Studies in Computational Intelligence 579

Maria Mach-Król
Celina M. Olszak
Tomasz Pełech-Pilichowski *Editors*

Advances in ICT for Business, Industry and Public Sector

Springer

Studies in Computational Intelligence

Volume 579

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Studies in Computational Intelligence" (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

More information about this series at http://www.springer.com/series/7092

Maria Mach-Król · Celina M. Olszak
Tomasz Pełech-Pilichowski
Editors

Advances in ICT for Business, Industry and Public Sector

Springer

Editors
Maria Mach-Król
Celina M. Olszak
Katowice University of Economics
Katowice
Poland

Tomasz Pełech-Pilichowski
AGH University of Mining and Metallurgy
Kraków
Poland

ISSN 1860-949X ISSN 1860-9503 (electronic)
ISBN 978-3-319-11327-2 ISBN 978-3-319-11328-9 (eBook)
DOI 10.1007/978-3-319-11328-9

Library of Congress Control Number: 2014952192

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

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

Preface

Information and Communication Technologies are widely used in business, industry, and public sector. They support core business, enable control of equipment, provide information or knowledge for decision makers, and they allow the creation/design of new software and hardware solutions. Ubiquitous Internet has forced the development of new services and it is still an inspiration to propose new tools, approaches, and paradigms.

The development of the information society (the knowledge era) is directly related to the growing importance of information. Large datasets processing give a possibility of obtaining relevant information and knowledge can be acquired from data. Thus, data processing, analyses, visualization are very important, practically essential for organizations, therefore decision makers.

Researchers' and business sector's interest in solutions for advanced processing and acquisition of information to be able to process knowledge. To this aim intelligent systems and software/hardware solutions are to be employed. High potential for Business ICT have, in particular, Business Intelligence, reasoning systems, knowledge management, advanced signal/data processing, text mining techniques (including processing data available on the web sites/services and content processing), and big data—including new analyzing and visualization algorithms.

This contributed volume is a result of many valuable discussions held at ABICT'13 (4th International Workshop on Advances in Business ICT) in Krakow, September 8–11, 2013.

The workshop focused on Advances in Business ICT approached from a multidisciplinary perspective. It provided an international forum for scientists/experts from academia and industry to discuss and exchange current results, applications, new ideas of ongoing research, and experience on all aspects of Business Intelligence and big data. ABICT has also been an opportunity to demonstrate different ideas and tools for developing and supporting organizational creativity, as well as advances in decision support systems.

This book is an interesting resource for researchers, analysts, and IT professionals including software designers. The book comprises 11 chapters. Authors present research results on business analytics in organization, business processes modeling, problems with processing big data, nonlinear time structures, and nonlinear time ontology application, simulation profiling, signal processing (including change detection problems), text processing and risk analysis.

Maria Mach-Król
Celina M. Olszak
Tomasz Pełech-Pilichowski

Contents

A Multi-level Hierarchical Approach for Configuring Business Processes

Mateusz Baran, Krzysztof Kluza, Grzegorz J. Nalepa and Antoni Ligęza

Abstract Business Process configuration constitutes a powerful tool for expressing similarities between different Business Process models. Such models in the case of real life systems are often very complex. Configuration gives the opportunity to keep different models in a single configurable model. Another approach to manage model complexity is hierarchization, which allows for encapsulating process details into sub-levels, helps to avoid inconsistencies and fosters reuse of similar parts of models. In this paper, we present an approach for configuring Business Processes that is based on hierarchization. Our approach takes advantage of the arbitrary *n*-to-*m* relationships between tasks in the merged processes. It preserves similar abstraction level of subprocesses in a hierarchy and allows a user to grasp the high-level flow of the merged processes. We also describe how to extend the approach to support multi-level hierarchization.

1 Introduction

Business Processes (BP) define the way a company works by describing control flow between tasks. Design and development of such processes, especially more and more complex ones, require advanced methods and tools.

M. Baran · K. Kluza (✉) · G.J. Nalepa · A. Ligęza
AGH University of Science and Technology, Al. A. Mickiewicza 30,
30-059 Krakow, Poland
e-mail: kluza@agh.edu.pl

M. Baran
Cracow University of Technology, Ul. Warszawska 24,
31-155 Krakow, Poland
e-mail: matb@agh.edu.pl

G.J. Nalepa
e-mail: gjn@agh.edu.pl

A.Ligęza
e-mail: ligeza@agh.edu.pl

M. Mach-Król et al. (eds.), *Advances in ICT for Business, Industry and Public Sector*,
Studies in Computational Intelligence 579, DOI 10.1007/978-3-319-11328-9_1

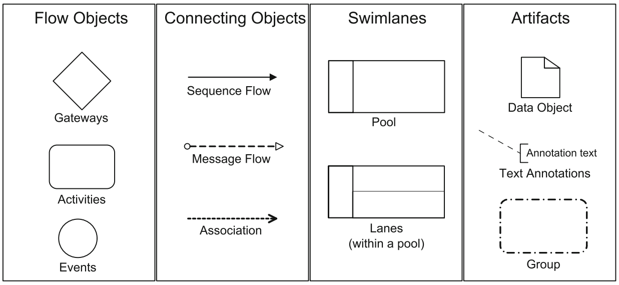


Fig. 1 BPMN core objects

Business Process Model and Notation (BPMN) [1] provides a visual language for modeling Business Processes. It consists of graphical elements denoting such constructs as activities, splits and joins, events etc. (see Fig. 1), which can be connected using control flow. Thus, a BPMN model provides a visual description of process logic [2]. Such a model is easier to understand than textual description and helps to manage software complexity [3].

Although the BPMN notation is very rich when considering the number of elements and possible constructs, apart from the notation rules, some style directions for modelers are often used [4]. To deal with the actual BP complexity, analysts use various modularization techniques. Mostly, they benefit from their experience and modularize processes manually during the design because these techniques are not standardized. Modularization issue is important in the case of understandability of models [5]. Thus, guidelines for analysts, such as [6], emphasize the role of using a limited number of elements and decomposing a process model.

In the case of large collection of processes [7], the models in the collection can be similar [8], but this similarity is lost when the models are kept separately.

Although modelers can take advantage of modularization techniques and design process models using some similar structures (especially using the same subprocesses), mostly process models are modeled by different analysts, and thus modularized in different ways on distinct granularity level [5].

Automatic hierarchization and configuration can help in preventing these problems. Hierarchization supports modeling at different abstraction levels and, properly developed, can ensure the same way of modularization for all the processes. Configuration, in turn, gives the opportunity of unification of processes and enables to keep different models in one configurable model.

This paper is an extended version of the paper [9] presented at the 4th International Workshop on Advances in Business ICT (ABICT 2013) workshop[1] held in conjunction with the FedCSIS 2013 conference in Krakow. The paper gives an overview of configuration methods describing various aspects of configuring and adapting BP models. As the original contribution, we proposed an extension of the hierarchical

[1] See: http://fedcsis.org/2013/abict.html.

approach for configuring business processes in order to support multi-level hierarchical configuration. Moreover, we compared our method to the existing configuration and adaptation methods used in BP modeling.

The rest of this paper is organized as follows: Sect. 2 presents motivation for our research. Sections 3 and 4 describe related works in the hierarchization and configuration areas. In Sect. 5, we present our configuration BP approach which takes advantage of hierarchization. We evaluated the proposed approach based on the issue tracker case study. The paper is summarized in Sect. 6.

2 Motivation

There are several challenges in Business Process modeling, such as:

- the *granularity modularization* challenge—how to model different processes similarly, especially in respect of abstraction layers [10].
- the *similarity capturing* challenge—it is not easy to grasp similarities in the collection of only partially similar models [11], especially if they are modularized differently.
- the *collection storing* challenge—how to store a large collection of models in some optimized way [12].

In our research, we address the first challenge by using automatic hierarchization that allows us to preserve similar abstraction level of subprocesses in a hierarchy. To deal with the two other challenges, we propose a new BP configuration technique, that allows us to express similarities between different BP models in a simple, but comprehensive way. It is worth noting that our configuration is based on the hierarchical model prepared previously.

The aim of this paper is to present and evaluate our approach for configuring Business Processes based on hierarchization. We present the automatic hierarchization algorithm that takes advantage of task taxonomy and the algorithm for configuration. In order to compare our solution with other ones, an overview of configuration methods describing various aspects of configuration and adaptation in BP models is given. Moreover, we present how to extend our approach to support multi-level hierarchization.

The proposed approach has several advantages. Thanks to the use of hierarchization, the obtained model structure incorporates similar activities, and the configuration step can be simplified. Thus, the configurable process diagram is easy to comprehend. Contrary to the existing configuration methods, in our approach we can bind not only one task with another, but also groups of tasks which were considered in hierarchization step. Such a configuration technique addresses the challenges mentioned above by managing both process model complexity and diversity.

3 Hierarchization Issues in Business Process Models

La Rosa et al. [13] distinguished 3 abstract syntax modifications that are related to modularization for managing process model complexity. These are:

1. vertical modularization—a pattern for decomposing a model into vertical modules, i.e. subprocesses, according to a hierarchical structure
2. horizontal modularization—a pattern for partitioning a model into peer modules, i.e. breaking down a model into smaller and more easily manageable parts, especially assigned to different users in order to facilitate collaboration.
3. orthogonal modularization—a pattern for decomposing a model along the cross-cutting concerns of the modeling domain, such as security or privacy, which are scattered across several model elements or modules.

Our approach is consistent with the vertical and horizontal patterns. Although we decompose a model into subprocesses, in fact we use some additional information to decompose it, such as task assignment or task categories, which is an example of the second pattern instance. The last one, orthogonal pattern, requires to extend the notation, as in [14]; thus, it is not our case.

It is important to notice that such decomposition has several advantages, i.e.:

- it increases their understandability of models by "hiding" process details into sub-levels [5],
- it decreases redundancy, helps avoid inconsistencies and fosters reuse by referring to a subprocess from several places [11, 15–17],
- it decreases the error possibility [18],
- it increases maintainability of such processes [6].

4 Business Process Configuration

Business Process configuration is a tool for expressing similarities between different Business Process models. There are mechanisms for managing and comparing processes in large repositories [19, 20], refactoring of such repositories [12] as well as automatic extraction methods for cloned fragments in such process model repositories [12, 16] in order to capture them as globally available subprocesses.

A comprehensive overview of a process of process model configuration can be observed in Fig. 2. Although it is based on Synergia [21], an open-source toolset providing end-to-end support for process model configuration, it can be also interpreted as a general workflow of process model configuration. In this approach, first domain experts create questionnaire models and modelers create process models. Based on domain configuration, the configurable process model is generated. Finally, the model can be individualized.

There are a few methods of extraction of configurable processes focusing on different goals. Analyzing such configurable Business Processes reveals high-level

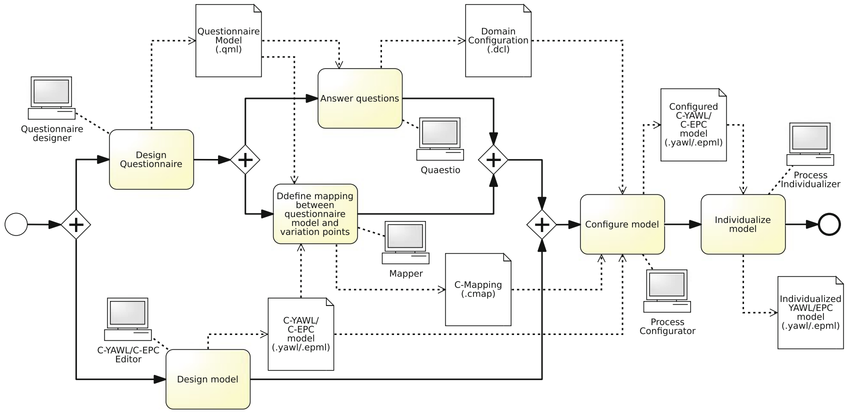

Questionnaire Model (.qml)
Domain Configuration (.dcl)
Questionnaire designer
Design Questionnaire
Answer questions
Quaestio
Configured C-YAWL/ C-EPC model (.yawl/.epml)
Process Individualizer
Ddefine mapping between questionnaire model and variation points
Mapper
Configure model
Individualize model
C-Mapping (.cmap)
Process Configurator
Individualized YAWL/EPC model (.yawl/.epml)
C-YAWL/C-EPC Editor
C-YAWL/ C-EPC model (.yawl/.epml)
Design model

Fig. 2 Configuration process (based on the Synergia toolset [21])

workflow that might not be apparent in particular models. In some cases, the structure of the processes can be partially lost. In our approach, we deal with the case: models merged into a configurable model [23]. Such models allow the analyst to notice a number of processes as special cases of one configurable model. This solution preserves all the details of the original models and at the same time emphasizes similarities between them.

A simple example of merging four models into a configurable model is presented in Fig. 3. Based on the four similar processes, composed of the particular BPMN

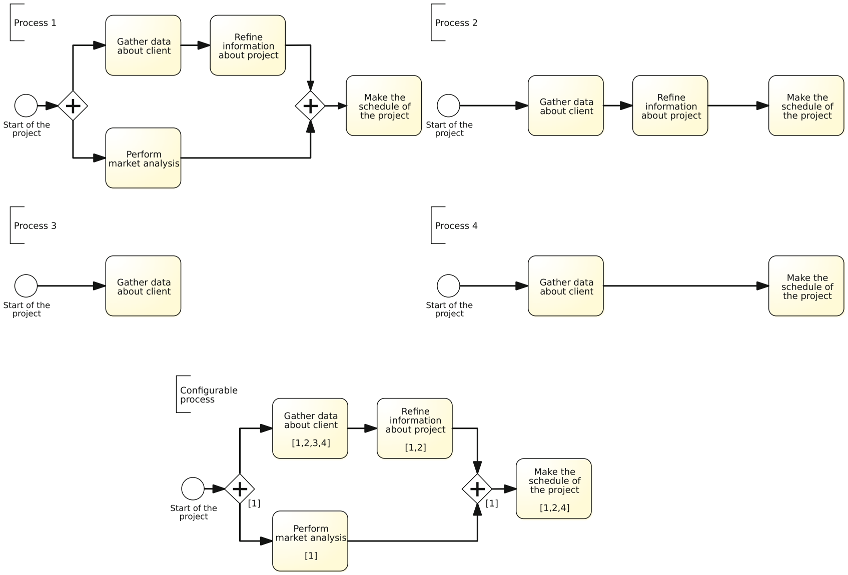

Process 1
Gather data about client
Refine information about project
Make the schedule of the project
Start of the project
Perform market analysis
Process 2
Start of the project
Gather data about client
Refine information about project
Make the schedule of the project
Process 3
Start of the project
Gather data about client
Process 4
Start of the project
Gather data about client
Make the schedule of the project
Configurable process
Gather data about client [1,2,3,4]
Refine information about project [1,2]
Start of the project
[1]
[1]
Make the schedule of the project [1,2,4]
Perform market analysis [1]

Fig. 3 An example of Business Process configuration [22]

elements that are either present in a model or not, a configurable process model is created. It generalizes all the four model parts into one configurable model. In order to obtain an individualized model from such a configurable model, its interpretation is as follows. The numbers in square brackets indicate in which process model a given element appears. If an element appears in all of the original models, it does not have to be marked with the numbers in square brackets. The numbers of models in Fig. 3 are presented in element labels, however for diagram readability they can be hidden in the *documentation* attributes of the BPMN elements. Thus, to obtain the particular individualized model, a user has to select the model number and all the elements with this number or without any number belongs to such an individualized model. If a particular element does not appear in the model, the control flow goes on to the next element.

Although in comparison to [24], it is a simplified approach, for some cases it is sufficiently expressive to use, e.g. such an approach has been successfully applied in [22] for recommending process model elements during modeling.

The area of configurable Business Processes is an active research field. In [25], Rosemann et al. describe an approach focused on hand-made diagrams for the purpose of reference modeling. Configurable process models constitute a good alternative to current reference model bases such as SAP. Instead of presenting the analyst a few exemplary models, a more general, configurable solution can be delivered. It makes producing final models faster and less error-prone [24]. Variant-rich process models were explored in PESOA project [26, 27] and by Tealeb et al. [28]. They enable process designer to specify a few variants of a task.

Our hierarchical approach for configuring Business Processes [9] is a specialized version of solution proposed by La Rosa in [29]. The hierarchization algorithm produces models with very specific structure and this fact is exploited in our approach. Our case differs from the existing approaches, because we do not take advantage of any directly visible similarity, but on previously defined taxonomy of states or roles in the processes etc. Moreover, the proposed hierarchization algorithm forces the generation of similar models as a result.

Döhring et al. [30] analyzed various methods of business process configuration by comparing it with the more general technique of process adaptation. In Table 1, the set of configuration and adaptation methods are compared according the following features [30]:

- *Control flow construction*—A (adaptation) or C (configuration). In the case of configuration, a configurable model is a superset of all models and a variant is obtained by removing elements not belonging to the case; in the case of adaptation the model is obtained by minimizing the number of operations (such as adding, removing or modifying the model) needed to produce the individual variants.
- *Modularization support*—modularization (or hierarchization) of the process model can occur at many levels. S—(single-element-based) indicates that the approach allows to define variability at a single element level; F—(fragment-based) indicates applying variability mechanism to multiple elements at once.

Table 1 Comparison of business process adaptation and configuration methods (based on [30])

Approach	Control flow construction	Modularization support	Runtime variant construction	Data flow variability	Resource perspective variability
ABIS [31]	A	F	–	–	–
AgentWork [32]	A	F	+	+	–
AO4BPMN [33]	A	F	–	–	–
C-EPC [24, 25]	C	S	–	+	+
C-MPSG [34]	C	F	–	–	–
C-YAWL [35]	C	S	–	–	–
Designby selection [36]	A	S	–	+	–
Multi-perspective variants [37, 38]	C	S	–	+	+
PESOA [27]	A	S	+	+	–
PROVOP [39]	A	F	+	–	–
VBPMN [40]	A	F	+	+	–
Our approach [9]	C	S	–	–	–

- *Runtime variant construction*—refers to the ability of changing the process during its execution in accordance with the general model.
- *Data flow and resource perspective variability*—a configurable process model can contain not only the control flow elements that are taken into account during configuration, but also data flow elements or resources.

5 Hierarchization-Based BP Configuration Approach

In this section we describe an approach for configuring Business Processes that relies on hierarchization for more expressive power and simplicity. The approach takes advantage of arbitrary n-to-m relationships between tasks in merged processes. These relationships are defined using taxonomy of tasks what makes the approach more flexible.

As a result of hierarchization, we obtain a simple model of a very specific type. Such models do not require general configuration algorithms that often produce complex and hard to analyze diagrams. A simple configuration algorithm that uses the hierarchical models is sufficient and is described in Sect. 5.2.

The general process of our approach is presented in Fig. 4.

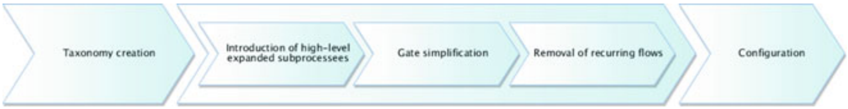

Taxonomy creation
Introduction of high-level expanded subprocessees
Gate simplification
Removal of recurring flows
Configuration

Fig. 4 General process of our configuration approach

5.1 Automatic Hierarchization Algorithm

Goal: To generate a hierarchical BP model based on the previously defined arbitrary n-to-m relationships between tasks.

Input:

- a BPMN model,
- a set of high-level BPMN tasks and an assignment of BPMN model's tasks to the high-level tasks (this can be achieved using a taxonomy).

Output: Two-level hierarchical diagram. The lower level contains one diagram for each high-level task. Higher level diagram contains high level tasks (with lower-level diagram as subprocesses). This is done in such way to maximize simplicity and preserve semantics of original model.

Algorithm steps:

1. **Introduction of high-level expanded subprocesses**
   In the first step, expanded subprocesses are introduced. Tasks are assigned to them according to given specification. Gateways are placed outside of all subprocesses unless all their incoming and outgoing flows lead to tasks of the same process. Intra-subprocess flows are kept. Inter-subprocess flows are replaced by:

   a. flow from source element to end event (in subprocess),
   b. OR-gateway right after subprocess (one per subprocess),
   c. flow from introduced gateway to target of initial flow with condition 'subprocess ended in event introduced in 1a'.

   This step is depicted in Fig. 5. After this step is performed, assumption 1 of configuration algorithm (Sect. 5.2) is fulfilled.
2. **Gateway simplification**
   The previous step introduced new gateways. The original diagram may contain unnecessary gateways too. Creating configurable diagram in proposed approach requires a very specific structure of high-level model. It can be achieved through gateway simplification.
   The process of gateway simplification is depicted in Fig. 6. In a simplified model one gateway G is placed after every subprocess $S(G)$ (unless it has only one outgoing flow which does not end in a gateway). Gateway G has outgoing flows to all

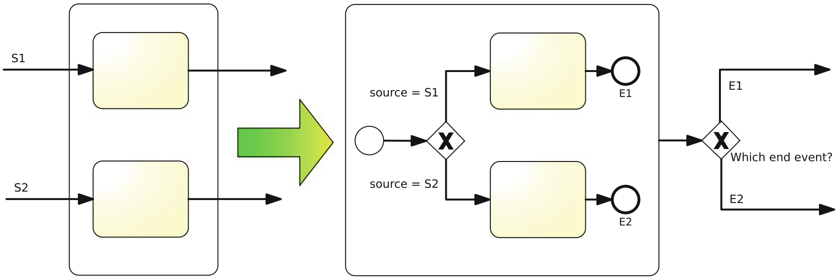


Fig. 5 First step of hierarchization algorithm

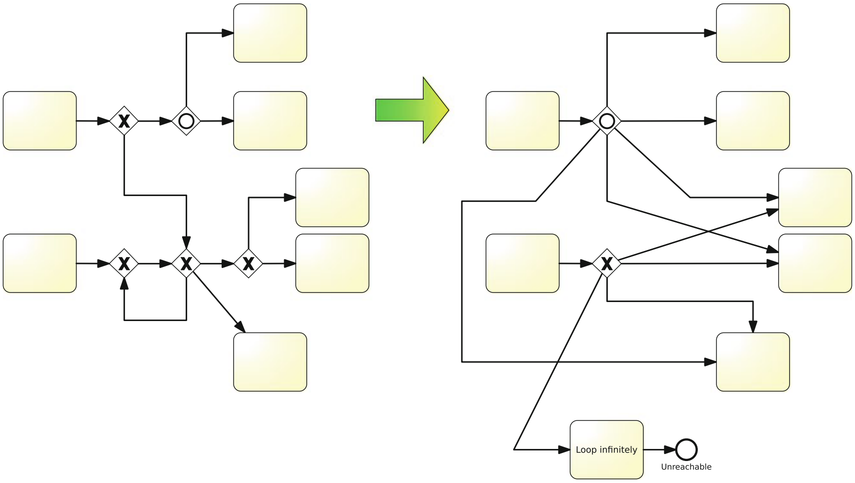


Fig. 6 Second step of hierarchization algorithm (gateway simplification)

subprocesses and end events reachable in original model from $S(G)$. Conditions labeling these flows are determined as follows.

Let $flow_N(G, T)$ and $flow_O(G, T)$ be the flows in the new and old graphs respectively from gateway G to target item T. Let $C_N(G, T)$ and $C_O(G, T)$ be the conditions on flow $flow_N(G, T)$ and $flow_O(G, T)$ respectively. Let $P(G, T)$ be the set of all paths in old graph (all gateways appear at most once) from G to P. Let $L(G)$ be the set of loops in gateway graph reachable from gateway G. Then the following hold:

$$\begin{aligned} &all((G_1, G_2, \ldots, G_k), T) \text{ holds iff } C_O(G_k, T) \text{ and} \\ &\quad \text{for all } i \in \{1, 2, \ldots, k-1\} \text{ such that } C_O(G_i, G_{i+1}) \\ &C_N(G, T) \text{ holds iff exists } p \in P(G, T) \text{ such that } all(p, T) \end{aligned}$$

Presented procedure works as long as graph of gateways is acyclic. Cycles need additional compensation for the fact that infinite looping is possible. We propose a solution where a new task “loop infinitely” is added and connected by flow from all gateways that allow looping. Condition on the new flow may be defined by analogy to the previous case:

$$\begin{aligned} &all((G_1, G_2, \ldots, G_k)) \text{ holds iff } C_O(G_k, G_1) \text{ and} \\ &\quad \text{for all } i \in \{1, 2, \ldots, k-1\} \text{ such that } C_O(G_i, G_{i+1}) \\ &C_N(G, T) \text{ holds iff exists } p \in L(G) \text{ such that } all(p) \end{aligned}$$

Figure 6 shows a graph of gateways before and after simplification. Simplification assures that requirements 2 and 3 from Sect. 5.2 are fulfilled.

3. **Removal of recurring flows**
   Last step of hierarchization is elimination of recurring flows. By this a flow from a gateway to activity preceding this gateway is meant (see Fig. 7).

Before each end event in the subprocess that can result in recurring flow a new XOR gateway is placed. It has two output flows: one to end event and one to task or gateway a recurring flow would lead to (see Fig. 7). The condition on the latter flow is created according to condition on original recurring flow. This step of hierarchization algorithm makes requirement 4 of configuration process fulfilled.

5.2 Process Configuration

Goal: To generate a configurable BP model based on the similar BP models.
Input: Similar BP models.
Output: A configurable BP model.

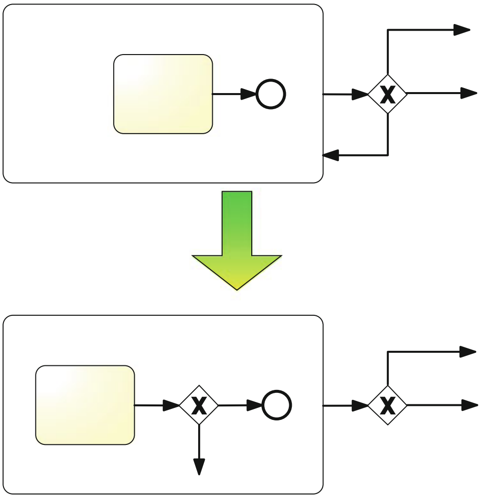


Fig. 7 The idea of third step of hierarchization

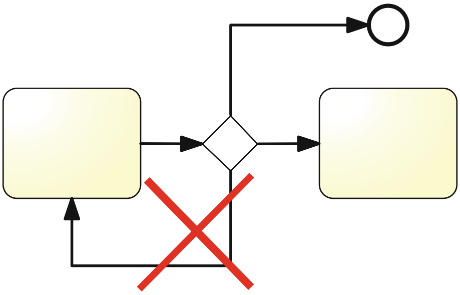


Fig. 8 Possible flows to and from a gateway

Algorithm:

Let us be given N BPMN models such that:

1. all of them share the same set of tasks,
2. flows outgoing from tasks or start event end in a different task, end event or an (XOR or OR) gateway (Fig. 8),
3. flows outgoing from gateways always end in tasks or an end event,
4. no flow outgoing from a gateway leads to a task that has a flow to this gateway.

Then the configurable model that entails all the given models can be defined as follows:

1. configurable diagram has one start event, all the specified tasks and all the end events from N given models,
2. for all tasks and start event (let i be the current item):
    a. If all diagrams have flow outgoing from i that ends in (the same) task or the only end event then the same flow exists in merged diagram.
    b. If in at least one model the flow f ends in a gateway, the merged model has a configurable gateway after i. It is a configurable type gateway if there are diagrams with two different types of gateways (or one without gateway).
    c. If and only if any input diagram gateway after i has a flow to an item, the configurable gateway has a flow to this item too. The flows are labeled with model number and condition from that model.

5.3 *Approach Evaluation*

To present our hierarchization and configuration approach, we chose 3 different BPMN models of bug tracking systems: Django,[2] JIRA[3] and the model of the issue tracking approach in VersionOne.[4]

[2] See: https://code.djangoproject.com/.

[3] See: http://www.atlassian.com/software/jira/.

[4] See: http://www.versionone.com/.

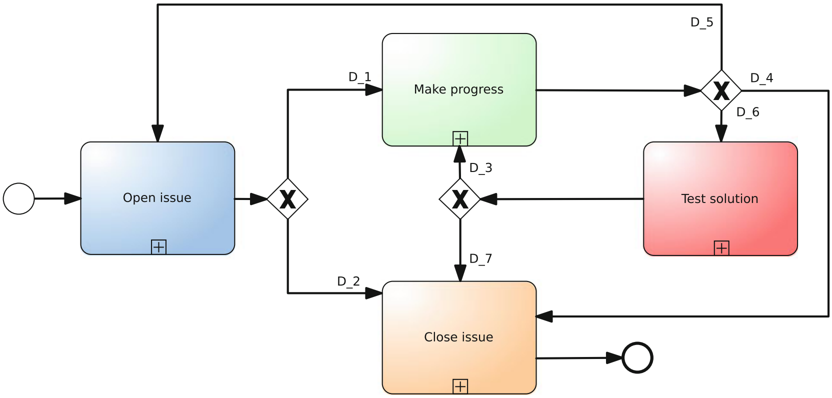


Fig. 9 High-level diagram of Django issue tracking

A bug tracking system is a software application which helps in tracking and documenting the reported software bugs (or other software issues in a more general case). Such systems are often integrated with other software project management applications, such as in VersionOne, because they are valuable for the company. Thus, apart from popularity, we selected such a case study because these kinds of processes have similar users assigned to similar kinds of tasks, the processes of different bug trackers present the existing variability, and such an example can be easily used to present our algorithm in a comprehensive way.

We tested our approach on the three issue tracking systems. The result of hierarchization of Django issue tracking process can be seen in Fig. 9. The obtained model is simple and comprehensible. Figure 10 compares initial and hierarchical versions of Django system. The three hierarchized models, simplified by the algorithm, can be simultaneously compared on high level and on the subprocess level. The final high level configurable model is presented in Fig. 11.

One of the drawbacks of our approach is that conditions on control flows outgoing from gateways may become complex after hierarchization. However, it is not an obstacle in understanding of high level flow in the process, which is the goal of the approach.

5.4 Multi-level Hierarchical Configuration

The described two-level configuration algorithm can be easily extended to allow more levels. The main change is the necessary extension of the first step of the algorithm. Next steps are then applied recursively to all levels of the diagram.

Firstly, a more general, tree-like, hierarchy of processes is needed. For each process we need to know its “super-process”—the process for which it is a subprocess, or that it is a top-level process. The hierarchy can be arbitrarily deep.

