is possible to describe a fuzzy subset A on a given set X very accurately by a membership function $\mu_A : X \to [0, 1]$, which assigns a membership value to each element of a set X. Iin practical applications it is sufficient to work with piecewise linear membership functions [6]. Figure 12.1 shows an example of a fuzzy production capacity. For a special plant the secure production capacity is p, but in ideal case a maximum capacity of $p + \pi$ units can be realized. Here, we have the special form of a triangular fuzzy number that the left-hand spread is equal zero. A triangular fuzzy numbers P can be abbreviated as $P = (p, \nu, \pi)$, where p is the mean value with the membership value 1 and ν, π are the left and right spreads. Here, we have the special form of a triangular fuzzy number where the left-hand spread is equal zero, therefore we can use the abbreviation $P = (p, 0, \pi)$.

The paper is organized as follows: In Sect. 12.2 the problem is formulated in detail. In Sect. 12.3 we demonstrate that the integer programming problem can adequately be solved by using an inter-active algorithm for solving the fuzzy linear programming system. A numerical example in Sect. 12.4 illustrates the efficiency of the proposed procedure. Finally, conclusions and possible extensions are presented in Sect. 12.5.

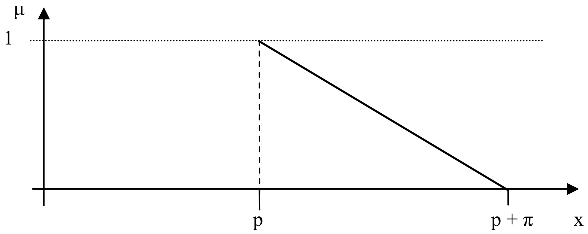


Fig. 12.1 Fuzzy production capacity

12.2 Problem Formulation

We deal with the production and transportation planning of a household appliances manufacturer HAM that has production facilities and central stores for resellers in several sites in Europe. Each store can receive products from all production plants and it is not necessary that all products are produced in all production units.

- The number of production and sales bases of HAM is N. Obviously, it is not necessary to differ between production plants and sales stores. A base without any demand is a pure production location and a base where nothing is produced is a pure sales shop.
- The number of products of HAM is K.
- The transport between any two bases is done by trucks. For simplicity we assume, that each truck has the same capacity of M EURO-pallets independent of the sort of products.
- For each product the unit is EURO-pallet.

In the production and transportation planning model, we use the following notations:

x_{ki} output of the product k at the base i,
y_{kij} total number of units of the product k transported from the base i to the base j,
s_{ki} total number of units of the product k for fulfilling the demand at the base i,
w_{ij} number of the trucks from base i to base j,
$\tilde{D}_{ki}$ demand for the product k at the base i,
$\tilde{P}_{ki}$ production capacity for the product k at the base i,
C_{ki} production cost of one unit of product k at the base i,
T_{ij} cost for transporting a truck from the base i to the base j, where $T_{ii} = 0$.

$$k \in \{1, \ldots, K\}, \quad i, j \in \{1, \ldots, N\}$$

The items $x_{ki}, y_{kij}, s_{ki}, w_{ij}, \tilde{D}_{ki}, \tilde{P}_{ki}$ are referred to the same time period, e.g. one week or one month.

Now, a production and transportation plan that minimizes the total sum of the production cost and the transportation cost can be determined by solving the following integer programming problem:

$$z(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{N} \sum_{k=1}^{K} C_{ki} x_{ki} + \sum_{j=1}^{N} \sum_{i=1}^{N} T_{ji} w_{ji} \to \text{Min}$$

$$\text{subject to} \tag{12.1}$$

$$x_{ki} \tilde{\leq} \tilde{P}_{ki}, \quad k=1,\ldots,K;\ i=1,\ldots,N$$

$$s_{ki} = x_{ki} - \sum_{\substack{j=1\\ j\neq i}}^{N} y_{kij} + \sum_{\substack{j=1\\ j\neq i}}^{N} y_{kji} \tilde{\geq} \tilde{D}_{ki}, \quad k=1,\ldots,K;\ i=1,\ldots,N$$

$$\sum_{k=1}^{K} y_{kij} \leq M w_{ij}, \quad i,j=1,\ldots,N$$

$$x_{ij}, y_{kij}, w_{ij} \in \mathbf{N} \cup \{0\}, \quad k=1\ldots,K;\ i,j=1,\ldots,N$$

$$\mathbf{x} = (x_{11},\ldots,x_{K1},x_{12},\ldots,x_{K2},\ldots,x_{1N},\ldots,x_{KN})$$

$$\mathbf{w} = (w_{11},\ldots,w_{N1},w_{12},\ldots,w_{N2},\ldots,w_{1N},\ldots,w_{NN})$$

The first constraint expresses that the output of the product k at the base i is smaller than or equal to its production capacity $\tilde{P}_{ki}$; the second constraint means that the supply s_{ki} of the product k at the base i is larger than or equal to the its demand $\tilde{D}_{ki}$; the third constraint indicates that the total amount of products transported from the base i to the base j is smaller than or equal to the transportation capacity of w_{ij} trucks; the fourth constraint means that the output of the product k at the base i, the number of products transported from the base i to the base j and the number of trucks from the base i to base j are nonnegative integer.

Concerning the imprecise right-hand sides $\tilde{P}_{ki}$ and $\tilde{D}_{ki}$ we assume that management is able to describe the production capacities in the different plants and the demand in the sales bases in form of fuzzy numbers $\tilde{P}_{ki} = (p_{ki}, 0, \pi_{ki})$ and $\tilde{D}_{ki} = (d_{ki}, \delta_{ki}, 0)$. Here, p_{ki} and $d_{ki} - \delta_{ki}$ are the production capacities and the demands respectively that are expected in any case. Furthermore the management estimate the maximal production capacities and the maximal demands as $p_{ki} + \pi_{ki}$ and d_{ki} respectively.

12.3 Solution Process

As shown in [9], fuzzy integer programming LP-problem can be effectively solved by interactive algorithms for solving fuzzy LP-systems. If we ignore for the moment that we look for a partial integer solution, the present system (12.1) is a relative simple model with one crisp objective function and soft constraints. Therefore we can use a special form of the algorithm FULPAL (**FU**zzy **L**inear **P**rogramming Based on **A**spiration **L**evels) for a stepwise calculation of an efficient compromise solution of the system (12.1).

At first, we have to calculate the smallest total cost $\underline{z}$ and the highest total cost $\bar{z}$ by solving the two crisp LP-systems:

$$
\begin{aligned}
&\underline{z} = \text{Min}\left(\sum_{i=1}^{N}\sum_{k=1}^{K} C_{ki}x_{ki} + \sum_{i=1}^{N}\sum_{j=1}^{N} T_{ji}w_{ji}\right)\\
&\text{subject to} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (12.2)\\
&\quad x_{ki} \le p_{ki} + \pi_{ki}, \quad k = 1,\ldots,K;\ i = 1,\ldots,N\\
&\quad s_{ki} = x_{ki} - \sum_{\substack{j=1\\ j\ne i}}^{N} y_{kij} + \sum_{\substack{j=1\\ j\ne i}}^{N} y_{kji} \ge d_{ki} - \delta_{ki},\\
&\qquad k = 1,\ldots,K;\ i = 1,\ldots,N\\
&\quad \sum_{k=1}^{K} y_{kij} - Mw_{ij} \le 0, \quad i, j = 1,\ldots,N\\
&\quad x_{ij}, y_{kij}, w_{ij} \ge 0, \quad k = 1\ldots,K;\ i, j = 1,\ldots,N
\end{aligned}
$$

and

$$
\begin{aligned}
&\bar{z} = \text{Min}\left(\sum_{i=1}^{N}\sum_{k=1}^{K} C_{ki}x_{ki} + \sum_{i=1}^{N}\sum_{j=1}^{N} T_{ji}w_{ji}\right) \qquad (12.3)\\
&\text{subject to} \quad x_{ki} \le p_{ki}, \quad k = 1,\ldots,K;\ i = 1,\ldots,N
\end{aligned}
$$

(In case of $\sum_{i=1}^{N} p_{ki} < \sum_{i=1}^{N} d_{ki}$ the capacities p_{ki} must be increased up to the existence of an feasible solution of (12.2), starting with the plants that have the highest production costs; $k = 1,\ldots,K$.)

$$
\begin{aligned}
&s_{ki} = x_{ki} - \sum_{\substack{j=1\\ j\ne i}}^{N} y_{kij} + \sum_{\substack{j=1\\ j\ne i}}^{N} y_{kji} \ge d_{ki}, \quad k = 1,\ldots,K;\ i = 1,\ldots,N\\
&\sum_{k=1}^{K} y_{kij} - Mw_{ij} \le 0, \quad i, j = 1,\ldots,N\\
&x_{ij}, y_{kij}, w_{ij} \ge 0, \quad k = 1\ldots,K;\ i, j = 1,\ldots,N
\end{aligned}
$$

In accordance with FULPAL, the objective function and the soft constraints of the system (12.1) are transformed in utility functions, where $z^A[r]$, $p_{ki}^A[r]$, $d_{ki}^A[r]$ with $r = 1$ are the crisp aspiration levels that are specified by the management for the time being. The process FULPAL strives for a satisfactory solution in the sense of bounded rationality, see [1, 10]. The stepwise calculation of optimal solutions serves only for creating possible solutions that fulfils the aspiration levels.

In the course of the inter-active solution process the management can change the aspiration levels step by step. For getting an effective comparability of the utilities, the same utility (membership degree) λ_A is assigned to all crisp aspiration levels. The goal of the management is to get a solution that satisfies all aspiration levels. If we abbreviate the aspiration levels of the step r with $z^A[r]$, $p^A_{ki}[r]$, $d^A_{ki}[r]$, we get the following membership functions:

$$\mu_z(\mathbf{x},\mathbf{w}) = \begin{cases} 1 & \text{if } z(\mathbf{x},\mathbf{w}) < \underline{z} \\ 1 - \frac{z(\mathbf{x},\mathbf{w}) - \underline{z}}{z^A[r] - \underline{z}} \cdot (1-\lambda_A) & \text{if } \underline{z} \le z(\mathbf{x},\mathbf{w}) < z^A[r] \\ \lambda_A + \frac{z(\mathbf{x},\mathbf{w}) - z^A[r]}{\bar{z} - z^A[r]} \cdot (1-\lambda_A) & \text{if } z^A[r] \le z(\mathbf{x},\mathbf{w}) \le \bar{z} \\ 0 & \text{if } \bar{z} < z(\mathbf{x},\mathbf{w}) \end{cases} \tag{12.4}$$

$$\mu_{xki}(x_{ki}) = \begin{cases} 1 & \text{if } x_{ki} < p_{ki} \\ 1 - \frac{x_{ki} - p_{ki}}{p^A_{ki}[r] - p_{ki}} \cdot (1-\lambda_A) & \text{if } p_{ki} \le x_{ki} \\ \lambda_A + \frac{x_{ki} - p^A_{ki}[r]}{p_{ki} + \pi_{ki} - p^A_{ki}[r]} \cdot (1-\lambda_A) & \text{if } p^A_{ki}[r] < x_{ki} \\ 0 & \text{if } p_{ki} + \pi_{ki} < x_{ki} \end{cases} \tag{12.5}$$

$$\mu_{ski}(s_{ki}) = \begin{cases} 0 & \text{if } s_{ki} < d_{ki} - \delta_{ki} \\ \frac{s_{ki} - (d_{ki} - \delta_{ki})}{d^A_{ki}[r] - (d_{ki} - \delta_{ki})} \cdot \lambda_A & \text{if } d_{ki} - \delta_{ki} \le s_{ki} \le d^A_{ki}[r] \\ \lambda_A + \frac{s_{ki} - d^A_{ki}[r]}{d_{ki} - d^A_{ki}[r]} \cdot (1-\lambda_A) & \text{if } d^A_{ki}[r] < s_{ki} \le d_{ki} \\ 1 & \text{if } d_{ki} < s_{ki} \end{cases} \tag{12.6}$$

$$k = 1, \ldots, K; \; i = 1, \ldots, N$$

For calculating a compromise solution of the multi-objective system

$$\big(\mu_z(\mathbf{x},\mathbf{w}), \mu_{x11}(x_{11}), \ldots, \mu_{xKN}(x_{KN}), \mu_{s11}(s_{11}), \ldots, \mu_{sKN}(s_{KN})\big) \to \text{Max}$$

subject to (12.7)

$$x_{ki} \le p_{ki} + \pi_{ki}, \quad k = 1, \ldots, K; \; i = 1, \ldots, N$$

$$s_{ki} = x_{ki} - \sum_{\substack{j=1 \\ j \ne i}}^{N} y_{kij} + \sum_{\substack{j=1 \\ j \ne i}}^{N} y_{kji} \ge d_{ki} - \delta_{ki},$$

$$k = 1, \ldots, K; \; i = 1, \ldots, N$$

$$\sum_{k=1}^{K} y_{kij} - M w_{ij} \le 0, \quad i, j = 1, \ldots, N$$

$$x_{ij}, y_{kij}, w_{ij} \ge 0, \quad k = 1 \ldots, K; \; i, j = 1, \ldots, N$$

we use the compromise objective function

$$\mu = \text{Min}\big(\mu_z(\mathbf{x}, \mathbf{w}), \mu_{x11}(x_{11}), \ldots, \mu_{xKN}(x_{KN}), \mu_{s11}(s_{11}), \ldots, \mu_{sKN}(s_{KN})\big) \tag{12.8}$$

In multi-objective programming an ideal solution, which fulfils all objective function at best, does usually not exist. Therefore, the total set of solutions consists of so-called pareto-optimal solutions. Pareto-efficient solutions are defined by the characteristic that no other solution exists, which is better concerning all objectives.

Then, we get a pareto-optimal compromise solution of (12.7) by solving the crisp mathematical programming system (12.9).

$$\begin{aligned}
&\lambda \rightarrow \text{Max}\\
&\text{subject to} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (12.9)\\
&\quad \lambda \leq \mu_z(\mathbf{x}, \mathbf{w})\\
&\quad \lambda \leq \mu_{xki}(x_{ki}), \quad k = 1, \ldots, K;\ i = 1, \ldots, N\\
&\quad \lambda \leq \mu_{ski}(s_{ki}), \quad k = 1, \ldots, K;\ i = 1, \ldots, N\\
&\quad x_{ki} \leq p_{ki} + \pi_{ki}, \quad k = 1, \ldots, K;\ i = 1, \ldots, N\\
&\quad s_{ki} = x_{ki} - \sum_{\substack{j=1\\ j\neq i}}^{N} y_{kij} + \sum_{\substack{j=1\\ j\neq i}}^{N} y_{kji} \geq d_{ki} + \delta_{ki},\\
&\qquad k = 1, \ldots, K;\ i = 1, \ldots, N\\
&\quad \sum_{k=1}^{K} y_{kij} - M w_{ij} \leq 0, \quad i, j = 1, \ldots, N\\
&\quad x_{ij}, y_{kij}, w_{ij} \geq 0, \quad k = 1 \ldots, K;\ i, j = 1, \ldots, N
\end{aligned}$$

As the management is only interested in a solution that satisfies all aspiration levels, it is sufficient to solve the following crisp linear LP-system, where $r = 1$; for details see [7, 8].

$$\begin{aligned}
&\lambda \rightarrow \text{Max}\\
&\text{subject to} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (12.10)\\
&\quad \big(z^A[r] - \underline{z}\big)\lambda + (1 - \lambda_A)\left(\sum_{i=1}^{N}\sum_{k=1}^{K} C_{ki} x_{ki} + \sum_{i=1}^{N}\sum_{j=1}^{K} T_{ji} w_{ji}\right) \leq z^A[r] - \lambda_A \underline{z}\\
&\quad \big(p_{ki}^A[r] - p_{ki}\big)\lambda + (1 - \lambda_A) x_{ki} \leq p_{ki}^A[r] - \lambda_A p_{ki},
\end{aligned}$$

$$k = 1, \ldots, K; \; i = 1, \ldots, N$$

$$\big(d_{ki} - d_{ki}^{A}[r]\big)\lambda - (1 - \lambda_A)\left(x_{ki} - \sum_{\substack{j=1 \\ j \neq i}}^{N} y_{kij} + \sum_{\substack{j=1 \\ j \neq i}}^{N} y_{kji} \right) \leq \lambda_A d_{ki} - d_{ki}^{A}[r]$$

$$k = 1, \ldots, K; \; i = 1, \ldots, N$$

$$\sum_{k=1}^{K} y_{kij} - M w_{ij} \leq 0, \quad i, j = 1, \ldots, N$$

$$\lambda, x_{ij}, y_{kij}, w_{ij} \geq 0, \quad k = 1 \ldots, K; \; i, j = 1, \ldots, N$$

According to the criteria defined above, a solution of the system (12.10) with $\lambda \geq \lambda_A$ is a pareto-optimal solution of the systems (12.7) and (12.1) and fulfils all aspiration levels $z^A[1]$, $p_{ki}^A[1]$, $d_{ki}^A[1]$.

Moreover, if λ is greater than λ_A, the management can increase some of the aspiration levels to $z^A[2]$, $p_{ki}^A[2]$, $d_{ki}^A[2]$ and can calculate a new pareto-optimal solution by means of the revised system (12.10), and so on. When the management is satisfied with the non-integer solution, it is time for looking for an integer solution. Due to the soft constraints several integer solutions will exist in the neighborhood of the non-integer solution.

12.4 Numerical Example

The manufacturer HAM produces four products P1, P2, P3 and P4 in three product plants in **E**ssen, **K**rakow and **L**yon. The numerical description of the current situation is given in Tables 12.1–12.6.

Using the LP-systems (12.2) and (12.3), we get the minimal total cost $\underline{z} = 864{,}100$ € and the maximal cost $\bar{z} = 1{,}091{,}700$ €. With all these information the management of HAM specifies for the total cost the aspiration level $z[1] = 950{,}000$ € and for the outputs and the demands the aspiration levels in Tables 12.7 and 12.8.

With all these information we calculate by means of the LP-system (12.10) the first production and transportation plan, see Table 12.9. As $\lambda = 0.53 > \lambda_A = 0.5$, this solution fulfils all aspiration levels. E.g. the total cost of this plan is 945,082.50 €.

Moreover this result indicates that it is possible to improve the aspiration levels.

For simplification we assume that the management is only interested in lower cost and decides to reduce the total cost to $z[2] = 940{,}000$ €.

Even the solution of the revised system (12.10) fulfils all aspiration levels. We get $\lambda = 0.5077$, $z = 938{,}712.20$ €.

Due to soft constraints, integer solutions usually exist in the neighborhood of non-integer solutions that have similar objective values. Therefore, it is very simple to derive the following integer plan that fulfils all aspiration levels and actually leads to lower total cost $z = 937{,}100$ €, see Table 12.10

Table 12.1 Production costs in € per unit

	P1	P2	P3	P4
E	200	400	330	500
K	180	350	280	
L	220		300	550

Table 12.2 Transportation costs in € per truck

	E	**K**	**L**
E		1000	600
K	1000		1200
L	600	1200	

Table 12.3 Secure production capacities p_{ki}

	p_{1i}	p_{2i}	p_{3i}	p_{4i}
E	600	450	250	250
K	350	200	250	
L	250		400	150

Table 12.4 Maximal production capacities $p_{ki} + \pi_{ki}$

	$p_{1i} + \pi_{1i}$	$p_{2i} + \pi_{2i}$	$p_{3i} + \pi_{3i}$	$p_{4i} + \pi_{4i}$
E	750	500	300	300
K	500	300	300	
L	300		500	200

Table 12.5 Minimal demands $d_{ki} - \delta_{ki}$

	$d_{1i} - \delta_{1i}$	$d_{2i} - \delta_{2i}$	$d_{3i} - \delta_{3i}$	$d_{4i} - \delta_{4i}$
E	650	300	300	150
K	450	100	200	80
L	150	150	250	100

Table 12.6 Maximal demands d_{ki}

	d_{1i}	d_{2i}	d_{3i}	d_{4i}
E	750	400	350	200
K	350	150	200	120
L	200	200	400	130

Table 12.7 Aspiration levels $p_{ki}[1]$

Output	$p_{1i}[1]$	$p_{2i}[1]$	$p_{3i}[1]$	$p_{4i}[1]$
E	720	480	280	280
K	450	270	280	
L	280		470	180

Table 12.8 Aspiration levels $d_{ki}[1]$

Demand	$d_{1i}[1]$	$d_{2i}[1]$	$d_{3i}[1]$	$d_{4i}[1]$
E	670	350	330	180
K	280	120	200	100
L	180	170	300	110

Table 12.9 Production and transportation plan

x1e	x1k	x1l	x2e	x2k	x3e
671.05	280.92	180.26	372.24	269.08	250.79
x3k	x3l	x4e	x4l	wek	wel
279.61	301.32	279.61	111.18	4.97	1.08
wke	wkl	wle	wlk	y1ek	y1el
3098	7043	0	0.05	0	0
y1ke	y1kl	y1le	y1lk	y2ek	y2el
0	0	0	0	0	21.58
y2ke	y2kl	y3ek	y3el	y3ke	y3kl
0	148.68	0	0	79.61	0
y3le	y3lk	y4ek	y4el	y4le	y4lk
0	0	99.34	0	0	0.92

According to this result the optimal production plan is in Table 12.11.

For transporting the products between the three bases, we need 5 trucks from Essen to Krakow, 1 truck from Essen to Lyon, 4 trucks from Krakow to Essen and 8 trucks from Krakow to Lyon.

By these transports the trucks deliver 20 units P2 from Essen to Lyon, 150 units P2 from Krakow to Lyon, 80 units P3 from Krakow to Essen and 100 units P4 from Essen to Krakow.

12.5 Conclusions

Fuzzy mathematical programming systems offer the possibility to model real problems as precisely as a decision maker is able to describe them. In any case the solution should be determined step by step in an interactive process, in which additional information out of the decision process itself or from outside should be used. In do-

Table 12.10 Integer production and transportation plan

x1e	x1k	x1l	x2e	x2k	x3e
670	280	180	370	270	250
x3k	x3l	x4e	x4l	wek	wel
280	300	280	110	5	1
wke	wkl	wle	wlk	y1ek	y1el
4	8	0	0	0	0
y1ke	y1kl	y1le	y1lk	y2ek	y2el
0	0	0	0	0	20
y2ke	y2kl	y3ek	y3el	y3ke	y3kl
0	150	0	0	80	0
y3le	y3lk	y4ek	y4el	y4le	y4lk
0	0	100	0	0	0

Table 12.11 Production per unit

	P1	P2	P3	P4
E	670	370	250	280
K	280	270	280	
L	180		300	110

ing so inadequately modeling of the real problem can be avoided and information costs will, in general, be decreased [9].

Another advantage of fuzzy models is the fact that (mixed) integer programming problems can be solved very easily because the boundaries are not crisp but fuzzy and points with integer variables in the neighborhood of an optimal solution are feasible in general.

In this chapter we have assumed that only the right-hand sides of some constraints are not known exactly. The model can be extended to the case that additionally coefficients of the objective function or of the constraints are described in form of fuzzy intervals. Moreover, it is possible to allow several objective functions. All these systems can adequately be solved by means of the inter-active algorithm FULPAL, see [7–9].

In all, this chapter is only one application that demonstrates the advantages of fuzzy systems for modelling and solving real world problems. An example how fuzzy systems may be helpful in the field of planning and budgeting is given in [5].

References

1. Becker, S.W., Siegel, S.: Utility of grades: level of aspiration in a decision theory context. J. Exp. Psychol. **55**, 81–85 (1958)
2. Bhutta, K.S.: International facility location decisions: a review of the modelling literature. Int. J. Integr. Supply Manag. **1**, 33–50 (2004)
3. Chopra, S., Meindl, P.J.: Supply Chain Management: Strategy, Planning and Operation. Pearson/Prentice Hall, Upper Saddle River (2007)
4. Kouvelis, P., Rosenblatt, M.J., Munson, C.L.: A mathematical programming model for global plant location problems: analysis and insights. IIE Trans. **36**, 127–144 (2004)
5. Rausch, P., Rommelfanger, H., Stumpf, M., Jehle, B.: Managing uncertainties in the field of planning and budgeting – an interactive fuzzy approach. In: Proceedings of the 32nd SGAI Conference, Cambridge (2012)
6. Rommelfanger, H.: Fuzzy Decision Support-Systeme – Entscheiden bei Unschärfe. Springer, Berlin (1994)
7. Rommelfanger, H.: FULPAL 2.0 – an interactive algorithm for solving multicriteria fuzzy linear programs controlled by aspiration levels. In: Scheigert, D. (ed.) Methods of Multicriteria Decision Theory, pp. 21–34 (1995). Pfalzakademie Lamprecht
8. Rommelfanger, H., Slowinski, R.: Fuzzy linear programming with single or multiple objective functions. In: Slowinski, R. (ed.) Fuzzy Sets in Decision Analysis, Operations Research and Statistics, pp. 179–213. Kluwer Academic, Norwell (1998)
9. Rommelfanger, H.: The advantages of fuzzy optimization models in practical use. Fuzzy Optim. Decis. Mak. **3**, 295–309 (2004)
10. Simon, H.A.: Behavioral model of rational choice. Q. J. Econ. **69**, 99–118 (1955)

Chapter 13
Design and Automation for Manufacturing Processes: An Intelligent Business Modeling Using Adaptive Neuro-Fuzzy Inference Systems

Alaa F. Sheta, Malik Braik, Ertan Öznergiz, Aladdin Ayesh, and Mehedi Masud

Abstract The design and automation of a steel making process is getting more complex as a result of the advances in manufacturing and becoming more demanding in quality requirements. It is essential to have an intelligent business process model which brings consistent and outstanding product quality thus keeping the trust with the business stakeholders. Hence, schemes are highly needed for improving the nonlinear process automation. The empirical mathematical model for steel making process is usually time consuming and may require high processing power. Fuzzy neural approach has recently proved to be very beneficial in the identification of such complex nonlinear systems. In this chapter, we discuss the applicability of an Adaptive Neuro-Fuzzy Inference System (ANFIS) to model the dynamics of the hot rolling industrial process including: roll force, roll torque and slab temperature. The proposed system was developed, tested as well as compared with other existing systems. We have conducted several simulation experiments on real data and the results confirm the effectiveness of the ANFIS based algorithms.