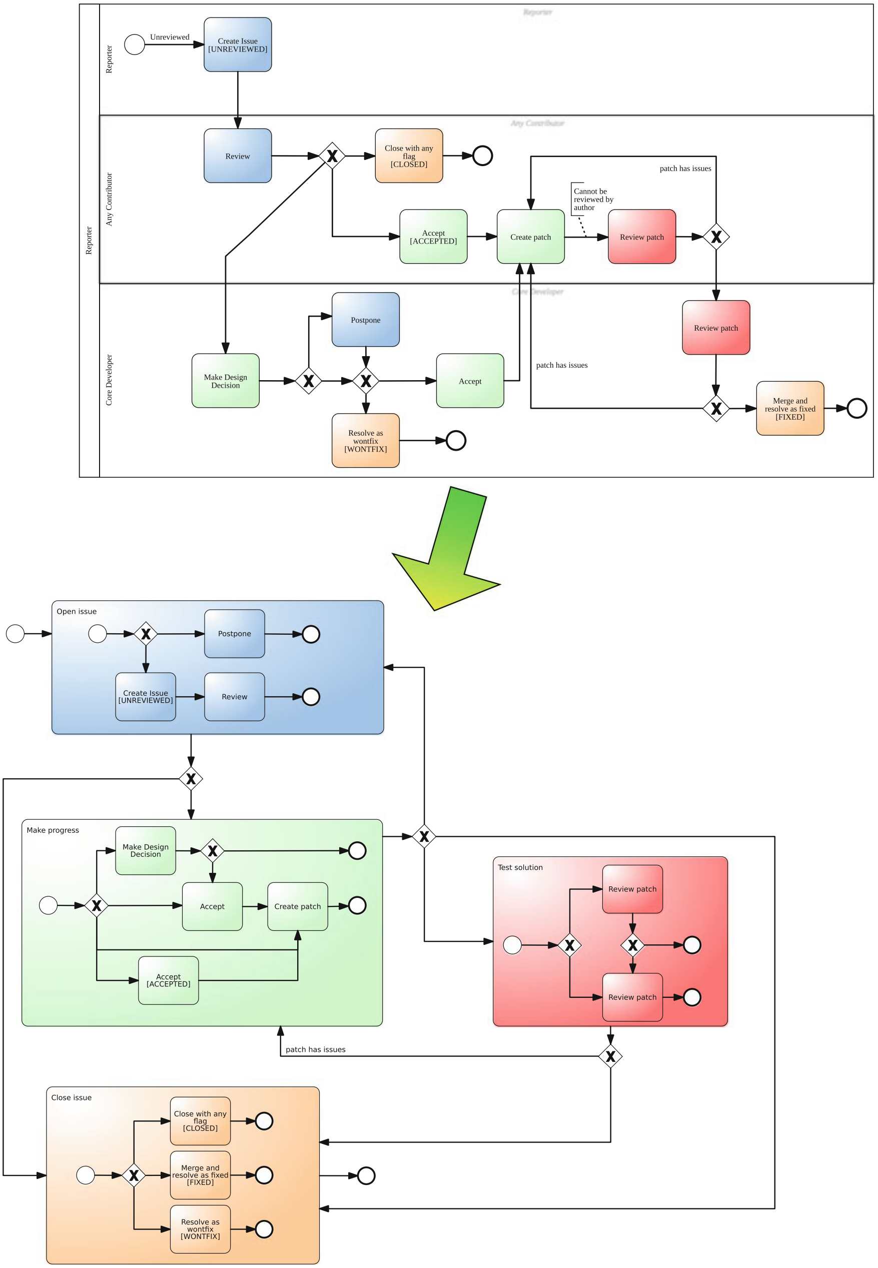


Fig. 10 Comparison of initial diagram and its hierarchical version

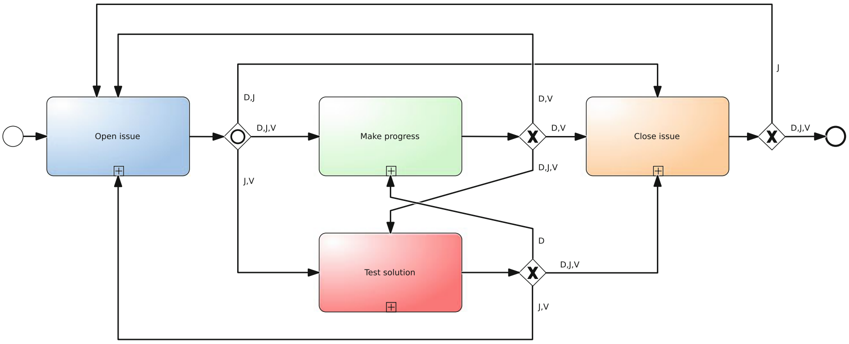


Fig. 11 Result of the proposed algorithm (after configuration)

For example, let us divide process "Open issue" into two processes: "Open issue for the first time" and "Reopen issue". In case of diagram in the Fig. 10 we would have the following hierarchy:

```
- Open issue
  - Open issue for the first time
    - Create issue
    - Review
 - Reopen issue
    - Postpone
- Make progress
  - Make design decision
  - Accept
  - Create patch
  - Accpet [ACCEPTED]
- Test solution
  - Review patch (Any Contributor)
  - Review patch (Core Developer)
- Close issue
  - Close with any flag [CLOSED]
  - Merge and resolve as fixed [FIXED]
  - Resolve as wontfix [WONTFIX]
```

The hierarchy presented above would result in "Open issue" subprocess to look like presented in Fig. 12.

There are two possible approaches to recursive application of the algorithm: top-down and bottom-up. In the first one the top-level process models are introduced first. This was the approach used in the example: the lower-level hierarchy was introduced later.

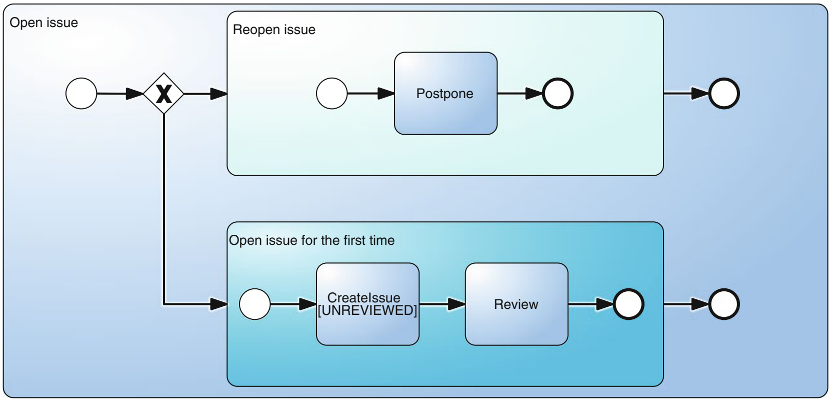


Fig. 12 Multilevel configuration version of Open issue subprocess

In the bottom-up approach the "Open issue for the first time" and "Reopen issue" subprocesses would be introduced first. For gateway simplification and removal of recurring flows elements introduces by introduction of high-level expanded subprocesses and other items are treated equally.

6 Conclusion and Future Work

In this paper, we present automatic hierarchization algorithm that takes advantage of task taxonomy and allows us to preserve similar abstraction level of subprocesses in a hierarchy. The paper describes a Business Process configuration technique that is based on the hierarchization result. This approach allows for expressing similarities between different BP models in a simple but comprehensive way. Thanks to this, a user can grasp the high-level flow of the merged processes. In comparison to other approaches, our hierarchization algorithm supports arbitrary n-to-m relationships between tasks in the merged processes. Moreover, we present how to extend our approach to support multi-level hierarchization.

To get a proof of concept of our approach, we narrowed our attention to the subset of BPMN, similarly expressive to EPC. Thanks to the use of the taxonomy shared by the three models and the hierarchization algorithm, the configuration approach is straightforward.

Our research addresses three challenges in Business Process modeling, which we distinguished in Sect. 2. These are granularity modularization, similarity capturing and collection storing challenges.

In future work, we consider to extend the approach in several ways, e.g. to integrate it with Business Rules [41], especially expressed in the XTT2 representation [42, 43], in order to use control flow as inference flow [44]. Moreover, automatic generation of taxonomy using some process metrics [8, 45] is also considered, as well as automatic assignment of tasks to subprocesses based on Natural Language Processing.

Acknowledgments The authors wish to thank the organizers of the ABICT 2013 workshop on the FedCSIS 2013 conference for providing an ample environment for presenting and discussing our research.

References

1. OMG: Business Process Model and Notation (BPMN): Version 2.0 specification. Technical report formal/2011-01-03, Object Management Group (2011)
2. Allweyer, T.: BPMN 2.0. Introduction to the Standard for Business Process Modeling. BoD, Norderstedt (2010)
3. Nalepa, G.J., Kluza, K.: UML representation for rule-based application models with XTT2-based business rules. Int. J. Softw. Eng. Knowl. Eng. (IJSEKE) **22**(4), 485–524 (2012). http://www.worldscientific.com/doi/abs/10.1142/S021819401250012X
4. Silver, B.: BPMN Method and Style. Cody-Cassidy Press (2009)
5. Reijers, H., Mendling, J., Dijkman, R.: Human and automatic modularizations of process models to enhance their comprehension. Inf. Syst. **36**(5), 881–897 (2011)
6. Mendling, J., Reijers, H.A., van der Aalst, W.M.P.: Seven process modeling guidelines (7PMG). Inf. Softw. Technol. **52**(2), 127–136 (2010)
7. Yan, Z., Dijkman, R., Grefen, P.: Business process model repositories—framework and survey. Inf. Softw. Technol. **54**(4), 380–395 (2012)
8. Dijkman, R., Dumas, M., van Dongen, B., Käärik, R., Mendling, J.: Similarity of business process models: metrics and evaluation. Information Systems **36**(2), 498–516 (2011)
9. Baran, M., Kluza, K., Nalepa, G.J., Ligęza, A.: A hierarchical approach for configuring business processes. In: Ganzha, M., Maciaszek, L.A.,Paprzycki, M. (eds.) Proceedings of the Federated Conference on Computer Science and Information Systems—FedCSIS 2013, Krakow, Poland, 8–11 September 2013, pp. 931–937. IEEE (2013)
10. Nuffel, D.V., Backer, M.D.: Multi-abstraction layered business process modeling. Comput. Ind. **63**(2), 131–147 (2012)
11. Pittke, F., Leopold, H., Mendling, J., Tamm, G.: Enabling reuse of process models through the detection of similar process parts. In: La Rosa, M., Soffer, P. (eds.) Business Process Management Workshops. Lecture Notes in Business Information Processing, vol. 132, pp. 586–597. Springer, Berlin (2013)
12. Weber, B., Reichert, M., Mendling, J., Reijers, H.A.: Refactoring large process model repositories. Comput. Ind. **62**(5), 467–486 (2011)
13. La Rosa, M., Wohed, P., Mendling, J., ter Hofstede, A., Reijers, H., Van der Aalst, W.M.P.: Managing process model complexity via abstract syntax modifications. Ind. Inf. IEEE Trans. **7**(4), 614–629 (2011)
14. Cappelli, C., Leite, J.C., Batista, T., Silva, L.: An aspect-oriented approach to business process modeling. In: Proceedings of the 15th Workshop on Early Aspects. EA'09, pp. 7–12. ACM, New York (2009)
15. Dumas, M., Garcia-Banuelos, L., Rosa, M.L., Uba, R.: Fast detection of exact clones in business process model repositories. Inf. Syst. **38**(4), 619–633 (2013)
16. Uba, R., Dumas, M., Garcia-Banuelos, L., Rosa, M.: Clone detection in repositories of business process models. In: Rinderle-Ma, S., Toumani, F., Wolf, K. (eds.) Business Process Management. Lecture Notes in Computer Science, vol. 6896, pp. 248–264. Springer, Berlin (2011)
17. Haddar, N.Z., Makni, L., Ben Abdallah, H.: Literature review of reuse in business process modeling. Softw. Syst. Model. **121**, 1–15 (2012)

18. Mendling, J., Neumann, G., Aalst, W.: Understanding the occurrence of errors in process models based on metrics. In: Meersman, R., Tari, Z. (eds.) On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS. Lecture Notes in Computer Science, vol. 4803, pp. 113–130. Springer, Berlin (2007)
19. Dijkman, R., Rosa, M.L., Reijers, H.A.: Managing large collections of business process models—current techniques and challenges. Comput. Ind. **63**(2), 91–97 (2012)
20. Kunze, M., Weske, M.: Metric trees for efficient similarity search in large process model repositories. In: Muehlen, M., Su, J. (eds.) Business Process Management Workshops. Lecture Notes in Business Information Processing, vol. 66, pp. 535–546. Springer, Berlin Heidelberg (2011)
21. La Rosa, M., Gottschalk, F.: Synergia-comprehensive tool support for configurable process models. In: Proceedings of the Demo Track of the 7th International Conference on Business Process Management (BPM'09), vol. 489. CEUR (2009)
22. Bobek, S., Baran, M., Kluza, K., Nalepa, G.J.: Application of bayesian networks to recommendations in business process modeling. In: Giordano, L., Montani, S., Dupre, D.T. (eds.) Proceedings of the Workshop AI Meets Business Processes 2013 co-located with the 13th Conference of the Italian Association for Artificial Intelligence (AI*IA 2013), Turin, Italy, December 6, 2013 (2013). http://ceur-ws.org/Vol-1101/
23. van der Aalst, W.M.: Business process management: a comprehensive survey. In: ISRN Software Engineering 2013 (2013)
24. La Rosa, M., Dumas, M., ter Hofstede, A.H., Mendling, J.: Configurable multi-perspective business process models. Inf. Syst. **36**(2), 313–340 (2011)
25. Rosemann, M., van der Aalst, W.M.P.: A configurable reference modelling language. Inf. Syst. **32**(1), 1–23 (2007)
26. Puhlmann, F., Schnieders, A., Weiland, J., Weske, M.: Variability Mechanisms for Process Models. PESOA-Report TR 17/2005, Process Family Engineering in Service-Oriented Applications (pesoa). BMBF-Project. Report, Hasso Plattner Institut, Postdam, Germany (2005)
27. Schnieders, A., Puhlmann, F.: Variability mechanisms in e-business process families. In: Proceeedings of the International Conference on Business Information Systems (BIS 2006), pp. 583–601 (2006)
28. Tealeb, A., Awad, A., Galal-Edeen, G.: Context-based variant generation of business process models. In: Bider, I., Gaaloul, K., Krogstie, J., Nurcan, S., Proper, H.A., Schmidt, R., Soffer, P. (eds.) Enterprise, Business-Process and Information Systems Modeling, no. 175 in Lecture Notes in Business Information Processing, pp. 363–377. Springer, Berlin (2014). http://link.springer.com/chapter/10.1007/978-3-662-43745-2_25
29. La Rosa, M., Dumas, M., Uba, R., Dijkman, R.M.: Business process model merging: an approach to business process consolidation. ACM Trans. Softw. Eng. Methodol. (TOSEM) **22**(2), 11 (2013)
30. Döhring, M., Reijers, H.A., Smirnov, S.: Configuration vs. adaptation for business process variant maintenance: an empirical study. Inf. Syst. **39**, 108–133 (2014)
31. Weidmann, M., Koetter, F., Kintz, M., Schleicher, D., Mietzner, R.: Adaptive business process modeling in the internet of services (ABIS). In: ICIW 2011, The Sixth International Conference on Internet and Web Applications and Services, pp. 29–34 (2011)
32. Müller, R., Greiner, U., Rahm, E.: Agentwork: a workflow system supporting rule-based workflow adaptation. Data Knowl. Eng. **51**(2), 223–256 (2004)
33. Charfi, A., Müller, H., Mezini, M.: Aspect-oriented business process modeling with ao4bpmn. In: Modelling Foundations and Applications, pp. 48–61. Springer, Heidelberg (2010)
34. Rastrepkina, M.: Managing variability in process models by structural decomposition. In: Business Process Modeling Notation, pp. 106–113. Springer (2011)
35. Gottschalk, F., Van Der Aalst, W.M., Jansen-Vullers, M.H., La Rosa, M.: Configurable workflow models. Int. J. Co-op. Inf. Syst. **17**(02), 177–221 (2008)
36. Awad, A., Sakr, S., Kunze, M., Weske, M.: Design by selection: a reuse-based approach for business process modeling. In: Conceptual Modeling-ER 2011, pp. 332–345. Springer, Berlin (2011)

37. Meerkamm, S.: Configuration of multi-perspectives variants. In: Business Process Management Workshops, pp. 277–288. Springer, Heidelberg (2011)
38. Meerkamm, S.: Staged configuration of multi-perspectives variants based on a generic data model. In: Business Process Management Workshops, pp. 326–337. Springer, Heidelberg (2012)
39. Hallerbach, A., Bauer, T., Reichert, M.: Capturing variability in business process models: the provop approach. J. Softw. Maint. Evol. Res. Pract. **22**(6–7), 519–546 (2010)
40. Döhring, M., Zimmermann, B.: vBPMN: Event-Aware Workflow Variants by Weaving BPMN2 and Business Rules. In: Enterprise, Business-Process and Information Systems Modeling, pp. 332–341. Springer (2011)
41. Nalepa, G.J.: Proposal of business process and rules modeling with the XTT method. In: V. Negru, et al. (eds.) Symbolic and numeric algorithms for scientific computing, 2007. SYNASC Ninth international symposium. September 26–29, pp. 500–506. IEEE Computer Society, IEEE, CPS Conference Publishing Service, Los Alamitos, California; Washington; Tokyo (2007)
42. Ligęza, A., Nalepa, G.J.: A study of methodological issues in design and development of rule-based systems: proposal of a new approach. Wiley Interdisciplinary Rev. Data Mining Knowl. Discov. **1**(2), 117–137 (2011). doi:10.1002/widm.11
43. Nalepa, G.J., Ligęza, A., Kaczor, K.: Formalization and modeling of rules using the XTT2 method. Int. J. Artif. Intell. Tools **20**(6), 1107–1125 (2011)
44. Nalepa, G., Bobek, S., Ligęza, A., Kaczor, K.: Algorithms for rule inference in modularized rule bases. In: Bassiliades, N., Governatori, G., Paschke, A. (eds.) Rule-Based Reasoning, Programming, and Applications. Lecture Notes in Computer Science, vol. 6826, pp. 305–312. Springer, Berlin / Heidelberg (2011)
45. Kluza, K., Nalepa, G.J.: Proposal of square metrics for measuring business process model complexity. In: M. Ganzha, L.A. Maciaszek, M. Paprzycki (eds.) Proceedings of the Federated Conference on Computer Science and Information Systems—FedCSIS 2012, Wroclaw, Poland, 9–12 September 2012, pp. 919–922 (2012). http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6354395

Profiling Simulation Performance: The Example of the German Toll System

Tommy Baumann, Bernd Pfitzinger and Thomas Jestädt

Abstract The execution performance of simulation model is an issue for large models, models used to give predictions in real-time and when the configuration space needs to be explored. These use cases apply to an existing large-scale simulation model of the German toll system which we use to identify typical workloads by profiling the run-time behavior. Performance hot spots are identified in the model and the execution environment and are related to the real-world application. In a benchmark approach we compare the observed performance to different simulation frameworks.

1 Introduction

Software evolution is a fact of life: The complexity, interconnectedness and heterogeneity of systems increases continually. Modeling and simulation techniques are one way to address the challenge of engineering, integrating and operating such systems. Especially in the design and specification stage executable models deliver tremendous value by lowering system design uncertainty and increasing specification speed and quality [1]. Hence, current system design approaches like Simulation Driven Design [2] are characterized by applying executable models to a large extend. In this context the performance in defining and executing models becomes vital due to the increased complexity of systems and processes as well as the customer requirement

T. Baumann
Andato GmbH and Co. KG, Ehrenbergstraße 11, 98693 Ilmenau, Germany
e-mail: tommy.baumann@andato.com

B. Pfitzinger (✉) · T. Jestädt
Toll Collect GmbH, Linkstraße 4, 10785 Berlin, Germany
e-mail: bernd.pfitzinger@toll-collect.de

T. Jestädt
e-mail: thomas.jestaedt@toll-collect.de

B. Pfitzinger
FOM Hochschule Für Oekonomie and Management, Zeltnerstraße 19, 90443
Nürnberg, Germany

M. Mach-Król et al. (eds.), *Advances in ICT for Business, Industry and Public Sector*,
Studies in Computational Intelligence 579, DOI 10.1007/978-3-319-11328-9_2

to create holistic, integrated, high accuracy models up to real world scale. Several use cases of simulations are only possible once the simulation performance is 'good enough': simulating the long-term dynamic behavior, iterative optimization loops, automatic test batteries, real-time models (higher reactivity to market demands and changes), and automated specification and modeling processes (including model transformation/generation) [3].

In this article we analyze the simulation performance of a large-scale Discrete Event Simulation (DES) [4] model of the German toll system implemented in MSArchitect [5]. The overall simulation performance emerges as the combination of the simulation framework (kernel) and the application domain simulation model. The outline of the article is as follows: Sect. 2 gives an overview of the automatic German toll system, the corresponding simulation model and typical simulation results. Section 3 introduces the simulation framework architecture and performance measurement techniques. Section 4 analyzes and evaluates the simulation performance of several DES kernels using small test models. Section 5 analyzes the performance within the application domain. Section 6 summarizes the results and describes future improvements and applications of our simulation model.

2 Executable Model of the German Toll System

For the application domain we use an existing simulation model of the German toll system [3, 6–8], a large-scale autonomous toll system [9] operated by Toll Collect GmbH. The tolls for heavy-goods vehicles (HGVs) driving on federal motorways—a total of 4.36 bn€ in 2012 [10]—is collected by the toll system, more than 90 % fully automatic using the more than 765,000 on-board units (OBUs) deployed in the HGVs.

In the Toll Collect example simulations are an additional method to ensure the service quality [11]. From a business perspective the simulation model is used to predict the dynamic system behavior for standard operational procedures (e.g. planned changes and updates), to determine the appropriate system sizing and to explore the system behavior close to the specification limits. Within the software development process the simulation model is used in the early design stages to implement prototypes in the context of the overall system allowing to gauge the dynamic system behavior already at the start of the software development cycle. The simulation model includes all subsystems necessary for the automatic tolling processes (Fig. 1) and for delivering updates to the OBU software, geo and tariff data as well as a model of the user interaction:

Fig. 1 High-Level simulation model of the Toll Collect system (*upper half*) and the model for the user interaction (*driving patterns*)

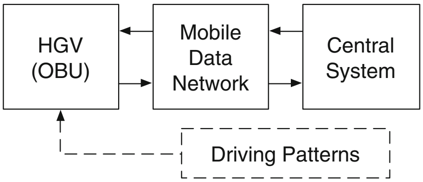


Table 1 Comparing fleet-wide updates (simulation results vs. data from April 2012 to January 2013, [13])

	Correlation
Software	0.99963
Geo data	0.99572
Tariff data	0.99475

- The vehicle fleet at a scale of 1:1 consisting of 765,000 instances of the OBU class responsible for tolling (collecting and forwarding to the central system) and for initiating the download of updates.
- The central systems necessary to receive the toll data and to serve download requests. A typical DMZ is used for authentication and accounting and the model includes a number of different resource constraints observed in the real-world system.
- The German mobile data networks provide the IP-based communication (via GPRS) between the OBUs and the central systems.
- The user interaction is modeled as statistically generated driving patterns (e.g. the points in time of HGV power cycles, toll events and mobile data network availability).

From the process perspective the simulation model covers business and system processes differing at least 7 orders of magnitude in time: All major technical processes with durations of 1 s and longer are included in the model aiming to predict the dynamic system behavior of fleet-wide updates In fact, the model includes some processes with higher temporal resolution (down to 50 ms for the connection handling in the DMZ) and is used to simulate all updates occurring over a whole year. Using the Pearson correlation as metric to compare the simulation results with the observed update rates between April 2012 and January 2013 we find the correlation to be above (better than) 0,994 (see Table 1). The current investigation is to validate the simulation model using additional metrics and a time-scale of 1 h [13] (instead of one day).

However, a further improvement of the simulation results requires the use of an optimization algorithm to search the parameter space (an example is given in Fig. 2). A technical challenge even when using a scaled-down simulation model (1:1000): Even on the application level the user interaction (scenario generator) creates a large number of events to be processed by the simulation logic. On average each OBU will be powered-on for 16 % of the time and process tolls for 32,000 km annually ([14], one toll event per 4.2 km on average [15]) spread across some 1,300 power cycles (including three times as many periods of mobile data network). Of course, many more events are created from within the application logic.

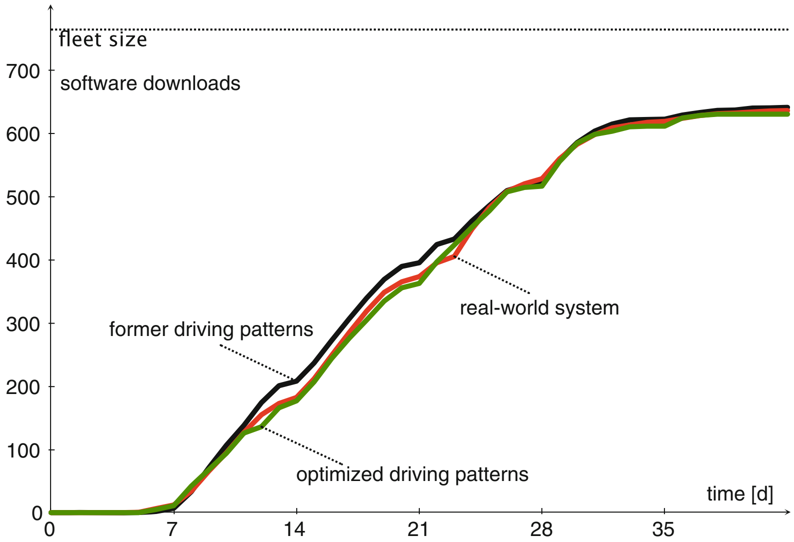


Fig. 2 Simulated software update before and after optimization [12]. The optimization algorithm is expected to improve the manually adjusted parametrization used previously (Table 1)

3 Simulator Architecture and Performance Measurement

This section describes the architecture of the simulator and different possibilities to measure and evaluate simulation performance. The simulation model of the German toll system is implemented using the general purpose system design toolkit MSArchitect. The toolkit supports high performance DES and parallel DES (PDES), automated model reduction [16], hierarchical simulation models with the concept of executable simulation missions, performance probe and visualization capabilities and the generation and transformation of models during run-time.

MSArchitect distinguishes between atomic models and composite models (as most actor-oriented DES/PDES tools) to capture the behavior and structure of systems and processes. The composite models provide the structural composition of models whereas the atomics interact with the simulation kernel and are typically compiled into binary code. Combining both modeling types results in a hierarchical model tree with atomic models as leafs and composite models as nodes, as shown in Fig. 3.

The definition of atomic models and composite models depend on the application programming interface (API) to the simulation kernel. The following enumeration explains the most important components of the kernel interface [17]:

- A port represents a communication interface from or to another model entity by a logical communication channel (connection). Ports are characterized by data-type, direction, priority, and delay.

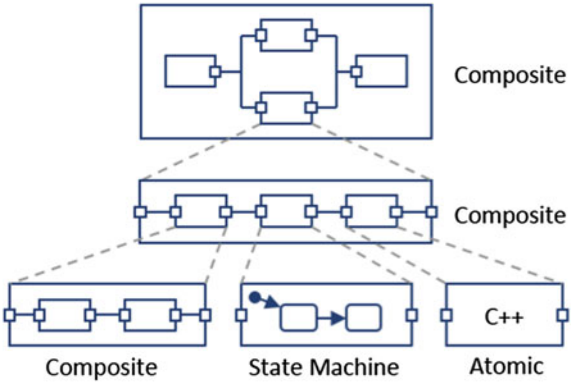


Fig. 3 Model hierarchy: Composite models and Atomics

- A state represents a data entity (memory) containing part of the model state during simulation. Depending on the visibility a state can be either only accessed locally or shared with other simulation entities.
- A parameter holds a data value, which is defined at modeling time and remains constant during simulation. A parameter is defined by data-type and value can be calculated from other parameters.
- A method encapsulates a user-defined functionality, which processes data or affect states. Methods are defined by signature and visibility.

Based on the kernel interface definition and the possibilities of the model description four technical representation layers can be differentiated in MSArchitect (Table 2, in principle applicable to all DES tools) with different factors impacting the performance on each layer. The *DES Composition Layer* describes the interconnection of models within a composite model and allows analyzing structural dependencies and properties [17]. On this layer, much of the performance depends on the application level behavior and its implementation as abstract simulation model. Starting with the *DES Atomic Interface Layer* the performance becomes independent of the application domain and is determined by the simulation toolset ([18], which describes the interface of atomic models as well as restrictions on features available

Table 2 Technical representation layers used in MSArchitect

Technical layer	Vocabulary
DES composition layer	Interconnection of ports, states and, parameters
DES atomic interface layer	Ports, states, parameters, lifecycle methods, inheritance, model/data types
Code layer	C++ types, variables, instructions, method calls, Control/data-flow, inheritance
Machine layer	Object code, register, memory, instructions (arithmetic, Jump, call)

for the functional description). The *Code Layer* contains the functional implementation of all atomics in the form of lifecycle methods and custom code sections. The performance is determined by the kernel event scheduling and the code generated by the compiler. The *Machine Layer* contains preprocessed and compiled code for model execution and references to external runtime libraries to be executed on a given CPU architecture.

The layers define the information available for profiling the run-time behavior, collected either by the simulation kernel, simulation logs or (external) performance profiling tools. This analysis aims first to assess the performance of simulation tools (i.e. DES Atomic Interface Layer and below). In a further analysis we take one application level example to benchmark a real-world simulation.

A kernel benchmark consists of several small test cases, which stress different aspects of a technical process. Thus they are useful to analyze low-level kernel mechanisms. The results are typically weighted according to their importance for a certain application domain. In Sect. 4 we evaluate the performance of several simulation frameworks by kernel benchmarking. Application benchmarks are characterized by use of real examples from the application domain. The selected applications should exhibit different characteristics and represent typical challenging workloads. In our case we rely on one application (the model of the German tolling system) and provide an in-depth performance analysis in Sect. 5.

4 Benchmark of Simulation Kernel Performance

In this section we perform a kernel benchmark based on test models. The models have been kept simple in order to assure universality regarding different kernels/tools and to avoid possible side effects. With regards to the technical representation layers in Table 2 the test models are defined on the *DES Composition Layer* and the *DES Atomic Interface Layer*. We included the most important performance influencing factors in our tests: The Future Event List (FEL) management, memory and data type management, pseudo-random number generator performance, and arithmetic operations performance [19]. Thereby the DES performance is characterized using the wall clock time of simulation runs, which represents the CPU time for executing a simulation model, as well as the peak memory allocated during the simulation run. Doing so, the simulation performance can be expressed as average events per second (ev/s) and memory usage. In the following we present five test models [7]. These models are applied later to investigate and compare DES kernel performance.

- Runtime scaling: A simulation model with several hierarchy levels is simulated with an increasing total number of events, while the size of the future event list remains fixed. This test determines whether the FEL performance is affected by the FEL size.
- FEL Size Scaling: The second test uses the simulation model with a delay-function. The delay is used to easily configure the (average) number of events waiting in the

future events list with minimal variance, while the total number of events processed remains constant. This test examines the overall performance of the FEL algorithm and data management. We used a uniform distribution of events in the FEL list.

- FEL Adaption: This test extends the future events list size during one test run by changing the parameter of the delay-function dynamically during the simulation execution. The test extends the previous test model by additional single events used to change the parameter of the delay-function. As a consequence the size of the FEL changes during the test run forcing the simulation kernel to adapt the FEL size (e.g. allocating and freeing memory) during the simulation run.
- Data Type Management: The test creates large data arrays of different sizes and passes the data through the simulation model in sequential or parallel order. When executing the model the memory management of the simulation kernel should recognize the passing of unmodified data and use references to this data. Ideally only on datum should be created and sent as reference through the model. As long as the delays are set to zero, no difference between serial and parallel passing should be recognizable.
- Random Number Generation: A large number of random values is generated using different distributions. The model uses a constant function as a reference to measure relative performance of the built-in pseudo-random-number generators.

Each test model is simulated with a set of simulation parameters using different system design tools. Currently more than 80 tools listed for DES [20]. We selected six system design tools for evaluation: Ptolemy II, Omnet++, AnyLogic, MLDesigner, SimEvents and MSArchitect. All tools were run in serial mode on an Intel Core i7 X990 at 3.47 GHz with 24 GB RAM using either Windows 7 Enterprise (64 bit) or openSuse 11.4 (32 bit, kernel 2.6.37.6). Performance data was recorded with Perfmon on Windows and sysstat and the Gnome System Monitor on the Linux system.

Figure 4 gives the results of the Runtime Scaling test. The upper chart shows the event processing performance for different simulation lengths and the bottom chart the private memory consumed during simulation. The tests show that neither the memory consumption nor the event processing performance is affected by increasing the simulation runtime. Looking at the sensitivity of running the tests with additional hierarchy levels we find that only OMNeT++ is sensitive to the additional hierarchy levels. However, from the test results it is already obvious that the different tools vary in event processing performance by an order of magnitude: MSArchitect provides the highest performance. MLDesigner, AnyLogic, and OMNeT++ provide 25 % of the speed (compared to [7] the MSArchitect performance improved by more than 30 %). Ptolemy II is twenty times slower. Looking at the memory usage during the simulation the difference between the tools is again more than an order of magnitude—the slowest tool using the most memory and the fastest tool using the least. Two of the tools (Ptolemy II and AnyLogic) are based on the Java programming language, where explicit memory deallocation is not possible. Apparently the Ptolemy II test run triggers the JVM garbage collection during the simulation run and is able to free 90 % of its memory.

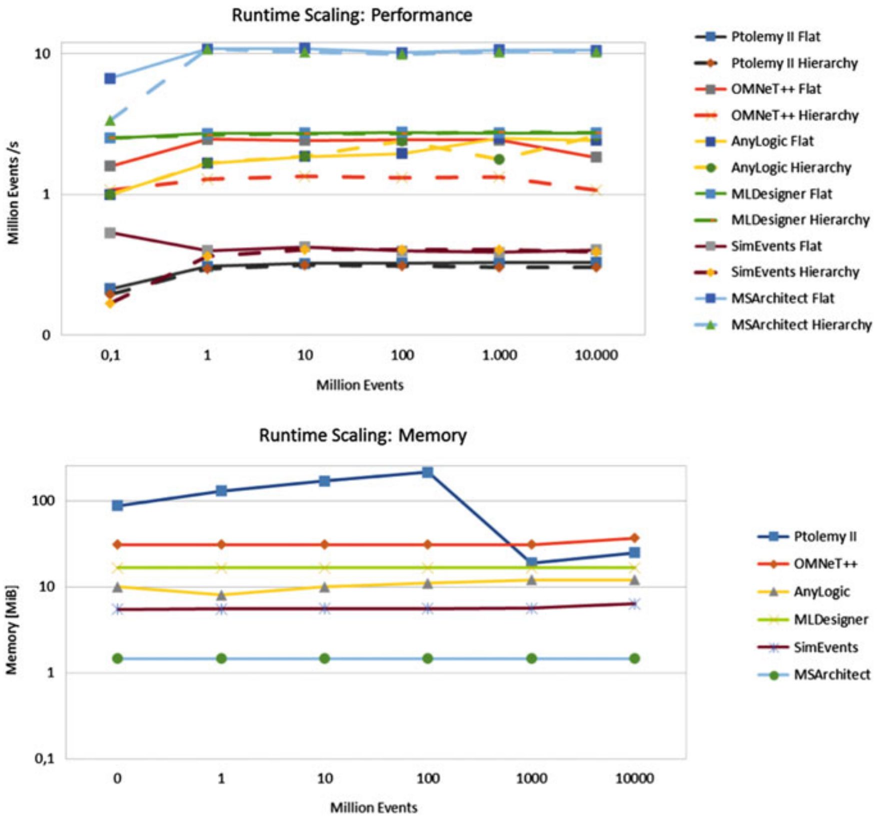


Fig. 4 Runtime performance (*top*) and memory usage (*bottom*) scaling of different DES tools

Figure 5 gives the results of the FEL Size Scaling test. As expected, a systematic performance reduction can be observed with increasing FEL size, due to the increasing overhead for FEL management. Most of the tools tested initially start with relative constant performance (on a log-log scale). With increasing FEL size three of the five tools develop drastic performance degradation. This coincides with a rapid grow of memory consumption with increasing FEL size. We propose that the performance reduction is correlated with increased FEL memory usage due to a performance penalty of calendar queue based schedulers for large queue sizes.

In the FEL Adaption test the simulation kernel is subjected to a varying demand to its FEL. Beyond the run-time needed for the test the main result is the memory consumption during the test run as given in Fig. 6. The simulation took 2,276 s with MLDesigner, 673 s with AnyLogic, 192 s with MSArchitect, 395 s with OMNeT++ and over 2 h with Ptolemy II. During that time the dynamically changing memory usage varies widely between the different tools. Most tools tend to allocate memory in chunks visible as steps in Fig. 6.

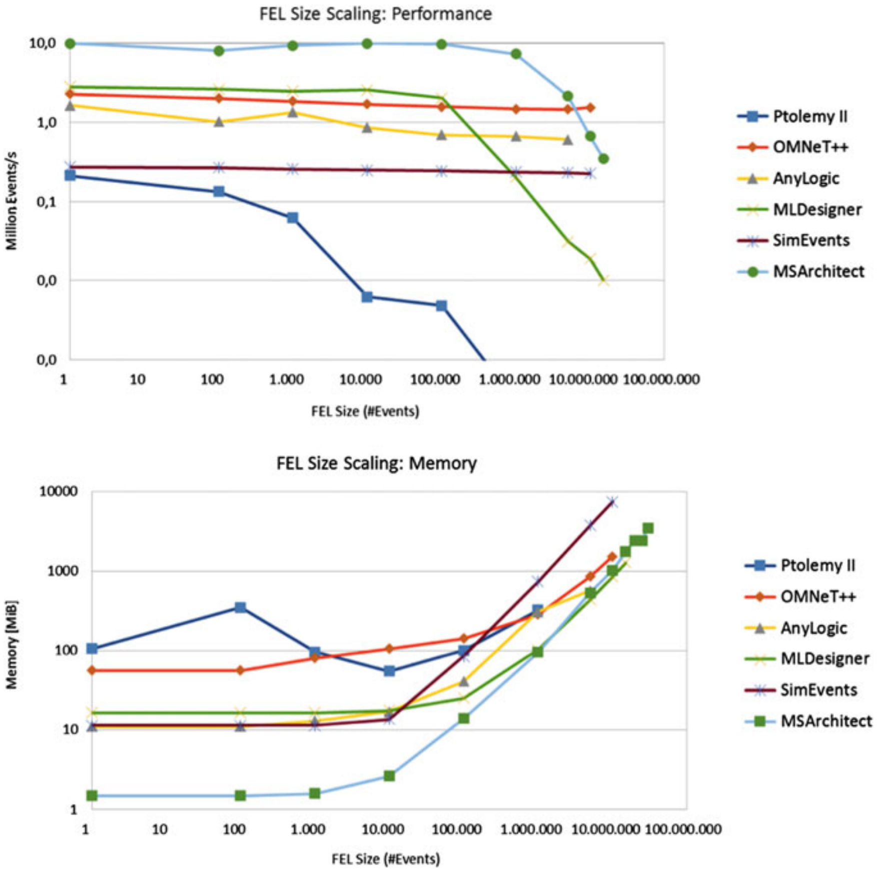


Fig. 5 Future event list size scaling test results for runtime performance (*top*) and memory usage (*bottom*) scaling of different DES tools

The Data Type Management test passes large arrays of data through the simulation running either in a parallel or serial configuration. The memory consumption during the test run is shown in Fig. 7. Most tools handle serial and parallel passing of token data in a different way, which can be recognized by the gap in memory consumption between both serial and parallel versions. Ptolemy II and MSArchitect do not show a difference between the parallel and serial version, only references are passes when only delays are used. However, Ptolemy II requires more memory and shows a different behavior according to the memory allocation: a large portion of the memory is allocated at initialization time with standard modeling elements.

The last test is the Random Number Generation test. As depicted in Fig. 8 the performance does not depend on the type of the generated distribution. Since this test relies on the correct implementation of the PRNG, i.e. we do not check the

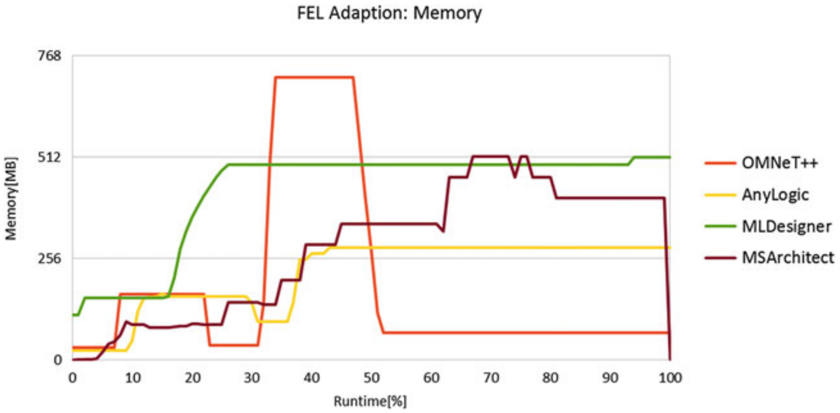


Fig. 6 Test results for future events list adaption

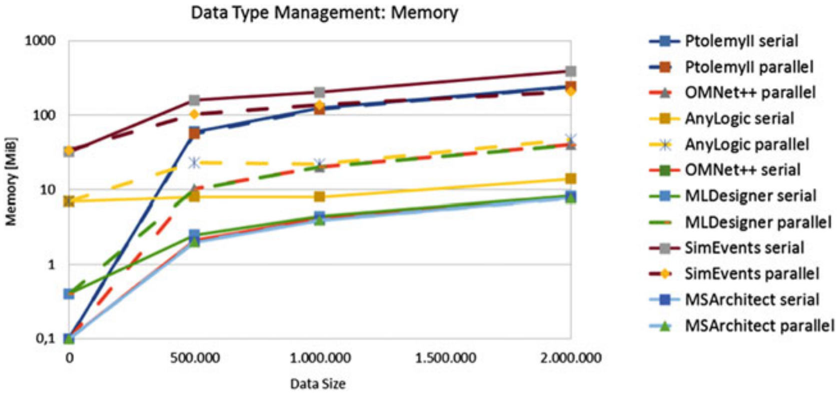


Fig. 7 Test results for the memory consumption during Data Type Management test case

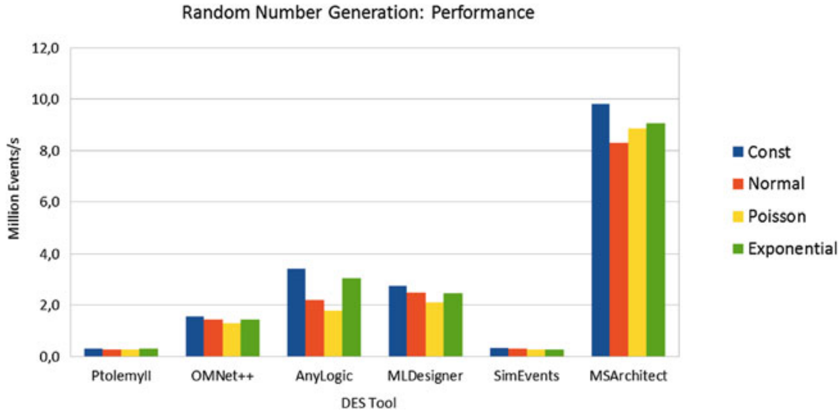


Fig. 8 Test results for random number generator scaling

statistical quality of the random numbers generated, the test results might not be fair if one of the tools were to use low-quality but high-speed generators.

It can be concluded, that Ptolemy II is the slowest tool in all simulation kernel benchmarks performed. MLDesigner is equal to or better than AnyLogic in all categories but FEL adoption. Due to the utilization of the JVM, AnyLogic requires more memory in equivalent models and therefore scales worse with increasing FEL size. Both, MSArchitect as well as OMNeT++ show the best performance in some categories. MSArchitect is the fastest simulation kernel in most categories and requires least memory for data handling.

5 Profiling of the Simulation Model

To evaluate the application-level simulation performance of our model of the German toll system, we use both the kernel logging capabilities of MSArchitect and an external profiling application (Intel VTune). Kernel logging allows to count the number of calls of atomic models as well as the total number of samples (corresponding to a processor cycle). The external profiler allows measuring the space complexity (memory), the time complexity (duration, CPU time), and the usage of particular instructions of a target program by collecting information on their execution. The most common use of a profiler is to help the user evaluate alternative implementations for program optimization. Based on their data granularity, on how profilers collect information, they are classified into event based or statistical profilers [21]. We chose the statistical profiler Intel VTune Amplifier XE and connected it to the generated C++ run-time representation of our model. As test environment, an Intel Core i7 K875 at 2.93 GHz with 8 GB RAM and Windows 7 Professional (64 bit) installed has been used.

To profile the simulation model we take the simulation scenario used to verify the simulation model against real-world data. We take statistical data to model the user interaction (driving patterns) of the vehicle fleet (700,000 OBUs used in the simulation run) to simulate the tolling system and to predict the daily rates of updates to the fleets' software, geo and tariff data. In the simulation run each OBU is treated individually and has on average 1,300 power cycles annually, the majority being very brief (i.e. less than 30 min) but some taking up to several days. The number of toll events created for the driving patterns is calibrated to correspond to the total toll amount collected in 2012 (≈32,000 km annually per OBU, using an average segment length of 4.2 km [15] and average speed of close to 80 km/h). The less-than-perfect availability of the mobile data networks is taken into account by adding (many brief) periods of network unavailability to the driving patterns. On average each power cycle has 3 such periods corresponding to almost 10 % of the power-on time (comparable to numbers seen in nation-wide network comparisons [22, 23]). In our simulation we calibrate the power-on time using the real-world update rates and choose a total power-on time of 16 % (of the whole year on average per OBU) of which 29 % are spent driving on toll roads.

Taking this scenario we use the hardware setup described above to perform the simulations with pre-calculated user interaction (driving patterns, stored as ASCII file on disk). The simulation model reads and parses the driving patterns (containing the events for power cycles, tolling and network unavailability at a granularity of 1 s) and feeds the events into the simulation model. The simulation model in turn creates many more events (with a time resolution below 1 s) to be processed by the simulation kernel. Using a single CPU core the simulation run encompassing a fleet of 700,000 OBUs and a simulated time period of 45 weeks takes less than 10 h to compute.

5.1 Profiling with MSArchitect

Applying the kernel logging capabilities of MSArchitect resulting in a file with profiling information at the end of the simulation run. Table 3 shows an excerpt of the file, containing all atomic blocks relevant to analysis (15 out of 65). Since during simulation all composite blocks are resolved to directly communicating atomic blocks (see Fig. 3), the table only contains atomic blocks and is therefore related to the *DES Atomic Interface Layer*. For each atomic block the table shows the number of calls, the accumulated count of samples, the time required in relation to other atomic blocks, and the samples needed for one call.

First of all, the atomic block "AccessSessionStateSwitch" is striking, since it consumes a large amount of time due to the high number of calls. The block is

Table 3 Kernel logging results

Atomic block	Calls (M)	Samples (G)	Time (%)	Samples per call
AccessSessionStateSwitch	19,980	10,760	10,89	539
ExternDStxt	0,0007	9,565	9,68	13,665 M
StaHandling	482	6,503	6,58	13,483
EinzelbuchungsHandling	4,660	6,442	6,52	1,382
IpAutomat	7,323	5,749	5,82	785
Delay (Standard)	8,874	5,522	5,59	622
CheckComponentState	7,363	3,859	3,91	524
NetzverlustHandling	3,020	3,479	3,52	1,152
AccessSessionStateWrite	5,841	3,215	3,26	551
MfbSwitch	5,525	3,196	3,24	579
Nutzdaten	3,563	2,694	2,73	756
TcmessageCopy	2,030	2,010	2,03	990
TcpAutomat	1,291	1,533	1,55	1,187
TimedAllocate	2,570	1,484	1,50	578
SimOutObuVersions	0.017	1,509	1,53	89 M

responsible for switching OBU data structures in response to its state to one of the output ports. As the block switches between 34 states, 539 samples per call are acceptable. Nevertheless the number of calls could be reduced for performance improvement by changing the model architecture—especially once the model is ported to the parallel DES core, it is an obvious block for introducing parallelism (considering that more than 700,000 instances are running simultaneously).

The next conspicuous atomic block is “ExternDStxt”, reading the pregenerated files provided by the scenario generator model as ASCII file. The block consumes 13,665 M samples/call and is rarely executed 700 times (twice per simulated day) resulting in a time consumption of 9,681 %. In order to reduce the load, scenarios should be computed on the fly. The atomic block “StaHandling” is responsible for generating and controlling status requests, which may result in update processes. The block consumes 6,581 % of simulation time. We see potential for improvements in changing the implementation (e.g. conversion of formulas to save operations, replacing divisions by multiplications with reciprocal and using of compare functions from standard libraries). With 4,660 M calls “EinzelbuchungsHandling” is a frequently executed atomic block. After analyzing the implementation we find 1,382 samples/call acceptable. The block depends on the random number generator and would benefit from faster random number generation algorithms. The atomic block SimOutObuVersions cyclically writes the software, region, and tariff version of all OBUs to an output file. In our scenario we simulate 50 weeks and write data every 30 min, resulting in 16,801 calls. 89 M samples/call seems to be quite costly and offers room for improvement.

In summary the simulation of the scenario took 98,811,263 M calls. Of these, the model components consumed 84,51 % and the simulation kernel (logical processor) 15,49 %.

5.2 *Profiling with Intel VTune*

In the second step we apply the profiling application Intel VTune [24], which operates at functional level resp. Code Layer (see table 4). The external profiler catches the activities of both the simulation kernel and the simulation model (denoted as “K” or “M” in Table 4).

An excerpt of the results is shown in Table 4. For each function the CPU time in percent, the amount of needed instructions (instructions retired), the estimated instruction call count, the instructions per call on average, and the last level cache miss rate (0,01 means one out of hundred accesses takes place in memory) is shown.

Most of the CPU time is consumed by kernel functions responsible for data transport. These functions are grouped by component (resp. namespace `msa.sim.core`, denoted as “K” in the first column of Table 4), as `Port.send`, `EventManager.(en|de)queueEvent` and `LogicalProccesor.mainLoopFast`. In sum the functions consume 61,1 % of the CPU time. Conspicuous is the relative high last level cache miss rate of function `EventManager.enqueueEvent` with

Table 4 VTune profiling results for simulation kernel (K) and model (M) ordered by CPU time

	Function	Time (%)	IR (G)	eCC (M)	IPC	MR (%)
K	Port.send	9,0	44	689	65	0,4
K	EventManager.enqueueEvent	7,7	21	92	237	3,2
K	LogicalProcessor.mainLoopFast	7,0	17	7	2,379	0,3
K	EventManager.dequeueEvent	6,3	104	2,517	41	1,1
K	big._mul<unsigned int>	5,0	103	2,611	40	0,3
M	StaHandling.Dice	4,8	12	11	1,097	0,1
K	EventManager.scheduleEvent	3,5	53	1,286	42	0,2
K	Any.extractToken	3,0	70	1,805	39	1,7
K	Pin.popFrontToken	2,8	49	1,234	40	0,2
K	EventManager.bucketOf	2,7	17	327	55	0,0
K	Any.operator=	2,5	54	1,403	39	0,2
K	Any.create	2,3	64	1,689	38	0,4
K	random.tr1.UniformRng.getNextV	2,2	39	961	41	0,2
K	Any.doClear	2,1	22	497	45	0,2
K	Tokenizer.nextToken	1,8	22	606	36	0,3
K	TemplatePort<Tcmessage>.re-ceiveToken	1,7	29	726	40	0,3
K	TemplateTypeInfo<EventData> .createToken	1,6	84	2,326	36	2,1
M	AccessSessionStateSwitch.run	1,5	13	287	48	0,2
M	EinzelbuchungsHandling.run	1,4	5	66	87	7,2
M	IpAutomat.run	1,4	7	103	70	3,4
K	Pin.popFront	1,3	6	89	74	0,3

Shown are the CPU instructions retired (IR), estimated call count (eCC), instructions per call (IPC) and last level cache miss rate (MR)

3,2 % and the needed instructions per call of function `LogicalProcessor.mainLoopFast` with 2379. However, the number of calls depends on the dispatch of data within atomic model components, which are grouped in form of user libraries. In our model we have two user libraries: GPRSSimulation (`GPRSSimulation.Components.Atomics`, denoted as "M" in the first column of Table 4) and Standard (`msa.Standard.Control`). The latter is a support library included in MSArchitect. Combined they are responsible for 20,1 % of CPU time consumption. Performance critical and starting point for improvement is function `StaHandling.Dice` with 1,097 instructions per call and a CPU time consumption of 4,80 %. This function is part of the update process where each OBU determines randomly (hence 'Dice') whether to check for any available updates.

Both, kernel logging and profiling showed that most of the resources are utilized by functions responsible for data input/output (data mining), and functions responsible for transmission and processing of tolling information. By doing the analysis we located multiple components with potential for optimization, e.g. `AccessSessionStateSwitch` and `StaHandling`. Furthermore we came to the conclusion to generate scenarios on the fly since the reading of pre-generated scenario files is as time consuming. Relating the resource utilization of model components to real-word applications we could recognize a weak correlation. Model components like `STAHandling`, `IPAutomat` and `Einzelbuchungs-Handling` are abstractions of important real word system applications and crucial to performance in both worlds.

6 Summary

Extending [7, 8] we have shown how to analyze the performance of DES simulations: Generic benchmark test-cases allow a simple and direct comparison of different simulation tools. Not surprisingly the tools differ vastly as to their time and memory consumption. Yet the real-world impact depends on the application: The workload generated in typical simulation runs and the necessity of high-performance simulations.

Particularly in real-time scenarios and optimization runs the use of large-scale simulations depends critically on the performance achieved. We have shown the benefit of profiling to identify bottlenecks and to choose the appropriate tool. MSArchitect, the simulation kernel used in the application benchmark, is currently extended to allow the automatic model reduction and (semi-) automatic parallelization of simulation runs. The single-core benchmark performed here will be the baseline to measure the improvements against.

References

1. Baumann, T.: Automatisierung der frühen Entwurfsphasen verteilter Systeme. Südwestdeutscher Verlag für Hochschulschriften, Saarbrücken (2009)
2. Baumann, T.: Simulation-driven design of distributed systems. SAE Technical Paper, no. 2011–01-0458 (2011)
3. Pfitzinger, B., Baumann, T., Jestädt, T.: Network resource usage of the german toll system: Lessons from a realistic simulation model. In: 46th Hawaii International Conference on System Sciences (HICSS), pp. 5115–5122 (2013)
4. Lee, V., Messerschmitt, D.G.: Static scheduling of synchronous data flow programs for digital signal processing. IEEE Trans. Comput. **C 36**(1), 24–35 (1987)
5. Andato GmbH & Co. KG: MSArchitect. http://www.andato.com/. Accessed 12 May 2013
6. Baumann, T., Pfitzinger, B., Jestädt, T.: Simulation driven development of the German toll system—simulation performance at the kernel and application level. Advances in Business ICT, pp. 1–25. Springer, Heidelburg (2014)
7. Baumann, T.: Simulation driven design of the German toll system—evaluation and enhancement of simulation performance. In: 2012 Federated Conference on Computer Science and Information Systems (FedCSIS). IEEE, pp. 901–909 (2012)
8. Baumann, T.: Simulation driven design of the German toll system—profiling simulation performance. In: 2013 Federated Conference on Computer Science and Information Systems (FedCSIS). IEEE, pp. 923–926 (2013)
9. CEN: ISO/TS 17575–1:2010 electronic fee collection—application interface definition for autonomous systems—part 1: Charging (2010)

10. Bundesministerium der Finanzen: Sollbericht 2013, Monatsbericht des BMF, 2, pp. 6–22, Feb. 2013. http://www.bundesfinanzministerium.de/Content/DE/Monatsberichte/2013/02/Downloads/monatsbericht_2013_02_deutsch.pdf?_blob=publicationFile&v=4. Accessed 20 Mar 2013
11. Pfitzinger, B., Jestädt, T.: Service governance from the perspective of service quality: Notes from the German toll system. In: Smart Services and Service Science, vol. 36, pp. 141–150 (2012)
12. Baumann, T., Pfitzinger, B., Macos, D., Jestädt, T.: Using parameter optimization to calibrate a model of user interaction. In: 2014 Federated Conference on Computer Science and Information Systems (FedCSIS). IEEE (2014)
13. Pfitzinger, B., Jestädt, T., Baumann, T. Simulating the German toll system: Choosing “good enough” driving patterns. In: Proceedings of the mobil.TUM 2013 - International Conference on Mobility and Transport (2013)
14. Bundesamt für Güterverkehr. Mautstatistik Jahrestabellen 2011, Bonn, Feb. 2012. http://www.bag.bund.de/SharedDocs/Downloads/DE/Statistik/Lkw-Maut/Jahrestab_11_10.pdf?_blob=publicationFile. Accessed 10 Mar 2012
15. Dettmar, M., Rottinger, F., Jestädt, T.: Achieving excellence in GNSS based tolling using the example of the German HGV tolling system. ITS Europe, Dublin (2013)
16. Pacholik, A., Baumann, T., Fengler, W., Rath, M.: A model reduction approach for improving discrete event simulation performance. In: Proceedings of the 6th International ICST Conference on Simulation Tools and Techniques. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, Belgium, pp. 101–108 (2013)
17. Zhou, Y., Lee, E.A.: Causality interfaces for actor networks. ACM Trans. Embed. Comput. Syst. (TECS) **7**(3), 29 (2008)
18. Pacholik, A., Baumann, T., Fengler, W., Grüner, D.: Discrete event simulation performance—benchmarking simulators. In: International Simulation Multi-Conference (SummerSim), Genoa, Italy (2012)
19. Fishman, G.S.: Discrete-Event Simulation: Modeling, Programming and Analysis. Springer, Berlin (2001)
20. Albrecht, M.C.: Introduction to discrete event simulation. http://www.albrechts.com/mike/DES/Introduction (2010)
21. Graham, S.L., Kessler, P.B., Mckusick, M.K.: Gprof: a call graph execution profiler. In: ACM Sigplan Notices—Proceedings of the 1982 SIGPLAN Symposium on Compiler Construction, vol. 17, no. 6, pp. 120–126, ACM (1982)
22. Theiss, B.: Der große Netztest in Deutschland (2010)
23. Theiss, B.: Mobilfunk in Deutschland—Der Netztest 2012 (2012)
24. Intel: Intel VTune amplifier. http://software.intel.com/en-us/intel-vtune-amplifier-xe. Accessed 12 May 2013

Analysis of Content of Posts and Comments in Evolving Social Groups

Bogdan Gliwa, Anna Zygmunt and Piotr Bober

Abstract Data reflecting social and business relations has often form of network of connections between entities (called social network). In such network important and influential users can be identified as well as groups of strongly connected users. Finding such groups and observing their evolution becomes an increasingly important research problem. Analyzing the evolution of communities is useful in many applications such as marketing, politics or public security domains. One of the significant problems is to develop method incorporating not only information about connections between entities but also information obtained from text written by the users. Method presented in this chapter combine social network analysis and text mining in order to understand groups evolution. Presented approach to the group evolution process takes many aspects of the group analysis into consideration. Due to proposed method the subjects discussed within the groups are known. We noticed that subjects discussed within groups play significant roles in group evolutions.

1 Introduction

In today's world it's hard to imagine doing business without the use of virtual reality: many companies have moved there the whole business, others try to combine running traditional business with virtual. However, all companies are trying to analyze the behavior of their customers, watching their activity in social media. Since writing

B. Gliwa · A. Zygmunt (✉) · P. Bober
Department of Computer Science, AGH University of Science and Technology,
30-059 Kraków, Poland
e-mail: azygmunt@agh.edu.pl

B. Gliwa
e-mail: bgliwa@agh.edu.pl

P. Bober
e-mail: bober@student.agh.edu.pl

M. Mach-Król et al. (eds.), *Advances in ICT for Business, Industry and Public Sector*,
Studies in Computational Intelligence 579, DOI 10.1007/978-3-319-11328-9_3

blogs or comments on someone else's posts, participating in discussions on forums, exchanging our opinions on fanpages of for example telecommunications companies and banks whose products we use, we everywhere leave traces of our activity, which can be analyzed and ably combined with each other.

Trading companies and banks might be interested in finding active or influential people in their environment and to offer them a new product in hope that it will be proposed by them to many others. Identification on time disgruntled people on banks or telecommunications companies fanpages will allow to respond quickly and prevent the spread of discontent.

The data about different types of dependencies can be modeled as a network of relationships and its structure can be analyzed using Social Network Analysis methods (e.g. for finding important or influential nodes). Such a network, however, is not homogeneous and one can distinguish groups of people, for example, who more often exchange opinions. Such groups frequently are formed around important individuals and, for various reasons, the groups continue to exist or not, grow, shrink or can be joined with other groups. To understand causes of such events, which significantly affect the behavior of groups, it is important to include—besides the fact that someone is connected with someone—information that we can extract from the content of opinions or comments that are left behind (e.g. what they are talking about). If the group is talking about the same themes (topics), does it affect its longer duration? Or, perhaps has a variety of discussed topics stronger impact on the duration of groups? How the themes of discussion are changing in a group? Is a small group with a strong leader more durable than a large one with few strong individuals?

Such knowledge derived from open sources can be combined, for example, with information about the history of bank transfers or loans as well as data about phone calls. To which of our client should be directed attractive offer? Certainly to the one who has a lot of friends, participates in many long-lasting discussions in different groups of people, and so ensures that our offer reaches a large group of potential customers. Proposed methods and algorithms have been tested on one of the largest and highly dynamic polish blogosphere: salon24.pl, in which the main topic of discussion are political issues. However, they can be applied to other social media such as, for example, different blogs, microblogs (e.g., Twitter), opinion websites (e.g. Epinions).

The structure of the rest of chapter is as follows. In Sect. 2 the state of the art concerning the blogosphere and social network analysis, groups and their dynamics and text mining in the context of social network analysis is given. Section 3 describes the main idea and detailed solutions of the process of the group analysis with subjects identification. Dataset, description of experiments and obtained result are shown in the Sect. 4. Finally, the Sect. 5 concludes and presents plans of future works.

2 Related Work

2.1 Blogosphere and Social Network Analysis

Internet social media such as online social networking (e.g., Facebook,[1] Myspace,[2]) blogging (e.g., HuffingtonPost,[3]) forums, media sharing systems (e.g., YouTube,[4] Flickr,[5]) microblogging (e.g., Twitter,[6]) wikis, social news (e.g., Digg,[7] Slashdot,[8]) social bookmarking (e.g., Delicious,[9]) Opinion, Review, and Ratings Websites (e.g., Epinions[10]) has revolutionized the Internet and the way of communication between people. Users stopped being only the consumers of information, becoming the creators of information. Among them, blogs play a special role in creating opinions and information propagation. According to [22], blog is a journal-like website for users, to contribute textual and multimedia content, arranged in reverse chronological order.

Blog can be treated as a kind of an Internet diary, where an author gives opinions on some themes or describes interesting events and readers comment on these posts. A typical entry on a blog can consist of text, photos, films and links to other blogs or web pages. Posts can be categorized by tags. A very important element of blogs is the possibility of adding comments, which enables discussions. Access to blogs is generally open, so everybody can read the posts and comments. Nowadays, many blogs are conducted by experts in the field, so they have become a reliable source of information.

Basic interactions between bloggers are writing comments in relation to posts or other comments. Generally, the relationships between bloggers change over time and are temporal: lifetime of a post is very short [1]. Of course, it depends strongly on the domain of discussion: the more controversial discussions, the more intensive but generally shorter.

Based on blogs, posts and comments, we can build a network where nodes are bloggers or posts and there is an edge between two nodes if one user writes the comment to the other user. Such network can be analysed by Social Network Analysis (SNA) methods [8]. The SNA approach provides measures (SNA centrality measures) which make it possible to determine the most important or influential nodes (bloggers) in the network. Around such bloggers, the groups are forming, sharing similar interests.

[1] http://www.facebook.com.

[2] http://myspace.com.

[3] http://www.huffingtonpost.com.

[4] http://www.youtube.com.

[5] http://www.flickr.com.

[6] https://twitter.com.