A.F. Sheta (✉) · M. Masud
Department of Computer Science, College of Computers and Information Technology,
Taif University, Taif, Saudi Arabia
e-mail: asheta@tu.edu.sa

M. Masud
e-mail: mmasud@tu.edu.sa

M. Braik
Electronic, Electrical and Computer Engineering Department, University of Birmingham,
Birmingham, Edgbaston, UK
e-mail: MSB158@bham.ac.uk

E. Öznergiz
Marine Engineering Operations Department, Faculty of Naval Architecture and Maritime,
Yildiz Technical University, Istanbul, Turkey
e-mail: oznergiz@itu.edu.tr

A. Ayesh
Faculty of Technology, De Montfort University, Leicester LE1 9BH, UK
e-mail: aayesh@dmu.ac.uk

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_13, © Springer-Verlag London 2013

13.1 Introduction

Due to the increasing quality requirements of steel products over the past few years, with very strict limits to meet the market demands, steel production system demands more accurate and speedy automated systems. A major stage in steel production is hot rolling. In order to produce a good quantitative description of the industrial operation, the automation of hot rolling processes requires the development of several mathematical models to identify the simulation process including system parameters and variables [1]. Although empirical mathematical models demand high processing power and more computation time, but still give poor performance. Providing fast, reliable, and accurate models are of great importance for predicting the roll force, roll torque and slab temperature. These models are significantly useful for a hot rolling process in order to generate pass schedules on-line [2]. Section 13.3 presents a more detailed description of the hot rolling process.

The advances in Fuzzy Logic (FL) and Artificial Neural Networks (ANNs) research have opened avenues for new advances in system modeling and identification. An application of fuzzy neural to model and control nonlinear industrial processes has been intensively studied in recent years [3, 4]. Moreover, ANNs have been used to assist in building a reasonable model structure for physical nonlinear systems to serve for process control [4]. FL has been used to develop a mathematical model for many industrial processes and showed significant improved results [5, 6]. Many researchers have focused their works on developing new automation techniques to meet the required quality of hot rolling processes [1, 7–9].

Complex generated models based on neural networks and fuzzy logic make it possible to work with higher performance of rolling forces, larger reductions and better flatness control [10, 11]. Application for quality monitoring in hot rolling process based on ANNs and fuzzy logic was presented in [12]. Accordingly, there is opportunity of further research to model the rolling process both theoretically and empirically based on the data measured in experimental or industrial rolling operations [6, 13, 14]. To improve the prediction ability of the rolling force model, many researchers have focused on evolving more effective physical models for the rolling force prediction [1]. Many important requirements of the hot rolling process including roll force, roll torque and slab temperature, should be affirmed in modeling a plate hot rolling process. Perhaps the most important requirement that should be exactly determined is the temperature of the slab at the entry of each pass in the rolling schedule, because of the fact that the strength of hot steel is highly dependent on temperature [15]. The prediction accuracies for the rolling force and torque are still vital issues from the point of the physical limitations of the rolling mill, and hence there are other requirements that must be emphasized, too [16].

In this chapter, we present an Adaptive Neuro-Fuzzy Inference System to model the structure of a hot rolling manufacturing process which consists of three subsystems. Three models for force, torque and slab temperature in the plate mill are precisely developed. The obtained results are compared to the earlier observed results based on fuzzy logic [6], conventional mathematical models [2] and FeedForward (FF) neural network models.

The methodologies which will be outlined in this chapter can also be transferred to other problem domains. They can help to analyse data which, for instance, is provided by a data warehouse. Besides, the approaches presented in the following sections can also be valuable for performance management, for instance in the field of business activity monitoring.

13.2 Motivation to use ANFIS?

Neural Networks (NNs) have been used successfully in identification of nonlinear systems. However, the conventional NN techniques for locating a suitable mathematical model from the input–output dataset of a system follow three familiar factors:

1. The first factor is related to the experimental data being driven, where the model qualities are mostly influenced by the quality of data being used.
2. The second factor is concerned with the network architecture, whereas different network architectures may result in a different estimation performance.
3. The third factor is directly correlated with the model size and its complexity. This factor is strongly dependent on the network training, which may be the most important factor, as it holds as an identification task to the model parameters that must fit with the given data.

Actually, a small network may not be able to represent the real situation of the model estimation, due to its limited capability, while the results of a large network may be distorted by noise or over fitting in the training data, which fails to provide a good generalization. Because of these reasons, search for adaptive modeling techniques are being pursued. Two observations have been made. First, the behavior of most dynamic systems is nonlinear. Second, Fuzzy Neural approach is suitable for system identification task, and it avails working with nonlinear systems. Adaptive Neuro-Fuzzy Inference System (ANFIS) [17] is a well-established method of combining Fuzzy and Neural approaches. ANFIS has been shown to be effective in identifying a model for a sufficient input–output data driven [3, 4], and is able to approximate any continuous function to an arbitrary accuracy. Considering the observations, we present in this chapter the use of ANFIS models in dealing with hot rolling plants automation as part of steel making systems.

13.3 Hot Rolling Plants: Problem Domain

13.3.1 Process Description

The processes involved in mill plants and production of steel have become more complicated. In this aspect, the hot rolling industrial process was considered as a

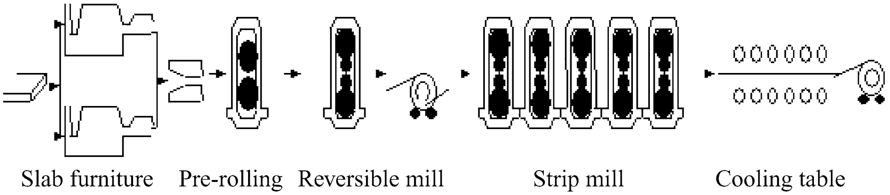


Fig. 13.1 Diagram of the hot rolling mill plant at the Ereĝli Iron and Steel Factory in Turkey

Table 13.1 The chemical composition of the low carbon steel (ASTM A53-96 Gr. A)

Carbon	Manganese	Sulfur	Phosphorus	Silicon
0.12 %	0.25 %	0.2 %	0.025 %	0.05 %

plant-wide problem. Ereĝli Iron and Steel factory in Turkey used the hot rolling mill plant (see Fig. 13.1) to provide the experimental data. Hot rolling process [13, 14] is based on an actual system developed in Ereĝli Iron and Steel factory. The system is a well-posed problem for analysis and identification design of a nonlinear rolling process. A large number of interacting processes and manipulated variables are incorporated into the model, making it a truly significant plant-wide problem. The plant consists of two slab furnaces, pre-rolling mill, edger, reversible mill, seven strip rolling stands, a cooling system, a hot leveller, and a shearing system. The plant has also a data acquisition and a computer control system modified by General Electrics.

Data acquisition and computer control systems are normally controlled by operators to achieve certain system performance goal. In manufacturing process, a mechanical or electrical controller is used to adjust the cooling system and the hot leveller for certain level. Steel strips with a thickness of 15–16 mm can be produced in the rolling mill plant. In a normal production cycle, each slab passes five times in forward and backward directions in the reversible mill. In this plant, the dimensions of slabs are monitored continuously during every passes with X-ray system, the temperature of slab with a pyrometer, roll force and torque with four load cells placed along the mill. But averages of these measured values for each pass are used for identification. The nonlinear dynamics of the rolling mill plant are mainly due to the chemical composition of the low-carbon steels within the hot rolling. The chemical composition of the low-carbon steels used, in this study, is given in Table 13.1.

13.3.2 Hot Rolling Process

The Ereĝli Iron and Steel Factory has four rolling mill plants: two are cold and two are hot. Cold rolling plants have a total capacity of 2.3 million tons per year. Hot rolling mill plants have 540.000 tons per year capacity. This corresponds to a total product capacity of 2.84 million tons per year. The dataset that describes the behavior of the rolling process was collected in order to measure different outputs of the rolling process and how it responds to various inputs. The dataset consists of 640 points and was generated from 128 different slabs by a General Electric's data acquisition system. The thickness and width distribution of the data ranged from 31.68 mm to 168.6 mm and 948.76 mm to 1457.26 mm, respectively.

13.4 System Identification Process

It is a common practice in engineering modeling for control system design is to first create and test a model offline for the system. This is called indirect control modeling. Identifying and modeling the hot rolling process requires some procedures.

The system identification process consists of constantly adopting a class of model structures, picking up the best model in the structure, and testing the model's performance whether it is acceptable. We use it for modeling the three sub-processes of the hot rolling process. The succession can be summarized as follows:

1. Experimental design: Collect input–output data from the process to be identified.
2. Pre-processing the data: Clean it to remove trends and outliers, and data scaling can be applied.
3. Select a class of models: Define a set of candidate systems in which a solution can be found.
4. Select a model structure: Pick the best model in the model structure set according to the input–output data and the selected performance criteria.
5. Model estimation: Estimate the model parameters and check the developed model's properties.
6. Model validation: If the model passes a given criterion, then stop; otherwise go back and try another model set.

13.4.1 Experimental Design

The collected measurements were practically measured for roll force, roll torque and slab temperature of the rolling process. We emphasis on modeling the hot rolling process using a neuro-fuzzy approach. The process is divided into three sub-processes: (i) the force f, (ii) the torque G and (iii) the slab temperature T. Each of the sub-process has six input variables. The force f and the torque G have the same input variables; u_1, u_2, u_3, u_4, u_5 and u_6. These inputs are standing for Entry Temperature (T_i), Width (W_s), Carbon Equivalent (C_e), Gauge (h_i), Draft (i) and Roll diameter (R), respectively. The six input variables u_1, u_2, u_3, u_4, u_5 and u_6 for the slab temperature T are the T_i, W_s, C_e, h_i, Torque (G_i), Power (E_i), respectively. The output of each sub-process is stated as $y(k)$. Figure 13.2 shows the six main inputs for both the force and the torque, while Fig. 13.3 shows the six main inputs for each of the slab temperature of the rolling mill plant.

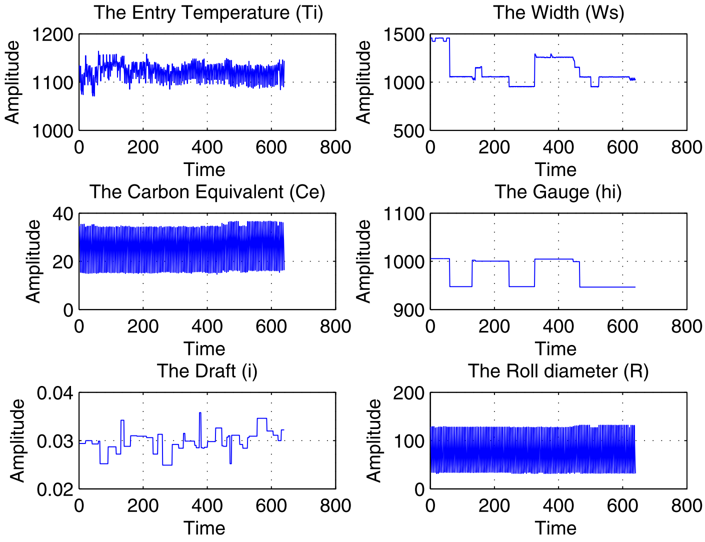


Fig. 13.2 The data input for the roll force/torque models

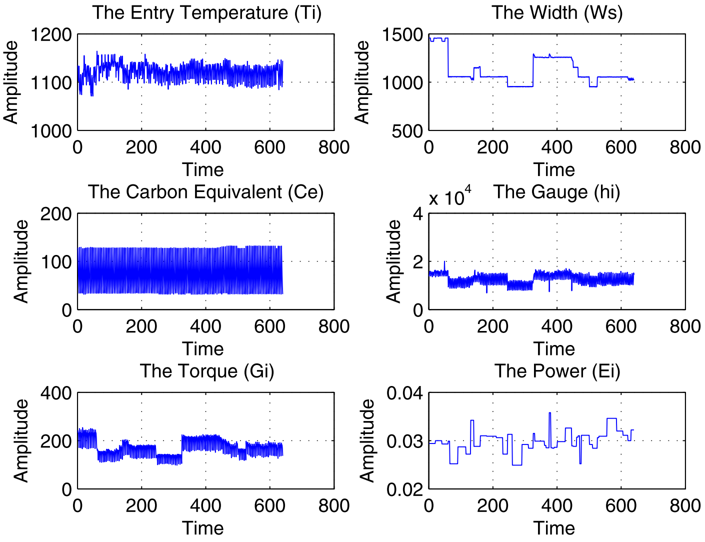


Fig. 13.3 The data input for the slab-temperature model

13.4.2 Pre-processing the Data

Model development based on the neuro-fuzzy approach requires some necessary preparation stages, which must be completed first to provide a good modeling process. These stages include data collection, preparation and suitable system modeling. The dataset is split into two parts: (i) the training dataset, which is used to train the neuro-fuzzy model and (ii) a testing dataset, which is used to verify the accuracy of the developed neuro-fuzzy model. The training dataset consists of approximately 78 % of the total dataset and the testing dataset consists of the remaining 22 %.

The quality and quantity of the training data is an important issue for neural networks training and for the accuracy of the fuzzification process. Usually, the success of neural network performance relies heavily on large amounts of data, but this demands more computing time for training. In order to reduce the amount of data whilst maintaining the model quality, the data used must be carefully selected to ensure that they are sufficiently 'rich'. A major concern with the high precision of neuro-fuzzy is data pre-processing, the scaling of data is needed to prevent data with larger magnitude from overriding the smaller, and impeding the premature learning process. In our study, the input and output data are scaled in the range of (0.1–0.9).

13.4.3 Select a Class of Models

The known empirical modeling techniques were unable to compensate the changes in size and chemical components. Therefore, carbon equivalent (C_e) and width (W_s) were added as inputs to the neuro-fuzzy model of the roll force and the roll torque. It is also difficult to get the correct force and torque values from the mathematical model itself. This is because a part of slight changes there would exist a highly nonlinear and complex structure, and with some various conditions some un-measurable parameters appear, such as, friction coefficient, yield stress, and disturbances. Moreover, reduction, chemical composition and temperature factors cannot be well suited and are not being considered in the mathematical model [2]. Also, a typical adaptation of these mathematical models to such a process would be a kind of fragmented look-up table. This approach has some drawbacks arising from the large size of the look-up tables due to the large product variety.

We consider the neuro-fuzzy approach in our study because it has a capability for good quality prediction, and it can produce a pure algebraic relationship between outputs and inputs. This means the predictor will be stable even if the system is not [3]. So we can overcome the instability of immeasurable parameters in the cases of torque and force. The ANFIS was used for evaluating and testing the neuro-fuzzy output error between the actual and the estimated outputs. ANFIS uses Fuzzy Inference System (FIS) structure such that relationships between train data and test data are adjusted until the specified inputs yield the desired output. Through these activities, the ANFIS learns the correct input–output response behavior. Thus, the ability of neuro-fuzzy models to model the hot rolling process can be improved.

13.4.4 Select a Model Structure

ANFIS is one of the most successful schemes which combines the benefits of both neural and fuzzy paradigms into a single channel [17]. ANFIS works by applying neural learning rules to identify and tune the parameters and the structure of FIS. The ANFIS is a multilayer feedforward network which uses ANN learning algorithms and fuzzy reasoning to characterize an input space to an output space. The architecture of the employed ANFIS is developed in the form of a zero-order Takagi-Sugeno-Kang (TSK) fuzzy inference system [18, 19].

ANFIS uses a hybrid learning algorithm to identify the membership function parameters of single-output, Sugeno type FIS. The architecture of ANFIS has been suggested by Roger Jang [17], which can be used for tuning the membership functions (i.e. the membership functions bounds) leading to improved performance. ANFIS requires the antecedent MFs and fuzzy rules in the training phase to initialize the neuro-fuzzy system; the MFs should be specified before the training. This employed training process is stopped whenever the designated iteration number is reached, or the training error goal is achieved. In this research, an ANFIS model is conducted to predict the future actions of the hot rolling process in an effort to formalize an identification task.

13.4.5 Model Estimation

The testing and validation processes are among the important steps in developing an accurate process model. The validation was performed by calculating some of the measurement criteria to evaluate the proposed models. The testing stage includes a criterion of fit and an iterative search algorithm. A neuro-fuzzy approach was used for estimating the hot rolling process, because it provides a rapid convergence and generally can be considered a very robust approach. The capability of neuro-fuzzy approach to emphasize the model validity was assured using the Mean Square Error (MSE) criterion between the actual and the estimated outputs. The relationship between the fuzzy model input and output is represented by what is called the Membership Function (MF).

In principle, the model validation should not only validate the accuracy of the model, but also verify whether the model can be easily interpreted to give a better understanding of the modeled process. It is therefore important to combine data-driven validation, aiming at checking the accuracy and robustness of the model, with more subjective validation, concerning the interpretability of the model. Takagi-Sugeno (TS) fuzzy models, which is based on fuzzy rules with crisp conclusions, were suitable to model a large class of dynamic datasets as stated in [20–22]. Once the model structure and parameters have been identified, it is necessary to validate the quality of the resulting model.

13.4.6 Model Validation

The performance of the ANFIS models in both training and testing data are evaluated, and the best training/testing data set is selected according to MSE and Variance-Account-For (VAF) [20]. VAF is computed to measure how close the measured values are to the developed values. The VAF is defined in Eq. (13.1).

$$\mathrm{VAF} = \left[1 - \frac{var(y - \hat{y})}{var(y)})\right] \times 100\ \% \tag{13.1}$$

where, y and $\hat{y}$ are the actual output and the estimated neuro-fuzzy model output, respectively.

13.5 Experimental Setup and Algorithms

13.5.1 Data Preparation

We consider the length of training and testing data set of 78 % and 22 % of the samples to improve the generalization properties of the adopted ANFIS. For each case of the rolling process, two ANFIS models of the same size, but different in initialization weights, were trained to study the stability and robustness of the each model. The best weights, which give the minimum MSE of two different training sessions over each input/output training set, were chosen as the final ANFIS models. Overall, the neuro fuzzy approach based on the adopted ANFIS model was accepted for modelling the hot rolling process, since it gives reasonable results and the prediction output is closely related to the actual output.

13.5.2 Learning Methodology

The neuro fuzzy system with the learning capability of neural network and with the advantages of the rule-base fuzzy system can improve the performance significantly. It can also provide a mechanism to incorporate past observations into the classification process. This approach uses neural networks for the membership function and mapping between fuzzy sets that are utilized as fuzzy rules. In the training process, a neural network adjusts its weights in order to minimize the MSE. According to the neuro-fuzzy approach, a neural network is utilized to implement the fuzzy system and to automatically tune the system parameters. The ANFIS structure is accomplished by defining, adapting and optimizing the topology and the parameters of the corresponding neuro-fuzzy network. The neuro-fuzzy models are trained based on ANFIS training approach. The checking data method uses the validation data to prevent over fitting of the training dataset that has the same format as the training

Table 13.2 Training parameters of the Hot Rolling ANFIS model

The ANFIS parameter type	Value
TSK Type	Zero-order
Number of iterations	200
Training error goal	0
Initial step size	0.001
Number of inputs	6
Number of MFs Gaussian	6
Step increasing rate	1.5
Step decreasing rate	0.1
Total fitting parameters	104
Consequent (linear) parameters	80
Premise (nonlinear) parameters	24
Number of nodes	55
Number of rules	64

data. The toolbox function ANFIS in MATLAB constructs the FIS whose membership function parameters are tuned. The configuration parameters of the employed ANFIS for modeling the rolling process are shown in Table 13.2.

A neuro fuzzy system is a combination of neural network and fuzzy systems combined in such a way, in which neural networks are used to determine the parameters of the fuzzy system, with a kind of automatic tuning method. To be more precise, optimizing the parameters, which are linearly related in a nonlinear way, neural networks and nonlinear optimization can be employed. A fuzzy model can be seen as a layered structure network similar to neural networks.

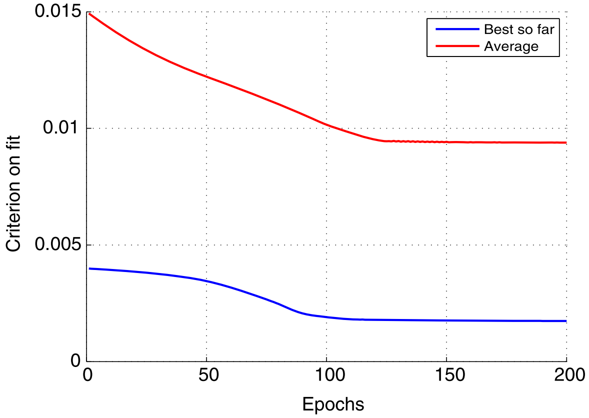
13.6 Experimental Results

13.6.1 Developed Neuro-Fuzzy Models

In this chapter, a neuro-fuzzy is proposed as a compensator of both fuzzy logic and neural networks to build a suitable model structure for three subsystems of the hot rolling process. It performs as a powerful method which has the ability to cover all the system variances for a typical rolling mill process. The neuro-fuzzy models are capable of producing estimated outputs similar to the actual outputs of each subsystem. This is accomplished by using the ANFIS method based on the MSE training errors.

After training and testing cases, the MSE became steady in all subsystems of the hot rolling system. We run 25 experiments to produce our results. The best and the average error are reported over all the experiments. For example, in the slab

Fig. 13.4 Convergence curve of the neuro-fuzzy slab-temperature model



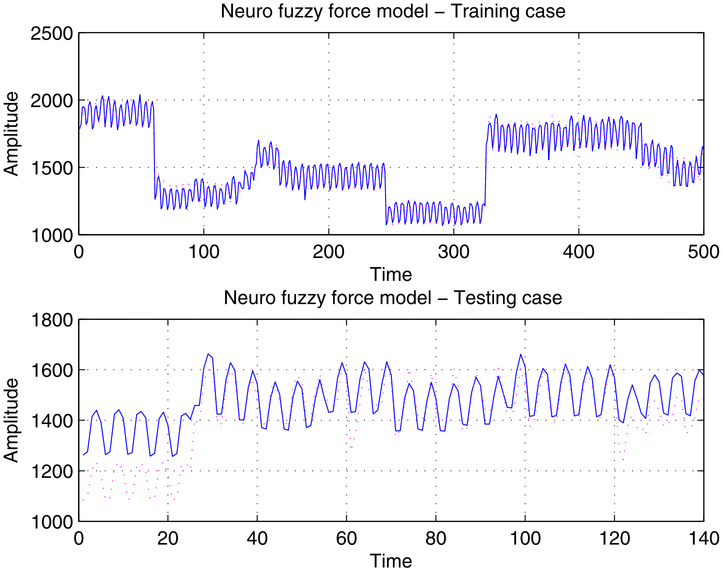


Fig. 13.5 Actual and estimated force response of the ANFIS model in both training and testing cases

temperature subsystem, the convergence training results of best and average are shown in Fig. 13.8.