[7] http://Digg.com.

[8] http://slashdot.org.

[9] http://delicious.com.

[10] http://www.epinions.com.

2.2 Groups and Their Dynamics

Social networks, built on the basis of the interactions in social media, consist of groups (communities). One can say that the groups are the key structural characteristics of networks [13]. They reflect the structure of relationships in social media: the more frequently users are discussing among themselves on a blog, the higher probability that they will form a common group. It is difficult to find one definition of a group, but in most cases a group is treated as densely interconnected nodes (but loosely to other nodes in network).

Many methods of finding groups have been proposed [12, 29]; discovered groups may be separate (node belong only to one group) or overlapping (node can be a member of many groups). In the case of blogs, more natural approach is using algorithms finding overlapping groups, because given user can participate in many discussions not necessarily with the same people. Popular representative of this class of algorithms is Clique Percolation Method (CPM) [26, 27] which is based on finding in a graph k-cliques, which means a complete, fully connected subgraph of k vertices, where every vertex can be reached directly from all other vertices. Groups are treated as sets of adjacent k-cliques (having k-1 mutual vertices).

Due to the dynamic nature of the social media, a growing interest in developing algorithms for extracting communities that take into account the dynamic aspect of the network has been observed. A method of tracking groups over time was proposed in [21]. First, a division into time steps is carried out. At each step, the graph is created and groups are extracted. Groups from consecutive time steps are matched using the Jaccard index and defined a threshold above which the continuation of the group is assumed. Palla et al. in [25] identified basic events that may occur in the life cycle of the group: growth, merging, birth, construction, splitting and death. Asur in [3] introduced formal definitions of five critical events. Gliwa et al. proposed in [14] two additional events and gave formal definitions. In [17, 18] new tool GEVi for context-based graphical analysis of social group dynamics was proposed.

For further analysis of the different characteristics describing the communities and their transformation in time [31] are calculated, which concerns the comparison of the strength of internal relations of group members with their external connections with nodes outside the group, density of connections in the group or stability of the membership in time.

In [28] incremental community mining approach is introduced: instead of extracting the groups at each time slot independently, they introduce L-metric which allows measure the similarity between groups in the subsequent time slots directly.

Analyzing the evolution of communities is useful in many applications such as marketing, politics or public security domains.

2.3 Text Mining in Social Networks

Aggarwal and Wang in [2] analysed a set of text mining methods in terms of usefulness in social networks analysis (SNA). Bartal et al. [4] proposed a method for link prediction in a social network with usage of text mining features. Velardi et al. [30] described content-based model for social network and edges in such a network were created based on similarity of text mining features between nodes.

One of the most important methods of text mining is Topic Modeling [23], a statistical technique that uncovers abstract "topics" in a collection of documents. Notion of "topic" can be defined as a set of words that co-occur in multiple documents, and, therefore, they are expected to have similar semantics. Two main branches of Topic Modeling [9] can be distinguished—algorithms based on Singular Value Decomposition such as Latent Semantic Indexing [23] and algorithms based on probabilistic generative process [5] such as Latent Dirichlet Allocation (LDA) [6, 7]. Topic modeling techniques aim to reduce dimensionality by grouping words with similar semantics together.

In [19, 20] the application of topic modeling to analysis of groups dynamics in social networks is presented and uncovered topics have manually assigned labels. Different approach is described in [24] where authors analyse topics in time and assignment of labels for them is conducted in an automatic way. Diesner et al. [11] used social network analysis methods to identify users in different roles assessing their attitude to change (i.e. roles: change agents and preservation agents) in the innovation diffusion network and topic modeling was utilized to describe their different characteristics. Cuadra et al. [10] presented new method of community detection in social networks based on similarity of users messages using topic modeling method.

3 General Concept

In this section we provide the concept of our method of analysis topics of groups and their impact on group behaviour. Analysis consists of the several steps. The introductory phase is gathering data from social media services (e.g. blogosphere) into repository (e.g. database). Collected data is divided into series of time slots, so each time slot contains static snapshot of the state of network from defined period of time. In each such slot the static groups are generated and their dynamics in time is inspected. So the further analysis comprises the following steps:

1. generating transitions between groups from neighbouring time slots (also finding stable groups) and identification of their type (group dynamics),
2. for each group assigning the set of topics discussed by group members (based on the content of messages exchanged between members of a group),
3. analyzing changes of topics during group transitions.

Measures that show very well the global user activity in blogosphere are: *lifetime of a post* and *reaction time for a post*. They can be used both in characterizing blogosphere and in determining the length of time slot and the length of overlapping.

3.1 Lifetime of a Post

The lifetime *lt* of a post *p* can be depicted in the following way

$$lt_p = \underset{i}{max}(t_{c_i}) - t_p \tag{1}$$

where t_p is the date when post *p* was published and t_{c_i} are dates of comments in the thread of post *p*.

In other words, lifetime of a post is the range of time between writing the post and the last comment for that post.

3.2 Reaction Time for a Post

The reaction time *rt* for a post *p* can be formulated as (symbols used below in the definition were explained above)

$$rt_p = \underset{i}{min}(t_{c_i}) - t_p \tag{2}$$

Reaction time for a post is the range of time between writing the post and the first comment for that post.

3.3 Groups Dynamics

In order to analyse groups dynamics, overall time range was divided into smaller periods of time (called *time slots*). Next step is to build static networks in each time slot and discover groups in them. To identify events between groups from the neighbouring time slots SGCI method [16, 17] was utilized, which comprise the following stages: identification of short-lived groups in each time slot, identification of group continuations (groups transitions), separation of the stable groups (lasting for at least certain time interval) and the identification of types for group changes (labels for transitions between the states of the stable group).

Identification of continuation between groups *A* and *B* (from neighbouring time slots) is performed using *MJ* measure

$$MJ(A,B) = \begin{cases} 0, & \text{if}\, A = \emptyset \vee B = \emptyset, \\ max\left(\frac{|A \cap B|}{|A|}, \frac{|A \cap B|}{|B|}\right), & \text{otherwise.} \end{cases} \tag{3}$$

and if the calculated value is above predefined threshold th (in experiments we set $th = 0.5$) and the ratio of groups size

$$ds(A,B) = max\left(\frac{|A|}{|B|}, \frac{|B|}{|A|}\right) \tag{4}$$

is below predefined threshold mh (in tests $mh = 50$), then we assumed that group B is a continuation of group A.

Let us define G_i as a set of groups in a time slot i. Then we can formally define transition $t_{m,n}$ between group g_m from slot k and group g_n from slot $k+1$ as

$$t_{g_m,g_n} : \exists g_m \in G_k \wedge \exists g_n \in G_{k+1} \wedge MJ(g_m, g_n) \geq th \wedge ds(g_m, g_n) < mh$$

The SGCI method identifies such event types as:

- **split**, takes place when group divides into several groups in next time slot,
- **deletion**, similar to split, but it happens when small group detaches from significantly bigger one (difference of sizes should be at least 10 times),
- **merge**, when several groups in the previous time slot join together and create larger group,
- **addition**, similar to merge, but it describes situation when small group attaches to significantly bigger group (difference of sizes should be at least 10 times),
- **split_merge**, when for the predecessor group the event is split and for the successor group of given transition the event is merge in the same time,
- **decay**, when analysed group does not exist in the next time slot,
- **constancy** means simple transition without significant change of the group size (in tests changes of group size should be smaller than 5 %),
- **change_size**—simple transition with the change of the group size.

3.4 Topics in Groups

Topics for groups were examined based on clusters found by LDA method. Presented method for analysis topics in groups was used by us in [17, 19, 20].

Firstly, we applied LDA method provided by mallet tool[11] for all posts and, as a result, the method identified 350 clusters of words. Next, we manually annotated each cluster by set of topics and joined similar clusters into bigger ones. The next step was inferring in every comment a set of topics that are referenced by analysed comment (the network is being built based on writing comments in response to other

[11] http://mallet.cs.umass.edu/.

message). Finally, we assigned for the group a set of topics discussed by members of this group (the following condition should be met in order to assign such topic for the group: a topic in a group should be present in at least 5 % of all interactions inside such group).

To describe the activity of topics in groups we formulated *topic exploitation* for given topic and group as a ratio between number of group messages on certain topic and all messages for this group:

$$topicExploitation_k = \frac{|T_k|}{\sum\limits_{i=1}^{n} |T_i|} \tag{5}$$

where: T_k—set of messages (posts and comments) for which topic with number k was inferenced, n—number of all topics, $|T_i|$—amount of elements in T_i.

3.5 Topics Changes in Groups

Topics changes during transition between groups are assessed using introduced following metrics:

- *Overall change in topic exploitation* for transition $t_{m,n}$ from m-th to n-th group is calculated as:
$$c_{m,n} = \sum_i |g_{m,i} - g_{n,i}| \tag{6}$$
where: i is a number of topic and $g_{m,i}$ is the topic exploitation of i-th topic for m-th group (the same denotations are used in definitions below).
- *Maximal positive change of single topic* (how much a topic gained) for transition $t_{m,n}$ from m-th to n-th group is calculated as:
$$mpc_{m,n} = \max\{g_{n,i} - g_{m,i}\} \tag{7}$$
- *Maximal negative change of single topic* (how much a topic lost) for transition $t_{m,n}$ from m-th to n-th group is calculated as:
$$mnc_{m,n} = \max\{g_{m,i} - g_{n,i}\} \tag{8}$$

Using above metrics we can analyse effect of different evolution types on topics change. Therefore, for each evolution type the average values of above defined measures for all groups are evaluated and we refer to them as *Average overall change in topic exploitation*, *Average maximal positive change of single topic* and *Average maximal negative change of single topic* respectively.

3.6 Migrations of Users Depending on Topics

To analyse difference in topics between given user and given group, we defined *topic divergence*, which has the following form:

$$m_t = t_{group} - t_{user} = \sum_{i=1}^{n} |(topic_{i,user} - topic_{i,group})|$$

where: n is a number of all aggregated topics in model, t_{group} is set of weights of each topic for given group, $topic_{i,group}$—weight of i-th topic for given group, t_{user} is set of weights of each topic for given user, $topic_{i,user}$ is weight of i-th topic for given user and weights for topics are determined based on topic exploitation measure for a group or an user.

The minimal value of m_t is 0.0 when user and a group has identical weight for every topic and maximal value is 2.0 when they are totally different. Maximum value of 2.0 is connected with the fact that group might cover topic X in 100 % and user might cover topic Y in 100 %, and therefore difference between group and user on topic X is 100 % and on topic Y is also 100 % which adds up to 200 %.

Using this measure, we carried out experiments focused on relations between *topic divergence* and migrations of users (leaving and joining to groups). For this purpose the following measures are utilized:

- *Probability of leaving the group*. We assumed that potentially any member can leave the group. This value is calculated as:

$$P_l(m) = \frac{|leavers_m \cap candidates_m|}{|candidates_m|}$$

  where: $leavers_m$ are users that in fact left any group and had the value of *topic divergence* measure equals m; $candidates_m$ are members of groups that have *topic divergence* = *m*.
- *Probability of joining the group*. We assumed that candidates for joining are all users that were active in the previous time slot. This value is calculated as:

$$P_j(m) = \frac{|joiners_m \cap candidates_m|}{|candidates_m|}$$

  where: $joiners_m$ are users that in fact joined any group and had the value of *topic divergence* measure equals m; $candidates_m$—users active in previous time slot with *topic divergence* = *m*.

During calculations of joiners and leavers sets we considered all group continuations to be a single group. The reason for that is to prevent *deletion* event to distort results—if a group splits into multiple small groups and we are assuming that anyone from the group can leave, then we will get very high accuracy from each event when huge group changes into a small group.

4 Results

Experiments are conducted on datasets obtained from Polish blogosphere. Blogosphere is analysed from different points of view: especially in terms of users activity, groups formation and dynamics, topics discussed by users in groups.

4.1 Data Sets Description and Characteristics

The dataset contains data from the portal *Salon24* [12] (Polish blogosphere). This portal comprises blogs from different subjects, but political ones constitute the largest part of them. The data from this dataset is from time range 1.01.2008–6.07.2013.

Detailed characteristics of the dataset is presented in Table 1. We can see that on average posts are much longer than comments (almost 10 times). Moreover, commenting others is highly popular which can be perceived by the average number of comments per person (193.09). We can also observe that in general threads are quite long (on average 18.65 comments per post).

The whole period of time is divided into overlapping time slots, each lasting 7 days with overlap equals 4 days. After this operation dataset contains 504 slot.

In every time slot a static network is built [15], where the users are nodes and relations between them are built in the following way: from user who wrote the comment to the user who was commented on or, if the user whose comment was commented on is not explicitly referenced in the comment (by using @ and name of author of comment), the target of the relation is the author of post.

4.2 Lifetime of Posts

Figure 1 shows *life time* of posts, i.e. time which elapsed from publish date and the moment when last comment to given post is written. As it can be seen life time of posts in Salon24 is very short—most of posts live up to one day.

4.3 Reaction Time for Posts

Figure 2 shows *reaction time*–time elapsed from post publish to the first comment. For Salon24 there is a tendency to write comments as soon as the post is published—most posts have comments within first hour. Both lifetime of a post and reaction time for posts show that this dataset is very dynamic and real world event have fast reflection in blogosphere.

[12] www.salon24.pl.

Table 1 Data description

	Salon24
Nr of posts	380,700
Nr of comments	5,703,140
Nr of tags	176,777
Avg nr of posts per author	37.58
Avg nr of comments per person	193.09
Avg nr of comments to a post	18.65
Avg post length (characters)	3,165.3
Avg comment length (characters)	350.72

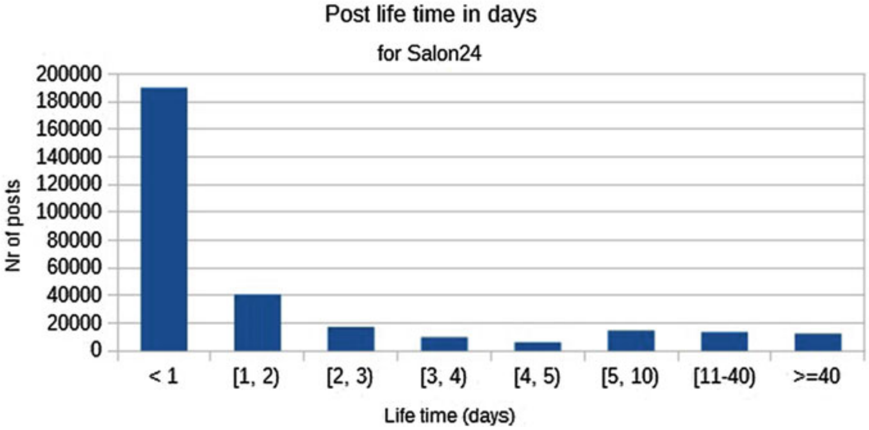


Fig. 1 Post life time

4.4 Groups and Their Dynamics

For group extraction we used CPM method (CPMd version which is designed to discover groups in directed networks) from CFinder[13] tool for k equals 3. Transitions between groups were assigned using our method SGCI described earlier.

Figure 3 presents how number of stable groups is influenced by group size. The general characteristics is the smaller group the bigger number of those groups.

Histogram of events taking places is shown in Fig. 4. Two very popular events are: change size, which seems natural because few users may join the group and some may leave, and decay—it is also expected as each group must finish at some point. Quite big values of addition and deletion may suggest that users of Salon24 are dynamic—they join discussions but also decide to leave threads.

[13] www.cfinder.org.

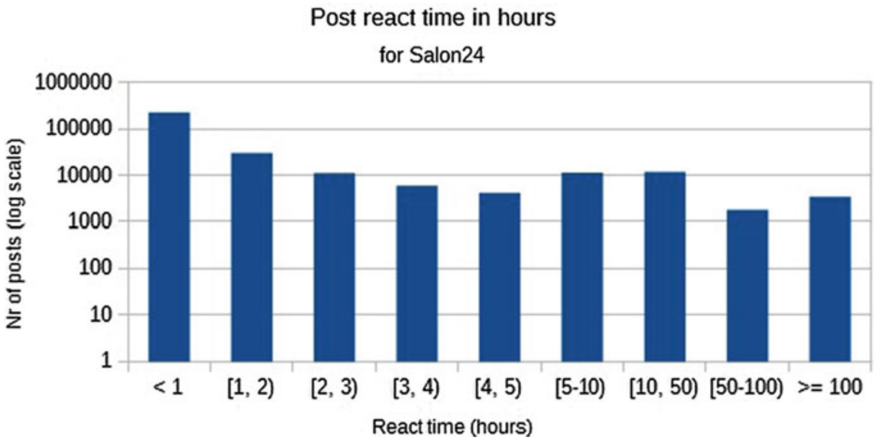


Fig. 2 Post reaction time

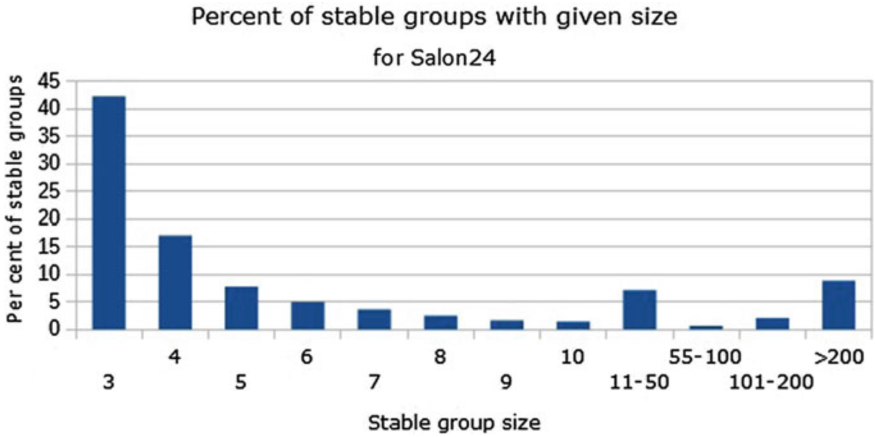


Fig. 3 Number of stable groups

4.5 Topics in Groups

Figure 5 shows categories of topics groups are talking about. Each group may discuss different topics belonging to different categories. As it can be seen, the most popular categories are *economics* and *politics*. Next are *other*—collection of topics which are difficult to put in one category. High value for *disaster* is most probably connected with Tu-154 crash in Smolensk—this crash is still widely commented in Polish media.

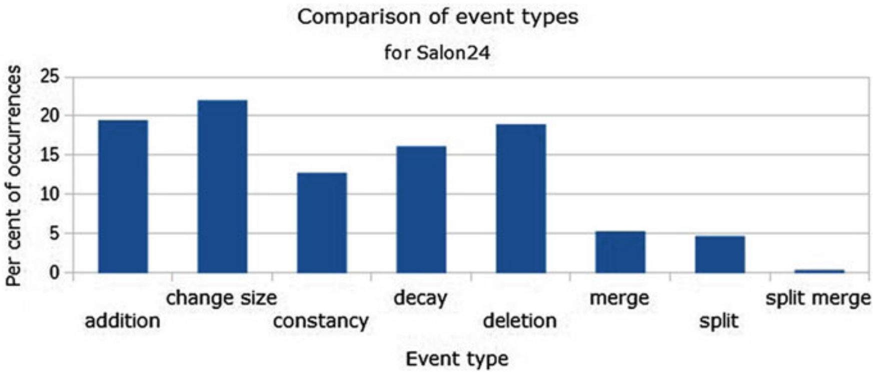


Fig. 4 Number of events found

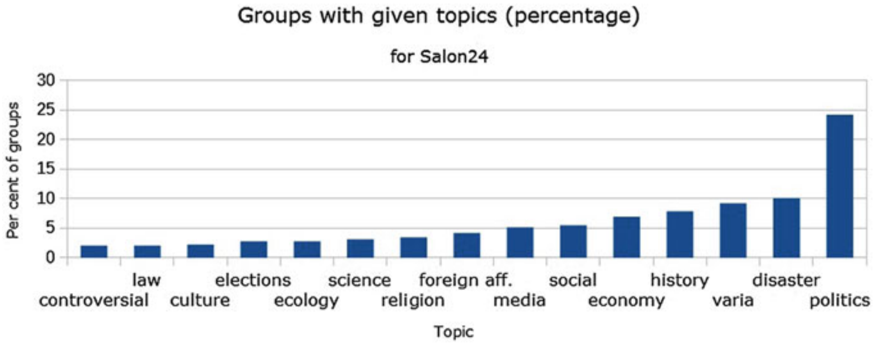


Fig. 5 Percent of groups talking about particular topics

Number of topics discussed by groups depending on their size is presented in Fig. 6. Generally, in Salon24 groups have quite a few topics. A tendency that bigger group means lower number of discussed topics can be explained by 5 % threshold the topic must achieve to be counted. In big groups there are more topics but with lower weight which leads to only few ones that exceed the limit.

Topic convergence between users and groups they belong to is depicted in Fig. 7. Convergence is calculated in cosine measure where 0 means total divergence and 1 total convergence. As it could be expected the lowest value is for low convergence. Definitely in more than 50 % cases the convergence is higher than 0.5. Lower value for convergence between 0.9 and 1 and extremely low for exactly 1 (total convergence) is obvious—users creating a particular group put their topics into the group, but do not share all of them.

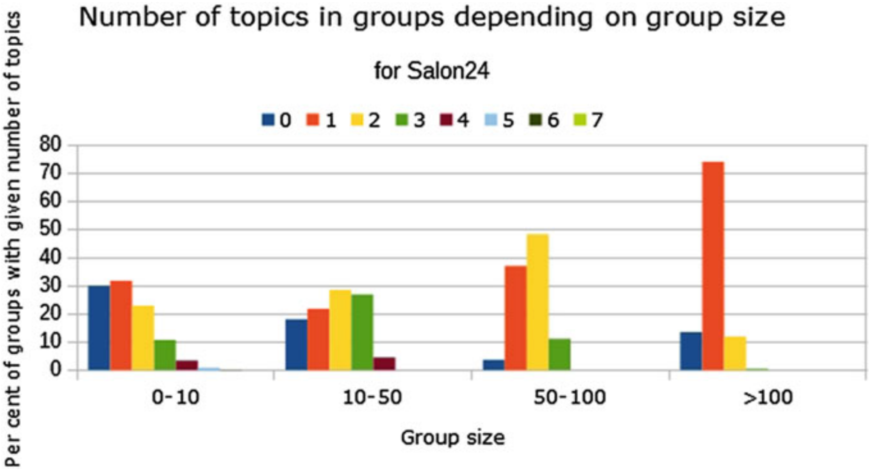


Fig. 6 Number of topics depending on group size for Salon24

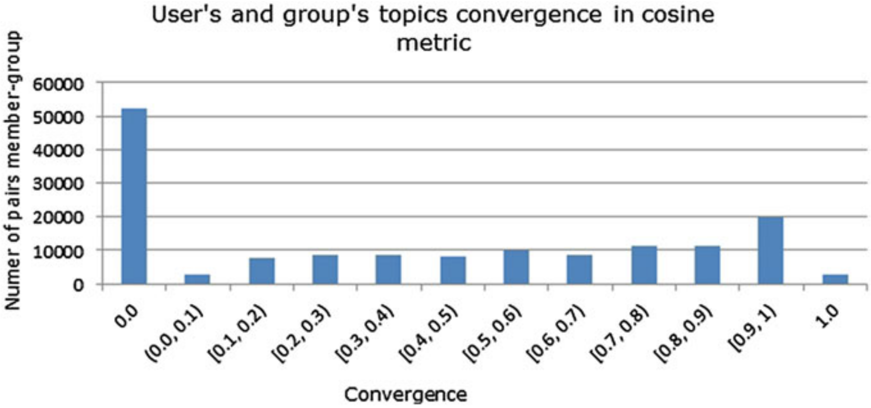


Fig. 7 Convergence between user's and group's topics for Salon24

4.6 Topic Changes

Figure 8 shows average topic change depending on event type in half-year periods. As expected, *constancy* and *change size* have similar topic change. *Addition* and *deletion* also have similar values of change and are the highest ones. *Merge* and *split* cause medium topic change. As we can see, the biggest changes of topics happen during interaction groups with big difference in size (events *deletion* and *addition*), so it may suggest that interests of people are quite stable and changes are mostly related with exchange of members in groups.

Figure 9 shows maximal positive topic change. One can observe that biggest topic gain is achieved during event *deletion*, and the smallest one-event *constancy*.

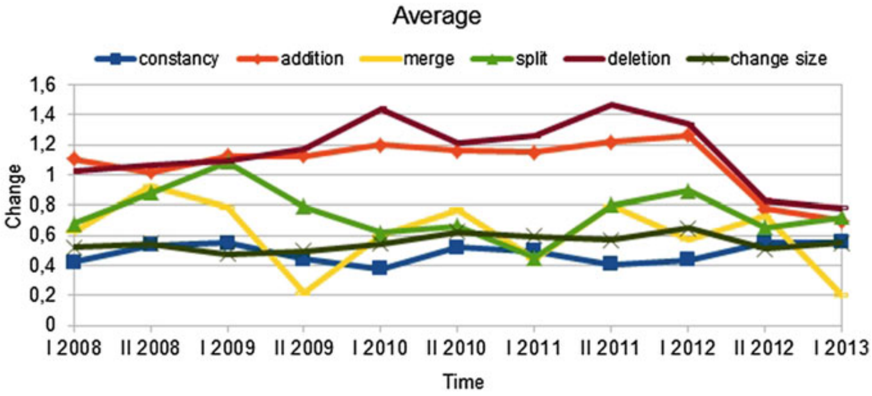


Fig. 8 Average topic change in half-year slots for Salon24

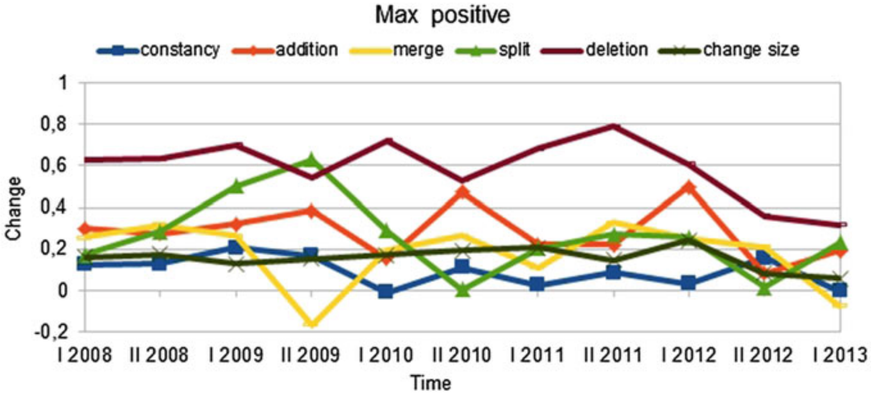


Fig. 9 Maximal positive topic change in half-year slots for Salon24

Figure 10 depicts maximal negative topic change. In this case the event that mostly changes topics is *addition*. Furthermore, changes caused by the events *merge* and *split* are quite similar (even more similar than these events for the maximal positive topic change).

Generally, *addition* is connected with: highest overall change in topics and highest negative change. Therefore, we can deduce that when a small group attaches to a single large group it is usually connected with significant drop of popularity of the main topic of the group, and small rise in popularity of different topics—presumably of main topic of the other group.

Deletion cause large overall topic change and large positive change. It means that splitting a group causes a rise of popularity of a single topic at the expense of all the others.

Change_size and *constancy* have the smallest effect on topics changes.

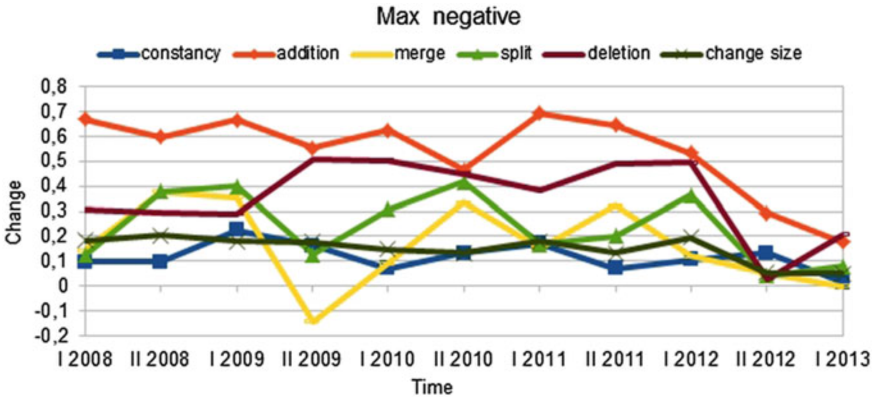


Fig. 10 Maximal negative topic change in half-year slots for Salon24

Split has higher values for average and maximal negative change than *merge*. Moreover, when *split* achieves high values in terms of maximal positive topic change then it has rather small values in terms of maximal negative topic change. Different behaviour is for *merge* event—in that situation high values for maximal positive topic change corresponds with high values of maximal negative change. It may suggest that *merge* has averaging behaviour—after merge one topic gain more and other one decrease more.

4.7 Users Migrations

Figure 11 presents probability of joining to a group and leaving it based on topic divergence and Fig. 12 shows numbers of people who actually did it. The value equals 0 for divergence means that a user talks about exactly the same topics as a group does and the value equals 2 means completely different topics. It is surprising that the highest convergence of topics does not mean high probability of joining [absolute number of such events is outstanding for total convergence (0)], but there are also some points (such as 0.05 and 0.4) where high convergence corresponds with higher probability of joining group, however, numbers of people with such values of convergence are rather small. We can observe different behaviour for leaving group by its members. Leaving the group is more probable when divergence is low and slightly increases for bigger divergence (however, for small values of topic divergence there are few members with such values of this measure, except the point 0 which means total convergence).

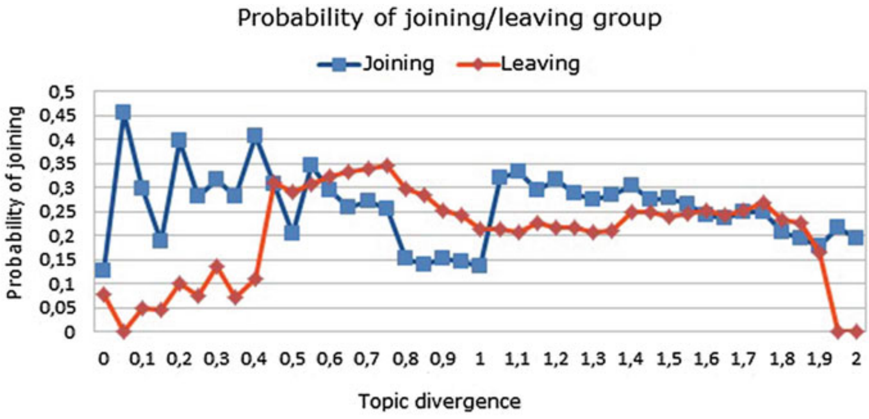


Fig. 11 Migrations of users depending on topic similarity between user and group—probability

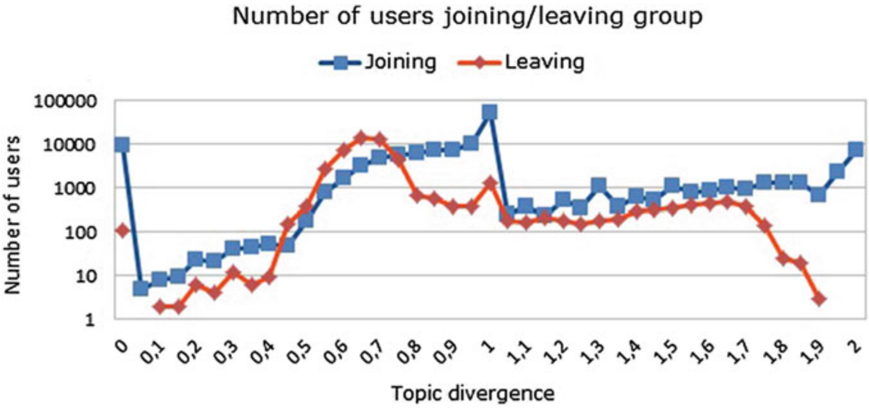


Fig. 12 Migrations of users depending on topic similarity between user and group—numbers

4.8 Sample Lifecycle of Groups

Figure 13 (screenshot from GEVi tool [18]—tool for visualisation of groups evolution) and Fig. 14 present sample group and its evolution in time. The notation in each boxes includes:

- slot (time period) and group name,
- 6 most important singular topics,
- 4 most important categories (aggregated and labeled topics),
- users belonging to the group.