It is observed from the developed error curves in Fig. 13.4 that the errors converged to optimum MSE values. The actual and estimated responses of the neuro-fuzzy models for the rolling process are shown in Figs. 13.5, 13.6 and 13.7. Figures 13.5, 13.6 and 13.7 show that the error between the actual and the predicted output of the model is very insignificant. This means that the neuro-fuzzy approach has learned to model the dynamics of a hot rolling process quite accurately. Over-

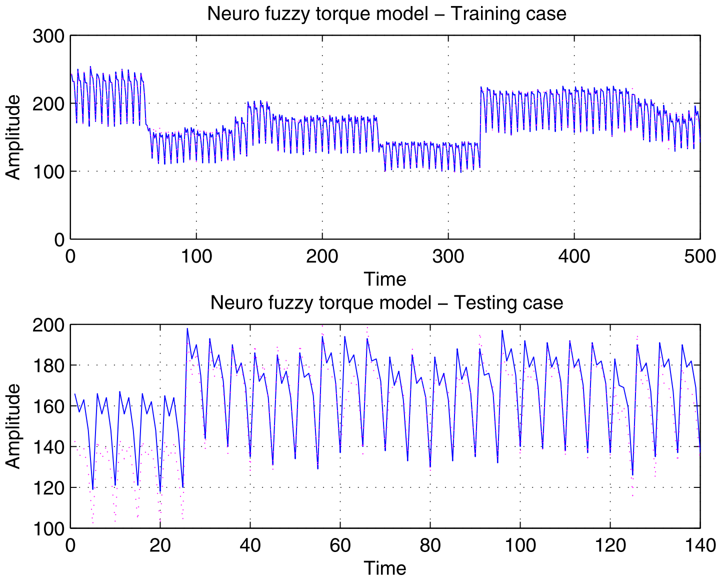


Fig. 13.6 Actual and estimated torque response of the ANFIS model in both training and testing cases

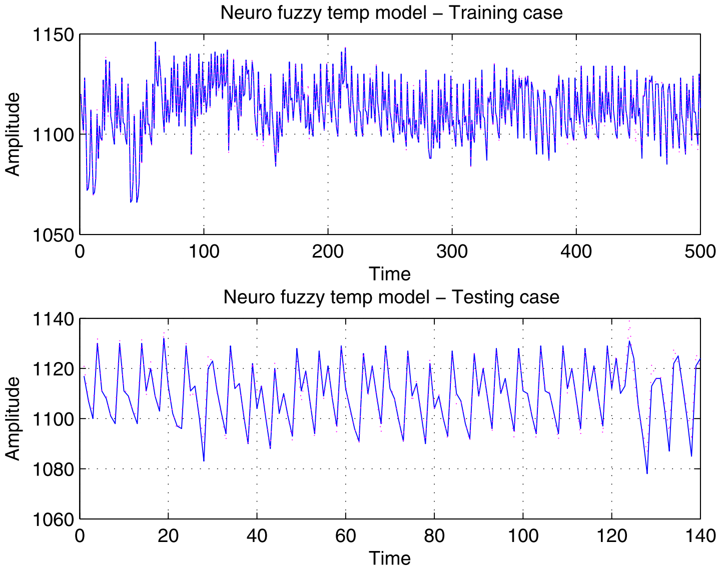


Fig. 13.7 Actual and estimated temperature response of the ANFIS model in both training and testing cases

all, all models display promising results in the training and testing sets for all the developed models.

13.6.2 Developed FeedForward Levenberg-Marquardt (FF-LM) Models

We avoid a restrictive mathematical model by selecting a particular class of models because of the limitations of the traditional model building approaches. The capability of neural networks to learn from examples seem to make it an ideal choice for modeling the hot rolling process. Multi-layer feedforward networks are the first to be used for the identification purposes. The neural networks of interest that we considered in this chapter are the feedforward backpropagation networks. The network model was trained based on Levenberg-Marquardt (LM) algorithm [23, 24]. The LM search algorithm was used as an iterative training algorithm, because it provides a rapid convergence. The MSE criterion was used for evaluating the ANN output using the LM optimization algorithm.

The network structure has an input layer, one hidden layer and an output layer with ten nodes in the hidden layer. The algorithm starts by assigning a random set of weights to the ANN, and the network adjusts its weights each time it comes across an input–output pair. The weights are adjusted in order to minimize the errors, where, the errors are propagated back to the connections preceding from the input nodes and the weights are adjusted accordingly. This type of model has been accepted for modeling the hot rolling process, as it produces a prediction output which is very close to the real actual output.

A feedforward back-propagation network approach based on LM algorithm was implemented to perform the same modeling task as in the case of the neuro-fuzzy approach. In Fig. 13.8, we show the error convergence when training a FF-LM neural network. The performances of the neural models in tracking the actual data in each sub-process of the hot rolling process are illustrated in Figs. 13.9, 13.10 and 13.11. The neural network models outputs follow the desired outputs quite closely. This indicates that the MSE reached the global minimum for all subsystems of the hot rolling system. As a result, the NN based LM method capable to learn the behavior of the industrial processes. The VAF values are considered for evaluating the experimental results.

13.6.3 Comparisons

In the neuro-fuzzy approach, an error tolerance of zero was used, and the ANFIS was trained with 200 epochs. The best error convergence curve in roll force achieved minimum MSE values at the last iteration of 0.0140 and 0.0187 for training and testing, respectively. While the roll torque achieved minimum MSE values to be 0.0134

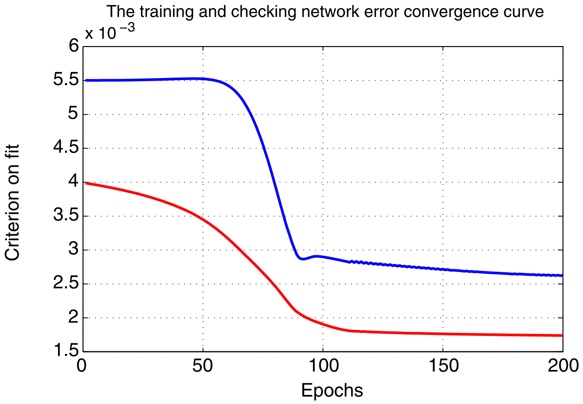


Fig. 13.8 Convergence curve of the neuro-fuzzy slab-temperature model

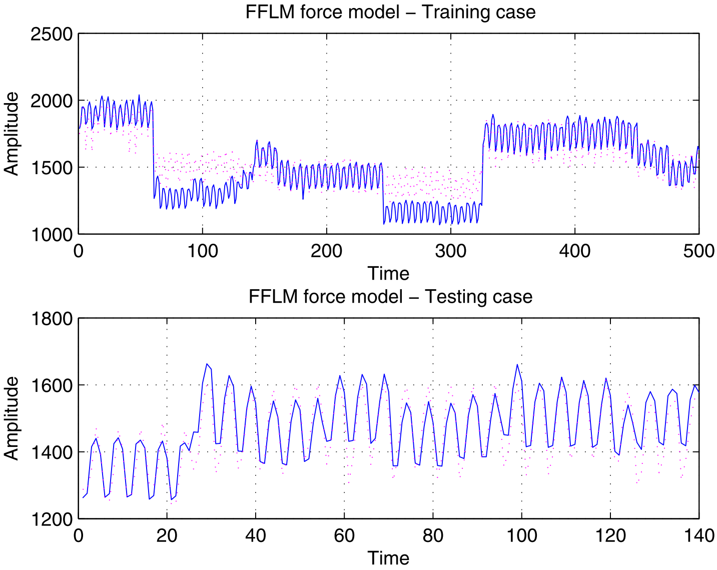


Fig. 13.9 Actual and estimated response of the force FF-LM model in both training and testing cases

and 0.0569, at the last iteration for training and testing cases, respectively. recorded best MSE values at the last iteration were 0.0017 and 0.0026 for training and testing, respectively. In Table 13.3, the VAF computed values using neuro-fuzzy, FF-LM, fuzzy logic and the empirical model as presented in [6] and [16], respectively are reported.

Roll torque and slab temperature models on the other hand, yield good results. This means that these models are highly stable towards the end of processing phase. Moreover, the performance of neuro-fuzzy for modeling the rolling process is slightly better than both of fuzzy logic and FF-LM, and shows better perfor-

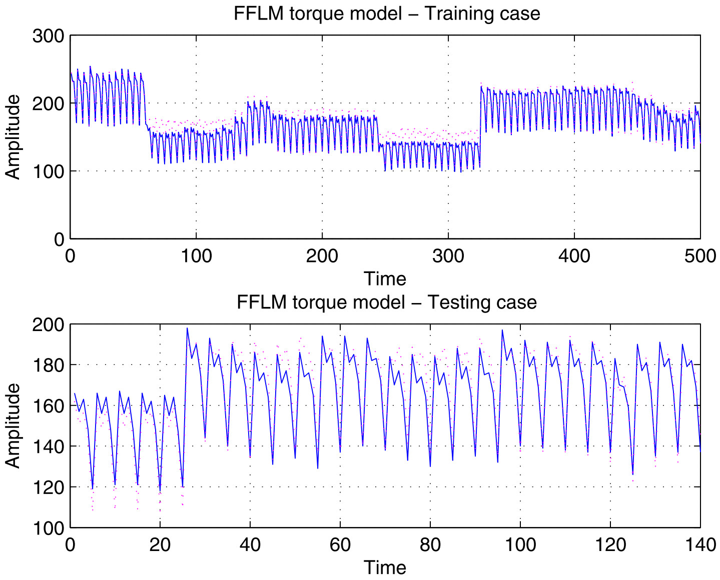


Fig. 13.10 Actual and estimated response of the torque FF-LM model in both training and testing cases

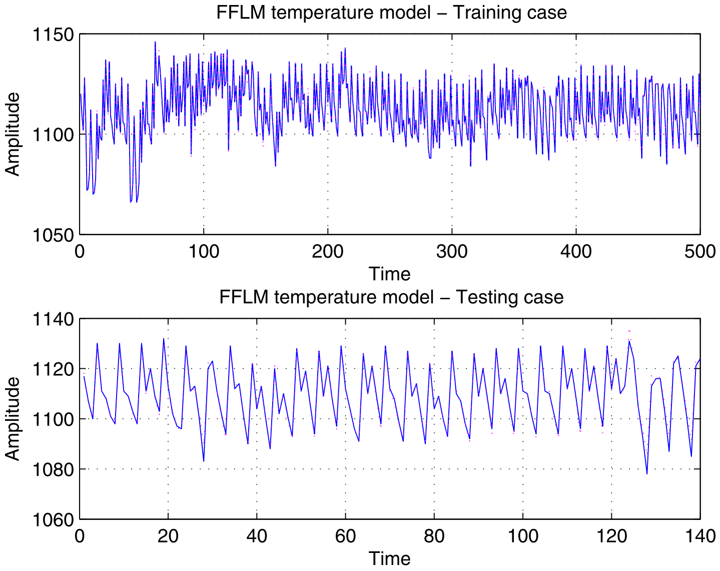


Fig. 13.11 Actual and estimated response of the temperature FF-LM model in both training and testing cases

Table 13.3 VAF values for the developed hot rolling models in %

VAF	Force *F*	Torque *G*	Temperature *T*
ANFIS training	99.22	99.47	99.15
ANFIS testing	84.99	95.76	98.48
FF-LM training	81.32	99.67	99.18
FF-LM testing	81.73	93.04	98.07
FL training	98.59	99.04	95.18
FL testing	81.72	95.20	97.73
Empirical	73.69	74.98	85.08

mance than the empirical model. In addition, the neuro-fuzzy and the fuzzy logic approaches have nearly the same VAF values. This confirms that both neuro-fuzzy and fuzzy logic approaches are adequate with a sufficient accuracy to model complex processes. Overall, VAF results reveal the fact that the proposed modeling methods reflect the nature of the hot rolling plant process quite well.

13.7 Conclusions and Future Work

This chapter explored the use of Takagi-Sugeno technique to develop adaptive neuro-fuzzy models for the hot rolling manufacturing process. Three models (i) roll force, (ii) roll torque, and (iii) slab temperature were implemented. A comparison of results between the neuro-fuzzy models and other developed models are presented successfully. The developed neuro-fuzzy models showed a distinct better performance with promising results. Due to adaptation and predictability, the proposed neuro-fuzzy models can be used to design model-based intelligent regulators especially in strip rolling in which the model parameters are updated online. The presented methodologies can also be applied to other problem domains. They offer an efficient way to analyse data and can support business activity monitoring in the field of performance management

Acknowledgements Authors would like to thank Prof. Dr. Can Özsoy and the Ereḡli Iron and Steel Factory for providing the required technical assistant and the data for this developed research work.

References

1. Kwak, W.J., Kim, Y.H., Park, H.D., Lee, J.H., Hwang, S.M.: Fe-based on-line model for the prediction of roll force and roll power in hot strip rolling. ISIJ Int., **40**(20), 1013–1018 (2000)
2. Öznergiz, E., Gülez, K., Ozsoy, C.: Neural network modeling of a plate hot-rolling process and comparision with the conventional techniques. In: International Conference on Control and Automation, vol. 1, pp. 646–651 (2005)

3. Al-Hiary, H., Braik, M., Sheta, A., Ayesh, A.: Identification of a chemical process reactor using soft computing techniques. In: 2008 IEEE International Conference on Fuzzy Systems (FUZZ 2008), pp. 845–853 (2008)
4. Sheta, A., Al-Hiary, H., Braik, M.: Identification of model predictive controller design of the Tennessee Eastman chemical process reactor using ANN. In: International Conference of Artificial Intelligence (ICAI'09), pp. 25–31 (2009)
5. Sheta, A.: Modeling the Tennessee Eastman chemical reactor using fuzzy logic. In: The ISE Book Series on Fuzzy System Engineering-Theory and Practice. Nova Science, New York (2005). ISBN: 3-540-25322-X
6. Sheta, A., Öznergiz, E., Abdelrahman, M.A., Babuška, R.: Modeling of hot rolling industrial process using fuzzy logic. In: CAINE-2009, San Francisco, CA, USA, 4–6 November 2009 (2009)
7. Kirihata, A., Siciliano, F. Jr., Maccagno, T.M., Jonas, J.J.: Mathematical modelling of rolling of multiply-alloyed mean flow stress during medium carbon steels. ISIJ Int. **38**(2), 187–195 (1998)
8. Cowling, P.: A flexible decision support system for steel hot rolling mill scheduling. Comput. Ind. Eng. **45**(2), 307–321 (2003)
9. Feng, X., Liu, Y., Luo, H., Tang, H., Liu, C.: Numerical simulation during expansion for hot rolling sheet strip. In: Proceedings of the First International Workshop on Knowledge Discovery and Data Mining, WKDD '08, Washington, DC, USA, pp. 310–313. IEEE Comput. Soc., New York (2008)
10. Cser, L., Gulyás, J., Szücs, L., Horváth, A., Árvai, L., Baross, B.: Different kinds of neural networks in control and monitoring of hot rolling mill. In: Proceedings of the 14th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA/AIE '01, pp. 791–796. Springer, London (2001)
11. Lee, D.M., Lee, Y.: Application of neural-network for improving accuracy of roll force model in hot-rolling mill. Control Eng. Pract. **10**(2), 473–478 (2002)
12. Bouhouche, S., Yahi, M., Hocine, B., Bast, J.: Soft sensor—based artificial neural networks and fuzzy logic: application to quality monitoring in hot rolling. In: Proceedings of the 10th WSEAS International Conference on Automatic Control, Modelling & Simulation, ACMOS'08. World Scientific and Engineering Academy and Society, pp. 149–154. Stevens Point, Wisconsin (2008)
13. Tarokh, M., Seredynski, F.: Roll force estimation in plate rolling. J. Iron Steel Inst. **208**, 694 (1970)
14. Özsoy, C., Ruddle, E.D., Crawley, A.F.: Optimal scheduling of a hot rolling process by nonlinear programming. Can. Metall. Q. **3**(31), 217–224 (1992)
15. Mandal, M., Pal, S.K.: Pseudo-bond graph modelling of temperature distribution in a through-process steel rolling. Math. Comput. Simul. **77**(1), 81–95 (2008)
16. Öznergiz, E., Özsoy, C., Delice, I.I., Kural, K.: Comparison of empirical and neural network hot-rolling process model. J. Eng. Manuf. **223**, 305–312 (2009)
17. Jang, J.S.R.: ANFIS, adaptive network based fuzzy inference systems. In: IEEE Transaction on Systems, Man and Cybernetics, vol. 23, pp. 665–684 (1993)
18. Jang, J.S.R., Sun, C.T., Mizutani, E.: Neuro-Fuzzy and Soft Computing. Prentice-Hall, Englewood Cliffs (1997)
19. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modeling and control. IEEE Trans. Syst. Man Cybern. **15**(1), 116–132 (1985)
20. Babuška, R., Roubos, J.A., Verbruggen, H.B.: Identification of MIMO systems by input–output TS fuzzy models. In: Proceedings of Fuzzy—IEEE'98, Anchorage, Alaska (1998)
21. Babuška, R., Verbruggen, H.: Identification of composite linear models via fuzzy clustering. In: Proceedings of European Control Conference, Rome, Italy, pp. 1207–1212 (1995)
22. Babuška, R., Verbruggen, H.B.: Applied fuzzy modelings. In: Proceedings of IFAC Symposium on Artificial Intelligence in Real Time Control, Valencia, Spain, pp. 61–66 (1994)

23. Subudhi, B., Jena, D.: Differential evolution and Levenberg Marquardt trained neural network scheme for nonlinear system identification. Neural Process. Lett. **27**(3), 285–296 (2008)
24. Asadi, S., Hadavandi, E., Mehmanpazir, F., Nakhostin, M.M.: Hybridization of evolutionary Levenberg-Marquardt neural networks and data pre-processing for stock market prediction. Knowl.-Based Syst. **35**, 245–258 (2012)

Chapter 14
How to Measure Efficiency in IT Organizations

Martin Kütz

Abstract IT systems are an essential part of firms and other organisations. If management is the brain then IT is the nervous system. As any other part of the organisation's IT is subject to the economic principle. It has to optimise IT efficiency, which is a major task in IT performance management. But if something has to be improved it has at first to be measured or to be made measurable. This chapter covers the aspect of measuring efficiency in IT organisations and provides an overview of selected methods. After a brief introduction of terms in the field of business administration approaches to measure efficiency are discussed. The focus is set on the aspect of measuring the output of IT by means of utility functions and the Analytical Hierarchy Process (AHP). A new AHP-based method to build scales for measurement is presented. Some fields of application are outlined.

14.1 Introduction

14.1.1 Preliminary Considerations

IT has become an indispensable resource for organisations and firms. Without IT an orderly operation is impossible. IT systems are a major part of the organisation's memory and they reflect all activities of the business. A great part of this detailed recording is claimed by laws and provisions.

However, from a business point of view IT is a resource and as any other resource it is subject to the economic principle. This principle states that either a given output has to be produced with a minimal consumption of resources or a maximum of output is to be generated with a given quantity of resources. Productivity is a specific occurrence of this economic principle.

To ask for IT productivity is easy, to measure IT productivity is a challenging task. The problem of measuring output and input is impeded by the problem of the

M. Kütz (✉)
Fachbereich Informatik und Sprachen, Hochschule Anhalt, Lohmannstr. 23, 06366 Köthen, Germany
e-mail: martin.kuetz@inf.hs-anhalt.de

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_14, © Springer-Verlag London 2013

measurability of IT output. This chapter considers methods to measure productivity respectively efficiency or to make it measurable.

14.1.2 Course of Investigation

Initially productivity and efficiency have to be defined. It will turn out that "traditional" metrics which are based on counting numbers of items of identical type, size and quality are of limited applicability. However, in daily operations it is often easier to measure changes of efficiency. Methods which use this property and thus enable the measurement of efficiency will be demonstrated. Finally, those methods will be used to build measures which are primarily usable within specific organisations but can be generalised to cross organizational measurement. Several examples for the measurement of efficiency will be given and discussed.

14.2 Basics and Definitions

14.2.1 Productivity

Productivity is the ratio of an output quantity and an input quantity, related to a specific object under management [2]. If this object creates n different outputs and consumes m different inputs then $n \cdot m$ productivities can be considered. It is a prerequisite that each output quantity and each input quantity can be measured, and the specific input quantity needed for each output can be identified.

Now let z be the output quantity and r the corresponding input quantity. If p denominates the productivity then $p = z/r$.

Typical examples of IT productivities are process productivities:

Example 14.1 Ratio of number of closed incidents and volume of manpower to process those incidents.

Example 14.2 Ratio of data center space and energy consumption for air conditioning.

The preceding definition of productivity is not as simple as it appears to be. The reason is that to produce a specific output quantity z a corresponding input quantity r is needed. But for many reasons in the field of IT the needed input is provided in bundles or packages. If an input quantity r is needed then an input quantity c (from: capacity) with $c > r$ has to be provided. This leads to an extended definition of productivity $p = (z/r) \cdot (r/c)$. In this formula z/r is the technical productivity related to the consumed input, and this technical productivity is mainly determined by the used technology. The ratio r/c is the utilisation of the provided input quantity resp. input capacity.

If IT management wants to increase productivity then they can try to increase the technical productivity or improve capacity utilization. Finally, they have to minimise idle capacities. Productivity is more or less a synonym for utilised capacity. In terms of cost accounting the machine represents an amount of fixed costs. If the resources are human beings then productivity depends on the learning curve of specific persons.

To this end consider Example 14.1 and assume a volume of 10 closed incidents per day and a consumption of 5 hours of working time of one employee. The technical productivity is 2 closed incidents per hour and with 8 working hours per day the capacity utilization is 62.5 %. The total productivity related to the provided input capacity is 1.25 closed incidents per hour. If the technical productivity in this example is increased to 2.5 then capacity utilisation will be reduced to 50 %. Firstly, management should assign additional tasks to that employee and, thus, reduce the capacity provided for incident processing to 5 hours per day. Afterwards it could try to increase the technical productivity, but subsequently the freed capacity again must be taken away from incident processing. Otherwise the total productivity would not be improved because of a (new) capacity utilisation of just 80 %.

Finally, the reader should note that productivity is considered for time intervals, and the defined ratios consider an average productivity for the considered time interval. To this end consider Example 14.1 again: Management is (normally) not interested in the specific productivity of each processed incident, because those values will vary extremely. But management wants to know and has to know how many incidents have been or can be processed with a given capacity or how much capacity was utilised or is needed to process a given number of incidents.

14.2.2 Efficiency

To allow multiple outputs and inputs the definition of productivity has to be generalised. This is usually done as follows: Let Z be the weighted sum of different output quantities and R be the weighted sum of different input quantities (resources) with non-negative weights. Then $P = Z/R$. Thus, efficiency is considered as a generalised productivity [2]. This definition is applicable to any object under IT management, for instance services, processes, systems or projects, total IT organizations or parts of it, etc. But there is always one essential limitation: Outputs and inputs must be measurable or at least be expressible in numbers.

Weighting and summation aggregate different output quantities respectively input quantities to one “virtual” output quantity and one general resource quantity. Often the weights are considered to be the prices of the resources and the output units. The prices of the consumed resources are easily obtained because resources have to be purchased on specific markets. The prices of the produced outputs are much harder to obtain, because for many IT outputs there is no external market since they are produced only for internal purposes of an organisation.

However, IT organisations must be able to assign a financial value to any output they produce. This value is the price they would take if they would sell this output to an independent third party. Those prices are called transfer prices and the determination of transfer prices is a daily business in all organisations which deliver services internationally to other subsidiaries within a group.

If the weights are financial parameter then the efficiency considered is the economic efficiency. More abstractly the weights are equivalence numbers to compare different objects.

14.3 Established Approaches to Efficiency Measurement

Based on the preceding definitions there are some well known ways to measure IT efficiency:

- Direct efficiency measurement
- Calculation of unit costs
- Calculation of OEE-type indicators

14.3.1 Direct Efficiency Measurement

If outputs and inputs can be measured then appropriate efficiency metrics can easily be established.

In the technical environment of IT, capacity utilisation will be a major subject of interest. Management attention is actually focused on peak utilisations and not on average utilisation. This may change in the future towards monitoring and measuring long-term utilisations when the basic load of infrastructure systems is covered with permanently installed (own) capacities and temporary peak loads are covered with (external) capacities provided by cloud computing technologies.

In the process management of IT with a focus on human or employee productivity it is of interest, how much output is generated in relation to working time, and whether and how these ratios can be improved. In this area productivity will be combined with quality. Only those output quantities are counted which fulfil predefined quality levels. If z denominates the applicable output quantity and q denominates the totally produced output with $z < q$ then $p = (z/q) \cdot (q/r)$ where q/r is the productivity ignoring the output quality, and z/q is the reached quality level. If IT management wants to improve employee productivity it has to work on quality and, if z is given, has to reduce the number of rejections $q - z$.

14.3.2 Calculation of Unit Costs

For different reasons unit costs play an important role in IT performance management. One reason is, that in many organisations the view onto IT is more or less cost driven. Top management states that IT is too expensive and not that IT performance is low. Another reason is the make-or-buy question. There is always to evaluate, whether it is better to produce an IT output within the own IT organisation or purchase it from an external supplier. To make a sound decision unit costs are compared with the prices of the external service providers. Furthermore unit costs are an important vehicle for IT benchmarking and they also must be evaluated if an organisation establishes an internal IT service cross charge system.

From an efficiency point of view the unit costs of a service or process are the reciprocal value of an efficiency ratio. As defined previously, efficiency is the ratio of a weighted output sum and a weighted input sum. The unit costs are the ratio of an input sum, weighted with purchase prices of the used resources and the number of output units, e.g. service volume or number of process realisations.

14.3.3 Calculation of OEE-Type Indicators

In the engineering environment the overall equipment efficiency (OEE) [1] is a wide spread indicator to measure efficiency of machines. In IT management OEE is not yet established. But it is recommended here to use it regularly for efficiency measurements of services and (application) systems due to following reasons: Firstly, OEE is simply a good indicator. This is proven by its wide spread usage in engineering. Secondly and because OEE is known by the production minded top management, its usage would ease and improve the communication of IT management with non-IT management. And thirdly, the OEE-approach seems to be suitable particularly for E-Business systems as it is shown in Example 14.3 below.