Red colour indicates element that appears in next slot, green—element that does not go to the next period. Underline is used for elements that are new in current

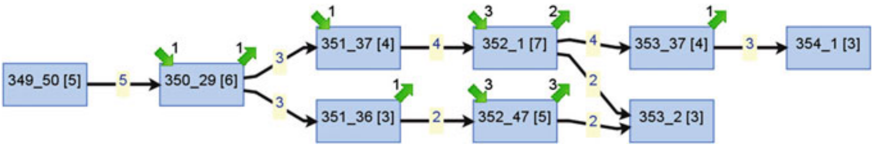


Fig. 13 Sample group transitions for Salon24 in GEVi

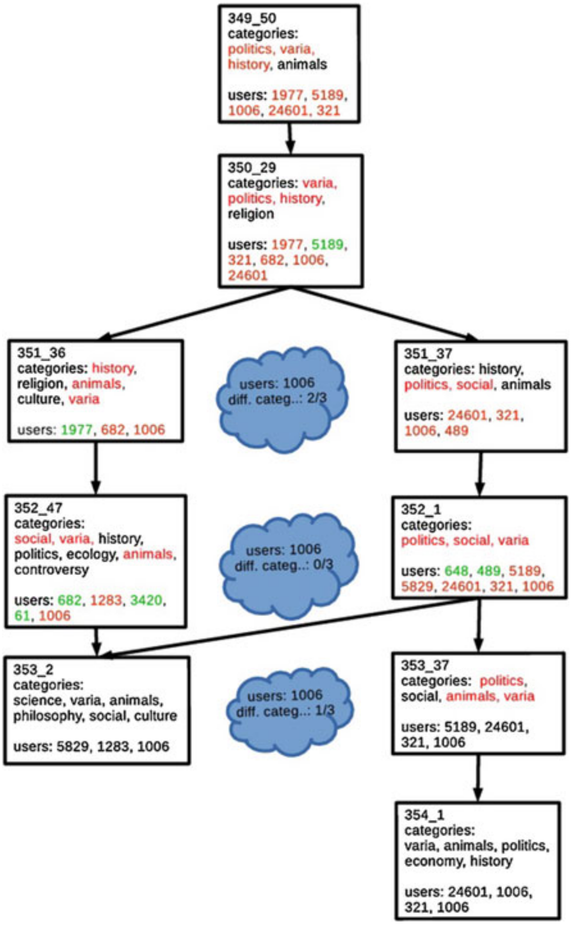


Fig. 14 Sample group transitions for Salon24

group (not used for topics and categories, since only most important are listed). Cloud presents: ids of common users and number of extra categories and topics each group has (e.g. 6/5 means that upper group has 6 more topics than common ones, and lower has 5 more).

There is one group of 5 users that split in two in a second slot. One user stays in both, split groups. There are no clear topic transitions from one slot to other, however categories suggest some consequence in transitions. At the beginning group discuss philosophy, media and politics. Next, controversy and sport become most important, but politics appears as well. After split, upper group's first category is history, whereas in lower is sport. It is important to notice, that history also appears but with less importance. In next slot there are more differences between those groups: upper begins talking about animals, sport, law and religion, and continue those categories in next slots (animals appears in next slot, religion in last). Lower group talks about economy, history and food, and mainly those categories are in its next slot. It can be seen that one member from upper group and two from lower merges in slot 353.

5 Conclusion

This paper presents analysis of communities and their topics for real-world data from blogosphere. We carried out experiments regarding relations between topics discussed by members of groups and the behaviour of groups. Most important findings in this work are some patterns of topics changes during different group evolution events. Moreover, we assessed convergence topic interests of users and groups they belong to. Furthermore, we examined influence of topic differences between an user and a group on users migrations. Example of scenario for analysis topics during groups evolution was also presented.

Future work can be performed in a few different directions. Firstly, we are planning to conducts experiments on different datasets (e.g. in English language) and compare results, especially in terms of the patterns of topics changes during groups evolution. The second direction of further research is the analysis of key persons in terms of discussed topics and their influence on changes of topics during discussions. The third direction is to incorporate knowledge about topics interest of user to our method of predicting group behavior [14].

References

1. Agarwal, N., Liu, H.: Modeling and Data Mining in Blogosphere. Moegan & Claypool Publishers, US (2009)
2. Aggarwal, C., Wang, H.: Social network data analytics. In: Aggarwal, C. (ed.) Text Mining in Social Networks, pp. 353–378. Springer, New York (2011)
3. Asur, S., Parthasarathy, S., Ucar, D.: An event-based framework for characterizing the evolutionary behavior of interaction graphs. ACM Trans. Knowl. Discov. Data **3**(4) (2009)
4. Bartal, A., Sasson, E., Ravid, G.: Predicting links in social networks using text mining and sna. In: Social Network Analysis and Mining, 2009. ASONAM '09. International Conference on Advances in, pp. 131–136 (2009). doi:10.1109/ASONAM.2009.12
5. Blei, D.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)

6. Blei, D., Lafferty, J.: Dynamic topic models. In: Proceedings of the 23rd international conference on machine learning, p. 113120 (2006)
7. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 9931022 (2003)
8. Carrington, P., Scott, J., Wasserman, S.: Models and Methods in Social Network Analysis. Cambridge University Press, Cambridge (2005)
9. Crain, S., Zhou, K., Yang, S., Zha, H.: Mining Text Data. In: Aggarwal, C., Zhai, C. (eds.) Dimensionality reduction and topic modelling: from latent semantic indexing to latent dirichlet allocation and beyond, pp. 129–162. Springer, New York (2012)
10. Cuadra, L., Rios, S., L'Huillier, G.: Enhancing community discovery and characterization in vcop using topic models. In: 2011 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 3, pp. 326–329 (2011). doi:10.1109/WI-IAT.2011.97
11. Diesner, J., Carley, K.: A methodology for integrating network theory and topic modeling and its application to innovation diffusion. In: 2010 IEEE Second International Conference on Social Computing (SocialCom), pp. 687–692 (2010). doi:10.1109/SocialCom.106
12. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
13. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
14. Gliwa, B., Bródka, P., Zygmunt, A., Saganowski, S., Kazienko, P., Kozlak, J.: Different approaches to community evolution prediction in blogosphere. In: ASONAM 2013: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining: Niagara Falls, Turkey (2013). Accepted for printing
15. Gliwa, B., Kozlak, J., Zygmunt, A., Cetnarowicz, K.: Models of social groups in blogosphere based on information about comment addressees and sentiments. In: Social Informatics—4th International Conference, Social Informatics, Lausanne, Switzerland, Lecture Notes in Computer Science, vol. 7710, pp. 475–488. Springer (2012)
16. Gliwa, B., Saganowski, S., Zygmunt, A., Bródka, P., Kazienko, P., Kozlak, J.: Identification of group changes in blogosphere. In: ASONAM 2012: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Turkey (2012)
17. Gliwa, B., Zygmunt, A.: Gevi: context-based graphical analysis of social group dynamics. Soc. Netw. Anal. Min. **4**(1), 1–15 (2014)
18. Gliwa, B., Zygmunt, A., Byrski, A.: Graphical analysis of social group dynamics. In: CASoN, pp. 41–46. IEEE (2012)
19. Gliwa, B., Zygmunt, A., Koźlak, J., Cetnarowicz, K.: Application of text mining to analysis of social groups in blogosphere. In: 5th Workshop on Complex Networks, CompleNet 2014, Bologna, Italy, 12–14 March 2014
20. Gliwa, B., Zygmunt, A., Podgórski, S.: Incorporating text analysis into evolution of social groups in blogosphere. In: Federated Conference on Computer Science and Information Systems, FedCSIS 2013, Krakow, Poland, 8–11 September 2013
21. Greene, D., Doyle, D., Cunningham, P.: Tracking the evolution of communities in dynamic social networks. In: Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM'10). IEEE (2010)
22. Gundecha, P., Liu, H.: Mining social media: A brief introduction. Tutorials in Operations Research 1,4, Informs. Arizona State University, US (2012)
23. Huang, Y.: Support vector machines for text categorization based on latent semantic indexing. Electrical and Computer Engineering Department, The Johns Hopkins University, Technical report (2003)
24. Nguyen, M., Ho, T., Do, P.: Social networks analysis based on topic modeling. In: IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), pp. 119–122 (2013). doi:10.1109/RIVF.2013.6719878
25. Palla, G., Barabsi, I.A., Vicsek, T., Hungary, B.: Quantifying social group evolution. Nature **446**, 664–667 (2007)

26. Palla, G., bel, D., Farkas, I.J., Pollner, P., Dernyi, I., Vicsek, T.: Handbook of large-scale random networks. In: Bollobs, B., Kozma, R., Mikls, D. (eds.) k-clique Percolation and Clustering. Springer, New York (2009)
27. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**, 814–818 (2005)
28. Takaffoli, M., Rabbany, R., Zaiane, O.R.: Incremental local community identification in dynamic social networks. In: J.G. Rokne, C. Faloutsos (eds.) ASONAM, pp. 90–94. ACM (2013)
29. Tang, L., Liu, H.: Community Detection and Mining in Social Media. Morgan & Claypool, US (2010)
30. Velardi, P., Navigli, R., Cucchiarelli, A., D'Antonio, F.: A new content-based model for social network analysis. In: IEEE International Conference, Semantic Computing, pp. 18–25 (2008). doi:10.1109/ICSC.2008.30
31. Xu, J., Marshall, B., Kaza, S., Chen, H.: Analyzing and visualizing criminal network dynamics: A case study. In: IEEE Conference on Intelligence and Security Informatics. Tuczon (2004)

Generation of Hierarchical Business Process Models from Attribute Relationship Diagrams

Krzysztof Kluza and Grzegorz J. Nalepa

Abstract Business Process models are mostly designed manually. However, they can be generated from other representations, especially from the system specification. Attribute-Relationship Diagrams (ARD) aim at capturing relations, especially dependency relation, between the attributes specified for a particular system. Originally, the ARD method was developed for prototyping rule bases in an iterative manner. We propose to apply this method to business processes. The paper examines the possibility of generating the rule-oriented BPMN model and enriching process models with rules from the ARD diagram. We describe a technique of automatic generation of a BPMN model with decision table schemes for business rule tasks and form attributes for user tasks from the ARD diagram. In our approach, processes and rules are generated simultaneously. Thanks to this, they are complementary and can be directly executed using our hybrid execution environment.

1 Introduction

Business Process (BP) models constitute graphical representations of processes in an organization. Such a process is composed of related tasks that produce a specific service or product for a particular customer [1]. When it comes to practical modeling, Business Process Model and Notation (BPMN) [2, 3] is a standard language for this purpose. The current version of the notation allows for modeling many aspects of business; nonetheless, it is not suitable for modeling some aspects of the enterprise, especially decision rules or constraints [4].

The paper is supported by the HiBuProBuRul Project funded from NCN (National Science Centre) resources for science (no. DEC-2011/03/N/ST6/00909).

K. Kluza (✉) · G.J. Nalepa
AGH University of Science and Technology, Al. A. Mickiewicza 30,
30-059 Krakow, Poland
e-mail: kluza@agh.edu.pl

G.J. Nalepa
e-mail: gjn@agh.edu.pl

M. Mach-Król et al. (eds.), *Advances in ICT for Business, Industry and Public Sector*,
Studies in Computational Intelligence 579, DOI 10.1007/978-3-319-11328-9_4

Recently, the Business Rules (BR) approach has been proposed as a new way of capturing the functional requirements and modeling system logic in a designer-friendly fashion. Moreover, the BR approach is a solution originated from Rule-Based Systems, that are a mature and well established technology. As processes and rules are related to each other [5, 6], BR are often, but not limited to, used for the specification of task logic in process models.

BR can be acquired based on some data using machine learning techniques [7] or generated from natural language specification [8]; however, they often have to be modeled manually based on the knowledge collected from the domain experts, as usually their knowledge is not written down anywhere. Similarly, processes are designed manually as well. However, the simplified process models can be generated using process mining tools [9] or acquired from natural language description using NLP techniques [10]. Other method of acquiring BPMN models is to transform the existing process models in other languages to the BPMN notation. From a researcher's point of view, this can be a challenge in the case that languages are of different paradigm or presents a different aspect of the system, e.g. UML use-case diagrams [11].

In this paper, we present work-in-progress research which is a part of our research concerning process and rules integration [12–14]. We examine the possibility of generation of the rule-oriented BPMN model as well as the possibility of enriching BP processes with rules from the ARD diagram.

As the ARD method allows a domain expert for gradual identification of the properties of a system being designed, we argue that having the system properties identified and described in terms of attributes, it is possible to generate executable BPMN model with the corresponding BR tasks as well as enrich the BP model with such BR tasks. Such an approach would allow for generating business processes and rule schemes for logic task specification at the same time. The generated rule prototypes comply with the XTT2 rule representation [15–18] from the Semantic Knowledge Engineering approach [19].

This paper is an extended version of the paper [20] presented at the 4th International Workshop on Advances in Business ICT (ABICT 2013) workshop[1] held in conjunction with the Federated Conference on Computer Science and Information Systems (FedCSIS 2013) conference in Krakow.

The paper is organized as follows. In Sect. 2 we present the motivation for our research. Section 3 provides a short overview of the related approaches. Section 4 presents the details of the ARD method. In Sect. 5, we give an overview of the proposed method for process model generation and enriching the BP model with BR tasks, and we describe how this approach can be extended in order to generate hierarchical process models. Section 6 presents the tools supporting our approach. Section 7 summarizes the paper.

[1] See: http://fedcsis.org/2013/abict.html.

2 Motivation

The complexity of software has been constantly increasing for the last decades. To deal with this growth, new design methods and advanced modeling solutions are required [21]. For this purpose, modern applications use business processes and business rules as business logic specification [22].

According to the BPMN 2.0 specification [3], the notation is not suitable for modeling such concepts as rules. Therefore, this reveals the challenges in modeling and executing processes with rules. It is so because processes and rules are developed or modeled separately and are not well matched.

Our research aims at developing the integrated method for modeling BP with BR to provide a consistent method for modeling business systems. Such a method will allow for modeling processes with rules in a straightforward way, and then for executing such a developed model.

The main contribution of this paper is a presentation of the possibility of generation of the rule-oriented BPMN model, which can be used for enriching the BP models with BR tasks.

3 Related Work

Several approaches can be considered as related to the method presented in this paper. As our method proposes automatic generation of a BPMN model, the approach can be compared with such approaches as: process mining [9], generating processes from text in natural language (based on NLP methods) [10], or finally transforming process from other notations to BPMN, especially from the notations that are not process-oriented, e.g. the UML use case diagrams [23].

The process mining methods [9] allow for very flexible process models generation, and in some cases this technique does not require any human activity. However, the result of the method is a very general process that is not suitable for direct execution. In order to be an executable BPMN model, it has to be significantly enhanced and refactored. In the case of our method, it is not as much flexible as process mining technique, but it produces a BPMN model which is executable and provides support for Business Rule tasks.

Generating processes from text description provided in natural language [10] can have practical results and allows for generating a high quality BPMN model. High quality models can also be obtained through translation from other representations, such as the UML use case diagrams [24, 25]. Unfortunately, a method based on the natural language description has to be supported by an advanced NLP system, thus practical applications of this method is very complex. Translation from other representations, in turn, requires process models designed using such representations, which often do not exist. In our approach, a process model is generated based on the carefully prepared ARD diagram. Although this requires the ARD diagram, it is very simple model and in some cases it can be obtained from text description using some mining technique to acquire attributes. This requires additional research yet. However, there has been trials of mining such attributes from text in natural language [26].

There are also other approaches, more similar to ours, such as generating business process models from Bill Of Materials (BOM) [27], Product Based Workflow Design (PBWD) [28–31], based on Product Data Model (PDM) [32], Product Structure Tree (PST) [33, 34] or Decision Dependency Design (D3) [35–37] which uses decision structures.

Bill of material (BOM) is a hierarchical assembly structure of parts that constitute a product. Van der Aalst [27] proposed to redesign a process by taking into account not an existing process itself, but a product and its specification like BOM. Reijers et al. [38] extended the method into Product Based Workflow Design. PBWD aims at improving process design based on a static description of a product. Vanderfeesten et al. [30] developed different algorithms for transforming Product Data Model into a workflow process.

Decision Dependency Design [37], in turn, is a method for structuring decisions and indicating underlying dependencies. Wu et al. designed different execution mechanisms from such decision structures.

Such solutions have been applied in different areas [39–41]. However, still better tool support is needed in order to make approaches like PBWD more suitable for practical use [29].

In the case of our approach, apart from the fact that it generates an executable BPMN model, it supports the rule prototypes generation for Business Rule tasks, what makes the BPMN model data consistent with the rule engine requirements for data. Therefore, we claim that our approach is partially rule-based [42].

It is important to mention that the presented approach can be used either for generating a new rule-oriented BPMN model or for enriching the existing process model with rules based on the corresponding ARD diagram.

Although the work presented here is work-in-progress research, the overview of the method reveals significant differences from the techniques mentioned above, especially in the case of method simplicity and support for rules in process models.

4 Attribute Relationship Diagrams

Attribute Relationship Diagram (ARD) [43] constitute a method which allows a user (especially a domain expert) for gradual identification of the system properties during design.

The goal of this method is to capture functional dependencies between attributes. The attributes are expressed in terms of Attributive Logic [19, 44, 45] and denote particular system properties identified by the domain expert. The identified dependencies form a directed graph in which properties are represented as nodes and dependencies are represented as transitions. In the following definitions, we present more formal description of ARD.

A typical atomic formula (or fact) takes the following form

$$a(p) = d$$

where a is an attribute, p is a property and d is the current value of a for p. More complex descriptions take usually the form of conjunctions of such atoms and are omnipresent in the AI literature [46].

An **attribute** $a_i \in A$ is a function (or partial function) of the form:

$$a_i : P \rightarrow \mathbb{D}_i$$

where

- P is a set of property symbols,
- A is a set of attribute names,
- $\mathbb{D}$ is a set of attribute values (the domains).

An example of an attribute can be the `carAge`, which denotes the age of a car, and the attribute value is within the domain $\mathbb{D}_{carAge} = [0, \inf]$.

A **generalized attribute** $a_j \in A$ is a function (or partial function) of the form:

$$a_j : P \rightarrow 2^{\mathbb{D}_j}$$

where $2^{\mathbb{D}_j}$ is the family of all the subsets of $\mathbb{D}_j$

An example of a generalized attribute can be the `ownedInsurances`, which is a set of the customer insurances, and the attribute value is a subset of the domain $\mathbb{D}_{ownedInsurances}$, which consists of the possible insurances that a particular customer can posses.

In the case of abstraction level, the ARD attributes and generalized attributes can be described either as conceptual or physical ones.

A **conceptual attribute** $c \in C$ is an attribute describing some general, abstract aspect of the system.

Conceptual attribute names are capitalized, e.g.: `BaseRate`. During the design process, conceptual attributes are being *finalized* into, possibly multiple, physical attributes.

A **physical attribute** $a \in A$ is an attribute describing a specific well-defined, atomic aspect of the system.

Names of physical attributes are not capitalized, e.g. `payment`. A physical attribute origins from one or more (indirectly) conceptual attributes and can not be further *finalized*.

A **simple property** $p_s \in P$ is a property described by a single attribute.

A **complex property** $p_c \in P$ is a property described by multiple attributes.

A **dependency** $d \in D$ is an ordered pair of properties (f, t), where $f \in P$ is the **independent property** and $t \in P$ is the **dependent property** that depends on f. For simplicity $d = (f, t) \in D$ will be presented as: $dep(f, t)$.

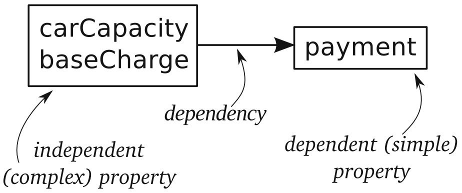


Fig. 1 An example of the ARD diagram

An **ARD diagram** R is a pair (P, D), where P is a set of properties, and D is a set of dependencies, and between two properties only a single dependency is allowed.

To illustrate the ARD concepts, an exemplary ARD diagram with properties and the dependency between them is presented in Fig. 1. The diagram should be interpreted in the following way: `payment` depends somehow on `carCapacity` and `baseCharge`.

The core aspects of the ARD method are diagram transformations, which regard properties and serve as a tool for diagram specification and development. Transformations are required to specify additional dependencies or introduce new attributes for the system. For the transformation of the diagram R_1 into the diagram R_2, the R_2 is more specific than the R_1.

Finalization *final* is a function of the form:

$$final : p_1 \to p_2$$

that transforms a simple property $p_1 \in P$ described by a conceptual attribute into a property $p_2 \in P$, where the attribute describing p_1 is substituted by one or more conceptual or physical attributes describing p_2, which are more detailed than the attribute describing a property p_1.

In Fig. 2, an exemplary finalization transformation is presented. It shows that the simple property `BaseRate` (described by a single conceptual attribute) is finalized into a new complex property described by two physical attributes `carCapacity` and `baseCharge`.

Split *split* is a function of the form:

$$split : p_c \to \{p^1, p^2, \ldots, p^n\}$$

where a complex property p_c is replaced by n properties, each of them described by one or more attributes originally describing p_c. Since p_c may depend on some other

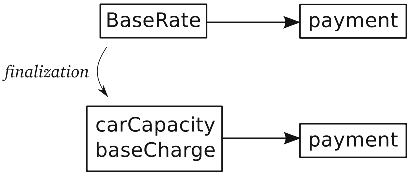


Fig. 2 An example of the ARD finalization transformation

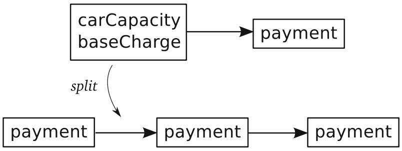


Fig. 3 An example of the ARD split transformation

properties $p_o^1 \dots p_o^n$, dependencies between these properties and $p^1 \dots p^n$ have to be stated.

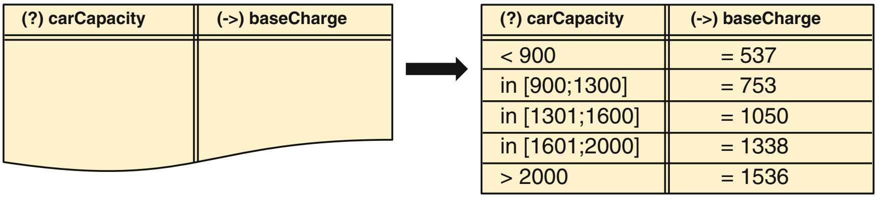
To illustrate this transformation, Fig. 3 shows the complex property described by two physical attributes (`carCapacity` and `baseCharge`), which is split into two simple properties described by these attributes.

Upon splitting and finalization, the ARD model is more and more specific. The consecutive levels of ARD forms a hierarchy of progressively detailed diagrams, which constitutes Transformation Process History (TPH) [47]. The implementation of this hierarchical model is provided through storing the lowest available, most detailed diagram level at any time, and additional information needed to recreate all of the higher levels. Such model captures information about changes made to properties at consecutive diagram levels.

Originally [48], the ARD method was proposed as a method for prototyping a knowledge base structure in a similar way as the relational data base structure (tables) is generated from the ERD diagrams, and as a simple alternative to the classic approaches [49]. Thus, based on the ARD dependencies, it is possible to generate structures (schemes) of decision tables [47]. Then, such schemes can be filled in with rules either manually by a domain expert or automatically mined from some additional data. Figure 4 presents an exemplary schema of a decision table generated from the identified ARD dependencies (specifically from the ARD dependency between two physical attributes `carCapacity` and `baseCharge`, see Fig. 3) and the same table filled in with rules.

4.1 Polish Liability Insurance Case Study

Let us now present an illustrative example of the Polish Liability Insurance (PLLI) case study. The example was developed as one of the benchmark cases for the Semantic Knowledge Engineering (SKE) approach for rule-based systems [19].



(?) carCapacity	(->) baseCharge

(?) carCapacity	(->) baseCharge
< 900	= 537
in [900;1300]	= 753
in [1301;1600]	= 1050
in [1601;2000]	= 1338
> 2000	= 1536

Fig. 4 Decision table schema (*left*) and the same table filled in with rules (*right*)

In the PLLI case study, the price for the liability insurance for protecting against third party claims is to be calculated. The price is calculated based on various reasons, which can be obtained from the domain expert. The main factors in calculating the liability insurance premium are data about the vehicle, such as the car engine capacity, the car age, seats, and a technical examination. Additionally, the impact on the insurance price have such data as the driver's age, the period of holding the license, the number of accidents in the last year and the previous class of insurance junction. Moreover, in the calculation, the insurance premium can be increased or decreased because of number of payment installments, other insurances, continuity of insurance or the number of cars insured.

The calculation of insurance premiums comprises the following steps. The first step is to determine the base rate based on the capacity of the car engine. The second step uses the so called "bonus malus" discounts which depend on the number of accidents in the last year. Insurance companies give up to 60 % discount on this account. The third step is to take into account other discounts or increases based on previously mentioned factors.

All these pieces of data, obtained from an expert, can be specified using the ARD method. The attributes that have been identified for this example are presented in Table 1. Using the ARD method, one can describe the system with the ARD diagram

Table 1 Attributes for the PLLI case study

Name of attribute	Type	Range	Description of attribute
`noAccident`	Integer	[0;inf]	Number of accidents in last 12 months
`class`	Integer	[−1;9]	A customer class
`carCapacity`	Integer	[0;inf]	Capacity of the car engine [cm^3]
`baseCharge`	Integer	[0;inf]	Base insurance charge [PLN]
`driverAge`	Integer	[16;inf]	Age of a driver (owner of the car)
`drLicAge`	Integer	[0;inf]	Period of holding a driving license
`driverDiscount`	Integer	–	Sum of `driverAge` and `drLicAge`
`carAge`	Integer	[0;inf]	Age of the car
`historiCar`	Boolean	[true; false]	Historic car
`noSeats`	Integer	[2;9]	Number of seats in the car
`technical`	Boolean	[true; false]	Current technical examination of the car
`carDiscount`	Integer	–	Sum of discounts: `carAge`, `historiCar`, `noSeats` and `technical`
`noRates`	Integer	[1;2]	Number of instalments
`contIns`	Boolean	[true; false]	Continuation of insurance
`noCarsIns`	Integer	[0;inf]	Number of insured cars
`otherIns`	Boolean	[true; false]	Other insurances
`insHistory`	Integer	[0;inf]	History of the driver insurance
`otherDiscount`	Integer	–	Sum of discounts: `noRates`, `contIns`, `noCarsIns`, `otherIns`, `insHistory`
`payment`	Float	[0;inf]	Charge for the car insurance

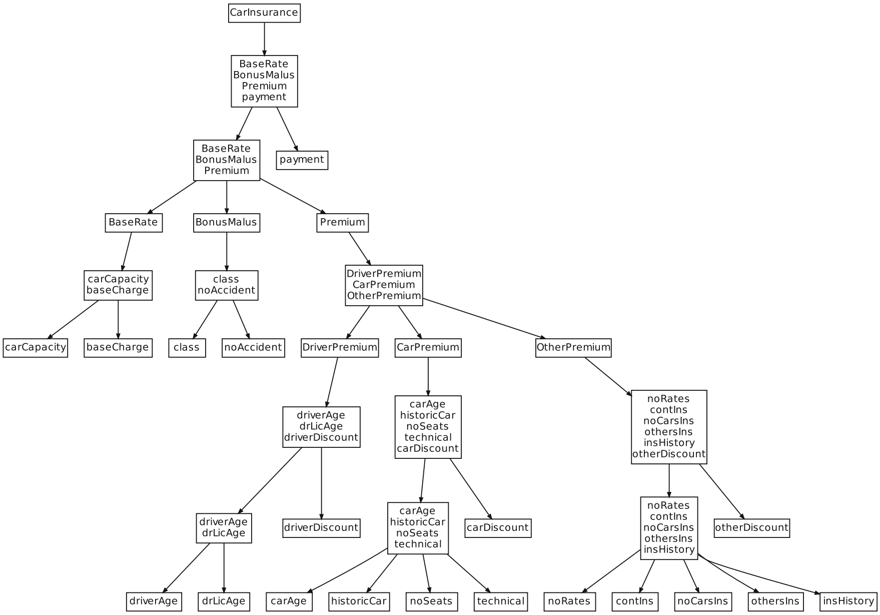

CarInsurance
BaseRate BonusMalus Premium payment
BaseRate BonusMalus Premium
payment
BaseRate
BonusMalus
Premium
carCapacity baseCharge
class noAccident
DriverPremium CarPremium OtherPremium
carCapacity
baseCharge
class
noAccident
DriverPremium
CarPremium
OtherPremium
driverAge drLicAge driverDiscount
carAge historicCar noSeats technical carDiscount
noRates contIns noCarsIns othersIns insHistory otherDiscount
driverAge drLicAge
driverDiscount
carAge historicCar noSeats technical
carDiscount
noRates contIns noCarsIns othersIns insHistory
otherDiscount
driverAge
drLicAge
carAge
historicCar
noSeats
technical
noRates
contIns
noCarsIns
othersIns
insHistory

Fig. 5 An example of the TPH diagram for the PLLI case study

(see Fig. 6). As specification of ARD is an iterative process, the corresponding TPH diagram, presenting split and finalization transformations, can be easily depicted, as shown in Fig. 5.

Based on this simple case study, we will present how to take advantage of the ARD features and transform this representation in order to either generate the whole BPMN model with decision table schemes for BR tasks and form attributes for User tasks, or enrich the existing BPMN model with BR tasks based on the corresponding ARD diagram.

4.2 Advantages of ARD

There are several advantages of using the ARD method to specify the system. Firstly, this method describes the system in an attribute-oriented way, and thus it is easy to comprehend and generates a simple model.

ARD can be used even if there are not many pieces of information available. It is so because this method does not require anything apart from the specification of dependencies between attributes. It is important to mention that we do not specify the detail semantics of the dependency relationship; thus it is only claimed that one property depends on other property. Although this limitation of ARD can be seen as a drawback, the main focus of this method is on simplicity.