OEE is the product of three factors, namely OEE $= av \cdot pl \cdot ql$, where av denominates the availability of the system, pl its performance level and ql its quality level:

- Availability (av): Which percentages of the total service window do those time intervals have where the equipment was up and running? This is the ratio of the time span where the equipment was up and running and the time span where the equipment should have been up and running. Availability can be considered in terms of output quantity, namely the ratio of the planned output quantity z_p and the output capacity z_c of the equipment.
- Performance level (pl): Which percentage of the maximum possible output quantity related to the availability time span has been really produced? This is the ratio of the output quantity z_m, which has been really produced within the availability time span and the output quantity z_p which should have been produced in the availability time span.
- Quality level (ql): Which percentage of the really produced output quantity has been free from faults? This is the ratio of the output quantity z_{fpy} which has been produced without faults ($fpy =$ first pass yield) and the really produced output quantity z_m.

Table 14.1 Calculation of productivity and OEE values (Example 14.3)

z_c/r	r/c	$av = z_p/z_c$	$pl = z_m/z_p$	$ql = z_{fpy}/z_m$
10,000/800	800/1,000	9,500/10,000	8,550/9,500	8,379/8,550
12.5	80 %	95 %	90 %	98 %
10		OEE = 83.79 %		
$p = 8.379$				

Combining productivity and OEE leads to $p = (z/r) \cdot (r/r_{\max})$, then $p = (z_c/r) \cdot (z_p/z_c) \cdot (z_m/z_p) \cdot (z_{fpy}/z_m) \cdot (r/r_{\max})$ and subsequently $p = (z_c/r) \cdot \mathrm{OEE} \cdot (r/r_{\max})$. If $z_c/r = p_{opt}$ then finally $p = p_{opt} \cdot \mathrm{OEE} \cdot (r/r_{\max})$. This consideration is valid and makes sense as resource consumption is of fix cost type. Most IT activities are of that type.

Example 14.3 Consider an online shop system which is able to handle up to 10,000 visits per day ($z_c = 10{,}000$). The input is an abstract system usage unit. A component has been installed with a capacity of 1,000 units ($c = 1{,}000$). The availability (av) is assumed to be 95 %, that means $z_p = 9{,}500$. The number of really incoming visits has been 8,550 and 2 % of those visits broke down due to software bugs. At the background of the used technology the technical productivity is 12.5, which results in a capacity consumption of 800 system usage units. The results of the efficiency calculation are presented in Table 14.1.

If it turns out that the number of really incoming visits is always lower than planned then it might be possible to reduce the installed system capacity to minimise costs and improve efficiency.

14.4 New Approaches to Efficiency Measurement

So far, the efficiency measurement approaches require some important prerequisites. Firstly: All outputs and inputs must be measurable. In the IT environment (and general in the service environment) the measurability of output is often not possible or traditional measurement causes too much effort. Secondly: The weights imply a specific structure of the benefits, namely that each single output has a unique value for the organisation independently from the values of the other output categories. And the weighted sum also assumes implicitly that the increase of a single output quantity has always the same value for the organisation whatever the value was where the increase started from.

In this section two methods are presented which can help to assess efficiency without those prerequisites:

- Concept of utility functions: This is a general approach of decision theory. A heuristic method which is based on that concept is presented. It is TOPSIS

(= Technique for order preference by similarity to ideal solution) [5]. We use a modified version of TOPSIS which standardises input data indifferently from the original method and fits better to the general approach of decision theory.

- Analytical Hierarchy Process (AHP): This heuristic method was introduced by Thomas L. Saaty in 1980 [6]. AHP has found a great variety of applications in different areas, e.g. market research, forestry. In IT it is still widely unknown.

14.4.1 Concept of Utility Functions

Starting point is a group of objects wherein each object is represented by an output vector $\underline{z} = (z_1, z_2, \ldots, z_n)$ and an input vector $\underline{r} = (r_1, r_2, \ldots, r_m)$. Both vectors are concatenated as follows: $\underline{x} = (z_1, z_2, \ldots, z_n, -r_1, -r_2, \ldots, -r_m)$. In this view the value of the object which is represented by the vector $\underline{x}$ increases/decreases, if one or more values of the coordinates of the vector $\underline{x}$ increase/decrease. Given any two objects the decision maker will prefer the object with a higher value assigned.

Given the vectors $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_s$ representing s different objects with their outputs and inputs then two additional objects with the corresponding vectors $\underline{x}_{\max} = (\max z_{i1}, \max z_{i2}, \ldots, \max z_{in}, \max -r_{i1}, \max -r_{i2}, \ldots, \max -r_{im})$ and $\underline{x}_{\min} = (\min z_{i1}, \min z_{i2}, \ldots, \min z_{in}, \min -r_{i1}, \min -r_{i2}, \ldots, \min -r_{im})$ for $1 \leq i \leq s$ are generated. The index i is the number which uniquely identifies each of the s objects and this object is represented by the vector $\underline{x}_i$. The value of any object $\underline{x}_i$ is from the decision maker's point of view higher than the value of $\underline{x}_{\min}$. And accordingly the value of any object $\underline{x}_i$ is lower than the value of $\underline{x}_{\max}$. It makes sense to normalise the vectors in the sense, that $\max z_{ij} = 1$ and $\min -r_{ik} = -1$ for $1 \leq i \leq s$, $1 \leq j \leq n$ and $1 \leq k \leq m$.

If one proceeds according to standard decision theory there is now a lottery for each $\underline{x}$ with two possible results [3, pp. 166–169]: You can either win $\underline{x}_{\max}$ or $\underline{x}_{\min}$. Now the assessor has to make a decision. Will he/she take $\underline{x}$ for sure or a lottery where he/she wins $\underline{x}_{\max}$ with a probability v or $\underline{x}_{\min}$ with a probability $1 - v$. $\underline{x}$ is because of the definition better than $\underline{x}_{\min}$ and worse than $\underline{x}_{\max}$.

Thus, there must be a probability v_0 where from the assessor's point of view $\underline{x}$ and the lottery are equivalent alternatives. He/she is not able to say whether $\underline{x}$ or the lottery is his/her preferred alternative. v_0 is considered to be the benefit or value of $\underline{x}$. It is $v_0 = 1$ for $\underline{x}_{\max}$ and $v_0 = 0$ for $\underline{x}_{\min}$.

The assessment of all objects $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_s$, however, has to fulfil some consistency conditions. If $\underline{x}_g > \underline{x}_h$ in the sense, that for one or more coordinates the inequality $x_{gi} > x_{hi}$ holds and the other coordinates have equal values ($x_{gi} = x_{hi}$) then $v_0(\underline{x}_g) > v_0(\underline{x}_h)$ must hold. The meaning of this inequation is that for the decision maker $\underline{x}_g$ has a higher value than $\underline{x}_h$.

From a theoretical point of view, the presented method is clear. But is it applicable in real-world management due to its weaknesses? Firstly: The determination of the values of the utility function of the decision maker is cumbersome. But this could be made easier by means of a software tool. Secondly: Each decision maker

Table 14.2 Preparation of data for TOPSIS evaluation (Example 14.4)

		Output	Input	Vector
1	Object 1	(20; 10)	(5; 5)	(20; 10; −5; −5)
2	Object 2	(10;15)	(8; 3)	(10; 15; −8; −3)
3	Object 3	(15; 5)	(4; 10)	(15; 5; −4; −10)
4	Object Max	(20; 15)	(4; 3)	(20; 15; −4; −3)
5	Object Min	(10; 5)	(8; 10)	(10; 5; −8; −10)
$\underline{6}$	$\underline{x}_{\max} - \underline{x}_1$	(0; 5; 1; 2)	$\underline{x}_1 - \underline{x}_{\min}$	(10; 5; 3; 5)
$\underline{7}$	$\underline{x}_{\max} - \underline{x}_2$	(10; 0; 4; 0)	$\underline{x}_2 - \underline{x}_{\min}$	(0; 10; 0; 7)
$\underline{8}$	$\underline{x}_{\max} - \underline{x}_3$	(5; 10; 0; 7)	$\underline{x}_3 - \underline{x}_{\min}$	(5; 0; 4; 0)
9	$\|\underline{x}_{\max} - \underline{x}_1\|_e$	$\sqrt{30}$	$\|\underline{x}_1 - \underline{x}_{\min}\|_e$	$\sqrt{159}$
10	$\|\underline{x}_{\max} - \underline{x}_2\|_e$	$\sqrt{116}$	$\|\underline{x}_2 - \underline{x}_{\min}\|_e$	$\sqrt{149}$
11	$\|\underline{x}_{\max} - \underline{x}_3\|_e$	$\sqrt{174}$	$\|\underline{x}_3 - \underline{x}_{\min}\|_e$	$\sqrt{41}$
12	$\|\underline{x}_{\max} - \underline{x}_1\|_{\max}$	5	$\|\underline{x}_1 - \underline{x}_{\min}\|_{\max}$	10
13	$\|\underline{x}_{\max} - \underline{x}_2\|_{\max}$	10	$\|\underline{x}_2 - \underline{x}_{\min}\|_{\max}$	10
14	$\|\underline{x}_{\max} - \underline{x}_3\|_{\max}$	10	$\|\underline{x}_3 - \underline{x}_{\min}\|_{\max}$	5

has a personal utility function and will come to another result than other decision makers. But this is an integral and usually hidden part of decision making. It can be circumvented if a given utility function has to be used by all decision makers in an organisation as it is done e.g. for NPV calculations (NPV = Net Present Value). And thirdly: The same decision maker might come to different results according to the specific order in which he considers the different objects. Though this is a strong counter-argument, it is also an integral feature of decision making and decision makers have to master it daily, e.g. in the recruiting process when new employees have to be selected from a long list of applications.

For day-to-day management an organisation should provide decision making tools which ensure that all members of the organisation ideally come to the same result if they have to make the same decision. But this requires the identification of the "right" utility function before deciding. Thus, decision making is and will remain a complex task.

TOPSIS is based on the concept of a given utility function. Here the utility function is explicitly given. In that $m + n$-dimensional space a metric d is taken and the utility function for the efficiency is defined as follows: $v = d(\underline{x}, \underline{x}_{\min})/(d(\underline{x}, \underline{x}_{\min}) + d(\underline{x}, \underline{x}_{\max}))$. For $\underline{x} = \underline{x}_{\min}$ the result is $v = 0$ and for $\underline{x} = \underline{x}_{\max}$ the result is $v = 1$. This is demonstrated by *Example* 14.4 in Table 14.2.

If now the Euclidian metric (Table 14.2; rows 9–11) is used in TOPSIS then the result is $v(\underline{x}_1) = 0.697, \ldots,\ v(\underline{x}_2) = 0.531\ldots$ and $v(\underline{x}_3) = 0.327\ldots$ If the maximum metric (Table 14.2; rows 12–14) is used then the result is $v(\underline{x}_1) = 2/3$, $v(\underline{x}_2) = 0.5$ and $v(\underline{x}_3) = 1/3$. This shows the influence of the used metric. In the given example the ranking of objects according to their value did not change when the metric was changed but even the ranking of objects can change if another metric is used.

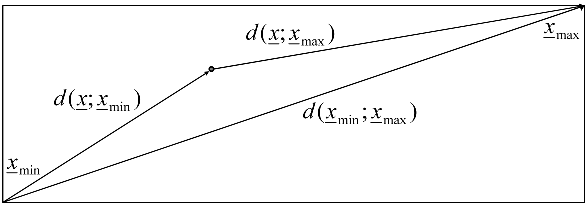

$d(\underline{x};\underline{x}_{\max})$
$\underline{x}_{\max}$
$d(\underline{x};\underline{x}_{\min})$
$d(\underline{x}_{\min};\underline{x}_{\max})$
$\underline{x}_{\min}$

Fig. 14.1 TOPSIS utility function (geometrical presentation)

Geometrically this can be interpreted as follows: $\underline{x}_{\text{min}}$ and $\underline{x}_{\text{max}}$ span a $n+m$-dimensional cuboid. All $\underline{x}$ lie within this cuboid and it is $0 < v < 1$ there. $d(\underline{x}, \underline{x}_{\text{min}}) + d(\underline{x}, \underline{x}_{\text{max}})$ is the length of the traverse line from $\underline{x}_{\text{min}}$ to $\underline{x}_{\text{max}}$ via $\underline{x}$. The value of v describes which part of the whole traverse line is accomplished if one has started in $\underline{x}_{\text{min}}$ and has reached $\underline{x}$ (see Fig. 14.1).

The advantage of this method, compared to the ratio with weighted output sum and weighted input sum is, that no weights have to be defined. However, there is a strong "weighting" through the geometry of the selected metric d.

The first disadvantage of this method is that all outputs and inputs have to be measurable. The second disadvantage is that v depends on the specific $\underline{x}_{\text{min}}$ and $\underline{x}_{\text{max}}$. If $\underline{x}_{\text{min}}$ or $\underline{x}_{\text{max}}$ changes then the values of v change accordingly. This could be healed, if upper and lower boundaries for the coordinates of $\underline{z}$ and $\underline{r}$ would be given. That means that $\underline{x}_{\text{min}}$ and $\underline{x}_{\text{max}}$ are given. For example an organisation could choose $\max -r_{ij} := 0$ and $\min z_{ik} := 0$ for arbitrary i, $1 \leq j \leq n$ and $1 \leq k \leq m$. For the maximal values of the coordinates of $\underline{z}$ and $\underline{r}$ it would choose values which will not be exceeded under normal circumstances. This is reasonable for the input because resource capacities will be strongly limited in every organisation. If an output limit is exceeded then satisficing [4], a replacement of the output value by the given boundary value, would deteriorate the calculated efficiency.

An alternative of that method could be: Apply TOPSIS to output vectors and input vectors separately. Then build $p = v(\text{output})/v(\text{input})$. However, advantages and disadvantages of this alternative are similar to the original alternative.

14.4.2 Analytical Hierarchy Process (AHP)

This method starts from following observation. Let a list of positive numbers $w_1, w_2, \ldots, w_n$ be given. Build a matrix $A = (a_{ij})$ with $a_{ij} = w_i/w_j$. This matrix has some interesting properties: $a_{ij} = 1/a_{ji}, a_{ii} = 1, a_{ik} = a_{ij} \cdot a_{jk}$. The largest eigen-value of this matrix is n and the eigen-vector corresponding to the largest eigen-value is $(w_1, w_2, \ldots, w_n)$.

If the measures $w_1, w_2, \ldots, w_n$, which could e.g. represent sizes or complexities of software modules, are not given, but the relations $a_{ij} = w_i/w_j$ can be measured or estimated, is it then possible to deduce the values w_i from the matrix A? Saaty has considered the series of matrices A, A^2, $A^3, \ldots$ and he showed that the eigen-vector of A^k can be approximated by positive numbers $w_i(k)$ and the matrix elements of A^k

Table 14.3 Matrix A to start the AHP calculus (Example 14.5)

1	3	1	5	0,33333	1	0,33333
0,33333	1	1	5	0,2	0,33333	0,14286
1	1	1	1	0,14286	0,2	0,11111
0,2	0,2	1	1	0,11111	0,33333	0,14286
3	5	7	9	1	1	1
1	3	5	3	1	1	0,11111
3	7	9	7	1	9	1

Table 14.4 Matrix A^4 in the AHP calculus and approximated values of the weights

1	1,89758	2,42764	3,68284	0,39363	0,9195	0,27879	0,10278069
0,53448	1,02172	1,29926	1,98057	0,212	0,49062	0,14828	0,05514226
0,41561	0,78865	1,02431	1,53806	0,16544	0,38694	0,11692	0,04301221
0,28611	0,54524	0,69646	1,07781	0,11306	0,26294	0,08081	0,02969422
2,38633	4,54104	5,84721	8,87484	0,94838	2,20157	0,66904	0,24694971
1,21615	2,33198	3,00714	4,56696	0,48684	1,1561	0,34526	0,12712279
3,81686	7,25003	9,24546	14,3225	1,4924	3,55477	1,08591	0,39529812

behave like $w_i(k)/w_j(k)$ [6]. Values of $w_i(k)$ therefore can be calculated arbitrarily exact from the matrices of the sequence. The sum of the $w_i(k)$'s is always equal to 1.

An impression of the AHP calculus is given by *Example* 14.5 presented in Tables 14.3 and 14.4. Table 14.3 shows the reciprocal matrix A of the original assessments of the pairwise relationships between 7 objects with the scale 1, 3, 5, 7, 9 respectively the reciprocal values. Table 14.4 shows the matrix A^4 after some normalisation (first seven columns starting from the left side) and in the 8th column the approximate values of the weights $w_i(4)$ are given.

The advantage of this approach is that no usual measurement is needed but only a judgement about the pairwise relationships between the considered objects. Such assessments can be executed by experts. Also objects can be assessed which are complex, amorphous or diffuse.

The disadvantages of AHP are even stronger compared to the second disadvantage of TOPSIS. If any of the considered objects is changed then its relationships to all other objects have to be re-assessed. If an object is added then its relationships to all other objects have to be assessed. If an object is added, changed or removed the AHP calculus has to be repeated and all values of w_i will change.

14.4.3 Discussion

All methods considered have some serious disadvantages.

- *Application of the concept of utility functions*: All coordinates of the considered objects must be measurable. It is not easy to ensure consistency of the judgement about the utility function v. Different assessors will come to varying results according to their personal utility function. The judgement of one single assessor will (slightly) change in time. The execution of an assessment is elaborate. Theoretically every decision maker must be able to work out his utility function (values) for the considered set of objects. But it is questionable whether he/she is able to do so, if he/she is under pressure and has to decide quickly.
- *Application of TOPSIS*: This method avoids the cumbersome assessment because the utility function is given by the method. Compared to the weighted sum approach TOPSIS eases the job of the decision maker because it is not necessary to determine $m + n$ weights. As long as $\underline{x}_{\min}$ and $\underline{x}_{\max}$ are unchanged then (with a given metric d) each object will get a specific value of v. But the geometry of the decision space is hidden in the metric d and thus it is less transparent than the weighted sum approach.
- *Application of AHP*: Different assessors will come to varying results according to their personal utility function. The judgement of one single assessor will (slightly) change in time. Primarily the assessment of the pairwise relations between the given objects is elaborate because with n objects this has to be done $n \cdot (n - 1)/2$ times. To ensure consistency among all single assessment results is almost impossible (however, the AHP calculus works even with inconsistent input). AHP is of interest for the support of decision making because some powerful software packages based on the method are available on the market.

Can these methods be applied in IT and do they have any benefits for assessing efficiency in IT organizations? This will be discussed in the following section.

14.4.4 Construction of Metrics with AHP

One important field of application of AHP is the determination of equivalence numbers. If weighted sums have to be built then these weights can be determined by means of this approach, because for the responsible persons it is usually easier to compare objects than to directly assign values to single objects. Also, when objects or outputs are not (easily) measurable, a comparison can be made by experts, and the relation between any two objects can be expressed as a numeric value. But there must always be a set of objects and the total value of all considered objects always is 1. Consider *Example* 14.5 given in Tables 14.3 and 14.4. If the calculated weights would be the weights in a benefit value analysis the mentioned standardization of the weights makes sense. That AHP considers a set of objects seems to make it inappropriate for measuring a single output or input.

But what is a measure? It is nothing else than the numerically expressed relation of a considered object to a defined reference object, the scale. Having this in mind a metric can be defined by means of AHP as follows: Carefully select n objects and assess the relationships between all pairs in this set of objects. Build the matrix

$A = (a_{ij})$. Those n objects are denominated as reference objects. Now, select a test object which has to be measured and is (of course) of the same category as the reference objects. Add this object to the matrix and assess its relationships to the n reference objects.

Running the AHP algorithm leads to the values $w_1, w_2, \ldots, w_n$ of the reference objects and w_{n+1} of the test object. The sum of the $n + 1$ values is 1. Because the reference objects are not changed, the total value of them is always the same. This leads to a measurement for the test object: Its value for the decision maker is expressed by $v = w_{n+1}/(1 - w_{n+1})$.

With this method weak, amorphous or diffuse objects can be measured. Prerequisite is, that there are n such objects, whose pairwise relationships can be assessed solidly. Primarily this is a method for internal measurement, but if two or more partners agree on the reference objects this can also be used as an inter-organisational measurement. It is even conceivable that some idealised objects are used as the reference objects which are fully independent from specific organisations. This would make that metric appropriate for benchmarking purposes.

14.5 Areas of Efficiency Measurement

The measurement of efficiency in the IT area is mainly a problem of measuring the output. In many cases the output is a complex mix and not fully quantifiable. It is not only the question of quantity. It is also the question of creativity and quality.

One of the most difficult but fascinating examples is the development of software. Assumed that the number of function points, use case points etc. can be measured, there are of course differences between the software volumes and the software quality developed by different programmers. Experts know that and they are able to compare and value the different outputs.

Sometimes a developer may come to the conclusion not to produce a new piece of software but to solve a problem by modifying a business process appropriately. In this case, the output has, of course, value for the business, but the value of the actual software portfolio, however it has been defined, did not change. This shows that the output of a software management team sometimes is not software (but process engineering) and thus pure software measures are not enough to measure the output and efficiency of (internal) software organisations.

Sometimes it is laborious to measure a complex output. Then the AHP-based metric could be an alternative which implies lower effort than other methods. The matrix calculus which is an essential element of the AHP method can be done by appropriate software. Thus, the main effort in using the AHP-based metric lies in the assessments of the relations between the reference objects and the test object.

14.5.1 IT Services

The usual approach to measure the efficiency of a single service is the determination of the unit costs. If unit costs increase/decrease then efficiency decreases/increases. Prerequisite is that the output remains unchanged over time. It is an (often forgotten) prerequisite that costs have to be assignable to the specific service. But most of the costs will be indirect costs and have to be broken down to the services according to some specific rules. The distribution key to allocate indirect costs as a percentage of total indirect costs to the different services can be evaluated with the original AHP method. The calculated eigen-vector is just the list of percentages asked for.

If a service is changed then the new "size" of the modified service has to be measured. This can be done with the AHP-based metric which is shown by the following example: Consider the great variety of desktop software packages. If an organisation has to conduct an internal cross charging or just wants to add some kind of financial value to each software package then it could take 5 to 6 carefully selected reference packages and take them as the basis for the AHP-based metric. Then for each software package the AHP-based metric can be applied. If all desktop software packages have been assessed in this way the organisation can take the resulting list of equivalence numbers as a basis for pricing. (If a reference package has the price x and the weight w_x and another software package has got the weight w_y then a proposal for the price y would be $y = (w_y/w_x) \cdot x$.)

Now, the reader might ask why the organisation should not consider all software packages at the same time (in several organisations this could be a set of several hundreds of desktop software packages) with the original AHP method but consider each packages separately with the AHP-based metric. A first reason is to avoid a matrix with hundreds of columns and rows. A second and crucial reason is that the catalogue of desktop software packages changes over time and the AHP-based metric allows to conduct the assessment for each single software package and to conduct this assessment step-by-step.

14.5.2 IT Service Catalogue

If the total costs of the service production are available then the original efficiency metric as the ratio of output value and input value can be applied. As a prerequisite all service quantities must be available because the total service value is the sum of the service quantities weighted with the equivalence numbers, prices or cross charge rates.

If different service catalogues are to be compared, e.g. in an IT benchmarking situation, all participants have to use the same output prices for those services which are produced by more than one participant. Otherwise a service unit would have different values according to the respective producer.

Another question is whether the input weights have to be the same for all participants. The answer depends on the particular objectives. If a more technical efficiency is considered then the weights should be identical. If the considered input is cost then the question is how much output value is generated with one cost unit, namely currency unit of costs. If this is considered then the input weights are purchase prices and prices can be different between the participants. This coincides with the consideration that if two participants produce the same output quantities and consume the same input quantities then that organisation with lower input prices is, of course, more efficient from an economics point of view.

14.5.3 IT Processes

Many processes are clearly defined and delimited. The number of process executions can be measured. If a process generates just one output then the determination of the output value is very easy. It will be the number of the process executions or a multiple. If costs are assignable (see the problem of indirect costs) then again the traditional ratio approach can be applied.

However, many processes are not precisely delimitable. Good examples are the IT management processes as they are given in the COBIT framework. Consider as an example the COBIT process "PO6 Communicate management aims and direction" or "DS5 identify and allocate costs" (both from the COBIT 4.1 framework). What is the measurable output? What is the number of process executions?

A similar example on the operational level is the ITIL (V3) service asset and configuration management process. This process has some sub-processes which are very suitable for measuring the output, e.g. changing a configuration item. The output of the total process could be the weighted sum of all outputs of the various sub-processes. The number of added, changed or deleted configuration items would be an important part of the output.

However, if nothing is changed then the performance of this process is, of course, not zero and thus the number of assets and configuration items under management will be a partial measure for the process output. But now another question emerges: Does the process output increase if the number of versions of configuration items or number of configurations increases inexorably? Isn't it an expected output from this process to keep the number of managed objects manageable? The output question of "big" processes is hard to answer.

Considering pure management processes, the basis for measuring output could be the number of decisions which have been made. But this would imply that all (relevant) decisions are documented. However, the size and complexity of these processes could be assessed with the AHP metric.

Measuring IT processes leads to another challenge. Even if the process executions can be measured very easily (e.g. incident management) there is another problem to be mastered, namely the differing "size" of the process executions and process results. Each incident differs from any other incident. Thus the incidents processed have to be weighted. Practically, the organisation will introduce a categorisation (e.g. very small, small, medium, big, very big) and the service desk agents will have to categorise those incidents. The output of the incident management then is the weighted sum of all processed incidents. Those weights can be evaluated with the help of the AHP method.

This approach can be similarly applied to other processes like request fulfilment, event management, problem management, change management, etc.

14.5.4 IT Projects

IT organisations manage a lot of projects. Thus they are interested in measuring the efficiency of their projects and improving it continuously. But each project is somehow unique. Projects differ in size and scope. The efficiency of a project must be assessed on the basis of the project result. This is the output. The project result is a complex set of different objects, hardware components, software components, documentation, training, etc.

The project output is an interesting application area for the AHP measure. Each project will be assessed against a set of reference projects. Those reference projects have to be carefully selected and the values a_{ij} in the AHP matrix have to be carefully evaluated. The different tasks of a project could be evaluated similarly.

14.5.5 IT Teams

Similarly to projects the output of single persons (e.g. developers), teams (e.g. consulting teams) or total IT organisations can be assessed with the help of the AHP metric. But the assessment of single persons might be legally restricted, e.g. in Germany the person itself must accept the assessment and also the worker's council has to agree.

14.5.6 IT Applications

Of course, the size of an application can be measured by the function point method or its derivatives. But this is very costly and time consuming. Here again the AHP metric offers a method to get helpful information with limited efforts.

A set of applications has to be selected as reference applications. It does not make sense to include huge applications in that process, e.g. ERP systems. Normally, an organisation will have only one or a few of those big packages. They should be addressed by an own category for the assessment.

For those unique applications the efficiency question is more or less concentrated on the question, whether the efficiency can be continually improved. This requires on the one hand a follow-up on the operating costs and on the other hand an assessment of changing functionality and complexity.

Specific classes of applications are information systems or data warehouses. The output of those systems is complex and depends on various parameters like size (number of data records or stored objects), structure and complexity, user interfaces and tools (usability), support by competent experts, etc. These output dimensions can be complemented by aspects like user or customer satisfaction. All these output dimensions allow to measure the provision efficiency as the ratio of output value and corresponding input value.

The provision efficiency does not allow any judgement on the value of such a system for the business. An information system may contain many data and have an excellent usability, but at the same time its value for the business may be low. But what is the value for the business and how can it be measured or assessed? Of course, there is a high business value if an information taken from the system helps to make a "right" decision whereas the rightness may be expressible in higher revenues, profit, customer satisfaction, etc.

The problem with this view is that it can (theoretically) be measured a posteriori and not a priori. In terms of this aspect, the value of an information can't be determined before the consequences of a decision occurred. But if the decision turns out to be wrong (however this is meant), does this change the value of that information? It has also to be asked, whether such considerations have to do anything with IT efficiency. If the answer would be "yes" then IT would be responsible for the information contained in the system. Is this the IT department's job?

In this context, three questions have to be stated and should be answered with corresponding measurements:

- What is the total capability of the system?
- To what degree was this capability used?
- To what degree was the used information helpful to the decision makers?

The value of the system's total capability can be interpreted as the information capacity of the system and can be measured as it was previously discussed. The degree of usage can be measured by the number of visits, queries, downloaded documents, etc. Finally, the value for the business can be addressed by means of specific categories of user or customer satisfaction.

Corresponding to these measures, different types of efficiency have to be considered. Provision efficiency is of central importance. At second, there must be efficiency with respect to the usage of the system. The approach would be similar to the resource consumption discussed in the beginning of this chapter. The utilization of the stored information is measured. At third, the benefit of the used information must be assessed by the users of the system. They must evaluate how helpful the information was in the specific decision making situation. This is different from the subsequent evaluation whether the used information has led to the "right" decision or not. The subject is complex and should be investigated separately. Here only some introductory thoughts could be exposed.

Nevertheless, the AHP method and the AHP-based metric are very relevant to measure complex objects.

14.5.7 Application in Real-World IT Management

IT performance management is on the agenda of many organisations. Productivity as the ratio of an output quantity and an input quantity is considered mainly in IT service operations and mostly applied to employee productivity. The usage of general efficiency ratios is not common with one exception, namely unit costs. Unit cost calculation is a standard tool in IT cost accounting, performance management and benchmarking.

More sophisticated methods like TOPSIS and AHP are sporadically known by IT managers and IT performance experts. Both methods are not yet contained in standard tool boxes of IT performance management.

The author has used TOPSIS in an IT benchmarking for more than six years. TOPSIS provided a top benchmark for the participating IT organisations. The major challenge was that all participants had to accept the "theoretical" approach.

The AHP-based metric is a quite new development and has already been communicated to some sophisticated IT performance managers. This has led to interesting and inspiring discussions. But up to now, in IT organisations no operational use of that method is known by the author.

14.6 Summary

Measurement of efficiency can be reduced to the measurement of outputs and inputs. In IT performance management the input measurement is well established. One (traditional) problem is the indirect cost nature of most resource consumptions. But the real problem is output measurement. If repeatability can be assumed, output measurement is relatively easy. However, because IT is a service provider, the same type of output may differ extremely in size and complexity.

But many outputs of IT organisations are of occasional nature and include a big portion of creativity or quality. Those outputs can be assessed by IT experts and, thus, can be made measurable. One approach which was shown in this chapter is based on the AHP method. It uses a set of reference objects and the test objects are measured against that reference set which delivers the scale implicitly.

It would be not an easy task, but it seems possible to define reference objects e.g. for projects or applications which could be used by any IT organisation and, thus, lead to better benchmarking opportunities compared to the applied instruments today.

The tool is available. Now, it has to be applied. It may be not the ultimate solution but it could be a further step forward to better and more effective output measurement and subsequent efficiency calculation. Let's give it a try.

References

1. Koch, A.: OEE für das Produktionsteam. Das vollständige Benutzerhandbuch. CETPM-Publishing, Ansbach (2011)

2. Krause, H.-U., Arora, D.: Controlling-Kennzahlen. Oldenbourg Verlag, München (2008)
3. Laux, H.: Entscheidungstheorie. Springer, Berlin (2007) (7th, revised and extended edition)
4. Peters, M.L., Zelewski, S.: Efficiency analysis under consideration of satisficing levels for output quantities. In: Proceedings of the 17th Annual Conference of the Production and Operations Management Society (POMS), 28.04.–01.05.2006, Boston (Mass.), pp. 1–18 (2006)
5. Peters, M.L., Zelewski, S.: TOPSIS als Technik der Effizienzanalyse. WiSt. Wirtschaftswiss. Stud. **1**, 9–15 (2009)
6. Saaty, T.L.: Fundamentals of Decision Making and Priority Theory with the Analytical Hierarchy Process. RWS Publications, Pittsburgh (2006) (2nd edn. 2nd printing)

Part V
Technologies

Chapter 15
Business Activity Monitoring (BAM)

Werner Schmidt

Abstract Providing reliable and timely management information is crucial for supporting the agility of organizations. Business Activity Monitoring (BAM) describes a concept and technology that complements periodic ex-post analysis of process execution by permanently identifying particular situations at runtime and reacting to them by setting alerts or triggering actions with no or low latency. Complex Event Processing (CEP) has emerged as a basic technology for an effective BAM environment. In an integrated BAM/CEP architecture enterprise applications and workflow engines can serve both as event producers and consumers, while event processors filter and transform events, find patterns among them and derive new events. The ladder are consumed e.g. by workflow or enterprise resource planning systems causing new processes or dashboard solutions displaying management information as it arises.

15.1 Introduction

15.1.1 Performance Management and Process Monitoring

Due to global dynamic markets with high competition in many industries enterprises need to proactively act within or at least quickly adapt to rapidly changing environments in order to achieve and sustain competitive advantages while balancing opportunities and risks.

This agility many enterprises strive for depends on their capability of timely making changes both on strategy level and on operational level where business processes help implementing the strategy and their single instances form the daily business.

This capability requires a sound management of business performance, as proposed in the closed-loop approach of performance management (see Fig. 15.1).

The upper part depicts the strategy level on which management instruments like Balanced Scorecard (BSC) and IT systems for business planning are used. Design-

W. Schmidt (✉)
University of Applied Sciences Ingolstadt, Esplanade 10, 85049 Ingolstadt, Germany
e-mail: werner.schmidt@haw-ingolstadt.de

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_15, © Springer-Verlag London 2013

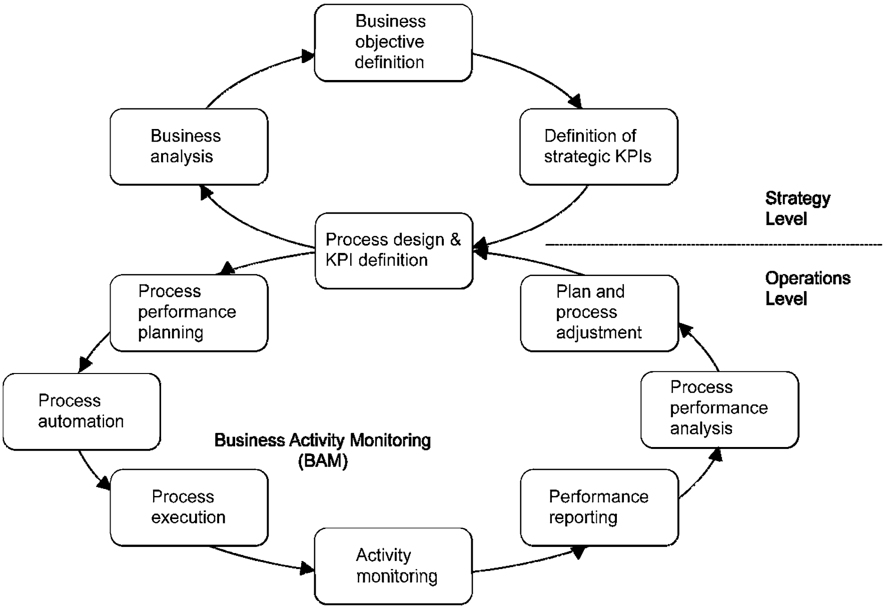

Business objective definition
Definition of strategic KPIs
Business analysis
Strategy Level
Process design & KPI definition
Operations Level
Process performance planning
Plan and process adjustment
Process automation
Process performance analysis
Business Activity Monitoring (BAM)
Process execution
Performance reporting
Activity monitoring

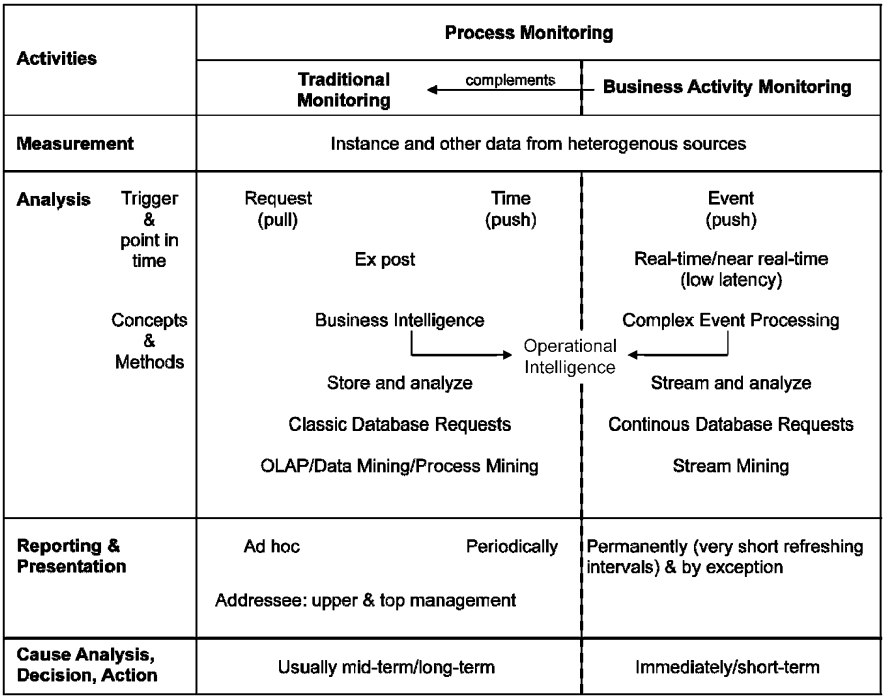
Fig. 15.1 Closed-loop approach to Performance Management (reprinted from [2] with kind permission of Springer Science+Business Media)

ing processes as means to implement strategies and defining key performance indicators (KPIs) for them links to the operational level of performance management. As illustrated in the lower part of the figure this level is represented by the Business Process Management (BPM) lifecycle which includes analyzing, designing, implementing, executing, monitoring and optimizing of business processes [19].

Within the BPM lifecycle Process Monitoring forms an essential part. It assesses the performance of process instances being executed in the key dimensions quality, time and cost and helps identifying weaknesses and opportunities for improvement [10]. IT systems hereby can support by measuring and recording metrics and computing values for KPIs, by comparing them with preset target values and by reporting and presenting results, to defined addressees. Monitoring thus provides basic data for analysis by human decision makers as well as for BI solutions and can also trigger subsequent actions by themselves.

15.1.2 Process Monitoring and Business Activity Monitoring

Traditional process monitoring is time-driven or request-driven and therefore delivers results with a more or less significant time-lag. This is of disadvantage, because process execution often comes along with unexpected events which need to be dealt with in a timely manner (e.g. shortage of work force on short notice due to illness).



Activities		Process Monitoring	
		Traditional Monitoring ← complements	Business Activity Monitoring
Measurement		Instance and other data from heterogenous sources	
Analysis	Trigger & point in time	Request (pull) Time (push) Ex post	Event (push) Real-time/near real-time (low latency)
	Concepts & Methods	Business Intelligence Store and analyze Classic Database Requests OLAP/Data Mining/Process Mining	Complex Event Processing Stream and analyze Continous Database Requests Stream Mining
		→ Operational Intelligence ←	
Reporting & Presentation		Ad hoc Periodically Addressee: upper & top management	Permanently (very short refreshing intervals) & by exception
Cause Analysis, Decision, Action		Usually mid-term/long-term	Immediately/short-term

Fig. 15.2 Characteristics of business activity monitoring in the context of process monitoring

This is where Business Activity Monitoring (BAM) as the event-driven complement of traditional monitoring comes into play (see Fig. 15.2).

The term is not precisely defined, but there is a common understanding among researchers, practitioners, software vendors etc. Gassman for instance summarizes BAM as ‘processes and technologies that provide real-time situation awareness, as well as access to and analysis of critical business performance indicators, based on event-driven sources of data’ [7]. This statement paraphrases many attributes often used to characterize BAM and its objectives (see also [5, 12, 17, 24]). Hence Business Activity Monitoring . . .

- means continuous and simultaneous real-time monitoring of IT systems and services which process business process instances and other information from inside or outside the enterprise (e.g. ERP, CRM, process engines, stock price tracking etc.). It aggregates data incurred by performing business activities in applications, analyses it using predefined rules like thresholds and displays results in real-time on dashboards.
- fosters awareness and provides real-time visibility of business processes with key performance indicators (KPIs) and other information like current status, elapsed time, projected completion time of instances etc.
- is based on Complex Event Processing (CEP) (see Sect. 15.2.1) and thus reduces the so-called IT blindness. This metaphor describes the phenomenon, that the

IT environment in the enterprise permanently creates a high number of single events without any semantics, which makes it impossible to understand their impact (as patterns) on processes, policies and business goals [17, 24]. Aggregating low-level events to complex ones by applying CEP methods improves business process insights.
- helps recognizing significant business events like bottlenecks or missed targets (e.g. time delays) as they occur and timely sets alerts or initiates specified escalation procedures (e.g. instantiating other business processes like problem workflows).
- allows for better understanding the consequences of events and acting adequately by putting them into their current, predictive and historical context (what is happening now, what happened in similar situations in the past and what is likely to happen in future?)

Regarding the action part adding to the sole observation of process execution the meaning of BAM can be extended from Business Activity Monitoring to Business Activity Management like proposed by [11].

This contribution emphasizes BAM with process instance related metrics and KPIs (throughput time etc.) rather than traditional monitoring with KPIs like revenue or margin (overall, by product group, region, customer etc.). In Sect. 15.2 we present an integrated BAM/CEP architecture, starting with an overall picture, followed by explanations of the building blocks using examples as well as by some requirements for event processing solutions. We conclude in Sect. 15.3 with the relationship between BAM/CEP and Business Intelligence, talk about typical application domains and give some perspectives of the importance and challenges of BAM/CEP.

15.2 Integrated BAM/CEP Architecture

15.2.1 Events and Complex Event Processing as a Basis

An event is 'anything that happens, or is contemplated as happening' [14]. A so-called simple, single or low-level event is an observed fact like a purchase with credit card at a certain location and time. Events can be dependent from each other (causality). Many low-level events are termed as an event cloud, or, if ordered chronologically as an event stream. An event cloud can consist of many streams from different sources.

Correlating and combining such low-level events allows for recognizing patterns and dependencies and can create a so-called high-level or complex event, which summarizes, represents or denotes a set of other events (aggregation) [14, 15]. An example could be an 'assumed credit card fraud event' derived from credit card purchases at different locations far apart each other (e.g. Munich and Los Angeles) within a short time span (e.g. 3 hrs).

In business and many other settings it is important to early conclude the probability of the occurrence of a high-level event after the occurrence of single events in order to proactively act to avoid or limit negative consequences [5, 21].

Complex Event Processing offers concepts and solutions to cope with situations like this. It provides computational methods, techniques, and tools to control event-driven systems. It enables to identify and process relevant events emanating from multiple, heterogeneous sources as they occur, i.e. with low latency [1, 3–5, 21, 24]. Processing events includes reading, creating, transforming, deleting and distributing them in a distributed computing environment and exploring relationships (semantic, causal, temporal) between them in a timely manner [4, 21].

15.2.2 Architecture Overview

A CEP environment supporting BAM usually consists of event producers (event sources), event consumers (event sinks), an intermediary event processing logic (Event Processor) and adapters for translating event data between different formats (see Fig. 15.3) [4, 11]. The event processing logic can be structured in several software modules, called event processing agents (EPAs). Together with the event producers and consumers they form an event processing network (EPN), in which events are exchanged by communication mechanisms [4]. At runtime events are fed into the system and analyzed and processed based on definitions of events, metrics, KPIs, event patterns, rules and conditions specified at design time. Processing the events might also require access to data which is external to the application. These are represented by global state elements.

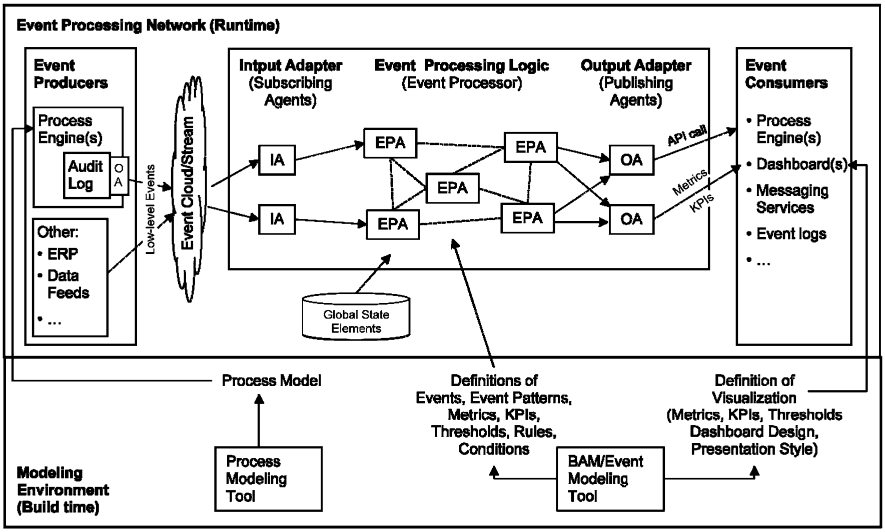

Event Processing Network (Runtime)
Event Producers
Process Engine(s)
Audit Log
OA
Other:
• ERP
• Data Feeds
• ...
Low-level Events
Event Cloud/Stream
Intput Adapter (Subscribing Agents)
Event Processing Logic (Event Processor)
Output Adapter (Publishing Agents)
IA
IA
EPA
EPA
EPA
EPA
EPA
OA
OA
API call
Metrics
KPIs
Event Consumers
• Process Engine(s)
• Dashboard(s)
• Messaging Services
• Event logs
• ...
Global State Elements
Process Model
Definitions of Events, Event Patterns, Metrics, KPIs, Thresholds, Rules, Conditions
Definition of Visualization (Metrics, KPIs, Thresholds Dashboard Design, Presentation Style)
Modeling Environment (Build time)
Process Modeling Tool
BAM/Event Modeling Tool

Fig. 15.3 Integrated BAM/CEP architecture

Event processing platforms supporting this architecture consist of a language to express and tools to design, test and execute the event processing logic. They also need mechanisms for distributing events.

An event processing language (EPL) is used for expressing the event processing logic. There is a good number of commercial or open source EPLs available. They can be grouped in rule-oriented languages (e.g. DROOLS), stream-oriented languages (e.g. extensions of SQL like Esper EQL or CQL) and imperative programming languages (e.g. Prova) (for more details see [4]).

15.2.3 Event Producers (Event Sources)

Sources for single events as defined in Sect. 15.2.1 can be manifold. Events are created by hardware like physical sensors, probes, RFID transmitters or by software applications like Enterprise Resource Planning (ERP) systems or Process Engines or data feeds (e.g. RSS feeds from the Internet). Combinations of both also including human interaction can be producers as well (e.g. an application creates a low stock alert or a user places an order in a web shop) [4].

As shown in Sect. 15.1.2 BAM looks at running business processes. Executing process instances causes state transitions and thus all sorts of low-level business events. These are usually recorded by process engines as a part of Business Process Management Systems (BPMS) and transactional systems for Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) or Supply Chain Management (SCM) etc.

Examples for producing a single event by an ERP system are writing a record for the sale of a product to a customer, for the purchase of material from a supplier or for a payment received from a customer.

The process engine as the major event producer in a BAM environment controls the flow of execution and invokes resources according to the process model designed at build time [6] (see Fig. 15.3). It logs its activities and the state changes as single events in the audit trail. This means it writes timestamps for the creation and the termination of a process instance or single execution steps, for calling and releasing of applications and services or for invoking of human interaction etc.

All sorts of event producers introduce raw events into the event processing system [11] suggest to have an output adapter in place on the producer side which filters the events to be published and processed in order to reduce the number of events and to prevent from sending sensible data into the event cloud or stream.

15.2.4 Event Processing Agents as Event Processing Logic

Before processing can start input adapters need to transform the events fed into the system into the internal format of the event processing application.

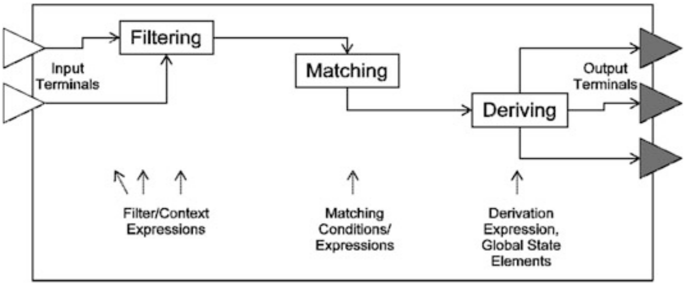

Filtering
Matching
Deriving
Input Terminals
Output Terminals
Filter/Context Expressions
Matching Conditions/ Expressions
Derivation Expression, Global State Elements

Fig. 15.4 Logical EPA functions

Raw events from a process engine like occurrence times of state transitions might not be of importance per se, but can be used to derive meaningful metrics, combine those to KPIs and evaluate them by the event processing logic.

15.2.4.1 Logical Functions of Event Processing Agents

The event processing logic, also called event processor (EP) or CEP processor/engine is structured by event processing agents (EPAs). The generic logical functions of EPAs are filtering, deriving/transforming events as they occur and detecting patterns in their occurrence (matching), based on predefined rules [4] (see Fig. 15.4). Filtering means selecting the events to be processed. The matching function finds event patterns and creates sets of events satisfying the same pattern. These are input for the transforming function which derives new events and sets their content. Many concrete EPAs just filter and derive.

As results of their processing the agents compose single or complex events and feed them back into the CEP system, trigger applications (e.g. process engine) or human interaction, invoke IT services, set alerts etc.

Each EPA has input and output terminals as interfaces to receive events from and emit events to other entities.

15.2.4.2 Types of Event Processing Agents and How They Work

With regard to the logical functions a number of EPA types can be specified. Figure 15.5 lists and describes them and indicates the functions they perform (O means optional) [4].

In the following we explain the three basic EPA types using examples that include metrics and KPIs, indicate the way they can be computed and what the result can be, although the consequences are subject of event consuming (see Sect. 15.2.5).

Filter agents work in several logical steps:

(1) The agent applies a filter expression on attribute values of the incoming events at the input terminal. If the comparison result is ‘false’ the events are not even accepted for further processing. This filtering option can be used by each EPA.