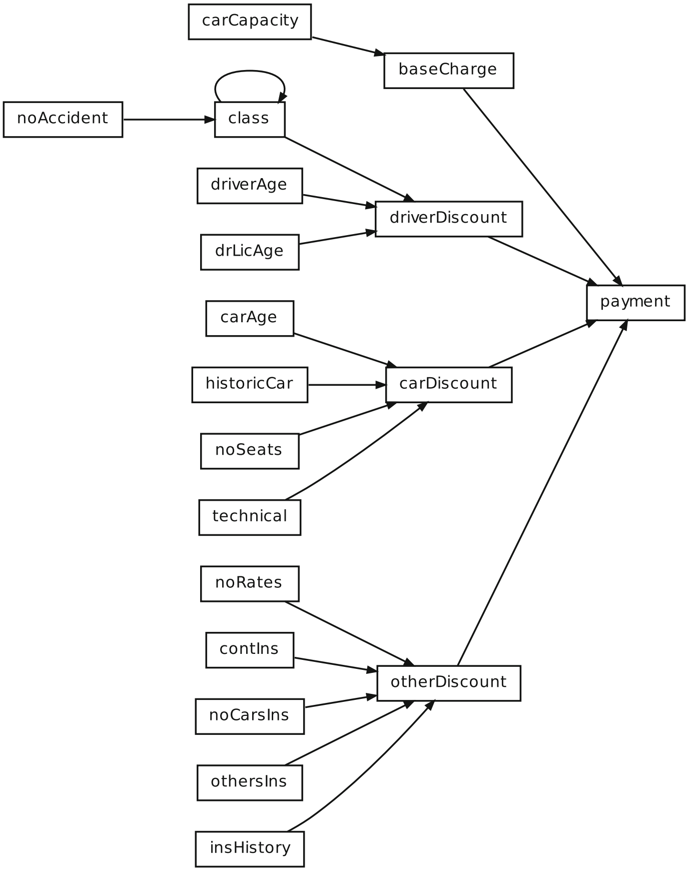


Fig. 6 A complete ARD model for the PLLI example

The ARD method can also be extended, e.g. there can be used some mining technique to acquire attributes and dependencies among them. However, this requires additional research tasks yet. There has been trials of mining such attributes from text in natural language [26].

Applying ARD as a design process allows a domain expert to identify attributes of the modeled system and refine them gradually, as well as generates rule prototypes based on the identified attributes. Thanks to storing the history of transformations, it is possible to refactor such a system [50].

In the following section, we give a short overview of the proposed method of generating process models and enriching the BP models with BR tasks.

5 Business Process Models Generation

In our approach for Business Process models generation [20], we consider two cases presented in the following sections.

5.1 Generating Business Process Model based on ARD

Input:

- Attribute-Relationship Diagram (ARD),[2]
- Transformation Process History (TPH).[3]

Output:

- Business Process model with User tasks containing the list of attributes and BR tasks containing rule prototypes in a very general format:

```
rule:
condition attributes | decision attributes
```

Goal:
To automatically build a BPMN process model on the basis of the ARD diagram (optionally supported by the TPH diagram). The algorithm will generate both User Tasks with form attributes for entering particular pieces of information and Business Rule Tasks with prototypes of decision tables.
Sketch of the algorithm:

1. Generate BR tasks from ARD based on the modified version of the algorithm for generating the XTT2 representation from ARD (detailed description of this part is presented below).
2. Generate proper User tasks which acquire necessary information from the user.
3. Generate proper User/Mail tasks to communicate process results to the user.
4. Complete the diagram using control flow with additional flow objects, such as start and end events, and gateways.

In [51], we presented the algorithm for generating BP models based on ARD.

[2] We consider here the most detailed ARD+ diagram, i.e. a diagram with all of the physical attributes identified. (However, the algorithm can be applied to higher level diagrams as well, and used for generating rules for some parts of the system being designed).

[3] TPH has to correspond to ARD. The diagram may be regarded as optional, but it improves the process of generating labels for the BPMN elements. It is required in the case of hierarchical approach.

5.2 Enriching Business Process Model with Rules based on ARD

Input:

- BPMN Business Process model,
- Attribute-Relationship Diagram (ARD),[4]
- Transformation Process History (TPH).[5]

Output:

- Business Process model with additional Business Rule tasks containing rule prototypes (decision table schemes).

Goal:

The goal of this approach is to automatically enrich a BPMN process model with rule tasks on the basis of the ARD diagram (optionally supported by the TPH diagram). The algorithm will support refactoring of the process model to rule-oriented way by proposing new BR tasks for the process model.

5.3 Generating Business Rule Tasks with Rule Prototypes

In both aforementioned cases, the most important aspect is to generate Business Rule tasks with rule prototypes for a process model. This can be done using the modified version of the algorithm for generating the XTT2 representation from ARD (described in [47, 48]). A draft of the algorithm for generating Business Rule tasks with rule prototypes for a process model is as follows:

1. Prepare data:
   a. Choose a dependency $d \in D : dep(f, t),\ f \neq t$, where D is a set of dependencies in the ARD diagram.
   b. Select all independent properties (other than n) that m depends on. Let $F_t = \{f_t^i : dep(f_t^i, t),\ f_t^i \neq f\}$.
      Remove the considered dependencies from the set:
      $D := D \setminus \{d_{f_t^i,t}\}$.
   c. Select all dependent properties (other than t) that depend only on f. Let $T_f = \{t_f^i : dep(f, t_f^i),\ t_f^i \neq t,\ \nexists f_x : (dep(f_x, t_f^i),\ f_x \neq f)\}$.
      Remove the considered dependencies from the set:
      $D := D \setminus \{d_{f,t_f^i}\}$.
2. Create BR tasks based on F_t and T_f:
   a. if $F_t = \emptyset, T_f = \emptyset$, create a BR task determining the value of the t attribute and associate the task with the following decision table schema: $f \mid t$.

[4] ARD has to correspond to the BPMN model.

[5] TPH has to correspond to ARD.

b. if $F_t \neq \emptyset, T_f = \emptyset$, create a BR task determining the value of the t attribute and associate it with the following decision table schema: $f, f_t^1, f_t^2, ... \mid t$.
c. if $F_t = \emptyset, T_f \neq \emptyset$, create a BR task determining the value of the $T_f \cup \{t\})$ attributes and associate it with the decision table schema: $f \mid t, t_f^1, t_f^2, ...$.
d. if $F_t \neq \emptyset, T_f \neq \emptyset$, create two BR tasks determining the value of the T_f and t attributes and associate them with the following decision table schemes respectively: $f, f_t^1, f_t^2, ... \mid t$ and $f \mid t_f^1, t_f^2, ...$.

3. Go to step 1 if there are any dependencies left ($D \neq \emptyset$).

The result of application of the BR task generation for the Polish Liability Insurance case is presented in Figs. 7, 8, 9 and 10. Next, User tasks which acquire necessary information from the user and User/Mail tasks to communicate process results to the user have to be generated (see Fig. 11). It is important to mention that the appropriate ARD attributes are associated with User tasks. These are important for generating suitable forms in the runtime environment (see an example in Fig. 12).

Finally, the model have to be completed using control flow with additional flow objects, such as start and end events, and gateways. The resulting diagram can be observed in the Activiti-based environment presented in Fig. 14.

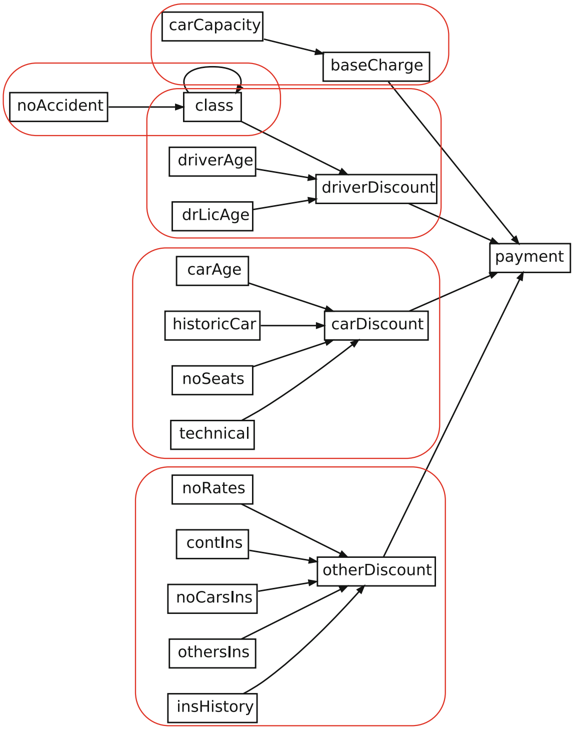


Fig. 7 Selecting ARD dependencies with input attributes

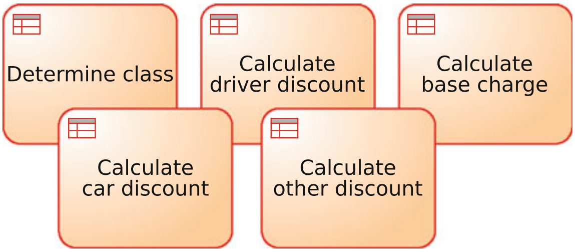


Fig. 8 BR tasks generated by algorithm from the ARD dependencies from Fig. 7

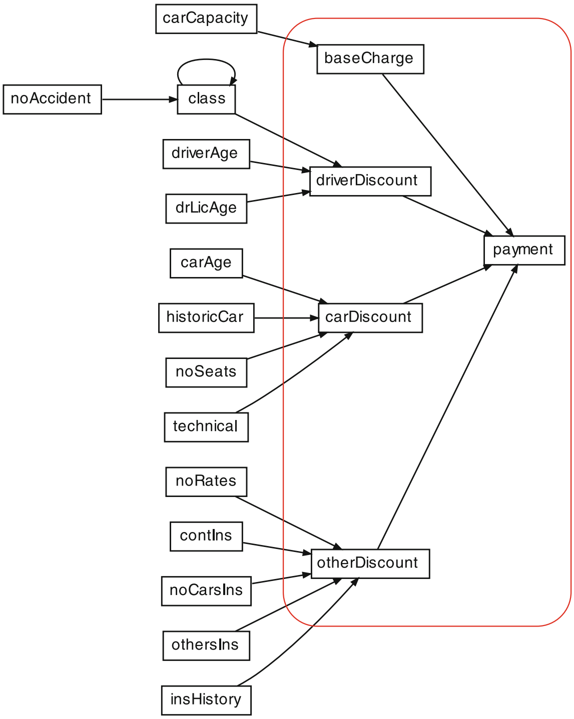


Fig. 9 Selecting ARD dependencies without input attributes

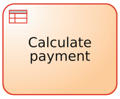


Fig. 10 A BR task generated from the ARD dependencies from Fig. 9

5.4 Hierarchical Extension of the Approach

Modularization helps in managing process model complexity as well as constitute an important issue in improving understandability of models. Vertical modularization (hierarchization) allows for encapsulating process details into sub-levels with hierarchical structure. In order to extend the approach to generate hierarchical Business

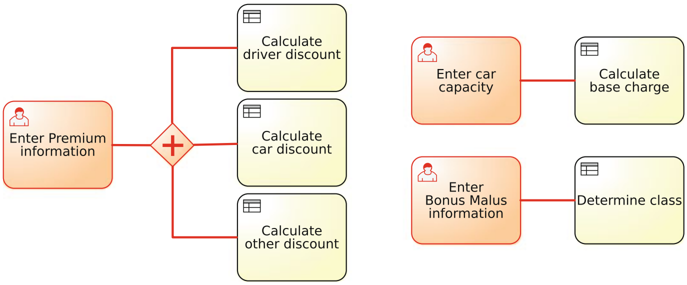
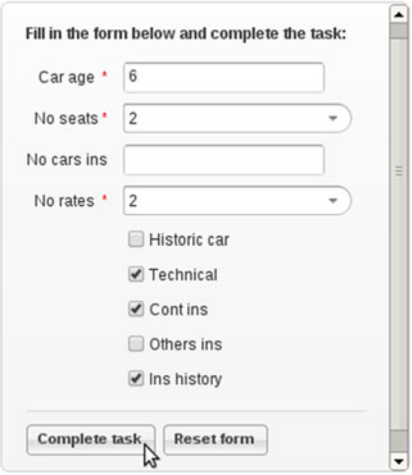


Fig. 11 User tasks generated by algorithm

Fig. 12 An Activiti form for the “Enter Premium Information” user task



Fill in the form below and complete the task:

Car age * 6

No seats * 2

No cars ins

No rates * 2

□ Historic car

☑ Technical

☑ Cont ins

□ Others ins

☑ Ins history

Process models, i.e. process models containing suprocesses, the extension of our approach can take advantage of a TPH model.

As a TPH model captures information about properties with subsequent abstraction levels, it is possible to use these levels for hierarchization purpose in the following way: If some attributes used for generating BR tasks have in TPH a common parent complex property composed of these attributes, they can be incorporated into a subprocess (and on their current level can be replaced by this subprocess).

In the case of PLLI example, the `driver discount`, `car discount`, `other discount` have a common parent property: `Premium`. Thus the 3 BR tasks and a User task associated with these properties can be grouped into a subprocess (see Fig. 13).

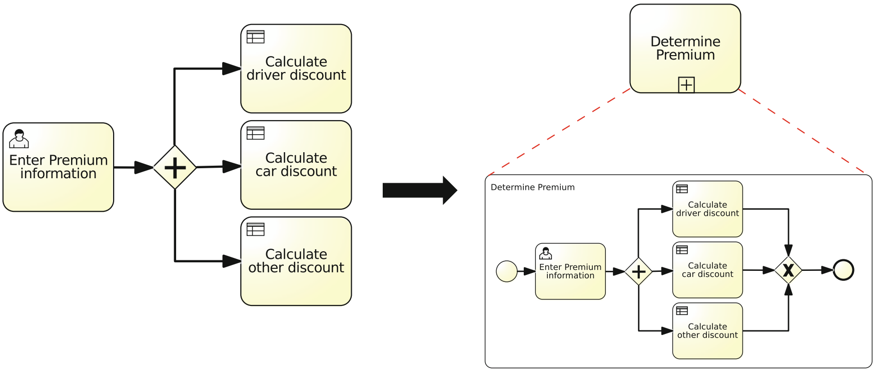


Fig. 13 Hierarchization example

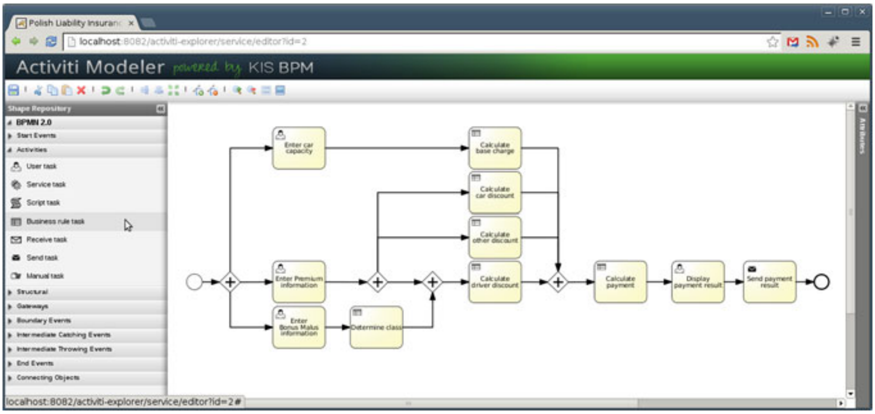


Fig. 14 A prototype Activitiy-based environment for modeling and executing processes with rules

6 Tool Support

VARDA (Visual ARD Rapid Development Alloy) [50, 52, 53] is a rapid prototyping environment for the ARD+ method. It allows for designing ARD, capturing the whole process in TPH as well as visualization and prototyping a rule base.

As a BPMN model generated from ARD constitutes an executable specification of a process, it can be executed in the process runtime environment. However, for complete execution of the model, i.e. execution of the Business Rule task logic, a process engine, such as jBPM [54] or Activiti [55], has to delegate rule execution to the business rule engine. As decision table schemes are generated automatically, the created decision tables have to be complemented with rules. Decision table can be filled in with rules using a dedicated editor [17] or a dedicated plugin for the process modeler [12]. Then, our prototype hybrid execution environment [14, 56], can serve as a specific execution example for this approach.

7 Concluding Remarks

The aim of this paper is to present the possibility of generating the rule-oriented BPMN model and enriching process models with rules based on the ARD diagram. We give an overview of the method for process model generation and present a sketch of the algorithm for automatic generation of rule-oriented BPMN process models from Attribute Relationship Diagram. In the algorithm, BR tasks with corresponding decision table schemes are generated and the resulting model can be executed in the hybrid execution environment.

The presented approach can be used either to generate the whole BPMN model based on the existing ARD diagram or to enrich the existing BPMN model with BR tasks based on ARD developed parallely to BP model or generated based on the process description. As the generated rule schemes are complementary to the process model, the solution addresses the two mentioned challenges: separation between processes and rules in the modeling phase and the problem of the execution of such separated data, which usually requires some additional integration or configuration in the execution environment.

As this paper presents a work-in-progress research, our future work will consist in refining and formalizing the presented approach. We plan to extend the approach with new patterns and some optimization elements. We consider also enriching the ARD diagram with selected relations from the similar methods [27, 29, 30, 36, 37], as well as take advantage of the Decision Model And Notation (DMN) [57].

Acknowledgments The authors wish to thank the organizers of the ABICT 2013 workshop on the FedCSIS 2013 conference for providing an ample environment for presenting and discussing our research [20].

References

1. Lindsay, A., Dawns, D., Lunn, K.: Business processes—attempts to find a definition. Inf. Softw. Technol. **45**(15), 1015–1019 (2003)
2. Allweyer, T.: BPMN 2.0. Introduction to the Standard for Business Process Modeling. BoD, Norderstedt (2010)
3. OMG: Business Process Model and Notation (BPMN): Version 2.0 specification. Technical Report-formal/2011-01-03, Object Management Group (2011)
4. Silver, B.: BPMN Method and Style. Cody-Cassidy Press, Aptos (2009)
5. Nalepa, G.J., Mach, M.A.: Conceptual modeling of business rules and processes with the XTT method. In: Tadeusiewicz, R., Ligęza, A., Szymkat, M. (eds.) CMS'07: Computer Methods and Systems. AGH University of Science and Technology, Cracow, Oprogramowanie Naukowo-Techniczne, Kraków, Poland, pp. 65–70, 21–23 Nov 2007
6. Nalepa, G.J., Mach, M.A.: Business rules design method for business process management. In: Ganzha, M., Paprzycki, M. (eds.) Proceedings of the International Multiconference on Computer Science and Information Technology, vol. 4, pp. 165–170. Polish Information Processing Society, IEEE Computer Society Press, Fremont (2009)
7. Mitchell, T.M.: Machine Learning. MIT Press and The McGraw-Hill companies, Inc., New York (1997)

8. Bajwa, I.S., Lee, M.G., Bordbar, B.: SBVR business rules generation from natural language specification. In: AAAI Spring Symposium: AI for Business Agility. AAAI (2011). http://www.aaai.org/Library/Symposia/Spring/ss11-03.php
9. van der Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes, 1st edn. Springer Publishing Company, Incorporated (2011)
10. Friedrich, F., Mendling, J., Puhlmann, F.: Process model generation from natural language text. In: Mouratidis, H., Rolland, C. (eds.) Advanced Information Systems Engineering. Lecture Notes in Computer Science, vol. 6741, pp. 482–496. Springer, Berlin (2011)
11. Sinha, A., Paradkar, A.: Use cases to process specifications in business process modeling notation. In: Proceedings of IEEE, International Conference on Web Services (ICWS), pp. 473–480 (2010)
12. Kluza, K., Kaczor, K., Nalepa, G.J.: Enriching business processes with rules using the Oryx BPMN editor. In: Rutkowski, L et al. (eds.) Proceedings of ICAISC 2012, 11th International Conference on Artificial Intelligence and Soft Computing: Zakopane, Poland, Lecture Notes in Artificial Intelligence, vol. 7268, pp. 573–581. Springer, 29 April–3 May 2012. http://www.springerlink.com/content/u654r0m56882np77/
13. Nalepa, G.J., Kluza, K., Ernst, S.: Modeling and analysis of business processes with business rules. In: Beckmann, J. (ed.) Business Process Modeling: Software Engineering, Analysis and Applications, Business Issues, Competition and Entrepreneurship, pp. 135–156. Nova Science Publishers, New York (2011)
14. Nalepa, G.J., Kluza, K., Kaczor, K.: Proposal of an inference engine architecture for business rules and processes. In: Rutkowski, L. et al. (eds.) Proceedings of 12th International Conference on Artificial Intelligence and Soft Computing, ICAISC 2013. Lecture Notes in Artificial Intelligence. Springer, Zakopane, Poland, vol. 7895, pp. 453–464, 9–13 June 2013. http://www.springer.com/computer/ai/book/978-3-642-38609-1
15. Ligęza, A., Nalepa, G.J.: A study of methodological issues in design and development of rule-based systems: proposal of a new approach. Wiley Interdisc. Rev. Data Min. Knowl. Discovery **1**(2), 117–137 (2011)
16. Ligęza, A., Szpyrka, M.: Reduction of tabular systems. In: Rutkowski, L., Siekmann, J., Tadeusiewicz, R., Zadeh, L. (eds.) Artificial Intelligence and Soft Computing–ICAISC 2004. Lecture Notes in Computer Science, vol. 3070, pp. 903–908. Springer, Berlin (2004)
17. Nalepa, G.J., Ligęza, A., Kaczor, K.: Formalization and modeling of rules using the XTT2 method. Int. J. Artif. Intell. Tools **20**(6), 1107–1125 (2011)
18. Szpyrka, M.: Exclusion rule-based systems—case study. In: International Multiconference on Computer Science and Information Technology, vol. 3, pp. 237–242. Wisła, Poland (2008)
19. Nalepa, G.J.: Semantic Knowledge Engineering. A Rule-Based Approach. Wydawnictwa AGH, Kraków (2011)
20. Kluza, K., Nalepa, G.J.: Towards rule-oriented business process model generation. In: Ganzha, M., Maciaszek, L.A., Paprzycki, M. (eds.) In: Proceedings of the IEEE, Federated Conference on Computer Science and Information Systems–FedCSIS 2013, Krakow, Poland, pp. 959–966, 8–11 Sept 2013
21. Nalepa, G.J., Kluza, K.: UML representation for rule-based application models with XTT2-based business rules. Int. J. Softw. Eng. Knowl. Eng. (IJSEKE) **22**(4), 485–524 (2012). http://www.worldscientific.com/doi/abs/10.1142/S021819401250012X
22. Nalepa, G.J.: Proposal of business process and rules modeling with the XTT method. In: Negru, V. et al. (eds.) Proceedings of IEEE, Ninth International Symposium, Symbolic and Numeric Algorithms for Scientific Computing–SYNASC. IEEE Computer Society, CPS Conference Publishing Service, Los Alamitos, California, Washington, Tokyo, pp. 500–506, 26–29 Sept 2007
23. OMG: Unified Modeling Language (OMG UML) version 2.2. superstructure. Technical Report-formal/2009-02-02, Object Management Group (2009)
24. Lubke, D., Schneider, K., Weidlich, M.: Visualizing use case sets as bpmn processes. In: Requirements Engineering Visualization, REV '08, pp. 21–25 (2008)

25. Nawrocki, J.R., Nedza, T., Ochodek, M., Olek, L.: Describing business processes with use cases. In: BIS, pp. 13–27 (2006)
26. Atzmueller, M., Nalepa, G.J.: A textual subgroup mining approach for rapid ARD+ model capture. In: Lane, H.C., Guesgen, H.W. (eds.) FLAIRS-22: Proceedings of the Twenty-Second International Florida Artificial Intelligence Research Society Conference, pp. 414–415, Sanibel Island, Florida, USA, 19–21 May 2009 (FLAIRS, AAAI Press, Menlo Park, California 2009)
27. van der Aalst, W.: On the automatic generation of workflow processes based on product structures. Comput. Ind. **39**(2), 97–111 (1999)
28. van der Aa, H., Reijers, H.A., Vanderfeesten, I.: Composing workflow activities on the basis of data-flow structures. In: Business Process Management, pp. 275–282. Springer, Berlin (2013)
29. Vanderfeesten, I., Reijers, H., Aalst, W.: Case handling systems as product based workflow design support. In: Filipe, J., Cordeiro, J., Cardoso, J. (eds.) Enterprise Information Systems. Lecture Notes in Business Information Processing, vol. 12, pp. 187–198. Springer, Berlin (2009)
30. Vanderfeesten, I., Reijers, H., Aalst, W., Vogelaar, J.: Automatic support for product based workflow design: generation of process models from a product data model. In: Meersman, R., Dillon, T., Herrero, P. (eds.) On the Move to Meaningful Internet Systems: OTM 2010 Workshops. Lecture Notes in Computer Science, vol. 6428, pp. 665–674. Springer, Berlin (2010)
31. Vanderfeesten, I., Reijers, H.A., Van der Aalst, W.M.: Product-based workflow support. Inf. Syst. **36**(2), 517–535 (2011)
32. Vanderfeesten, I., Reijers, H.A., van der Aalst, W.M.: Product based workflow support: dynamic workflow execution. In: Advanced Information Systems Engineering, pp. 571–574. Springer, Berlin (2008)
33. Li, S., Shao, X., Zhang, Z., Chang, J.: Dynamic workflow modeling based on product structure tree. Appl. Math. **6**(3), 751–757 (2012)
34. Li, S., Shao, X.D., Chang, J.T.: Dynamic workflow modeling oriented to product design process. Comput. Integr. Manuf. Syst. **18**(6), 1136–1144 (2012)
35. van der Aalst, W.M., Reijers, H.A., Liman, S.: Product-driven workflow design. In: Proceedings of the IEEE, The Sixth International Conference on Computer Supported Cooperative Work in Design, pp. 397–402 (2001)
36. Roover, W., Vanthienen, J.: On the relation between decision structures, tables and processes. In: Meersman, R., Dillon, T., Herrero, P. (eds.) On the Move to Meaningful Internet Systems: OTM 2011 Workshops. Lecture Notes in Computer Science, vol. 7046, pp. 591–598. Springer, Berlin (2011)
37. Wu, F., Priscilla, L., Gao, M., Caron, F., Roover, W., Vanthienen, J.: Modeling decision structures and dependencies. In: Herrero, P., Panetto, H., Meersman, R., Dillon, T. (eds.) On the Move to Meaningful Internet Systems: OTM 2012 Workshops. Lecture Notes in Computer Science, vol. 7567, pp. 525–533. Springer, Berlin (2012)
38. Reijers, H.A., Limam, S.: Product-based workflow design. J. Manage. Inf. Syst. **20**(1), 229–262 (2003)
39. Callahan, S.: Extended generic product structure: an information model for representing product families. J. Comput. Inf. Sci. Eng. **6**(3), 263–275 (2006)
40. Li, Y., Wan, L., Xiong, T.: Product data model for plm system. Int. J. Adv. Manuf. Technol. **55**(9–12), 1149–1158 (2011)
41. Reijers, H.A., Hee, K.M.: Product-based design of business processes applied within the financial services. J. Res. Pract. Inf. Technol. **34**(2), 110–122 (2002)
42. Goedertier, S., Vanthienen, J.: Rule-based business process modeling and execution. In: Proceedings of the IEEE EDOC Workshop on Vocabularies Ontologies and Rules for The Enterprise (VORTE 2005). CTIT Workshop Proceeding Series (ISSN 0929–0672), pp. 67–74 (2005)
43. Nalepa, G.J., Wojnicki, I.: Towards formalization of ARD+ conceptual design and refinement method. In: Wilson, D.C., Lane, H.C. (eds.) FLAIRS-21: Proceedings of the 21st International Florida Artificial Intelligence Research Society conference, Coconut Grove, Florida, USA, pp. 353–358. AAAI Press, Menlo Park, California, 15–17 May 2008

44. Ligęza, A.: Logical Foundations for Rule-Based Systems. Springer, Berlin (2006)
45. Ligęza, A., Nalepa, G.J.: Knowledge representation with granular attributive logic for XTT-based expert systems. In: Wilson, D.C., Sutcliffe, G.C.J. (eds.) FLAIRS-20: Proceedings of the 20th International Florida Artificial Intelligence Research Society Conference: Key West, Florida. Florida Artificial Intelligence Research Society, AAAI Press, Menlo Park, California, pp. 530–535, 7–9 May 2007
46. Hopgood, A.A.: Intelligent Systems for Engineers and Scientists, 2nd edn. CRC Press, Boca Raton (2001)
47. Nalepa, G.J., Wojnicki, I.: ARD+ a prototyping method for decision rules. Method overview, tools, and the thermostat case study. Technical Report (CSLTR 01/2009), AGH University of Science and Technology (2009)
48. Nalepa, G.J., Ligęza, A.: Software Engineering: Evolution and Emerging Technologies, Frontiers in Artificial Intelligence and Applications. Conceptual modelling and automated implementation of rule-based systems, vol. 130, pp. 330–340. IOS Press, Amsterdam (2005)
49. Vanthienen, J., Wets, G.: From decision tables to expert system shells. Data Knowl. Eng. **13**(3), 265–282 (1994)
50. Nalepa, G.J., Wojnicki, I.: VARDA rule design and visualization tool-chain. In: Dengel, A.R. et al. (eds.) KI 2008: Advances in Artificial Intelligence: 31st Annual German Conference on AI, KI 2008. Lecture Notes in Artificial Intelligence. Kaiserslautern, Germany, vol. 5243, pp. 395–396. Springer, Berlin, 23–26 Sept 2008
51. Kluza, K., Nalepa, G.J.: Automatic generation of business process models based on attribute relationship diagrams. In: Lohmann, N., Song, M., Wohed, P. (eds.) Business Process Management Workshops. Lecture Notes in Business Information Processing, vol. 171, pp. 185–197. Springer International Publishing (2014)
52. Nalepa, G.J., Wojnicki, I.: Hierarchical rule design with HaDEs the HeKatE toolchain. In: Ganzha, M., Paprzycki, M., Pelech-Pilichowski, T. (eds.) Proceedings of the International Multiconference on Computer Science and Information Technology, vol. 3, pp. 207–214. Polish Information Processing Society (2008) (Submitted to AAIA 2008)
53. Nalepa, G.J., Wojnicki, I.: An ARD+ design and visualization toolchain prototype in prolog. In: Wilson, D.C., Lane, H.C. (eds.) FLAIRS-21: Proceedings of the 21st International Florida Artificial Intelligence Research Society conference, pp. 373–374. AAAI Press, Coconut Grove, Florida, USA, 15–17 May 2008
54. The jBPM team of JBoss Community: jBPM User Guide, 5.2.0. Final edn (2011). http://docs.jboss.org/jbpm/v5.2/userguide/
55. Rademakers, T., Baeyens, T., Barrez, J.: Activiti in Action: Executable Business Processes in BPMN 2.0: Manning Pubs Co Series. Manning Publications Company, Greenwich (2012)
56. Kaczor, K., Kluza, K., Nalepa, G.J.: Towards rule interoperability: design of drools rule bases using the XTT2 method. Trans. Comput. Collect. Intell. XI **8065**, 155–175 (2013)
57. OMG: Decision model and notation. beta1. Technical Report-dtc/2014-02-01, Object Management Group (2014)

Nonlinear Time Structures and Nonlinear Time Ontology for Description of Economic Phenomena

Maria Mach-Król

Abstract The aim of the paper is to present possible time structures for the task of describing economic phenomena, and to propose an ontology for such time structures.The paper thus describes some possible time structures. Since a simple linear time structure is not enough for this task, use of more complex structures are proposed. The paper also deals with the problem of representing time while reasoning about changing economic domain. The modification of Hobbs and Pan time ontology is proposed and formalized. Also the resulting ontology is augmented with proposal of Hajnicz and McDermott to provide basis for representing both abstract branching time as well as the calendar one.

1 Introduction

Introducing the notion of time allows to perform inference about changing domains, including the economic one. It also allows a computer to simulate human inference, because people infer about time and change [4]. In particular, such notions as change, causality or actions are described in the context of time, therefore the proper representation of time, and proper temporal reasoning is so important in the field of e.g. artificial intelligence [14].

In order to represent properly temporal phenomena (or temporal knowledge about them) it is necessary to establish—prior to choosing the temporal formalism—a proper time structure. It is so because time structure determines the representation. For example, if one chooses dense time, it is not possible to represent knowledge in the situation calculus, because it is for only discrete phenomena [10]. In turn, the structure of time depends on the characteristics of the domain to be modeled.