Agent type		Description	Functions		
			Filtering	Matching	Deriving
Event Filtering Agent		Performs filtering only	X	-	-
Event Transformation Agents	Translate	Generates a single derived event from a single input event	O	-	X
	Enrich EPA	Uses a single input event to query a data from a global state element and creates a derived event with the (modified) attributes of the original event and new attributes	O	-	X
	Project EPA	Takes an input event and creates a single derived event with a subset of the attributes of the original event	O	-	X
	Aggregate EPA	Creates a single derived event from a collection of events by applying a function on them	O	-	X
	Split EPA	Creates a collection of events from a single input. Each new event can be a copy or a projection of the original event	O	-	X
	Compose EPA	Takes groups of events from two input terminals, applies a matching criterion and creates derived events for the events satisfying the criterion.	O	-	X
Event Pattern Detect Agent		Performs a pattern matching function on one or more input streams and emits one or more derived events if the defined pattern occurs in the input stream	X	X	O

Fig. 15.5 Types of event processing agents

(2) If necessary a context expression is used to group event instances by specifying conditions which allow temporal, spatial, state-oriented or segmentation-oriented partitioning. For each partition separate agent instance can be instantiated then in order to continue processing in parallel.

(3) Principal filtering applies filter criteria on the event instances having passed the input terminal and the optional partitioning and filters out those not meeting the constraints.

An example illustrates the described options: An ERP system as an event producer sends all sorts of record creation messages. The input terminal of an EPA filters for purchase requisitions only. The agent partitions them by using an ABC classification of the material requested, where A indicates a high importance for the company's operation while B stands for middle and C for low importance. For class A inputs the EPA filters all purchase requisitions with a desired delivery date less than three days from now. They are sent to a process portal as event consumer which alerts the responsible procurement manager so that he can talk in person to the supplier in order to have the order being handled with priority.

Transformation agents apply derivation methods (functions) on events depending on their specializations as they are explained in Fig. 15.5. Optional filtering as described above defines the set of events to be transformed. In our example a translating agent would enrich the event data by adding information about the sales person on the supplier side (name, e-mail address, phone number) to enable the procurement manager to immediately contact her or him. Providing this information requires the agent to obtain it from the ERP database with the material number as query parameter. This is an example for using a global state element (see Fig. 15.4).

Pattern detect agents usually also work on the results of several filtering steps. They select subsets of the filtered events by matching them with predefined patterns of event combinations. The matching result is then subject to derivation as outlined before. In our example a count pattern could help identify class A materials where the low stock threshold might be set too low. The pattern to look for could be that there need to occur more than 7 out of the last 10 purchase requisitions with a requested delivery time less than three days. If the pattern is satisfied the agent derives a new event 'threshold problem recognized' with the relevant information and publishes it in the EPN where suitable subscribers can process it. Such an event consumer could be the process portal again which alerts the responsible manager in order to make him adjust the threshold.

Besides those already given, the following examples should serve to show how combinations of EPA functionality can turn raw events into metrics and KPIs, interpret situations and make use of the results.

(1) Current states of instances of different processes (number of instances running, on hold, idle, completed etc.) can be computed by permanently filtering the relevant time stamp events and aggregating them by status and process. The metrics can be displayed on a dashboard.
(2) The moving average of throughput time can be determined by adding up the differences of end and start time and dividing the result by the accumulated number of instances. Each recalculation is triggered by the termination event of an instance. Its result can also be sent to a dashboard for presentation.
(3) Using the time stamps of an instance's entering and leaving its several execution steps allows for predicting its particular throughput time. The event processor can compare the values for an instance with the long-term average (dynamically updated) and compute the probability of being in time or late at the end of the process (predictive workflow [19]). If a delay is being detected in an early phase the information can be used to alter the execution of the running instance, e.g. by increasing the priority level in the process engine. This could lead to a reduction of wait time in the steps still to come in order to catch up time.
(4) By permanently filtering the start and end times of all instances of an ordering process a system can compute the number of instances currently active in a sliding time window (e.g. one hr.), for example updated every five minutes. If this number violates a threshold (event pattern) the application sends out an alert to the manager in charge, so that he can decide on whether to add staff to handle the high amount of orders without delay for the customers.

All examples apply some of the generic logical functions of EPAs described in Sect. 15.2.4.1. They filter relevant events, transform them into metrics and KPIs, look for and compare with patterns and produce results.

Output adapters finally change the results from the internal CEP engine format into formats that can be understood by the consumers and send them out. The engine usually does not store the event data itself. Results are not considered to be only events, but also 'metrics, messages or function calls' [11].

15.2.4.3 Event Processing Systems Requirements

The requirements CEP applications should meet include features for event input and output, filtering, projection, aggregation, transformation, pattern detection, context awareness, logging and analysis, prediction, learning, and distribution. Due to the objective of timely processing (low latency) and the high number of potentially relevant events solutions also require high computing capacity and performance up to some hundred thousands of events processed per second. A detailed list of both functional and non-functional requirements of CEP applications can be found in [21].

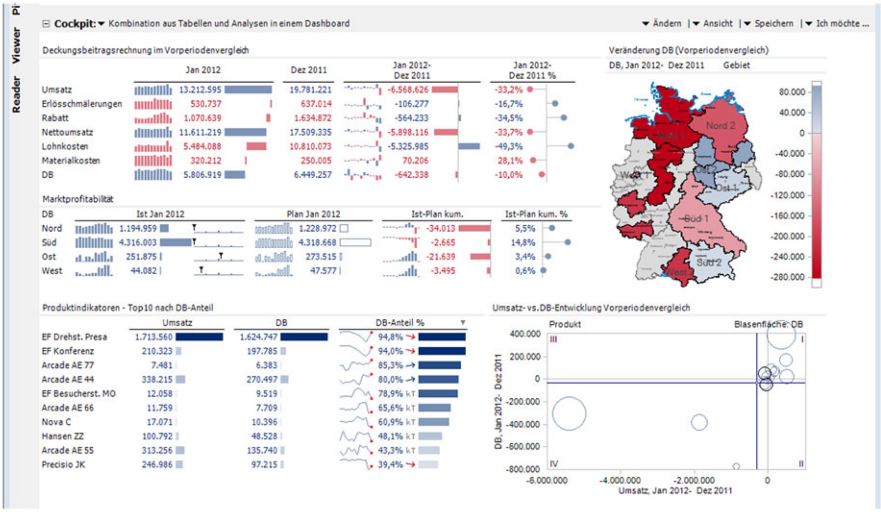
15.2.5 Event Consumers (Event Sinks)

Event consumers receive derived events from the event processing system after they have been transformed from the engine format into theirs by the output adapters. Like on the producer side there is a broad spectrum of event consumers, sometimes entities in both roles. For example there are hardware actuators like switches as well as software applications like ERP and data stores (e.g. event logs).

Regarding the types of results coming from event processing, major event consumers in the BAM context are process engines, dashboard applications (stand alone or as part of a BPMS) and message services.

A process engine can receive a function call via its API and as a consequence, depending on the type of call and the parameters coming with it, either creates a new instance of a process, alters the behavior of a running instance, sets it on hold or terminates it [6]. This means the process engine and the CEP environment are integrated through the engine's role both as event producer and consumer. This integration is termed Event-driven BPM [9, 18, 23].

Dashboard software can get fed with events which influence the information it visualizes to the user on its display. Dashboard or cockpits solutions, often integrated into process portals, serve to inform user with an information push without him to become active [5]. The metaphors refer the intuitive and quickly understandable display of a few, but important metrics and KPIs as parameters for control. Dashboards permanently visualize values like the number of instances currently in progress. Refreshing of the display is triggered by the recalculation which itself is caused by an incoming event. In order to improve information density on the limited dashboard space traditional display means like tachometers, colored traffic lights and pie charts are more and more replaced or at least mixed with intense word-sized graphics, the so-called micrographs. Examples are bullet graphs or sparklines which


Reader
Viewer
Pi
⊟ Cockpit: ▾ Kombination aus Tabellen und Analysen in einem Dashboard
▾ Ändern
▾ Ansicht
▾ Speichern
▾ Ich möchte ...

Deckungsbeitragsrechnung im Vorperiodenvergleich

	Jan 2012	Dez 2011	Jan 2012- Dez 2011	Jan 2012- Dez 2011 %
Umsatz	13.212.595	19.781.221	-6.568.626	-33,2%
Erlösschmälerungen	530.737	637.014	-106.277	-16,7%
Rabatt	1.070.639	1.634.872	-564.233	-34,5%
Nettoumsatz	11.611.219	17.509.335	-5.898.116	-33,7%
Lohnkosten	5.484.088	10.810.073	-5.325.985	-49,3%
Materialkosten	320.212	250.005	70.206	28,1%
DB	5.806.919	6.449.257	-642.338	-10,0%

Marktprofitabilität

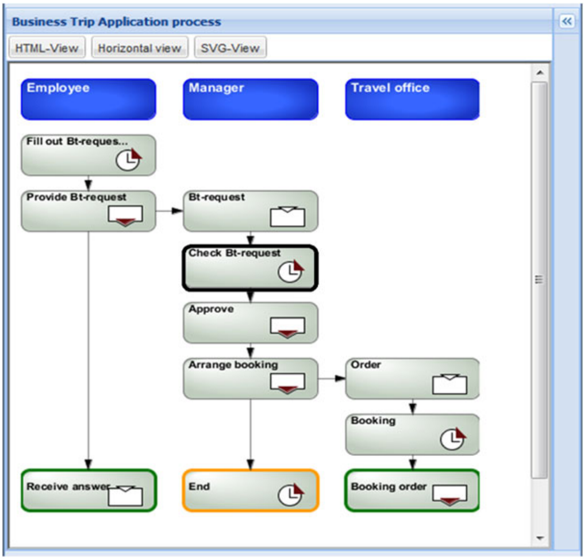
DB	Ist Jan 2012	Plan Jan 2012	Ist-Plan kum.	Ist-Plan kum. %
Nord	1.194.959	1.228.972	-34.013	5,5%
Süd	4.316.003	4.318.668	-2.665	14,8%
Ost	251.875	273.515	-21.639	3,4%
West	44.082	47.577	-3.495	0,6%

Produktindikatoren - Top10 nach DB-Anteil

	Umsatz	DB	DB-Anteil % ▾
EF Drehst. Presa	1.713.560	1.624.747	94,8% ↘
EF Konferenz	210.323	197.785	94,0% ↘
Arcade AE 77	7.481	6.383	85,3% →
Arcade AE 44	338.215	270.497	80,0% →
EF Besucherst. MO	12.058	9.519	78,9% kT
Arcade AE 66	11.759	7.709	65,6% kT
Nova C	17.071	10.396	60,9% kT
Hansen ZZ	100.792	48.528	48,1% kT
Arcade AE 55	313.256	135.740	43,3% kT
Precisio JK	246.986	97.215	39,4% ↘

Fig. 15.6 Dashboard example [25]

Fig. 15.7 Real-time instance execution path (reprinted from [5])



may signal the presence of special situations, such as exceeding the specified maximum processing time of an instance or show the development of a metric over time like the average execution time. The metrics and KPIs to be presented to the audience and the styles of visualization are specified in the design phase (see Fig. 15.3). Figure 15.6 depicts a dashboard example including bullet graphs and sparklines.

Some dashboards also provide the possibility to graphically track the execution of single instances at runtime like shown in Fig. 15.7. The illustration depicts an example for an instance of a business trip application process. The bold rounded rectangular around the state ‘Check Bt-request’ in the behavior of the manager indicates the current execution state.

In order to not only visualize exceptions and other events via dashboards it is helpful to be able to inform and alert people in charge (e.g. process owners) by messaging services like E-Mail, text message etc. In this case the CEP processor would also use the API of such services to call their functionality in order to let them send out messages.

15.2.6 Communication Mechanisms

Members of an event processing network communicate with each other via channels (for simplification reasons not included in Fig. 15.3). These receive input events from sending elements and route them to the dedicated addressees without any changes [4]. To realize this communication EPNs support protocol or API-based mechanisms like the Publish/Subscribe method of the Java Message Service API or the ‘notification service’ in CORBA [24]. Event producers and consumers like ERP, CRM, Process Engines etc. are usually part of an integrated IT environment interconnected by Enterprise Application Integration (EAI) middleware like Enterprise Service Buses (ESB). Also connecting the Event Processor to the ESB allows for using the features the bus offers for handling the event and message exchange in distributed systems.

15.3 Conclusion and Perspectives

Section 15.2 described Business Activity Monitoring or Management as a more or less loosely coupled combination of mainly CEP functionality, process engine and dashboard application. As indicated in Fig. 15.2 it complements traditional process monitoring and ex-post analysis and is therefore considered to be a real-time extension of Business Intelligence. The fusion of BAM/CEP and BI technologies is also referred to as Event-driven BI or Operational Intelligence [13, 20].

BI is useful to derive thresholds and target values for metrics and KPIs from historic data to be used in BAM. Results of Data Mining can trigger BAM, e.g. if a BI tool reports accounts suspected for fraud, the real-time monitoring can freeze those [13]. BI can also periodically or on request analyze the historical events to reveal the occurrence of certain process instances over time. For example for planning staff capacity it is important to know how 500 incoming loan requests per week are distributed over the different weekdays or even shorter time periods. A periodical analysis of the audit trail and transaction data written by the workflow engine and other applications when executing instances can help designing and optimizing processes. This BI-type analysis is called Process Mining and has three variants:

Discovery derives process models out of the log data. Conformance means to check whether real instances are processed as specified in the model. Enhancement with the subtypes repair and extension aims for modifying an existing model or adding new aspects to it, both initiated by data detected in the log files [8, 22].

Application domains of BAM/CEP are numerous. Some examples are [7, 16]:

- decision support in supply chain logistics
- baggage handling and predicting flight delays in airline industry
- fraud detection in financial service industry
- monitoring compliance

Applications like those listed above are build on standard software packages or individually implemented. A lot of major software vendors like SAP, Oracle, IBM, Microsoft, Software AG or Tibco as well as many specialized software producers offer tool environments to develop BAM/CEP solutions.

The importance of supporting business decisions with BAM/CEP permanently growing due to the rapidly increasing amount of data created in many areas, e.g. transactional systems, mobile communication and computing, social networks etc. This phenomenon is known as Big Data, and event processing with filtering, transforming, pattern detection, alerting, visualizing relevant information etc. offers valuable solutions to cope with it.

Some of the current and future challenges to tackle include the event modeling, interoperability and the runtime performance of solutions. Modeling events, event patterns, metrics, KPIs etc. to be processed requires a lot of domain knowledge and experience (like in process modeling). To support this activity [24] suggest to develop domain specific reference models. On the technology side the interoperability of heterogeneous system elements is limited due to missing standards like an event description. Another issue is assuring a satisfactory event processing performance even with high numbers of events per time unit. In-memory databases are a promising technology for that.

References

1. Chandy, K., Schulte, W.: Event Processing: Designing IT Systems for Agile Companies. New York (2010)
2. Dinter, B., Bucher, T.: Business performance management. In: Chamoni, P., Gluchowski, P. (eds.) Analytische Informationssysteme, 3rd edn., pp. 23–50. Springer, Berlin (2006)
3. Eckert, M., Bry, F.: Complex event processing (CEP). Inform. Spektrum **32**(2), 163–167 (2009)
4. Etzion, O., Niblett, P.: Event Processing in Action. Manning Publications, Stamford (2011)
5. Fleischmann, A., Schmidt, W., Stary, C., Obermeier, S., Börger, E.: Subject-Oriented Business Process Management. Springer, Berlin (2012)
6. Fleischmann, A., Schmidt, W., Stary, C., Strecker, F.: Nondeterministic events in business processes. In: Proceedings of the 6th International Workshop on Event-Driven Business Process Management (edBPM12), Co-located with BPM 2012, Tallinn/Estonia. Lecture Notes in Business Information Processing (LNBIP). Springer, Berlin (2012)

7. Gassman, B.: Business activity monitoring. In: Dixon, J., Jones, T. (eds.) Hype Cycle for Business Process Management. Gartner Inc., pp. 82–83 (2011)
8. Grob, H., Coners, A.: Regelbasierte Steuerung von Geschäftsprozessen – Konzeption eines Ansatzes auf Basis von Process Mining. WIRTSCHAFTSINFORMATIK **50**(4), 268–281 (2008)
9. Hermosillo, G., Seinturier, L., Duchien, L.: Using complex event processing for dynamic business process adaptation. In: Proceedings of the 7th IEEE International Conference on Service Computing (SCC), Miami, pp. 466–473 (2010)
10. Heß, H.: Von der Unternehmensstrategie zur Prozess-Performance – Was kommt nach Business Intelligence? In: Scheer, A.-W., Jost, W., Heß, H., Kronz, A. (eds.) Corporate Performance Management, pp. 7–29. Springer, Berlin (2005)
11. Janiesch, C., Matzner, M., Müller, O.: A blueprint for event-driven business activity management. In: Rinderle, S., Toumani, F., Wolf, K. (eds.) 9th International Conference on Business Process Management (BPM). LNCS, vol. 6896, pp. 17–28. Springer, Berlin (2011)
12. Kang, J., Kwan, H.: A business monitoring system supporting real-time business performance management. In: Proceedings of the 2008 Third International Conference on Convergence and Hybrid Information Technology (ICCIT '08), Washington DC, vol. 1, pp. 473–478 (2008)
13. Kochar, H.: Business intelligence and business intelligence, http://www.ebizq.net/topics/cep/features/6596.html?&pp=1, last Access 2012-07-25
14. Luckham, D., Schulte, R. (eds.): Event processing glossary Version 2.0/2011, Event Processing Technical Society, http://www.complexevents.com/2011/08/23/event-processing-glossary-version-2-0/, last Access 2012-07-25
15. Luckham, D.: The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems. Amsterdam (2002)
16. Luckham, D.: BAM providers as online banking fraud preventers, http://www.ebizq.net/topics/cep/features/4891.html?&pp=1, last Access 2012-07-25
17. Luckham, D.: The beginnings of IT insight: business activity monitoring, http://www.ebizq.net/topics/cep/features/4689.html?&pp=1, last Access pp. 2012-07-25
18. Paschke, A.: A semantic rule and event driven approach for agile decision-centric business process management. In: Abramowicz, W., et al. (eds.) ServiceWave 2011. LNCS, vol. 6994, pp. 254–267. Springer, Berlin (2011)
19. Schmidt, W., Fleischmann, A., Gilbert, O.: Subjektorientiertes Geschäftsprozessmanagement. HMD, Prax. Wirtsch.inform. **266**, 52–62 (2009)
20. Schulte, R., Sumic, Z.: Complex-event processing. In: Dixon, J., Jones, T. (eds.) Hype Cycle for Business Process Management. Gartner Inc., pp. 50–52 (2011)
21. The event processing manifesto, written by the participants of the 2010 Dagstuhl seminar on event processing, http://drops.dagstuhl.de/opus/volltexte/2011/2985/pdf/10201.SWM.2985.pdf, last Access 2012-07-25
22. Van der Aalst, W.: Process Mining. Springer, Berlin (2011)
23. Von Ammon, R., Ertlmaier, T., Etzion, O., Kofman, A., Paulus, T.: Integrating complex events for collaborating and dynamically changing business processes. In: Dan, A., Gittler, F., Toumani, F. (eds.) ICSOC/ServiceWave 2009. LNCS, vol. 6275, pp. 370–384. Springer, Berlin (2010)
24. Von Ammon, R., Silberbauer, C., Wolff, C.: Domain Specific Reference Models for Event Patterns—for Faster Developing of Business Activity Monitoring Applications. VIPSI, vol. 2007. Lake Bled (2007)
25. Created with DeltaMaster 5.5.3, see www.bissantz.com

Chapter 16
Scaling up Data Mining Techniques to Large Datasets Using Parallel and Distributed Processing

Frederic Stahl, Mohamed Medhat Gaber, and Max Bramer

Abstract Advances in hardware and software technology enable us to collect, store and distribute large quantities of data on a very large scale. Automatically discovering and extracting hidden knowledge in the form of patterns from these large data volumes is known as data mining. Data mining technology is not only a part of business intelligence, but is also used in many other application areas such as research, marketing and financial analytics. For example medical scientists can use patterns extracted from historic patient data in order to determine if a new patient is likely to respond positively to a particular treatment or not; marketing analysts can use extracted patterns from customer data for future advertisement campaigns; finance experts have an interest in patterns that forecast the development of certain stock market shares for investment recommendations. However, extracting knowledge in the form of patterns from massive data volumes imposes a number of computational challenges in terms of processing time, memory, bandwidth and power consumption. These challenges have led to the development of parallel and distributed data analysis approaches and the utilisation of Grid and Cloud computing. This chapter gives an overview of parallel and distributed computing approaches and how they can be used to scale up data mining to large datasets.

F. Stahl (✉)
The School of Design, Engineering & Computing, Poole House, Bournemouth University, Talbot Campus, Poole, Dorset BH12 5BB, UK
e-mail: fstahl@bournemouth.ac.uk

M.M. Gaber · M. Bramer
School of Computing, University of Portsmouth, Lion Terrace, Portsmouth, Hants PO1 3HE, UK

M.M. Gaber
e-mail: Mohamed.Gaber@port.ac.uk

M. Bramer
e-mail: Max.Bramer@port.ac.uk

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_16, © Springer-Verlag London 2013

16.1 Performance Challenges in Data Mining

There is a substantial commercial interest in developing and improving business intelligence and data mining applications in order to extract useful information in the form of patterns from very large data volumes. Computer systems capture our lives in the form of credit card transactions; loyalty reward systems record our shopping habits; CCTV cameras, GPS systems embedded in our smartphones and navigation systems record our movement and whereabouts; and the world wide web records our data through applications such as facebook, email, twitter and blogs [52]. In [22] the authors estimated that in the year 2020 the size of our digital universe will be 44 times as big as it was in the year 2009. Advances in storage technology make it possible to store all these data volumes at a very low cost, hence the ubiquitous challenge in data mining is the scalability of data mining techniques to these large data volumes. Many areas in science are confronted with the problem of scalability of data mining techniques also. For example in cosmology, researchers store terabytes of image data in massive databases such as in the Sloan digital sky survey [49, 51]. The bioinformatics community is just starting to be able to store and analyse molecular dynamics simulation data which can easily comprise hundreds of gigabytes for only one simulation experiment [3]; also in bioinformatics the human genome project stores the entire genetic blueprint of our bodies [35]. In the business area large and complex databases are reported, for example Amazon's two largest databases combine 42 terabytes of data and AT&T's largest database comprises 312 terabytes [34]. Loosely speaking, the scientific and business worlds are confronted with very large databases storing information. In order to extract meaningful patterns, analysts need to apply data mining techniques. Hence scalable data mining technologies are required.

A further complication to the mining of these massive amounts of data is the fact that organisations often store their data in geographically distributed locations in order to overcome bandwidth problems when transferring such large amounts of data to a central data repository. This generates further problems such as heterogeneous data base schemas and confidentiality issues when dealing with sensitive data. Some research has been conducted in order to overcome these bandwidth constraints such as in the DataMiningGrid.org project [48]. One of their approaches is to deploy individual data mining close to the data sources in a dynamic way and execute them remotely rather than downloading myriads of data. However this chapter is about scaling data mining algorithms to deal with large datasets rather than dealing with geographically distributed data. For a comprehensive reading list about Distributed Data Mining approaches that can be used to mine geographically distributed data sources the reader is referred to [4].

The rest of this chapter is organised as follows. Section 16.2 highlights parallel and distributed data mining approaches to tackling the problem of scalability of data mining techniques. Section 16.3 discusses approaches to scaling up data stream mining techniques in resource constraint environments. Section 16.4 provides a summary of successful applications, available open and commercial parallel data mining systems are discussed. This chapter closes with a discussion of remaining challenges in scaling up data mining to large datasets in Sect. 16.5.

16.2 Parallel and Distributed Data Mining Approaches and Frameworks

Parallel and distributed data mining approaches have been proposed in the past in order to tackle the challenge of scalability to large data sources. Whereas parallel data mining clearly refers to the parallelisation of a data mining task by executing data mining tasks concurrently, the term distributed data mining is used ambiguously in the data mining literature. Often the term 'distributed' is associated with data mining of geographically distributed datasets and is not necessarily concerned with the computational scalability. The parallelisation of a data mining task often follows a data parallel approach as the computational workload of data mining tasks is usually directly dependent on the amount of data that needs to be processed [43]. In data parallelisation the data is partitioned into smaller subsets and distributed to multiple processors on which the data mining tasks are executed concurrently. Unless stated otherwise this chapter uses both the terms parallel data mining and distributed data mining to refer to *data parallelism*.

16.2.1 Multiprocessor Computer Architectures

In order to execute data parallel algorithms a multiprocessor architecture is needed. The basic idea in data parallelism is to distribute or assign the workload to several processing units in the form of subsets of the dataset. There are two relevant multiprocessor architectures that are suited for this purpose, *tightly-coupled* architectures and *loosely-coupled* architectures. However hybrids between the two architectures are possible.