The economic domain encompasses—among others—the environment in which modern enterprises operate. It is characterized by discontinuity, changeability,

M. Mach-Król (✉)
Department of Business Informatics, University of Economics,
Ul. Bogucicka 3, 40-287 Katowice, Poland
e-mail: maria.mach-krol@ue.katowice.pl

M. Mach-Król et al. (eds.), *Advances in ICT for Business, Industry and Public Sector*,
Studies in Computational Intelligence 579, DOI 10.1007/978-3-319-11328-9_5

heterogeneity. Because of these features it is very difficult to percept changes in the environment, to define their causes, effects and directions, and how they affect enterprises' operations. The impact of these changes on enterprise's strategy is also seen.

The analysis of changes may not be performed apart from the aspect of temporality, because time is strictly connected with the notion of change. The temporal aspect is important, because in modern enterprises knowledge becomes a valuable asset. And a great part of knowledge is temporal. Therefore time becomes an important category for enterprises.

The above leads to a simple conclusion: a complete representation of economic knowledge must be a temporal one. As Sowa claims [2, 15], knowledge representation consists of three elements: ontology, logic, and computational procedures. Every logic assumes an ontology of the domain being depicted. Therefore the representation of knowledge about dynamic economic environment should be composed of ontology of time and domain, temporal logic and computational procedures (reasoning). From the other side, the ontology of time concerns basic temporal entities and time structure. If we assume that the choice of proper temporal entities is strictly dependent on time structure, it becomes obvious that the question of time structure is fundamental. And so is the question of temporal ontology.

The rest of the paper is organized as follows. In Sect. 2 we show how many possible time structures may be obtained from the basic ones. Section 3 provides basic assumptions for the time structure used for describing economic realm. In Sect. 4 the basic information on ontology is summarized. Section 5 contains formal description of time branching into the future (the left-linear one). In Sect. 6 we propose an ontology of point time, and combine it with dates in Sect. 7. Section 8 contains concluding remarks, and future research directions.

2 A Set of Possible Time Structures

In this section the basic time structures are presented, and the possible set of them is discussed.

The most commonly adopted model is the one of linear time, which can be graphically depicted as a straight line, while formally a time structure T is linear, if ([8, p. 20]):

$$\forall t_1, t_2 \in T : (t_1 < t_2) \vee (t_1 = t_2) \vee (t_2 < t_1) \tag{1}$$

The models of nonlinear time are: time branching into the future, time branching into the past, time branching in both directions (parallel time), cyclic time. A motivation for adopting the branching time structure is as follows: many different pasts ("routes") may have led to the present time, and from "now" may arise many different "routes" into the future. The formal definitions are (ibid., p. 21 and next):

A time structure T is branching into the future (left-linear), if

$$\forall \mathrm{t}_1,\ \mathrm{t}_2,\ \mathrm{t}_3 \in \mathrm{T}(\mathrm{t}_2 < \mathrm{t}_1 \wedge \mathrm{t}_3 < \mathrm{t}_1) \Rightarrow (\mathrm{t}_2 < \mathrm{t}_3) \vee (\mathrm{t}_2 = \mathrm{t}_3) \vee (\mathrm{t}_3 < \mathrm{t}_2) \quad (2)$$

A time structure T is branchingintothepast (right-linear) if

$$\forall \mathrm{t1},\ \mathrm{t2},\ \mathrm{t3} \in \mathrm{T}(\mathrm{t1} < \mathrm{t2} \wedge \mathrm{t1} < \mathrm{t3}) \Rightarrow (\mathrm{t2} < \mathrm{t3}) \vee (\mathrm{t2} = \mathrm{t3}) \vee (\mathrm{t3} < \mathrm{t2}) \quad (3)$$

A time structure T is parallel, if it is left- and right-linear, that is branching into both directions.

One more structure, discussed rarely in the literature, but interesting, is a cyclic time structure. A metric point time structure T is an ordered tuple⟨ *T*, *C*, <, <*, *δ*, *S* ⟩, where: *T*—set of time points, *C*—set of distances between points, <—a global order over *T*, <*—local order over *T*, *δ*—metrics over *T*, *S*—length of a semicircle.

For each time point $x \in T$ there exists exactly one point $x^* \in T$ that $\delta(x, x^*) = S$. These two points divide the time circle into two semicircles. The characteristics of a cyclic time structure are as follows (after ([6, p. 30])):

- completeness:

$$\forall x, y(x < y) \quad (4)$$

- local antisymmetry:

$$\forall x, y(x < *y \rightarrow \neg(y < *x)) \quad (5)$$

- local linearity:

$$\forall x, y((x \neq y \ \ \& \ \ x \neq y*) \rightarrow (x <^* y \vee y <^* x)) \quad (6)$$

- local transivity:

$$\forall x, y, z((\delta(x, y) + \delta(y, z) < S) \rightarrow (x <^* y \ \ \& \ \ y <^* z \rightarrow x <^* z)) \quad (7)$$

- coherence:

$$\forall x, y, z((\delta(x, y) + \delta(y, z) < S) \rightarrow (x <^* y \,\&\, y <^* z \rightarrow \delta(x, y) + \delta(y, z) = \delta(x, z))) \quad (8)$$

One may imagine such situations, in which classic time structures would be not enough. Consider a situation, when two enterprises join (perform a fusion), operate as one enterprise for a certain period of time, and divide again into two enterprises, but cooperating together (so it is justified to analyze them together, but not on one time axis). In this case we deal subsequently with time structures: right-linear one, linear one, and left-linear one. Formally this situation may be written as:

$$(t_1, t_2, t_3 < t_F) \Rightarrow (\forall t_1, t_2, t_3 \in T : (t_1 < t_2 \wedge t_1 < t_3) \Rightarrow (t_2 < t_3) \vee (t_2 = t_3) \vee (t_3 < t_2)) \tag{9}$$

$$((t_1, t_2 > t_F) \wedge (t_1, t_2 < t_P)) \Rightarrow (\forall t_1, t_2 \in T : (t_1 < t_2) \vee (t_1 = t_2) \vee (t_2 < t_1)) \tag{10}$$

$$(t_1, t_2, t_3 > t_P) \Rightarrow (\forall t_1, t_2, t_3 \in T(t_2 < t_1 \wedge t_3 < t_1) \Rightarrow (t_2 < t_3) \vee (t_2 = t_3) \vee (t_3 < t_2)) \tag{11}$$

where: t_F—the moment of fusion, t_P—the moment in which enterprise divides again.

This structure has the following properties:

- transivity:

$$\forall x, y(x < y \,\&\, y < z \rightarrow x < z) \tag{12}$$

- anti-reflexivity:

$$\forall x \neg (x < x) \tag{13}$$

- antisymmetry:

$$\forall x, y(x < y \rightarrow \neg (x < y)) \tag{14}$$

- discreteness:

$$\begin{aligned} &\forall x, y(x < y \rightarrow \quad \exists z(x < z \,\&\, \neg \exists u(x < u \,\&\, u < z))) \\ &\forall x, y(x < y \rightarrow \quad \exists z(z < y \,\&\, \neg \exists u(u < u \,\&\, u < y))) \end{aligned} \tag{15}$$

Of course, different structures may be combined together in different ways, according to the analytical needs. It seems that the easiest is combining branching structures with linear one. The combinations similar to the one shown above may be numerous, one can imagine e.g. chains composed of branching and linear time structures. On the other hand, combining the cyclic time structure with branching and/or linear ones seems difficult, or even impossible, because the cyclic structure is a closed one.

It is necessary to discuss, how many time structures can arise from basic ones? Before we answer this question, we have to make some assumptions:

- for simplicity, we consider only a basic structure ⟨T, <⟩, other axioms of this structure are omitted;
- we assume representation in 1st order predicate calculus; in the calculus generally the linearity axiom is manipulated, therefore we deal with a finite set: linearity axiom, left-linearity axiom, right-linearity axiom;
- we omit the question of time metrics, because this does not affect the linear or non-linear property of time structure.

Having the above assumptions, we may say that the set of possible time structures arising from combining basic time structures is a combination with repetitions of those structures. As it is commonly known, the number of *k*-element combinations with repetitions of a *n*-element set is given by a formula:

$$C_n^k = \frac{(k + n - 1)!}{k!\,(n - 1)!} \tag{16}$$

where

k—number of sequence elements,

n—number of set elements.

In the considered case, k = 1, 2, 3 or 4, and n = 4 (linear structure, left-linear structure, right-linear structure, parallel structure).

Therefore, the set of possible time structures is computed as:

$$S = 1 + \frac{(2 + 4 - 1)!}{2!\,(4 - 1)!} + \frac{(3 + 4 - 1)!}{3!\,(4 - 1)!} + \frac{(4 + 4 - 1)!}{4!\,(4 - 1)!} = 66 \tag{17}$$

It should be noted here that in case of using multiple time structures for economic realm description and in case of using a temporal intelligent system, we have to deal with a heterogenic time structure in the knowledge base of the system. Thus, we face a problem of unifying the structure. Is it necessary for performing reasoning by an intelligent system? This problem is similar to the one of the heterogeneous knowledge in a temporal intelligent system, described in detail in [9], but it is beyond the scope of this paper.

3 Time Structure to Describe Economic Phenomena—Basic Assumptions

We are convinced that for a temporal analysis of enterprise's environment it should be assumed that time is: discrete, branching into the future, finite in the past, but infinite in the future. We have chosen this structure because of several reasons:

- Discrete time—there are several elements of the environment that change in a continuous manner, but some of the elements (e.g. barriers to entry) change discretely. From a practical point of view, it is not possible to provide information to the temporal intelligent system continuously. Changes have to be registered discretely. Moreover, assuming continuous time would be linked with introducing a second order axiomatization [1, p. 36].
- Time branching into the future—the enterprise's environment is nondeterministic. Linear time assumes deterministic domain, while time branching into the future assumes a nondeterministic one. Also, introducing time branching into the future, when present actions may develop into several future ones, would allow to deepen

the analysis of the environment, allowing e.g. for “what-if” analyses. It is not the only possible structure. For example, if we take into account the differences of temporal aspects of different markets we may think of a parallel time structure, which enables e.g. analyzing different markets simultaneously. Also a right linear time structure (time branching into the past) could be adopted—in order to determine, which changes in the past on the markets are responsible for the present situation of an enterprise. Using nonlinear time structures for analyzing economic environment is surely an interesting research area.

- Time unbounded in the future—this assumption seems obvious: in a given moment, an enterprise is not able to define for how long it will be operating, therefore it is not possible to determine a moment in time, when the analysis will not be needed any more. However, as managerial practice shows, a time horizon longer than 5 years is not needed. For example, investment plans for more than 5 years are practically unreal. But it should be pointed out here, that a time point named “now” is different every day, moving into the future, and so moves the time horizon, even striving for infinity. Also obvious is assuming time bounded in the past: nor an enterprise, nor a temporal intelligent systems operate from “always”. Moreover, an analysis getting far into the past would not be useful, because the environment is changing and turbulent. It should be assumed a certain “past time horizon of analysis”, therefore bounding time in the past is justified.

Formally speaking, a time structure for modeling economic realm is a structure fulfilling the following conditions:
$\forall t_1, t_2 \in T$

$$t_1 < t_2 \Rightarrow \exists t_3 : (t_1 < t_3) \wedge \neg(\exists t_4 : t_1 < t_4 \wedge t_4 < t_3),$$

$$t_2 < t_1 \Rightarrow \exists t_3 : (t_3 < t_1) \wedge \neg(\exists t_4 : t_3 < t_4 \wedge t_4 < t_1) \tag{18}$$

$$\forall t_1, t_2, t_3 \in T(t_2 < t_1 \wedge t_3 < t_1) \Rightarrow (t_2 < t_3) \vee (t_2 = t_3) \vee (t_3 < t_2) \tag{19}$$

$$\neg(\forall t_1 \exists t_2 : t_2 < t_1) \tag{20}$$

It is a general assumption, but in specific situations the model may be broadened, because—as it has been already pointed out in the Introduction—the time structure has to be adjusted to the phenomenon being analyzed. In the next section different time structures are presented and discussed.

4 Ontology Basics

Having established a time structure for description of economic realm, one has to define a proper time ontology in order to represent and reason about this domain. Ontology is one of the main elements of knowledge representation, because every logic—including temporal ones—assumes an domain ontology. Many authors defined ontology. Here we cite only a few definitions.

Eder and Koncilia [3] define ontology as a conceptualization of a domain. It describes domain notions, their properties and relations. A similar view of the ontology is presented by Salguero et al. [13], who describe it as a specification of knowledge domain conceptualization. A controlled dictionary depicting formally objects and relations between them, and has a grammar (p. 126).

Grenon [5] writes that formal ontology is a branch of philosophy, that analyses and creates theories at the highest level of generality.

Perry et al. [12] add to the above that ontology assures the context or domain semantics (p. 147).

It seems that for time ontology the best definitions are these by [3, 13]. It is because time ontology has to encompass time elements—points or intervals or both—and their

relations. Moreover, it has to contain a "way" to manipulate these elements, so they become meaningful. The detailed description of time ontology adopted in the economic analysis is presented in Sect. 6.

5 Left Linear Time (Time Branching into the Future)—Formally

As it has been said in Sect. 3, for the economic analysis of enterprise's environment we assume the model of time branching into the future (left linear one).

The formal definition is given by Eq. (2).

We assume point structure of time as the basic one. Generally, in the ontology of time one may assume points, intervals or both as basic temporal entities. Different authors force different solutions to this question (see e.g. [16]). In the paper we assume point (discrete) time because of the application domain—the economic one (see Sect. 3).

Formally:

Time is composed of points and a precedence relation $<$, therefore a point structure T is an ordered pair $\langle \mathrm{T}, < \rangle$, where T—a nonempty set of points, $<$ the precedence relation.

The axiom of time discreteness is formulated as:

$$\forall x, y(x < y \rightarrow \exists z(x < z \,\&\, \neg\exists u(x < u \,\&\, u < z))) \tag{21}$$

$$\forall x, y(x < y \rightarrow \exists z(z < y \,\&\, \neg\exists u(z < u \,\&\, u < y))) \tag{22}$$

This structure has the following properties:

- transitivity

$$\forall x, y(x < y \,\&\, y < z \rightarrow x < z) \tag{23}$$

- anti-reflexivity

$$\forall x \neg(x < x) \tag{24}$$

- antisymmetry

$$\forall x, y(x < y \rightarrow \neg(x < y)) \tag{25}$$

Moreover, because we deal with time branching into the future (left/backward linear one), we add an axiom of back linearity:

$$\forall x, y(x < z \,\&\, y < z \rightarrow X < y \vee x = y \vee y < x) \tag{26}$$

6 The Ontology of Point Time

The point time ontology presented in this section extends and modifies the ontology by Hobbs and Pan [7], also adding time properties from [6], because of economic application domain, presented in Sect. 3.

Topological temporal relations.

Time points are a subclass of temporal entities:

$$\text{Instant(t)} \rightarrow \text{TemporalEntity(t)} \tag{27}$$

$$\forall(\text{T}) \text{ TemporalEntity(T)} \rightarrow \text{Instant(T)} \tag{28}$$

Predicates *begins* and *ends* are the relations between points and temporal entities:

$$\text{begins (t, T)} \rightarrow \text{Instant(t)} \wedge \text{TemporalEntity(T)} \tag{29}$$

$$\text{ends(t, T)} \rightarrow \text{Instant(t)} \wedge \text{TemporalEntity(T)} \tag{30}$$

Moreover

$$\text{Instant(t)} \equiv \text{begins(t,t)} \tag{31}$$

$$\text{Instant(t)} \equiv \text{ends(t,t)} \tag{32}$$

If exists a beginning and an end of a temporal being, it is unique:

$$\text{TemporalEntity(T)} \wedge \text{begins}(t_1, \text{T}) \wedge \text{begins}(t_2, \text{T}) \rightarrow t_1 = t_2 \tag{33}$$

$$\text{TemporalEntity(T)} \wedge \text{ends}(t_1, \text{T}) \wedge \text{ends}(t_2, \text{T}) \rightarrow t_1 = t_2 \tag{34}$$

Predicate *TimeBetween* is a relations between a temporal being and two points:

$$\begin{gathered}\mathrm{TimeBetween}\,(\mathrm{T}, \mathrm{t}_1, \mathrm{t}_2) \rightarrow \mathrm{TemporalEntity}(\mathrm{T}) \wedge \mathrm{Instant}(\mathrm{t}_1) \wedge \mathrm{Instant}(\mathrm{t}_2) \wedge \\ (\mathrm{Instant}(\mathrm{t}_1) < \mathrm{Instant}(\mathrm{t}_2) \vee \mathrm{Instant}(\mathrm{t}_2) < \mathrm{Instant}(\mathrm{t}_1) \vee \mathrm{Instant}(\mathrm{t}_1) = \mathrm{Instant}(\mathrm{t}_2))\end{gathered} \tag{35}$$

The condition in the above expression—$(Instant(t_1) < Instant(t_2) \vee Instant(t_2) < Instant(t_1) \vee Instant(t_1) = Instant(t_2))$-means that time points lie on the same time branch. It is necessary to add this condition, because only on the same time branch (time axis) the condition of strict linear order is fulfilled, which allows to compute the value of *TimeBetween* predicate. It is not possible to compare distances between points on different time branches.

7 Combining Time and Events

After presenting the basic ontology of left linear time it should be discussed how time is linked to events in the world. Hobbs and Pan propose to use 4 predicates: *atTime, during, holds, timeSpan* [7, p. 70]. We extend here the notion of an event. The classic definition (see e.g. [6, p. 4]) says that an event is a dynamic picture of the world, causing changes in facts. Following Hobbs and Pan, we will however understand events very broadly—as "anything that may be placed in time" ([7, p. 70]), so not only event by itself, but also a state, a process, a logical statement etc.

As said above, Hobbs and Pan propose four predicates. Because we adopt a discrete model of time, we do not use predicate *during*, concerning intervals, and the predicate *timeSpan* will have a narrower definition compared to the original one (see [7, p. 71]).

Predicate *atTime* link san event with a time point, therefore it is crucial for the ontology of discrete time. It says that an event happens, occurs at time point *t*.

$$\mathrm{atTime}(\mathrm{e}, \mathrm{t}) \rightarrow \mathrm{Instant}(\mathrm{t}) \tag{36}$$

Predicate *Holds* is generally a duplicate of predicate *atTime*. In the original approach by Hobbs and Pan it says that an event occurs at time point *t* or over a time interval *T*. As we assume time discreteness, we omit the second meaning of the predicate and in this way we duplicate the two predicates. We may write:

$$\mathrm{holds}(\mathrm{e}, \mathrm{t}) \equiv \mathrm{atTime}(\mathrm{e}, \mathrm{t}) \tag{37}$$

Finally, the predicate *timeSpan* links events with time points (or sequences of time points)—it is a narrowed version of the original predicate by Hobbs and Pan, which linked events also with intervals and sequences of intervals. This predicate is used for states or processes that adhere to each other, it shows the whole time span during which a process or a state holds. Formally:

$$\text{timeSpan}(T, e) \rightarrow \text{TemporalEntity}(T) \vee \text{tseq}(T) \tag{38}$$

where *tseq(T)* is a sequence of time points.

Moreover

$$\text{timeSpan}(t, e) \wedge \text{Instant}(t) \rightarrow \text{atTime}(e, t) \tag{39}$$

$$\text{timeSpan}(t, e) \wedge \text{Instant}(t) \wedge t_1 \neq t \rightarrow \neg\,\text{atTime}(e, t_1) \tag{40}$$

The predicate *atTime* links an event with a concrete time point, but this is not a direct linking of an event with the date of its occurrence. At the same time, dates are necessary in the description of economic reality. Therefore there is a question how to link time branching into the future with calendar time.

As McDermott pointed out [11], two dates cannot be placed on two different time branches, but one date (the same one) may be placed on many branches, as time branches are independent. Therefore in McDermott's opinion one should discuss a date line independent from the main time structure. In this way, the date line preserves a linear order.

If we adopt the solution proposed by McDermott, we will have to extend the time structure presented in Sect. 5 to the following one[1]:

$$T = \langle \mathrm{T}, \mathrm{D}, <_{tt}, <_{dd}, <_{td}, <_{dt} \rangle \tag{41}$$

where T—a set of time points, D—a set of dates, < tt—backward partial linear order over T, < dd—a linear order over D, < td and < dt are precedence relations linking the former two orders. In this situation we need to add a few new axioms to the ones presented in Sect. 5. Hajnicz [6] calls them the axioms of quasi-transitivity:

$$t_1 <_{tt} t_2 \wedge t_2 <_{td} d \rightarrow t_1 <_{td} d \tag{42}$$

$$d_1 <_{dd} d_2 \wedge d_2 <_{dt} t \rightarrow d_1 <_{dt} t \tag{43}$$

$$d <_{dt} t_1 \wedge t_1 <_{tt} t_2 \rightarrow d <_{dt} t_2 \tag{44}$$

$$d_1 <_{dt} t \wedge t <_{td} d_2 \rightarrow d_1 <_{dd} d_2 \tag{45}$$

$$t <_{td} d_1 \wedge d_1 <_{dd} d_2 \rightarrow t <_{td} d_2 \tag{46}$$

$$t_1 <_{td} d \wedge d <_{dt} t_2 \rightarrow t_1 \neq t_2 \wedge \neg(t_2 <_{tt} t_1) \tag{47}$$

[1] The solution presented in this section comes from [6, p. 24–25].

Adopting the extended structure of time and additional axioms, we have a time theory that is described by the notions of transitivity, anti-symmetry, backward linearity and quasi-transitivity. Together with the ontology of left linear time, we are able to place economic events in time.

8 Concluding Remarks

In the paper the motivation for temporal representation of economic knowledge has been presented, the possible time structures have been pointed out, and the possible combinations of them for economic realm description have been shown. In the paper we also presented the basic ontology of time branching into the future (left linear one). We presented motivation for the choice of this time structure. Next, we modified the ontology by Hobbs and Pan, extending it to the nonlinear time. The next step consisted of presenting, how to link such time with the line of dates. For this purpose, we combined the modified ontology with the proposal by McDermott and Hajnicz, thus proposing a complete model of time, allowing to temporally describe economic phenomena.

The main conclusion stresses out the variety of possibilities given by only 4 time structures, combined in different ways. Leaving linear time axiomatization apart, in order to take into account richer structures, will enable to better depict the economic realm, e.g. for building a knowledge base of a temporal intelligent system.

The second conclusion is that the classical structure of linear time is too simple and non-adequate to complex economic reality. However—and this is the third conclusion—the ontology of time by itself is also insufficient (even the ontology of nonlinear time). It is necessary to link with it dates, because economic activities are registered using the standard calendar. This was the reason for combining the proposal of Hobbs and Pan with the proposal by Hajnicz, introducing additional axioms.

There are several potential future research directions.

The first one is developing an ontology of a selected part of economic reality and then combining it with the ontology presented in this paper. A good example of the part of economic reality are barriers to entry to a marketspace. They are interesting, because they are a good exemplification of economic environment: they are heterogeneous, they change in time, they can be both qualitative and quantitative, dense or discrete.

The second research direction is the implementation of nonlinear time ontology in the temporal intelligent system. The need of using such systems for economic domain was suggested in [9]. In the temporal system it is assumed that the knowledge base is encoded using temporal logic. Therefore time is explicit. It is also explicit in the reasoning mechanism. It seems that implementing the ontology of nonlinear time would be useful at least in the reasoning layer of the system.

The problem of time structures heterogeneity in a temporal knowledge base arises while using more than one structure. It has to be discussed and checked, whether to perform reasoning in a temporal intelligent system it is necessary to unify these structures, and if so, how it should be done. This also will be the topic of future research studies.

References

1. Bennett, B., Galton, A.P.: A unifying semantics for time and events. Artif. Intell., **153**, 13–48 (2004)
2. Chua, C., Storey, V., Chiang, R.: Deriving knowledge representation guidelines by analyzing knowledge engineer behavior. Decis. Support Syst. **54**(1), 304–315 (2012)
3. Eder, J., Koncilia, C.: Interoperability in Temporal Ontologies. Springer, Porto (2005)
4. Galton, A.: Time and change for AI. In: Gabbay, D., Hogger, C., Robinson, J. (eds.) vol. 4. Epistemic and Temporal Reasoning. Clarendon Press, Oxford (1995)
5. Grenon, P.: The Formal Ontology of Spatio-Temporal Reality and its Formalization, AAAI (2003)
6. Hajnicz, E.: Time Structures. Formal Description and Algorithmic Representation. Springer, Berlin (1996)
7. Hobbs, J.R., Pan, F.: An ontology of time for the semantic web. ACM Trans. Asian Lang. Inf. Process. **3**(1), 66–85 (2004)
8. Klimek, R.: Wprowadzenie do logiki temporalnej. [Introduction to Temporal Logic] Kraków: Uczelniane Wydawnictwa Naukowo-Dydaktyczne AGH (1999)
9. Mach, M.A.: Temporalna analiza otoczenia przedsiębiorstwa. Techniki i narzędzia inteligentne. [Temporal Analysis of Enterprise's Environment. Intelligent Techniques and Tools]. Wrocław: Wydawnictwo AE (2007)
10. McCarthy, J., Hayes, P.: Some philosophical problems from the standpoint of artificial intelligence. Mach. Intell. **4**, 463–502 (1969)
11. McDermott, D.: A temporal logic for reasoning about processes and plans. Cognitive Sci. **6**, 101–155 (1982)
12. Perry, M., Hakimpour, F., Sheth, A.: Analyzing Theme, Space, and Time: An Ontology-Based Approach. ACM, New York (2006)
13. Salguero, A., Araque, F., Delgado, C.: Spatio-Temporal Ontology Based Model for Data Warehousing. WSEAS Press, Istanbul (2008)
14. Saracco, C., Nicola, M. and Gandhi, L.: A Matter of Time: Temporal Data Management in DB2 for z/OS. ftp://170.225.15.26/software/data/sw-library/db2/papers/A_Matter_of_Time_-_DB2_zOS_Temporal_Tables_-_White_Paper_v1.4.1.pdf (2012). Accessed 20 06 2014
15. Sowa, J.: Knowledge Representation. Brooks/Cole, Pacific Grove (2000)
16. Vila, L.: A survey on temporal reasoning in artificial intelligence. AI Commun. **7**(1), 4–28 (1994)

Business Intelligence and Analytics in Organizations

Celina M. Olszak

Abstract The paper concerns Business Intelligence (BI) issue and the opportunities of using BI applications in organizations. BI enables to better understand not only the internal business processes, but also the competitive environment through the systematic acquisition, collation, analysis, interpretation and exploitation of information. It is considered that BI transforms the information into strategic knowledge. This allows for the identification of the opportunities and threats, which may occur on the market, while cooperating with customers, suppliers and competition. The main goal of this paper is to present the basic assumptions underlying the idea of BI and BI maturity as well as to identify the level of BI using in selected organizations. The structure of this paper is organized as follows. Firstly, an overview of subject literature on BI has been conducted. Then, the idea of BI maturity has been described. Different BI maturity models have been presented. Using an in-depth interview method, the results from the analysis of twenty organizations applying BI systems have been described. Gartner's Maturity Model for Business Intelligence and Performance Management was used to assess the level of BI in surveyed organizations.

1 Introduction

The information, knowledge and intelligence have become very important resources of a contemporary organization [1, 3, 33, 44, 65]. It is highlighted that its success depends more and more from the ability to take advantage of all intangible resources [6, 15, 40, 43, 55]. This challenge becomes more difficult with the constantly increasing volume of information.

Recently, many organizations turn to Business Intelligence (BI) [13, 19, 35, 45, 54, 62]. It is reported that BI has become the critical component for the success of the contemporary organization [27, 64, 67, 68]. It enables to take competitive advantage

C.M. Olszak (✉)
Department of Business Informatics, University of Economics,
Ul. Bogucicka 3, 40-226 Katowice, Poland
e-mail: celina.olszak@ue.katowice.pl

M. Mach-Król et al. (eds.), *Advances in ICT for Business, Industry and Public Sector*,
Studies in Computational Intelligence 579, DOI 10.1007/978-3-319-11328-9_6

of all available information, provides actionable intelligence for effective business decision-making and business processes [2, 5, 12, 59, 60].

On the other hand, it is pointed that although, BI applications have become the most essential technologies to be purchased in the last years, the success from such applications is still questionable. According to some authors the practical benefits from BI are often unclear and some organizations fail completely in their BI approach. Organizations do not achieve the appropriate benefits [10, 27, 30, 62].

The analysis of the literature shows that the research studies focus mainly on BI using in large enterprises [4, 55, 68]. Unfortunately, the studies that investigate the BI using in small and local business contextare only partly addresses by existing research [47].

The main purpose of this paper is to explore the issue of BI and to investigate of BI using in the selected polish organizations. The reminder of the paper is organized as follows. Firstly, the issue of BI was explained. Next, different BI maturity models were explored. Finally, some selected findings from the survey, conducted in 20 purposefully selected organizations, have been presented. Some guidelines and recommendations were provided in order to improve the using of BI for the organizational success.

2 The Business Intelligence

Business Intelligence and analytics (BI&A) have become the significant research area in the domain of management information systems in the last years [11]. The roots of BI&A originate from decision support systems, which first emerged in the early 1970 s when managers used computer applications to model business decisions. Over the years, other applications, such as executive information systems (EIS), online analytical processing (OLAP), data warehousing, and data mining became important [15, 35, 41, 66, 68]. Today BI&A is compared to “an umbrella” that is commonly used to describe the technologies, applications, and processes for gathering, storing, accessing and analyzing data to help users to make better decisions [16, 59, 61].

It is noted that although BI is frequently defined in the literature, there is no universal explanation of BI [13]. The overview of different BI definitions is presented in Table 1.

BI&A is comprised of both technical and organizational elements [3, 18, 32, 46, 65]. From technical point of view BI&A is an integrated set of tools, technologies and software products that are used to collect heterogenic data from dispersed sources and then to integrate and analyze data to make them commonly available. The key BI&A technologies include: data warehousing, data mining and OLAP [29]. They are often called BI&A 1.0. In the last years, new techniques, such as: web mining, opinion mining techniques, mobile mining techniques and semantic processing are applied in BI&A applications. Such applications, focused on processing of semi-structured or un-structured data that originate mainly from Internet and social media, are named

BI&A 2.0. In turn, applications responsible for collecting and analyzing data from various mobile devices are called BI&A 3.0 [11, 21, 42, 49, 56].

From organizational perspective, BI&A means a holistic and sophisticated approach to cross-organizational decision support [30, 39]. Negash and Gray [41] argue that BI is responsible for transcription of data into information and knowledge. Also, it creates some environment for effective decision-making, business processes, strategic thinking, acting in organizations and taking the competitive advantage [2, 5, 12, 47, 60]. Many authors highlight that BI is predisposed to support decision-making on all levels of management [16, 26, 37, 39, 41]. On the strategic level, with the help of BI it is possible to set objectives precisely and follow the realization of such established objectives. BI allows for performing different comparative reports, e.g. on historical results, profitability of particular offers, effectiveness of distribution channels or forecasting future results on the basis of some assumptions. On the tactical level BI may provide some basis for decision-making within marketing, sales, finance, capital management, etc. BI allows for optimizing future actions and modifying organizational, financial or technological aspects of company performance appropriately in order to help enterprises to realize their strategic objectives more effectively. In turn, on the operational level, BI systems are used to perform ad hoc analyses and answer questions related to departments' ongoing operations, up-to-date financial standing, sales and co-operation with suppliers, customers [46].

It is indicated that BI&A facilitates the realization of business objectives through reporting of data to analyze trends, creating predictive models for forecasting and optimizing process for enhanced performance. The value of BI&A for business is predominantly expressed in the fact that such systems cast some light on information that may serve as the basis for carrying out fundamental changes in a particular enterprise. It is stated that BI&A has become the critical component for the success of the contemporary organization [13, 27, 64, 67, 68]. Wells [65] argues that BI is the "capability of an organization to explain, plan, predict, solve problems, think in an abstract way, understand, invent, and learn in order to increase organizational knowledge, provide information for the decision-making process, enable effective actions, and support establishing and achieving business goals".