(a) A *loosely-coupled* architecture comprises multiple standalone computers. Each computer comprises one processing unit and its local private memory. This architecture requires data distribution and accumulation mechanisms, and a communication network. An implementation of this architecture is also often referred to as *'Massively Parallel Processors'* (MPP). MPPs are illustrated in Fig. 16.1(a).
(b) A *tightly-coupled* architecture consists of multiple processors that share a common memory using a shared bus system. No data distribution is required as the processors do not have a private memory. An implementation of this architecture is also often referred to as *'Shared memory MultiProcessor machines'* (SMPs). SMPs are illustrated in Fig. 16.1(b).
(c) A hybrid approach between both architectures is possible by building a loosely-coupled system out of 'Shared memory Multiprocessor machines'. SMP/MPP hybrids are illustrated in Fig. 16.1(c).

A loosely coupled system uses a communication network in order to communicate data, whereas a tightly-coupled system uses the system bus. Regarding tightly coupled systems there is a bottleneck with the number of processors the system can

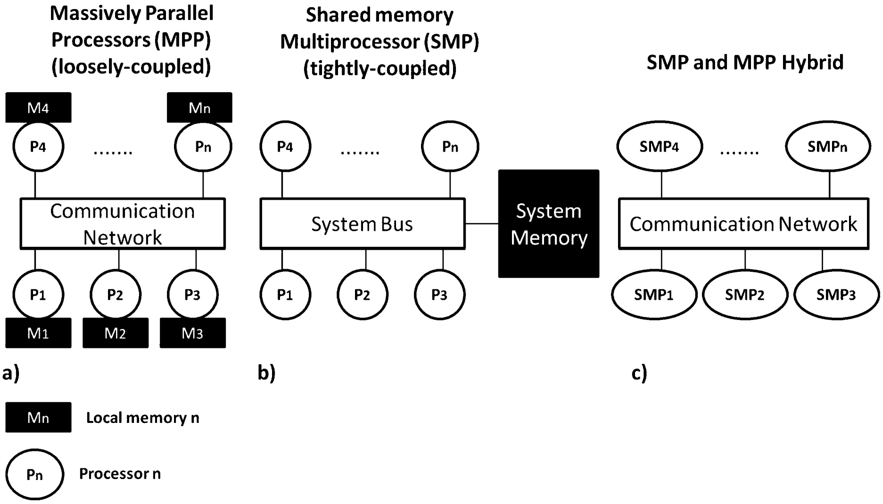


Fig. 16.1 Multiprocessor architectures. (**a**) Loosely-coupled multiprocessor architecture. (**b**) Tightly-coupled multiprocessor architecture. (**c**) A hybrid of loosely- and tightly-coupled architectures

support in the context of data mining applications. This is because the more processors share the system bus, the less bandwidth is available per processor. A loosely coupled system does not have this bottleneck as the processors do not share a common system bus. A further advantage of a loosely-coupled system is that the application components are hosted on different computers distributed in a network. This makes loosely-coupled systems more robust to hardware failures compared with tightly coupled systems as a failing processing node will not cause the entire application to fail. In a tightly-coupled system a failing processor usually represents a single point of failure. However a disadvantage of a loosely-coupled system is that it requires communication and collaboration between its computing nodes which introduces an additional overhead for the application. An advantage of a tightly-coupled system over a loosely-coupled system is that it is usually more efficient at processing data as it avoids data replication and does not need to transfer information between processing units. Yet loosely-coupled systems can be obtained at a relatively low cost compared with tightly coupled systems, if standard workstations are used as processing units. This allows modest sized organisations which cannot afford a SMP to harvest the computational power, memory, and storage of their whole local area network in order to execute parallel data mining tasks. Also a loosely-coupled system can be upgraded gradually by simply replacing old workstations with newer ones, whereas a tightly-coupled system needs to be replaced in its entirety.

Hybrid architectures of loosely-coupled SMPs are becoming common and most workstations nowadays can be seen as small MPPs as they utilise multicore processor technology. However, research in the area of data mining has just begun to utilise such hybrid systems efficiently. The data mining community is called on to investigate the advantages of such hybrids, their suitability and benefits for scaling up data mining systems to large data volumes.

16.2.2 Parallel Predictive Algorithms

One of the most important data mining tasks is classification rule induction, which can be categorised into 'divide and conquer' and 'separate and conquer' methods [52]. 'Divide and conquer' generates a decision tree by recursively partitioning the training data according to the variables and classifications, aiming to optimise a chosen metric such as the entropy [36]. 'Separate and conquer' generates a ruleset by specialising a general rule for a certain target class on the training data. After each rule induced the subset of the training examples that is covered by the rules induced so far is deleted and the next rule is induced until all training examples are covered by the ruleset. Examples of 'separate and conquer' classifiers are [13, 14] and examples of the Prism family of algorithms are [5, 6, 10].

Most recent research does not focus on the parallelisation of decision tree induction, but rather on concurrent execution of whole data mining tasks, we highlight general parallel decision tree induction approaches in this section as they give a valuable insight into issues that may occur when parallelising data mining tasks. There are two principal ways of parallelising decision tree classifiers: the *synchronous tree construction* approach highlighted in Fig. 16.2(a) and the *partitioned tree construction* approach highlighted in Fig. 16.2(b) [41].

In the 'synchronous tree construction' the training dataset is initially distributed between *n* processors. During the tree induction each processor holds an exact copy of the tree in its memory (assuming a MPP machine has been used). The processors cooperate in expanding the same tree node by gathering statistics of their portion of the data and sharing these statistics through communication. Eventually each processor will perform the same tree node expansion independently on their copy of the tree. Some of the most well-known parallel tree classifiers are based on 'synchronous tree construction' with vertical partitioning [39]. In the 'partitioned tree construction' different processors work on different parts of the tree and the training data. Initially only one processor is assigned to expand the root node. The resulting child nodes are then assigned to different processors, each independently expanding the subtree of its child node. This is done recursively until all processors are assigned to different subtrees.

The advantage of the 'synchronous tree construction' is that there is no communication of training data. However, the disadvantage of 'synchronous tree construction is that the communication of statistics increases as the tree grows. Also the workload between the processors is changing during the tree induction and may cause workload imbalances [41]. The advantage of 'partitioned tree construction' is that as each processor works independently no communication is needed. However, the disadvantage of the 'partitioned tree construction' approach is that initially a single processor has the entire workload [41]. Taking the aforementioned advantages

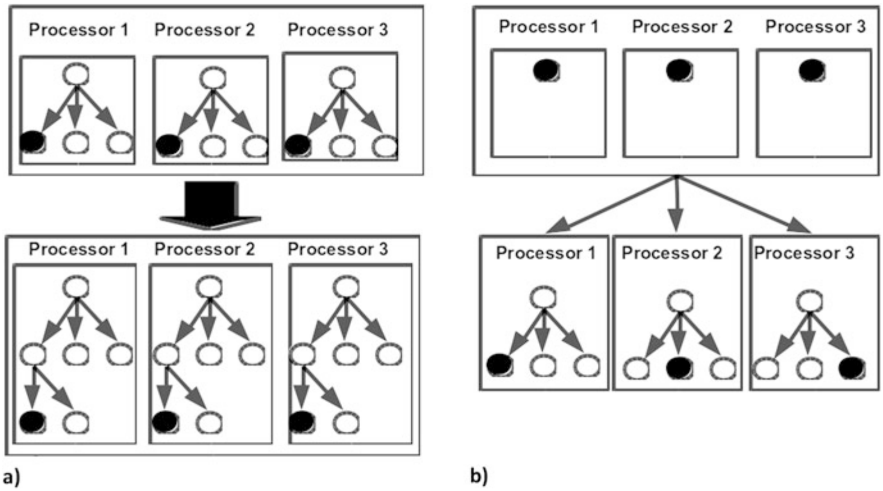


Fig. 16.2 Parallel decision tree induction approaches. The data is evenly distributed between all processors. (**a**) Shows the synchronous tree construction approach. All processors keep the same decision tree in memory and cooperate on expanding the same tree node. (**b**) Shows the partitioned tree construction approach. Different processors are assigned on to different subtrees (when possible) and expand these subtrees simultaneously and independently

of both approaches and minimizing the disadvantages has resulted in a hybrid approach [41]. This approach starts with the 'synchronous tree construction' until the communication overhead becomes too high and then switches to the 'partitioned tree construction', which removes the entire communication.

This illustration of different parallel decision tree based approaches shows the difficulties encountered when parallelising data mining algorithms, i.e. communication overheads, synchronisation overheads and workload imbalances. This has led to the development of frameworks that support the parallelisation of whole families of algorithms. The remainder of this section highlights frameworks and infrastructure technologies to assist with the parallelisation of data mining algorithms.

16.2.3 Parallel Formulations of Separate and Conquer Classification Rule Induction Using PMCRI

The Parallel Modular Classification Rule Induction (PMCRI) framework [44, 46] is a framework for parallelising algorithms of the Prism family whose members follow the 'separate and conquer' approach.

Figure 16.3 outlines PMCRI's basic architecture, which is based on a distributed blackboard system. A blackboard system is often illustrated using the metaphor of several experts, with expertise in different domains, that are gathered around a physical blackboard. The experts can solve a problem they have in common by using

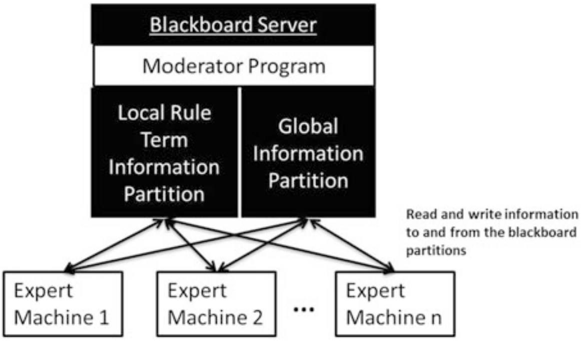

Blackboard Server
Moderator Program
Local Rule Term Information Partition
Global Information Partition
Read and write information to and from the blackboard partitions
Expert Machine 1
Expert Machine 2
...
Expert Machine n

Fig. 16.3 The PMCRI framework based on a distributed blackboard architecture

their own expert knowledge and information written on the blackboard in order to infer new knowledge and information. Experts share information by writing it on the blackboard. The software implementation of a blackboard system can be realised by a client server architecture [32]. Experts are client machines whose expertise is represented by the portion of the data they hold in memory and the blackboard is a communication server. PMCRI divides the blackboard into several logical partitions that have different meanings to the experts. PMCRI's blackboard is divided into two logical partitions: one to submit local rule term information and one to retrieve global information about the whole algorithm's status. PMCRI partitions the data vertically (according to the features) and distributes the subsets evenly amongst the expert machines. The experts induce each rule concurrently by inducing rule terms that are the 'best' terms in the local feature subspace to further specialise a rule. The experts use the blackboard in order to communicate which rule term is globally the best one and to assembly the final rule. Figure 16.3 also highlights a moderator program. The moderator is also implemented in the form of an expert machine. It coordinates the rule induction schedule and thus represents the underlying Prism algorithm. Just replacing the moderator by a different one allows one member of the Prism family to be changed to another.

16.2.4 The MapReduce Paradigm for Parallelisation Data Mining

Google's *MapReduce* paradigm of parallel programming [16] provides a means to simplify the development of parallel data mining techniques offering load balancing and fault tolerance. The actual source code regarding parallelisation and data communication is hidden from the programmer by limiting the parallel programming model to only the *map* and the *reduce* functions. Data mining applications parallelised using MapReduce make use of the Google File System (GFS) [23] which provides a means of storing data in a distributed manner and redundantly over a network of commodity workstations. Google's MapReduce is a proprietary software. However, Hadoop provides an open source implementation of Google's MapReduce paradigm based on its Hadoop Distributed File System (HDFS) which is Hadoop's

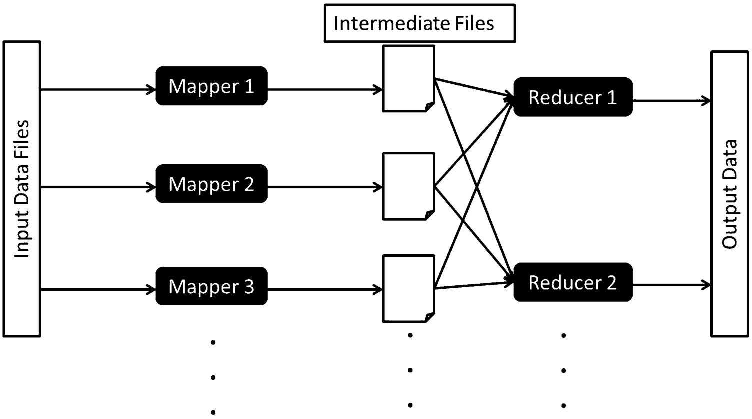


Fig. 16.4 A typical setup of a Hadoop computing cluster. A physical node in the computer can execute more than one Mapper and Reducer

implementation of GFS [25]. MapReduce splits an application into smaller parts called *mappers*. Each mapper can be processed by any of the workstations in the nodes in the cluster. A high reliability of the application is provided by the framework's ability to recover a failed mapper. Intermediate results produced by these mappers are then combined by one or more *reducer* nodes.

A typical Hadoop computing cluster is highlighted in Fig. 16.4. Large amounts of data are processed by splitting this data into smaller portions and storing it redundantly in the cluster using HDFS. Then the smaller portions of the data are loaded into the mapper machines and processed using a user defined function. The results computed by the mappers (intermediate files) are passed on the reducers where they are combined. The programmer is only required to implement the processing functions used by the mappers and the combining function by the reducers.

MapReduce's relevance in the data mining community has been demonstrated in countless projects. For example Google reported having utilised MapReduce in at least 900 projects [16]. However, this was in 2008 and it is very likely that there are many more projects by now. For example Google developed MiniHash clustering using the MapReduce paradigm in order to generate personalised recommendations for users of Google News [15]. Google also makes use of MapReduce to cluster billions of images in order to find new duplicates [30]. The authors of [33] used MapReduce to create the Parallel Learner for Assembling Numerous Ensemble Trees (PLANET) system. The PLANET system provides a scalable parallel decision tree classifier, regression trees and also parallel ensemble learners.

A data mining approach that lends itself to parallelisation using MapReduce is *ensemble learning*. In ensemble learning multiple models are generated from subsamples or bags of the same data and combined in order to achieve a better predictive performance. Ensemble learners lend themselves to parallelisation as the multiple models can potentially be generated concurrently using multiple processors. Recently published work that aims to parallelise ensemble learners is reported in [2, 33, 45, 53]. The authors of [12] adapted MapReduce in order to parallelise several learning algorithms: amongst others algorithms such as k-means, logistic regression, naive bayes, support vector machines etc. Also the partitioned tree construction approach as highlighted in Sect. 16.2.2 could be parallelised using MapReduce by assigning the construction of different subtrees to different mappers.

However as pointed out by [8], MapReduce is designed for use in a dedicated cluster assuming that each node is equally powerful, reliable and only dedicated to processing. The authors of [8] further propose to remove some of these assumptions and create a more ‘grid like’ version of MapReduce.

16.2.5 Grid and Cloud Computing for Parallel Data Mining

The grid and cloud computing paradigm has emerged as an attractive computing infrastructure for implementing complex and computationally costly applications. The difference between grid and cloud computing is often blurred in the literature. In grid computing users consume but also share computing resources whereas cloud computing is a rather commercial paradigm where computing resources are offered through services on demand against payment.

The computing grid is often described with the analogy of an electrical power grid that transparently provides electricity to the end user. According to this analogy the computing grid provides computing power in terms of CPU time, memory and storage. The grid aims to mass computing resources whilst hiding their specifications. Thus it provides a consistent interface for the end user to access high performance and/or high throughput computation [31]. A grid is organised in geographically distributed virtual organisations. Virtual organisations are collections of computational resources in terms of processing and storage. In general computing grids are constructed using a set of grid software libraries, the middleware, whose service-oriented architecture provides services to access virtual organisations for software applications that want to make use of the grid. An example for a widely used open source grid middleware is the Globus Toolkit [24].

Data mining in a grid environment is a special form of distributed data mining [47] as it is stimulated by sharing of resources using local and wide area networks [29]. Several projects concerned with high performance data mining using grid environments have been followed up in the past couple of years, such as the *DataMiningGrid.org* project which aims to integrate a wide variety of data mining applications and different application scenarios into a single grid framework [47, 48], or the gridminer project [7] which aims to integrate the entire knowledge discovery process in a service-oriented grid application.

Cloud computing refers to the on demand delivery of applications and hardware resources over the internet in the form of services. Three types of service can be accessed using a cloud infrastructure. Both grid computing and cloud computing aim to provide access to large computing and storage resources. However, the cloud additionally uses visualisation in order to provide access to the computing resources and thereby conceals physical heterogeneity, geographical distribution, and faults [37]. Compared with the cloud the grid provides services that enable the collaborative sharing of distributed computing resources. In this sense the grid is complementary yet independent from cloud computing [37]. In other words, Grid computing aims to solve computational problems whereas cloud computing provides software and computing services on demand. Because of this on demand principle cloud computing is usually used in the private sector whereas grid computing is used more in public sector research projects. Probably the best known commercial cloud system is *Amazon Web Services* [1], comprising Amazon's *S3* (Simple Storage Service) providing data storage, and Amazon's *EC2* service providing on demand computing capacity. However, the disadvantage of grid and cloud computing is that consumers give away partial control of security management to the grid partners or cloud service providers. For a discussion of this topic the reader is referred to [42].

Scaling up data mining is not only restricted to problems that deal with large datasets. Performing data mining tasks on resource-constrained devices like smartphones and small sensing devices stimulates the need for new strategies for scaling up data mining techniques. The following section is devoted to the discussion of strategies of dealing with this problem. In fact, this problem represents the other side of the coin, traditional scaling up of data mining techniques look at having high performance hardware and large data sets, while scaling up data mining techniques for resource-constrained devices looks at having possibly smaller data sets to be analysed using restricted processing capabilities.

16.3 Approaches to Scaling up Data Stream Mining in Resource Constrained Environments

We witness the era of handheld devices and small sensors performing tasks that in the near past required high performance systems. Data streams in and/or produced on such small devices can serve a number of extremely important applications in areas such as astronomy, stock market analysis and national security, to name a few. The following two important facts require the processing of data streams to be performed locally on-board small devices with low computational power. We shall refer to such devices in this chapter as resource-constrained environments.

1. It has been proven experimentally that local data processing is an energy efficient alternative to sending the data streams to a computational service with high power like the cloud [19]; and
2. current computational capabilities of resource-constrained environments do allow the performance of complex tasks, like data mining [11].

Despite the continuous advances in the computational capabilities of resource constrained devices, the demand of having increasingly complex computational tasks performed has also been continuous. Thus, we encounter what we can refer to as *relative resource constraints*, i.e. the advances in the hardware technologies fall short of addressing the demands of current application needs. The application needs are in fact coupled with the rise of large amounts of streaming data, which in turn led to the *big data* phenomenon. To address the issue of *relative resource constraints*, adapting the process to resource availability is needed. The *Algorithm Granularity* approach [18] is a generic framework that is able to adapt any data stream mining to data rate and resource availability. The following subsection gives an overview of the approach.

16.3.1 Algorithm Granularity Approach Overview

Algorithm Granularity approach has been introduced by Gaber with a comprehensive treatment of the subject reported in [17]. The approach relies on the concept of *resource consumption patterns*. Resource consumption patterns represent the change in resource consumption over a period of time, referred to as a time frame. The algorithm can change its settings from its three entry points: input, output, and processing. This could be applied to any data stream processing technique. However, the *Algorithm Granularity* approach was specifically designed for adapting data stream mining techniques. The three categories of settings of a mining algorithm are changed over time to cope with the availability of resources and current data rate. Definitions of these settings are as follows:

Algorithm Input Granularity (AIG) represents the process of changing the data rates that feed the algorithm. Examples of this include sampling, load shedding, and creating data synopsis. This is a common solution in many data stream mining techniques.

Algorithm Output Granularity (AOG) is the process of changing the output size of the algorithm in order to preserve the limited memory space. In the case of data mining, we refer to this output as the number of knowledge structures. For example the number of clusters or rules. The output size could be changed also using the level of output granularity which means the less detailed the output, the higher the granularity and vice versa.

Algorithm Processing Granularity (APG) is the process of changing the algorithm parameters in order to consume less processing power. Randomisation and approximation techniques represent the potential solution strategies in this category.

It has to be noted that there is a collective interaction among the above three categories. *AIG* mainly affects the data rate and it is associated with bandwidth consumption and battery life. Batteries tend to be drained rapidly when continuously sending or receiving data streams. On the other hand, *AOG* is associated with memory and *APG* is associated with processing power. The more memory is consumed, this implies that more knowledge structures are resident in the memory, and thus *AOG* is the suitable strategy. When the algorithm falls short of processing the incoming streams, algorithmic approximation and randomisation are used to speed the process up. Thus, *APG* appears as the suitable strategy. Figure 16.5 [21] shows the interaction among the three strategies and associated computational resources.

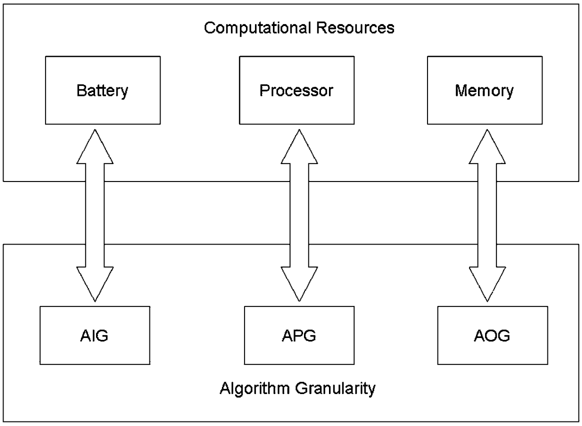


Fig. 16.5 The effect of algorithm granularity on computational resources

However, the change in any of the three affects the other resources. For example, approximation could be used to address the problem of high data rate by having less processing time per data record. It is important to note that the process of enabling resource awareness should be very lightweight in order to be feasible in a streaming environment characterised by its scarcity of resources.

16.3.2 Formalisation of Algorithm Granularity

The Algorithm Granularity requires continuous monitoring of the computational resources. This is done over fixed time intervals/frames that are denoted as *TF*. According to this periodic resource monitoring, the mining algorithm changes its parameters/settings to cope with the current consumption patterns of resources. These parameters are *AIG*, *APG* and *AOG* settings discussed briefly in the previous section. It has to be noted that setting the value of *TF* is a critical parameter for the success of the running technique. The higher the *TF* is, the lower the adaptation overhead will be, but at the expense of risking a high consumption of resources during the long time frame, causing the run-out of one or more of the computational resources.

The use of *Algorithm Granularity* as a general approach for mining data streams will require us to provide some formal definitions and notations. The following are definitions and notation that we will use in our discussion.

R: set of computational resources $R = \{r_1, r_2, \ldots, r_n\}$
TF: time interval for resource monitoring and adaptation.
ALT: application lifetime.
ALT': time left to last the application lifetime.
$NoF(r_i)$: number of time frames to consume the resources r_i, assuming that the consumption pattern of r_i will follow the same pattern of the last time frame.
$AGP(r_i)$: algorithm granularity parameter that affects the resource r_i.

According to the above, the main rule to be used to use the algorithm granularity approach is as follows:

IF $\frac{ALT'}{TF} < NoF(r_i)$
THEN SET $AGP(r_i)+$
ELSE SET $AGP(r_i)-$

Where $AGP(r_i)+$ achieves higher accuracy at the expense of higher consumption of the resource r_i, and $AGP(r_i)-$ achieves lower accuracy at the advantage of lower consumption of the resource r_i. For example, when dealing with clustering, it is computationally cheaper to allow incoming data instances in the stream to join an existing cluster with randomisation applied to which cluster the data instance would join. Ideally, the point should join a cluster that has sufficient proximity or a new cluster should be created to accommodate the new instance. This strategy has been applied to a technique by Gaber and Yu [20] termed *RA-Cluster*.

This simplified rule could take different forms according to the monitored resource and the algorithm granularity parameter applied to control the consumption of this resource. The *Algorithm Granularity* approach has been successfully applied to a number of data stream mining techniques. These techniques have been packaged in a java-based toolkit, coined *Open Mobile Miner* [28].

This section gives an overview of the *Algorithm Granularity* approach. For a comprehensive treatment of the subject area, the reader is referred to [17]. For interested readers in the data stream mining area including this approach, Gama's textbook [21] is our suggested source.

16.4 Software Tools and Applications

This section highlights some of the readily available parallel and distributed mining tools, and highlights several successful applications. The success in merging data mining with distributed computing technologies has lead to several distributed and parallel commercial as well as freely available data mining tools. Probably the most popular commercial system is Amazon's cloud [1], as mentioned in Sect. 16.2.5, it provides data storage, data analytics tools and on demand computing capacity. Amazon's cloud can only be remotely accessed whereas SAS's 'Analytics Infrastructure' can be deployed on site as well being used as a cloud service. SAS's 'Analytics Infrastructure' can be deployed on SMP or on MPP parallel architectures and comprises three means to scale up data analysis to large datasets. These are grid computing, moving computation close to the data in order to avoid data movement, and in memory analytics in order to reduce data access time to disk storage. A further commercial product is Microsoft's SQL Server which offers a scalable data warehouse implementation that can be hosted on MPP architectures. However, there are also free products such as the well known WEKA data mining software which allows parallel cross-validation calculations [9]. The Hadoop infrastructure [25] highlighted in Sect. 16.2.4 is a freely available technology, still it is widely used in commercial business intelligence applications.

With the rapid development of parallel and distributed data mining technology, several successful applications have been reported. For example Google's PLANET system [33] mentioned in Sect. 16.2.4 has been applied in the domain of 'computational advertising'. Two of the applications of the 'DataMiningGrid.org' project mentioned in Sect. 16.2.5 are in the automotive industry [47] for text mining as well as in bioinformatics for the distributed storage and analysis of very large quantities of Molecular Dynamics simulation data [50]. Parallel processing capabilities of grid architectures are the method of choice for the analysis of very large astronomical datasets such as in [54]. Probably the most popular application of grid computing is the SETI@home project. The project's goal is to detect intelligent extraterrestrial life through the analysis of massive radio telescope data [38].

16.5 Conclusions and Future Directions

This chapter presented techniques that can be and are used to scale up data mining tasks to large quantities of data. The approaches and systems highlighted in the previous sections along with already available commercial as well as free tools show the prosperity of this field of research. However, despite the recent successes in the scalability of data mining techniques, it remains an active research area, with unresolved issues and distinctly new innovative approaches.

Apart from the obvious improvements on hardware such as the usage of Graphics Processing Units (GPUs) [26], or better parallelisation frameworks and algorithms, the movement of data and the cost is an issue. Cloud computing is often presented as the method of choice if you urgently need a large amount of computing power to analyse a large amount of data. However, bandwidth is often the bottleneck as cloud systems are usually remotely located. In cases where the data becomes very large, local processing is still needed. A further open issue on the usage of the cloud and parallel data mining approaches is the cost/benefit relationship, which is hardly explored in the literature. Usage of cloud services or in-house parallel computing facilities is an investment decision that is only justified if the financial benefit of analysing large quantities of data outweighs the associated computational and hardware costs in financial terms. A further issue of grid and cloud computing that needs to be addressed more thoroughly is the security of confidential data, as outsourcing the data analysis also relies on trusting the grid/cloud service provider.

Whereas this chapter mainly discussed scalability issues associated with the size of the data, the type of method chosen may also inflict a computational bottleneck. For example the computational INtelligence platform For Evolving and Robust predictive systems (INFER) [55] provides a complex environment that automatically adapts to concept changes in the data. It basically evolves by training many different data mining models and returns the 'fittest model' either autonomously or with user assistance. Rather than the size of the training data, the training as such and the adaptation of many models in a concurrent way imposes a considerable computational bottleneck. Hence loosely coupled as well as tightly coupled parallelisation needs to be considered for improving the scalability of such systems.

The trend towards multi-core processors in standard workstations needs to be explored, as networks of such workstations are what we described as a hybrid architecture between loosely and tightly coupled architectures. Existing systems such as Hadoop rely on the workstations' operating system to balance the workload efficiently among the available cores.

A distinctly different approach to the analysis of large and complex datasets is emerging: 'Visual Analytics' (VA). VA describes the reasoning assisted by graphical visualisations in an interactive way. VA is based on the concept of information visualisation which aims at using the computational power of the human brain to process images and hence gain understanding of the data to be analysed [40]. VA extends this concept by interactively incorporating automatic analysis methods prior to and during the visualisation process [27]. The visual representations are not only used to visualise data and patterns for the user, but also to feed back information from the user to the analysis system. Using the combined computational power of silicon as well as biological hardware may result in well-scaling data analysis systems.

In general, parallel and distributed data analysis is still an open field of research. Past efforts to parallelise data mining techniques have come to fruition, but open issues outlined in this section remain to be addressed.

Acknowledgements The research leading to these results has received funding from the European Commission within the Marie Curie Industry and Academia Partnerships & Pathways (IAPP) programme under grant agreement n^{o} 251617.

References

1. Amazon. Amazon web services, 2012
2. Basilico, J.D., Munson, M.A., Kolda, T.G., Dixon, K.R., Kegelmeyer, W.P.: Comet: a recipe for learning and using large ensembles on massive data. CoRR, abs/1103.2068 (2011)
3. Berrar, D., Stahl, F., Goncalves Silva, C.S., Rodrigues, J.R., Brito, R.M.M.: Towards data warehousing and mining of protein unfolding simulation data. J. Clin. Monit. Comput. **19**, 307–317 (2005)
4. Bhaduri, K., Das, K., Liu, K., Kargupta, H., Ryan, J.: Distributed data mining bibliography (2008)
5. Bramer, M.A.: Automatic induction of classification rules from examples using N-prism. In: Research and Development in Intelligent Systems XVI, pp. 99–121. Springer, Cambridge (2000)
6. Bramer, M.A.: An information-theoretic approach to the pre-pruning of classification rules. In: Neumann, B., Musen, M., Studer, R. (eds.) Intelligent Information Processing, pp. 201–212. Kluwer Academic, Dordrecht (2002)
7. Brezany, P., Janciak, I., Tjoa, A.M.: GridMiner: An Advanced Support for E-Science Analytics, pp. 37–55. Wiley, New York (2009)
8. Cardona, K., Secretan, J., Georgiopoulos, M., Anagnostopoulos: A grid based system for data mining using mapreduce. Technical report, AMALTHEA TR-2007-02 (2007)
9. Celis, S., Musicant, D.R.: Weka-parallel: machine learning in parallel. Technical report, Carleton College, CS TR (2002)
10. Cendrowska, J.: PRISM: an algorithm for inducing modular rules. Int. J. Man-Mach. Stud. **27**(4), 349–370 (1987)

11. Chambers, L., Tromp, E., Pechenizkiy, M., Gaber, M.: Mobile sentiment analysis. In: Proceedings of the 16th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, September (2012)
12. Chu, C.T., Kim, S.K., Lin, Y.A., Yu, Y., Bradski, G.R., Ng, A.Y., Olukotun, K.: Map-reduce for machine learning on multicore. In: Schölkopf, B., Platt, J.C., Hoffman, T. (eds.) NIPS, pp. 281–288. MIT Press, Cambridge (2006)
13. Clark, P., Niblett, T.: The CN2 induction algorithm. Mach. Learn. **3**(4), 261–283 (1989)
14. Cohen, W.W.: Fast effective rule induction. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 115–123. Morgan Kaufmann, San Mateo (1995)
15. Das, A.S., Datar, M., Garg, A., Rajaram, S.: Google news personalization: scalable online collaborative filtering. In: Proceedings of the 16th International Conference on World Wide Web, WWW '07, pp. 271–280. ACM, New York (2007)
16. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Commun. ACM **51**, 107–113 (2008)
17. Gaber, M.: Data stream mining using granularity-based approach. Found. Comput. Intell. **6**, 47–66 (2009)
18. Gaber, M.: Foundations of adaptive data stream mining for mobile and embedded applications. In: Cairo International Biomedical Engineering Conference. CIBEC 2008, December, pp. 1–6. IEEE, Piscataway (2008). doi:10.1109/CIBEC.2008.4786099
19. Gaber, M.M., Röhm, U., Herink, K.: An analytical study of central and in-network data processing for wireless sensor networks. Inf. Process. Lett. **110**(2), 62–70 (2009)
20. Gaber, M.M., Yu, P.S.: A holistic approach for resource-aware adaptive data stream mining. New Gener. Comput. **25**(1), 95–115 (2006)
21. Gama, J.: Knowledge Discovery from Data Streams. Chapman & Hall/CRC, London (2010)
22. Gantz, J., Reinsel, D.: The digital universe decade, are you ready? IDC **2009**(May), 1–16 (2010)
23. Ghemawat, S., Gobioff, H., Leung, S.-T.: The Google file system. In: Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles, SOSP '03, pp. 29–43. ACM, New York (2003)
24. Globus. The globus toolkit (2012)
25. Hadoop. Hadoop mapreduce (2012). http://hadoop.apache.org/mapreduce/
26. Jian, L., Wang, C., Liu, Y., Liang, S., Yi, W., Shi, Y.: Parallel data mining techniques on graphics processing unit with compute unified device architecture (cuda). J. Supercomput., pp. 1–26. doi:10.1007/s11227-011-0672-7
27. Keim, D.A., Mansmann, F., Schneidewind, J., Ziegler, H.: Challenges in visual data analysis. In: Proceedings of the Conference on Information Visualization, IV '06, pp. 9–16. IEEE Comput. Soc., Washington (2006)
28. Krishnaswamy, S., Gaber, M., Harbach, M., Hugues, C., Sinha, A., Gillick, B., Haghighi, P., Zaslavsky, A.: Open mobile miner: a toolkit for mobile data stream mining. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June (2009)
29. Kumar, A., Kantardzic, M., Madden, S.: Guest editors' introduction: distributed data mining–framework and implementations. IEEE Internet Comput. **10**(4), 15–17 (2006)
30. Liu, T., Rosenberg, C., Rowley, H.A.: Clustering billions of images with large scale nearest neighbor search. In: Proceedings of the Eighth IEEE Workshop on Applications of Computer Vision, WACV '07, p. 28. IEEE Comput. Soc., Washington (2007)
31. Luo, P., Lü, K., Shi, Z., He, Q.: Distributed data mining in grid computing environments. Future Gener. Comput. Syst. **23**(1), 84–91 (2007)
32. Nolle, L., Wong, K.C.P., Hopgood, A.: DARBS: a distributed blackboard system. In: Proceedings of the Twenty-First SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence. Springer, Cambridge (2001)
33. Panda, B., Herbach, J.S., Basu, S., Bayardo, R.J.: Planet: massively parallel learning of tree ensembles with mapreduce. Proc. VLDB Endow. **2**, 1426–1437 (2009)
34. Compare Business Products. The 10 largest data bases in the world (2012)

35. Human Genome Project. Human genome project information (2012)
36. Quinlan, R.J.: Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
37. Rings, T., Caryer, G., Gallop, J.R., Grabowski, J., Kovacikova, T., Schulz, S., Stokes-Rees, I.: Grid and cloud computing: opportunities for integration with the next generation network. J. Grid Comput. **7**(3), 375–393 (2009)
38. SETI@home. About seti@home (2012)
39. Shafer, J., Agrawal, R., Metha, M.: SPRINT: a scalable parallel classifier for data mining. In: Proc. of the 22nd Int'l Conference on Very Large Databases, pp. 544–555. Morgan Kaufmann, San Mateo (1996)
40. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings of the 1996 IEEE Symposium on Visual Languages, VL '96, pp. 336–343. IEEE Comput. Soc., Washington (1996)
41. Sirvastava, A., Han, E., Kumar, V., Singh, V.: Parallel formulations of Decision-Tree classification algorithms. In: Data Mining and Knowledge Discovery, pp. 237–261 (1998)
42. Srinivasan, M.K., Sarukesi, K., Rodrigues, P., Sai Manoj, M., Revathy, P.: State-of-the-art cloud computing security taxonomies: a classification of security challenges in the present cloud computing environment. In: Proceedings of the International Conference on Advances in Computing, Communications and Informatics, ICACCI '12, pp. 470–476. ACM, New York (2012)
43. Stahl, F., Bramer, M.: Scaling up classification rule induction through parallel processing. Knowl. Eng. Rev. doi:10.1017/S0269888912000355 in press
44. Stahl, F., Bramer, M., Adda, M.: Pmcri: A parallel modular classification rule induction framework. In: Machine Learning and Data Mining in Pattern Recognition, pp. 148–162 (2009)
45. Stahl, F., Bramer, M.: Random prism: an alternative to random forests. In: Thirty-First SGAI International Conference on Artificial Intelligence, Cambridge, England, pp. 5–18 (2011)
46. Stahl, F., Bramer, M.: Computationally efficient induction of classification rules with the pmcri and j-pmcri frameworks. In: Knowledge-Based Systems (2012)
47. Stankovski, V., Swain, M., Kravtsov, V., Niessen, T., Wegener, D., Rohm, M., Trnkoczy, J., May, M., Franke, J., Schuster, A., et al.: Digging Deep into the Data Mine with Datamining-grid (2008)
48. Stankovski, V., Swain, M., Kravtsov, V., Niessen, T., Wegener, D., Kindermann, J., Dubitzky, W.: Grid-enabling data mining applications with dataminingrid: an architectural perspective. Future Gener. Comput. Syst. **24**(4), 259–279 (2008)
49. Sloan Digital Sky Survey. The sloan digital sky survey (2012)
50. Swain, M., Silva, C.G., Loureiro-Ferreira, N., Ostropytskyy, V., Brito, J., Riche, O., Stahl, F., Dubitzky, W., Brito, R.M.M.: P-found: grid-enabling distributed repositories of protein folding and unfolding simulations for data mining. Future Gener. Comput. Syst. **26**(3), 424–433 (2010)
51. Szalay, A.: The Evolving Universe. ASSL, vol. 231 (1998)
52. Witten, I.H., Eibe, F.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 2nd edn. Morgan Kaufmann, San Mateo (2005)
53. Wu, G., Li, H., Hu, X., Bi, Y., Zhang, J., Wu, X.: Mrec4.5: C4.5 ensemble classification with mapreduce. In: Fourth ChinaGrid Annual Conference, ChinaGrid '09, pp. 249–255 (2009)
54. Zhao, Q., Sun, J., Yu, C., Xiao, J., Cui, C., Zhang, X.: Improved parallel processing function for high-performance large-scale astronomical cross-matching. Transact. Tianjin Univ. **17**, 62–67 (2011)
55. Zliobaite, I., Bifet, A., Gaber, M., Gabrys, B., Gama, J., Minku, L., Musial, K.: Next challenges for adaptive learning systems. SIGKDD Explorations Newsletter **14**(1) (2012)

Part VI
From Past to Present to Future

Chapter 17
Evolution of Business Intelligence

W.H. Inmon

Abstract From the first simple report to data warehousing to BI tools to today's evolved state of intelligence, BI continues the evolution. Once BI was for structured data only. Now BI can operate on textual data, and in doing so BI can operate on the full spectrum of data found in the corporation.

17.1 Introduction

There are three constants in life—death, taxes, and change. Change occurs in the form of evolution. Some evolutions occur at glacial speeds. The formation of the continents—the grating of tectonic plates that cause earthquakes and volcanic eruptions—are recognizable today, although at very slow speeds. Other evolutions occur much more quickly. The evolution of the automobile began at the turn of the 20th century and continues today. From the black, boxy model T to today's Hybrid or Porsche, the automobile has evolved in front of our very eyes, in memorable history.

17.2 Evolution of the Cell Phone

Evolving even faster has been the evolution of the personal telephone. In the 1950's there was no cell phone. In those early days, people had black dial up phones with wall mounted cords for the most part. There were party lines, exchanges and expensive long distance phone calls. Then in the 1980 timeframe, there first appeared the cell phone. The initial purpose of the cell phone was to call home. Businesspeople on the road now had a means of staying connected to home and headquarters.

In the very early days of cell phones, the cost was high, the coverage was spotty, and the quality of the connection was questionable. But there was a very real marketplace and the evolution of the cell phone began. Over time the cost and size of

W.H. Inmon
Inmon Consulting Services, Castle Rock, CO, USA
url: http://www.inmoncif.com

P. Rausch et al. (eds.), *Business Intelligence and Performance Management*,
Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-4866-1_17, © Springer-Verlag London 2013

cell phones dropped. Over time the coverage of cell phones increased. Over time the quality of the service and the reliability of the connection improved. Today, the cellular industry looks vastly different than the cell phone industry of just a few short years ago. The evolution of the cell phone has come at a whirlwind pace.

But there was another most interesting aspect to the evolution of cell phones and cell phone technology. That interesting aspect was the increase in functionality of the cell phone device. Today calling home is only one rather limited aspect of what cell phones do. Today cell phones:

- can be used as a camera
- can be used as a recording device
- can be used for Instant messaging
- can be used for amusement, with all sorts of games
- can be used for storing and retrieving business information
- can be used for managing a calendar
- can be used for handling alerts
- can be used for many other purposes.

And amazingly all this diverse functionality is found in a single device, in a single place. In truth, handling phone calls is only one small function among many that has evolved over time in the cell phone. And all of this functionality is passed through one sophisticated, electronic device.

17.3 Evolution of Business Intelligence

Now consider the evolution of Business Intelligence (BI). The origins of BI go back to the humble application report, generated in the days of COBOL and assembler processing. The application report purported to tell the organization what was happening in the environment. While the humble COBOL or assembler report served a very real purpose, there were many problems with application reports. Application reports chewed up a lot of paper. Application reports took a long time to run. Application reports were usually out of date by the time they were printed. But application reports were a start.

Soon there were online transactions which also told the organization what was occurring, but in real time. Online transactions took no paper. Online transactions showed data that was up to date as of the moment of access. And online transactions were fast. But online transactions had their limitations, too. Online transactions were not good for showing large amounts of data. Online transactions did not leave an auditable trail. And the data behind online transactions often changed by the second. A person could do an online transaction only to have the information invalidated in the next second. And online transaction processing systems were expensive and fragile.

But the real problem behind online transactions was that online transactions showed only a limited type of data and showed data that was unique to an application. Given enough online systems, it was possible to find the same piece of data with multiple different values coming from multiple different applications. User A looks for some information and finds it. User B looks for the same information in another system, and finds it. However the values shown to user A are not the values shown to user B.

The value of integrating data into a corporate format began to be obvious. Decision making in a world where there was no definitive source of data became a challenging and dicey exercise. For large corporations, it became obvious that a historical, integrated, subject oriented source of data was needed for the corporation. Having multiple overlapping applications simply was not a sound basis for business decisions. Thus born was the data warehouse. With the data warehouse it was now possible to do a whole new style of reporting and analysis. Once there was definitive corporate data in a data warehouse, the world of BI began, at least as we know BI today.

17.4 Enter Business Intelligence

In the early days of data warehouse and BI, doing simple analysis was a breath of fresh air compared to the world that existed before data warehouse and BI. Simple reporting from data warehouses was the first infant step of BI. But people quickly discovered that there were many other possibilities for BI. Many other functions belonged to the purview of BI.

Some of the newly discovered functions included:

- graphical visualization of results
- looking at information over time
- looking at very large volumes of data
- doing statistical analysis of data
- looking at small subsets of data in a personalized fashion
- using spreadsheets to analyze data
- people building customized collections of information to suit their individual needs called data marts
- collecting information about what analysis has already been done
- collecting metadata so that an analyst knows where to begin in doing an analysis
- transforming data so that different sources of data can become integrated, and so forth.

Indeed once BI became a reality, all sorts of activities occurred. Unfortunately there was no coordination of these BI activities. One department would run a report. Another department would do a projection. One individual would create a spreadsheet. Another analyst would create metadata by building a data set. In many ways, the world of BI resembled the California gold rush. One individual or one organization worked in their own self interest with no concern for the work being done by others. The California gold rush and BI resembled a colony of bees in the springtime. There was little or no apparent coordination of people working in tandem with each other.

The activities of BI soon encompassed lots of different types of technology. There certainly was reporting. There was graphical software. There was ETL. There were spreadsheets. There was statistical processing, and so forth. And sitting on top of this beehive of activity was no organization or coordination of the different activities of BI.

17.5 The Evolution to Textual ETL

But the evolution of BI is ever evolving. There is another important aspect of the evolution of BI, and that aspect is the evolution to unstructured, textual data. Consider this—most organizations build data bases based entirely on the basis of what is termed structured data. Structured data refers to data that occurs repetitively. Consider banking transactions. For the most part all banking transactions are the same, insofar as the processing that occurs is concerned. The only real difference between one banking transaction is the date, the account number, the amount of the transaction, and the parties involved in the transaction. Other than those differences, there really isn't any difference between one banking transaction and the next. And transaction processing occurs everywhere. Airlines do transactions. Retailing does transactions. Insurance does transactions. Manufacturing does transactions, and so forth. In one way or the other, all businesses do transactions as a part of their day to day processing.

Modern data base management systems (dbms) are geared to handle repetitive transactions. Dbms lay out data and the repeated occurrences of data that are generated by a transaction are recorded by a record or a row created by the dbms. Dbms are designed to handle efficiently many, many transactions and many transaction types.

It comes as a surprise to many people that in most corporations the majority of data in the corporation is not repetitive transaction based data. It is estimated that approximately 80 % of the data in the corporation is unstructured textual data, not transaction based, repetitive data. (This is a surprise to the IT professional who has spent his/her entire life working with transaction based data.) So where is this unstructured textual data found? It is found in many places. Some of those typical places are:

- in email
- in contracts
- in human resource files
- in warranties
- in chat log/help log sessions
- in medical records
- in loan applications
- in customer responses, and so forth.

In short unstructured, textual data is found everywhere. It is—in a word—pervasive.

17.6 Textual Data Has Great Business Value

And there is much textual data that is very important to the corporation. Corporate contracts contain a wealth of valuable information to the corporation. Corporate contracts represent legal obligations either to the corporation or by the corporation. Chat log sessions represent the direct interface with the customer. In addition, chat log sessions contain invaluable information about products, services and defects or customer complaints. Medical information contains huge amounts of information about disease, treatments, therapies, medications, and so forth. Warranty claims contain important information about the quality of parts and products. Insurance claims contain important information about fraud. And the list goes on. In fact there is probably more important information about the corporation found in text than there is in transaction based information. Yet, because textual information is not repetitive, it does not fit well with dbms, and as a consequence is not used in the decision making of the corporation.

BI continues its evolution and an important part of the evolution is the ability to start to read textual data, transform it, and move the text into a standard relational data base. Once in a standard relational data base, text is able to be analyzed like any other source of data.

17.7 Integrating Text

The process of reading and analyzing text is called the process of "integrating text". Unlike a search engine, textual integration starts with the assumption that the raw text needs to be changed. Search engines and data mining processes make the basic assumption that raw text should either be changed not at all or should be changed only very lightly. Textual integration on the other hand starts with the assumption that raw text needs to be heavily changed before it is fit to be placed into a data base.

The process of integration is done by (patented) technology called "textual ETL" such as that sold and supported by Forest Rim Technology. Textual ETL reads raw text, does the integration, and produces the output into a standard relational data base such as Oracle, Teradata, DB2/UDB, SQL Server, or other dbms. Once the data has been placed in other BI technology, standard analytical processing can be done against the text.

17.8 Some Differences

There are some fundamental differences between processing raw text and processing repetitive transaction data. One of those differences is the ongoing nature of processing. Repetitive transactions require constant and ongoing processing. As long as the bank is doing transactions, those activities generate data that must be placed in a data warehouse. But much of textual ETL processing against text is of a one time nature. Once a contract is completely and thoroughly processed, there is no need to go back and reprocess it. Of course if the contract is updated, reprocessing is necessary. But if the contract is updated, it is in fact a new contract. As a rule people don't update or change documents. People create new documents, but it is not normal for a document to be updated. If an email says something incorrectly, a new email is written. If a contract specifies something improperly, a new contract is written. If a chat log session discusses information incorrectly, a new conversation is held.

Because updates of text are not the norm, there is no need to constantly process and reprocess the same document.

Another difference in transaction processing and document processing is the volume of data that must be processed. While there certainly are environments that must process large amounts of transactions, the transactions enter the system on a finite basis. There may be lots of transactions, but as a rule the transactions are relatively small.

When it comes to processing documents, there may be MASSIVE volumes of data. Take loan application portfolios, for example. A given loan may be up to 250 pages in length. The portfolio may contain 2,000 or more loans, and there may be many, many portfolios. In all, there may be multiple petabytes of text associated with loan applications to be processed.

Another factor when dealing with the textual ETL processing of text is the fact that text comes in many different forms. Classical text is in the form of proper English grammar. This paper (hopefully!) is written in proper English grammar. There are verbs, adverbs, nouns, adjectives, pronouns and so forth. Words are spelled properly. There are periods, question marks, and exclamation points. There is a proper structuring of the words in the sentence. Some parts of the text are capitalized. Those are all the earmarks of proper English sentences. And proper English sentences should certainly be able to be processed by textual ETL.

But text comes in other forms as well. There are doctor's notes. Doctor's notes have their own shorthand—both in terms of spelling and in terms of structure. "HA" may mean "headache". Bp 120/67 may mean "blood pressure" of 120 systolic over 67 diastolic. Doctors just don't have the time or take the time to write proper grammar. And as long as the information is for the doctor's personal usage, shorthand and comments are just fine.

Teenagers write in "IM" or instant messaging. "THX" may mean "thanks". "2" may mean "to". "U" may mean "you". And textual ETL needs to be able to handle ALL forms of text.

In addition, occasionally textual ETL comes in the form of different languages. An international banking organization may do business in Spanish, English, and Russian. But the data base analyst must have all of the data base written in the same language for the purpose of doing analytical processing.

There are then some major challenges in the handling of text. Text is not text. Text takes many different forms and textual ETL needs to be able to handle all of those forms.

One of the real values of textual ETL and the integration and assimilation of text into a data base is that—once committed to a data base—the textual data can be queried and analyzed along side of classical structured data. The ability to read and analyze both structured and unstructured data together is powerful. Entirely new kinds of analysis can be done that simply are not possible when there is only structured data that can be analyzed.

17.9 Summary

From the first simple report to data warehousing to BI tools to today's evolved state of intelligence, BI continues the evolution. Once BI was for structured data only. Now BI can operate on textual data, and in doing so BI can operate on the full spectrum of data found in the corporation.