Last time a new trend in BI, called "cloud BI" or "BI services on demand", has been appeared. Cloud BI presents a model that provides on demand access to software and hardware resources with minimal management efforts [57]. It is reported that cloud BI is a revolutionary concept of delivering business intelligence capabilities "as service" using cloud based architecture that comes at a lower cost yet faster deployment and flexibility [22]. Cloud BI solution has special interest for organizations that desire to improve agility while at the same time reducing IT costs and exploiting the benefits of cloud computing.

The evolution of different BI&A models has been presented in Table 2.

Table 1 The overview of BI definitions

Author	Definition
Adelman, Moss [1]	An umbrella term to describe the set of software products for collecting, aggregating, analyzing and accessing information to help organization make more effective decisions
Alter [3]	An umbrella term for decision support
Azvine et al. [4]	BI is all about capturing, accessing, understanding, analyzing and converting one of the fundamental and most precious assets of the company, represented by the raw data, into active information in order to improve business
Business Objects [8]	A system that provides different information and analysis for employers, customers, suppliers in order to make more effective decisions
Chung et al. [12]	Results obtained from collecting, analyzing, evaluating and utilizing information in the business domain
Power [50]	An umbrella term to describe the set of concepts and methods used to improve business decision-making by using fact-based support systems
Eckerson [18]	A system that takes data and transforms into various information products
Glancy, Yadav [20]	BI focuses on supporting a variety of business functions, using the process approach, and advanced analytical techniques
Hannula, Pirttimaki [24]	Organized and systematic processes which are used to acquire, analyze and disseminate information to support the operative and strategic decision making
Jordan, Ellen [31]	BI is seen as a critical solution that will help organizations leverage information to make informed, intelligent business decisions to survive in the business world
Jourdan et al. [32]	BI is both a process and a product, that is used to develop useful information to help organizations survive in the global economy and predict the behavior of the general business environment
Lonnqvist, Pirttimaki [36]	A managerial philosophy and tool that helps organizations manager and refine information with the objective of making more effective decisions
Moss, Atre [39]	An architecture and a collection of integrated operational as well as decision support applications and databases that provide the business community easy access to business data
Negash [40]	A system that combines data collection, data storage and knowledge management with analytical tools so that decisions makers can convert complex information into competitive advantage
Olszak, Ziemba [46]	A set of concepts, methods and processes that aim at not only improving business decisions but also at supporting realization of an enterprise' strategy
Reinschmidt, Francoise [52]	BI is an integrated set of tools, technologies and programmed products that are used to collect, integrate, analyze and make data available

(continued)

Table 1 (continued)

Author	Definition
Watson, Wixom [62]	BI describes the concepts and methods used to improve decision making using fact based systems
Wixom, Watson [68]	BI is a broad category of technologies, applications, and processes for gathering, storing, accessing, and analyzing data to help its users make better decisions
White [66]	An umbrella term that encompances data warehousing, reporting, analytical processing, performance management and predictive analytics
Williams, Williams [67]	A combination of products, technology and methods to organize key information that management needs to improve profit and performance

3 Business Intelligence Analysis in Organizations

The great expectations are set for BI. BI is constantly ranked as top priority global. Nearly 90 % of organizations across the globe have invested in a BI capability, bringing BI's annual global outlay to around USD$ 60 billion [13]. It is reported that the beneficiaries of BI include a wide group of users, representing trading companies, insurance companies, banks, financial sector, health sector, telecommunications, manufacturing companies, government, security and public safety. It is assumed that organizations use BI for [10, 15, 25, 44]:

- increasing the effectiveness of strategic, tactic and operational planning including first of all: (a) modelling different variants in the development of an organization; (b) informing about the realization of enterprise's strategy, mission, goals and tasks; (c) providing information on trends, results of introduced changes and realization of plans; (d) identifying problems and 'bottlenecks' to be tackled; (e) providing analyses of "the best" and "the worst" products, employees, regions; (f) providing analyses of deviations from the realization of plans for particular organizational units or individuals; (g) and providing information on the enterprise's environment;
- creating or improving relations with customers, mainly: (a) providing sales representatives with adequate knowledge about customers so that they could promptly meet their customers' needs; (b) following the level of customers' satisfaction together with efficiency of business practices; (c) and identifying market trends;
- analysing and improving business processes and operational efficiency of an organization particularly by means of: (a) providing knowledge and experience emerged while developing and launching new products onto the market; (b) providing knowledge on particular business processes; (c) exchanging of knowledge among research teams and corporate departments.

The most used BI analysis refer to (Table 3): cross selling and up selling, customer segmentation and profiling, parameters importance, survival time, customer

Table 2 BI models

Type BI&A	Function	Scope	Decision support level	Used techniques
Data Marts	Ad hoc analysis, comparative analysis, reporting	Narrow, limited to unit, department	Operational, well structured	Reporting, OLAP
Data warehouse	Multidimensional analysis	The whole enterprise	Operational, tactical, strategic	OLAP, data mining
BI with PA	Forecasting of different scenarios	Narrow, limited to unit, department	Operational, tactical, strategic	OLAP, AP
Real-time BI	Monitoring of current activities, discovering irregularities	Narrow, limited to unit, department	Operational, well structured	EII
Corporative BI	Corporative management, building loyalty strategy	All actors of value chain	Operational, tactical, strategic	ETL, data mining
BI portals	Content management and document management, group work	Selected communities	Operational, tactical, strategic	Internet, Web mining, CMS, work group, personalization techniques
BI nets	The building of expert' nets, social capital management	Global, various communities	Operational, tactical, strategic	Web mining, Web farming, cloud computing
BI for everyone BI for demand	The building of social nets, social capital management	Global	Operational, tactical, strategic	Mobile, social media, semantic Web, Web mining, cloud computing

Table 3 BI analysis

The type of BI analysis	Description
Cross selling and up selling	It involves selling products to specific customers taking their previous purchases into consideration
Customer segmentation and profiling	It is based on grouping customers in some homogenous segments. Segmentation and profiling of customers provide some knowledge that is useful while designing new products and addressing marketing campaigns appropriately, as well
Parameters importance	It allows for determination of the most important variables that describes products, processes and customers in the situation when there are different variables that describe analysed objects
Survival time	It evaluates customer's survival time length and a possibility that they leave during that time. The analysis describes a distribution of survival time for individuals of a given population, monitors strength of other parameters impact on the expected survival time, and additionally
Customer loyalty and customer switching to competition	It is strictly related to analyses of customer's switching to competition. That results in identifying customers who are inclined to leave a company and join competition
Credit scoring	It enables to determine financial risk that is related to particular customers. Such a process may be performed at the very moment a contract with a customer is concluded, and it is based on the data that come from application forms provided by a customer subject to analysis
Fraud detection	It means identification of suspicious transfers, orders and other illegal activities that target a company in question
Logistics optimisation	Logistics optimisation problem involves offering the best possible plan of logistics activities (including transportation or distribution), simultaneously taking already known limitations and available potential into consideration
Forecasting of strategic business processes development	Modelling of multidimensional forecasts based on historical, present and anticipated data. Analyses of time series make it possible to identify and analyze hidden trends and fluctuations
Web mining	Analysis and assessment of the performance of Internet services (web mining) helps to obtain knowledge who uses services, when, why and how
Web farming	It offers a possibility of constant analyzing of the Internet, finding important business information there, acquiring such information, saving it in data warehouse of a company and delivering processed information to adequate persons or departments in a company

loyalty and customer switching to competition, credit scoring, fraud detection, logistics optimizations. Others concern: the forecasting of strategic business processes development, analysis of web mining, and social media.

It should be pointed that although, BI&A applications have become the most essential technologies to be purchased in the last years, the success from such applications is still questionable. The practical benefits from BI&A are often unclear and some organizations fail completely in their BI&A approach. Organizations do not achieve the appropriate benefits [10, 27, 30, 55, 62]. It is said that about 60–70 % of business intelligence applications fail due to the technology, organizational, cultural and infrastructure issues [13, 24].

It is reported that the most important elements that decide on BI&A success in the organizations include: quality of data and used technologies, skills, sponsorship, alignment between BI and business, and BI use [15]. Others elements concern: organizational culture, information requirements, politics. According to Olszak and Ziemba [47] the biggest barriers that the organizations encounter during the implementation of BI systems have a business and organizational character. Among the business barriers, the most frequently mentioned are: the lack of well defined business problem, not determining the expectation of BI and the lack of relations between business and BI vision system. Whereas as the key organizational barriers the enterprises enumerate: the lack of manager's supporting, the lack of knowledge about the BI system and its capabilities, exceeded the BI implementation budget, ineffective BI project management and complicated BI project, the lack of user training and support.

4 Business Intelligence Maturity Models

The effective development of BI in organization should be based on the proven, scientific theories. It seems that the theory of maturity models gives good foundations. The term of maturity describes a "state of being complete, perfect or ready. To reach this a desired state of maturity, an evolutionary transformation path from an initial to a target stage needs to be progressed" [34]. Maturity models are used to guide this transformation process. They help define and categorize the state of an organizational capability [14, 63]. The maturity model for BI helps an organization to

Table 4 Overview of BI maturity models

Name of the model	Description
TDWI's Business Intelligence Model—Eckerson's Model Eckerson [18]	The model focuses mainly on the technical aspect for maturity assessment. It constitutes of 6 maturity levels and uses a metaphor of human evolution: prenatal, infant, child, teenager, adult and sage
Gartner's Maturity Model for BI and PM [7, 51]	The model is a mean to assess the maturity of an organization's efforts in BI and PM and how mature these need to become to reach the business goals. The model recognizes 5 maturity levels: unaware, tactical, focused, strategic, pervasive
AMR Research's Business Intelligence/ Performance Management [23]	The model is described by 4 maturity levels: reacting (where have we been?), anticipating (where are we now?), collaborating (where are we going?), and orchestrating (are we all on the same page?). It is used to assess the organization in the area BI and PM

(continued)

Table 4 (continued)

Name of the model	Description
Business Information Maturity Model Williams [28]	The model is characterized by 3 maturity levels. The first level answers the question "what business users want to access", the second "why the information is needed", the third "how information put into business use"
Model of Analytical Competition Davenport, Harris [15]	The model describes the path that an organization can follow from having virtually no analytical capabilities to being a serious analytical competitor. It includes 5 stages of analytical competition: analytically impaired, localized analytics, analytical aspirations, analytical companies, and analytical competitors
Information Evolution Model, SAS SAS [53]	The model supports organization in assessing how they use information to drive business, e.g., to outline how information is managed and utilizes as a corporate asset. It is characterized by 5 maturity levels: operate, consolidate, integrate, optimize, innovate
Model Business Intelligence Maturity Hierarchy [17]	The model was developed in knowledge management and constitutes of 4 maturity levels: data, information, knowledge and wisdom
Infrastructure Optimization Maturity Model [28]	The model enables a move from reactive to proactive service management. It aids in assessing different areas comprising the company infrastructure. The model is described by 4 maturity levels: basic, standardized, rationalized (advanced), and dynamic
Lauder of Business Intelligence (LOBI) [9]	The model describes levels of maturity in effectiveness and efficiency of decision making. IT, processes and people are assessed from the perspective of 6 levels: facts, data, information, knowledge, understanding, enabled intuition
Hawlett Package Business Maturity Model [58]	The model aims at describing the path forward as companies work toward closer alignment of business an IT organizations. It includes 5 maturity levels: operation, improvement, alignment, empowerment, and transformation
Watson's Model [63]	The model is based on the stages of growth concept, a theory describing the observation that many things change over time in sequential, predictable ways. The maturity levels include: initiation, growth, and maturity
Teradata's BI and DW MM [38]	Maturity concept is process-centric, stressing the impact of BI on the business processes The model has 5 maturity levels: reporting (what happened?), analyzing (why did it happen?), predicting (what will happen?), operationalizing (what is happing?), and activating (make it happen)

answer these questions: where the most of reporting and business analysis is done in an organization today?, who uses business reports, analysis and success indicators?, what drives BI in the organization?, which strategies for developing BI are in use today?, and what business value does BI bring? [28].

A high number of maturity models for BI has been proposed [18, 34, 63]—Table 4. One of the most popular is Gartner's Maturity Model for Business Intelligence and Performance Management [49]. It describes a roadmap for organizations to find out where they are in their usage of BI. It provides a path for progress by which they can benefit from BI initiatives. The model recognizes five levels of maturity: unaware, tactical, focused, strategic, and pervasive. The assessment includes three key areas: people, processes, metrics and technology [7, 28]. The first level is often described as "information anarchy". It means that data are incomplete, incorrect, inconsistent and organization does not have defined metrics. The uses of reporting tools are limited. The second level of BI maturity means that the organization starts to invest into BI. Metrics are usually used on the department level only. Most of the data, tools, and applications are in "silos". Users are often not skilled enough in order to take advantage of the BI system. At the third BI maturity level the organization achieves its first success and obtains some business benefits from BI, but it still applies to a limited part of the organization. Management dashboards are often requested at this level. At the strategic level, organizations have a clear business strategy for BI development. The application of BI is often extended to customers and suppliers. It supports the tactical and strategic decision making. Sponsors come from the highest management. At the last BI maturity level, BI pays pervasive role for all areas of the business and corporate culture. BI provides flexibility for adapting to the fast business changes and information demand. The users have access to information and analysis needed for creating a business value and influence business performance. The usage of BI is available to customers, suppliers, and other business partners.

Another interesting BI maturity model is the model introduced by Eckerson [18]. It includes six levels, called: "prenatal", "infant", "child", "teenager", "adult" and "sage". Maturity is being evaluated trough eight key areas: scope, sponsorship, founding, value, architecture, data, development and delivery. The prenatal phase lasts until a data warehouse is created. Reports are usually built into operational systems and limited to that individual system. At the child level the organizations bay their first interactive reporting tools, which are used to drill data. Regional data warehouse are build, but they are not linked to each other. The teenager level means that organization recognizes the value of consolidating regional data warehouse into centralized data warehouse. Such infrastructure enables to perform enterprise-wide analysis, bridging the border of individual departments gaining new knowledge. At this level customized dashboards are introduced. The main characteristics of the adult level are: centralized management of BI data sources, common architecture of the data warehouse, fully loaded with data, flexible and layered, delivery in time, predictive analysis, performance management, and centralized management. Key performance indicators and business performance are used to compare the actual state with the strategic goals of the organization. At the sage level, business and IT are aligned and cooperative. BI provides services with high added value, bringing high business value and competitive advantage. Highly customized reports and key performance indicators are applied. For faster development of different Bi solutions service oriented architecture (SOA) is used [28].

Interesting BI maturity model was presented by Davenport and Harris [15]. It explains the building the competitive advantage through BI and analytics. The model can be compared to some stages, illustrating the state and the capabilities of the organization to compete on the market through data. There are five stages of analytical competition, called: “analytically impaired”, “localized analytics”, “analytical aspiration”, “analytical companies”, and “analytical competitors”. The first stage means that “organizations have some desire to become more analytical, but thus far they lack both the will and the skill to do so”. They face some substantial barriers—both human and technical. They may also lack the hardware, software and skills to do substantial analysis. Additionally, senior executives are not enough interested to use BI systems and do not support BI initiatives. Data are not completed and inconsistent. The second stage “localized analytics” is characterized by reporting with pockets of analytical activity. The organizations undertake the first analytical activities, but they have no intention of competing on it. BI activities produce economic benefits but not enough to affect the company competitive strategy. The third stage called “analytical aspirations” is triggered when BI activities gain executive sponsorship. The organizations build the plan of using BI. The primary focus in “analytical companies” stage is building word-class analytical capabilities at the enterprise level. The organizations implement the plan developed in previous stage, making considerable progress toward building the sponsorship, culture, skills, strategic insights, data and technology needed for analytical competition. The last stage of analytical competition called “analytical competitors” means that analytics moves from being a very important capability for organizations to the key to its strategy and competitive advantage. Executive managers trust in BI and all users are highly educated in BI.

Regardless of the used model, moving from one maturity level to another requires changes in all of the characteristics that make up these stages. Achieving the highest BI maturity level is particularly complex and requires changes in management vision, founding, data management, and more [68].

5 Methodology

The study was based on: (1) a critical analysis of literature, (2) a observation of different BI initiatives undertaken in various enterprises, as well as on (3) semi-structured interviews conducted in polish enterprises in 2012. Some interviews, conducted in 20 polish enterprises, were held with over 80 responders: executives, senior members of staff, and ICT specialists They represented the service sector: telecommunications (T)-4, consulting (C)-4, banking (B)-4, insurance (I)-4, marketing agencies (MA)-4. All of them had at least 5 years of experience in BI. Interviewees were selected on their involvement in BI or on their ability to offer an insight based on experience in BI and related decision support systems. The survey was conducted in 2012 among purposefully selected firms (in Poland) that are considered to be advanced in BI. The research was of qualitative nature. Gartner’s Maturity Model for Business

Intelligence and Performance Management (described in the previous section), for the assessment of the BI maturity level in selected organizations, was used [48, 49].

6 Findings

The responders in surveyed organizations were asked about the answering different questions that concerned among others: the understanding the term of BI, using BI (what BI models are used and what areas are supported by BI), BI strategy, quality of data, motivation to use BI, sponsorship, BI skills and some benefits from using BI. Table 5 presents the selected answers for asked questions.

The collected and processed data were mapped onto Gartner's Maturity Model for Business Intelligence and Performance Management—Table 6 [49].

More detailed benefits from using different BI models in surveyed organizations are presented in Table 7.

7 Discussion

The obtained results [49] allow to state that among 20 surveyed organizations two organizations fall into the category of "pervasive" level. These were telecommunication company and marketing agency. Their analytical and BI competences are aimed at business benefits, like: acquiring new customers, launching new products and new channels of sale.

BI competences are treated by these organizations as their core competences that help them to compete on the market. Organizations achieve significant economic benefits and use BI for marketing analyses (sales profitability, profit margins, meeting sales targets, time of orders), customer analyses (time of maintaining contacts with customers, customer profitability, modeling customers' behavior and reactions, customer satisfaction), monitoring of competitors and current trends in the marketplace. The common analytical approach is used by the whole organization where broadly supported fact-based and learning culture is cultivated. The interviewees confirmed that the factors that help those organizations to stay at that high maturity level in BI, include strong support of CEO and all user's trust in BI.

An interesting group was made up of organizations classified at the fourth BI maturity level. Four organizations (1MA, 1T, 2B) in my study fit into the strategic level. They do not compete through BI, but they have high competences in using different BI analyses, like: financial analyses (reviewing of costs and revenues, calculation and comparative analyses of corporate income statements, analyses of corporate balance sheet and profitability), customer profitability, customer segmentation, improving marketing effectiveness. It seems that there is a very little to be done in order to use BI for making significant changes in running a business. Therefore, shifting these organizations from "strategic" to "pervasive" level requires more support from CEO

Table 5 Types of asked questions and selected answers

No	Asked questions during interviews	Answers (number of organizations)
1	How do you define BI?	Tools to manage information (9), data warehouse (5), analytical applications (4), new way of doing business (2)
2	What do you use BI for (reporting, ad-hoc reporting, analyzing, alerting, predictive modeling, operationalizing, optimization, activating, etc.)?	Reporting (15), ad-hoc reporting (9), analyzing (12), alerting (2), predictive modeling (2), optimization (3), activating (2)
3	Assess the quality of data used in your organization (complete, correct, consistent; high/medium/poor quality data, etc.)	High quality data (6), medium quality data (11), rather poor quality data (3)
4	Are you skilled enough in order to take advantage of BI systems?	Skilled enough (7), not skilled enough (8), poor skilled (5)
5	Do you use management dashboards?	Used management dashboards in limited scope (14), used management dashboards in whole organization (4), not used (2)
6	Is your BI (un)limited to the part/department of organization?	BI limited to the part of organization (15), unlimited (5)
7	Are you motivated to use BI (how)?	Users motivated by training (8), motivated by bonuses (6), not motivated (6)
8	Do you use BI for analyzing customers, suppliers, competitors and other business partners?	BI for analyzing customers (17), suppliers (14), competitors (5), other stakeholders (4)
9	What kind of BI software do you use?	Regional data warehouse (9), centralized data warehouse (5), operational data bases (6)
10	Describe some successes/failures from using BI	Success: acquiring new customers (14), acquiring new suppliers (11), increase of sale (8), fraud detection (6), launching new channels of sale (3), launching new products (3). Failures: not trust in BI (4), gap between BI/ business (12), users do not recognize their own data after it is processed (7), decision-making skills absent (6), BI is expensive (5)
11	Indicate some benefits from using BI	Better access to data (13), better decisions (12), improvement of business process (9), improved business performance (8), costs saving (7), transparency of information (5), new way of doing business (2)

and his/her real passion. The interviewees indicated the greater need for motivation of users for collecting, analyzing and using information.

The survey has shown that up to 9 organizations (2T, 1MA, 2C, 2I, 2B) use BI on the department level. Although they would be much more common in a random sample, and perhaps the largest group. BI in these organizations has not been

Table 6 Overview of BI maturity levels in surveyed organizations

Level	People	Process	Metrics and technology	Scope of benefits
Unaware	Users do not know their own data requirements or how to use them	Users do not know business processes; data are poor quality	Lack of appropriate hardware and software; the metrics are not defined; the use of reporting is limited	Almost none
Tactical 2I, 2C, 1MA	The users take the first BI initiatives; low support from senior executives	Identification of basic business processes	Regional data warehouses are built; analyzing trends and past data; first interactive reporting tools; metrics are usually used on the department level only	Low benefits limited to small group of users; better access to data and static reporting
Success factors: support from senior management, appropriate BI tools, quality of data, defined business processes and metrics				
Focused 2T, 1MA, 2C, 2I, 2B	Users try to optimize the efficiency of individual departments by BI	Standardization of business processes and building best practices in BI	Management dashboards are used; a centralized data warehouse is built; ad-hoc reporting, query drilldown	Benefits limited to departments and business units; improvement of internal business processes and decision making on operational level
Success factors: developing corporate culture based on facts, stating clearly BI strategy, implementing training system on BI				

(continued)

Table 6 (continued)

Strategic 1MA, 1T, 2B	Users have high BI capabilities, but often not aligned with right role	Business process management based on facts	High-quality data; have BI strategy; using more complex prediction and modeling tools; data mining	Benefits for the whole organization; integrated analysis for finance, logistics, production; improvement of decision making on all levels of management
Level	People	Process	Metrics and technology	Scope of benefits
Success factors: support from CEO, motivation of users for collecting, analyzing and using information				
Pervasive 1T, 1MA	Users have capabilities and time to use BI; skill training in BI; users are encouraged to collect, process analyze and share information; CEO passion and broad-based management commitment	Broadly supported, process-oriented culture based on facts, learning and sharing of knowledge	Enterprise-wide BI architecture largely implemented; customized reports; business and BI are aligned and cooperative	Benefits for the whole environment; competing in BI; new ways of doing business
Success factors: strong support of CEO, effective HRM and all user’s trust in BI				

Table 7 Used BI models and obtained benefits

Enterprises	Used BI models and BI analysis	Benefits from BI using
Telecommunication	Enterprise-wide BI architecture, BI-PA, customer profiling and segmentation, customer demand forecasting	(1) Determine high-profit product profiles and customer segments, provide detailed, integrated customer profiles, develop of individualized frequent-caller programs, determine future customer needs; (2) Forecast future product needs or service activity, provide basis for churn analysis and control for improving customer retention
Consulting	Data warehouse, BI-PA, data marts, analysis of parameters importance, identification of sales and inventory, optimization orders, marketing companions	(1) Reduction in the turnaround time for preparation of reports, direct and faster access to the data needed to support decision-making, analyze the flow of businesses across services, regions, clients, pricing, currencies, and market factors in time etc.; (2) Forecasting and estimating of customer demand (in short and long term); (3) Service and product distribution plans of a companies are in place to meet its customer expectations, inventory requirements are more accurately
Banking	Data warehouse, BI-PA, customer profitability analysis, credit management, branch sales	(1) Determinate the overall profitability of individual customer, current and long term, provide the basis for high-profit sales and relationship banking, maximize sales to high-value customers, reduce costs to low-value customers, provide the means to maximize profitability of new products and services; (2) Establish patterns of credit problem progression by customers class and type, warn customers to avoid credit problems, to manage credit limits, evaluate of the bank's credit portfolio, reduce credit losses; (3) Improve customer service and account selling, facilitate cross selling, improve customer support, strengthen customer loyalty

(continued)

Table 7 (continued)

Enterprises	Used BI models and BI analysis	Benefits from BI using
Insurance	Regional data warehouses, data mining, OLAP, data marts, claims and premium analysis, customer analysis, risk analysis	(1) Analyzing detailed claims and premium history by product, policy, claim type, and other specifics; (2) Developing marketing programs on client characteristics, improving client service; (3) Identification high-risk market segments and opportunities in specific segments, reducing frequency of claims
Marketing agencies	Regional data warehouses, OLAP, marketing companions, customer profiling and segmentation, customer demand forecasting	(1) Better understanding of customers, identification their place in a customer lifetime cycle and customer segments for marketing campaigns; (2) Providing analyses of customer transactions (what is selling, who is buying); (3) Monitor customer loyalty by evaluating which customers are loyal and which are likely to leave; (4) Identify which products are most profitable and monitor customer behavior in purchasing products. By closely tracking sales performance and consumer behavior, companies are able to set better marketing strategies and ensure proper allocation of marketing funds

playing a strategic role and benefits from it are limited. BI is used to perform ad hoc reporting and to answer questions related to departments' ongoing operations, up-to-date financial standing, sales and co-operation with suppliers and customers. BI and management are often not aligned. The observation and interviews with senior executives allow to state that the lack of appropriate knowledge about possibilities of BI among staff results in a relatively low use of it. Therefore, the main tasks for organizations include first of all: developing corporate culture based on facts and learning, stating clearly BI strategy and implementing training system on BI.

I found in my study that 5 organizations (2I, 2C, 1MA) are at the position of "tactical" maturity level. They use a traditional approach to management, focused more on the performing the basic tasks of departments than on business processes. The knowledge about BI in these organizations is rather low and identified mainly with regional data warehouses or databases. Only basic business processes are recognized and basic metrics are used. The interviewees indicated that many users have some problems with recognizing their own data after processing. The users have rather low experience with other types of management information systems. Their intellectual resources are not adequate in order to develop complex BI infrastructure and to use it for improvement of business processes and decision-making.

I wonder why organizations in a similar segment, with similar financial resources and comparable BI infrastructure, derive from BI such diverse benefits (e.g. in the studied case, telecommunications companies and advertising agencies). Seeking the answer to this question it should be noted that the organizations, that have been classified into the category of BI "pervasive" level, were highly determined to collect, process, analyze, and share information. Corporate culture based on facts and learning helped them to use chances offered by BI. The most important factors that decided on the success of BI initiatives refer not to the technology, but to the strong believe of all users in BI.

The conducted interviews in surveyed organizations allow to state that the organizations use BI systems first of all to optimize operational decisions, improvement of internal business processes and decision making on operational level and to better access to data and static reporting. The majority of the organizations apply information systems like: ERP, MRP II, CRM and operational systems. Unfortunately, only one organization (in conducted survey) in a professional way was able to monitor and process information about their competitors, suppliers and customers. Then, such information was converted into knowledge and consequently made use of gathering intelligence in decision process. Only a few enterprises indicated the benefits from the analysis of the whole environment that leads to competing on BI and new ways of doing business. The surveyed organizations do not build the social nets and manage social capital.

8 Conclusions

This paper has explored the possibility of BI using in organizations. Different BI models have been presented. It has been argued that BI maturity models provide a path for progress by which they can benefit from BI initiatives.

The main conclusion of this study is that BI may offer different possibilities for the organizations. They include first of all: making more effective decisions, improving business processes, and business performance. Unfortunately, conducted survey allows to state, that the local, rather small organizations very seldom use more advanced BI models e.g. for building expert' nets, social capital management, creating the active communities, and knowledge sharing. Most of them stay still at the age of traditional BI (with data warehousing, OLAP and reports). They are focused more on the internal business processes than on the environment: competition, users in social media etc.

The survey was confirmed that the factors that allow organizations to achieve business benefits with BI, include first of all: management leadership and support, corporate culture, expressed by effective information resources management, clearly stated strategy and objectives, and use of appropriate BI technologies. Additionally, the important factors were: clearly defined business processes, business performance measurement, incentive system to encourage collecting, analyzing information and knowledge sharing, appropriate resources (financial, intellectual), training and education on BI and knowledge management.

References

1. Adelman, S., Moss, L.: Data Warehouse Project Management. Addison-Wesley, NJ (2000)
2. Albescu, F., Pugna, I., Paraschiv, D.: Business intelligence & knowledge management—technological support for strategic management in the knowledge based economy. Rev. Inform. Econ. **4**(48), 5–12 (2008)
3. Alter, A.: A work system view of DSS in its forth decade. Decis. Support Syst. **38**(3), 319–327 (2004)
4. Azvine, B., Cui, Z., Nauck, D.: Towards real-time business intelligence. BT Technol. J. **23**(3), 214–225 (2005)
5. Baaras, H., Kemper, H.G.: Management support with structured and unstructured data—an integrated business intelligence framework. Inf. Syst. Manage. **25**(2), 132–148 (2008)
6. Barney, J.: Looking inside for competitive advantage. Acad. Manag. Exec. **9**, 49–61 (1995)
7. Burton, B.: Toolkit: Maturity Checklist for Business Intelligence and Performance Management, Gartner Research (2009)
8. Business Objects: About business intelligence (2007). http://www.businessobjects.com/businessintelligence/default.asp?intcmp=ip_company2. Accessed February 2012
9. Cates, J.E., Gill, S.S., Zeituny, N.: The Ladder of business intelligence (LOBI): a framework for enterprise IT planning and architecture. Int. J. Bus. Inf. Syst. **1**(1–2), 220–238 (2005)
10. Chaudhary, S.: Management factors for strategic BI success. In: Raisinghani, M.S. (ed.) Business Intelligence in Digital Economy. Opportunities, Limitations and Risks, pp. 191–206. IGI Global, Hershey (2004)
11. Chen, H., Chiang, R.H.L., Storey, V.C.: Business intelligence and analytics: from big data to big impact. MIS Quart. **36**(4), 1–24 (2012)
12. Chung, W., Chen, H., Nunamaker, J.F.: A visual framework for knowledge discovery on the web: an empirical study of business intelligence exploration. J. Manag. Inform. Syst. **21**(4), 57–84 (2005)
13. Clavier, P.R., Lotriet, H., Loggerenberger, J.: Business intelligence challenges in the context of goods-and service-domain logic. In: 45th Hawaii International Conference on System Science, IEEE Computer Society, pp. 4138–4147 (2012)
14. Cosic, R., Shanks, G., Maynard, S.: Towards a business analytics capability maturity model. In: 23rd Australasian Conference on Information Systems Business Analytics Capability Maturity, Geelong (2012)
15. Davenport, T.H., Harris, J.G.: Competing on Analytics. The New Science on Winning. Harvard Business School Press, Boston (2007)
16. Davenport, T.H., Harris, J.G., Morison, R.: Analytics at Work: Smarter Decisions. Better Results. Harvard Business Press, Cambridge (2010)
17. Deng, R.: Business intelligence maturity hierarchy: a new perspective from knowledge management, information management (2011). http://www.information-management.com/infodirect/20070323/1079089-1.html. Accessed September 2011
18. Eckerson, W.W.: The keys to enterprise business intelligence: critical success factors. The data warehousing institute (2005). http://download.101com.com/pub/TDWI/Files/TDWIMonograph2-BO.pdf. Accessed October 2011
19. Ferguson, M.: Architecting a big data platform for analysis. Intelligent business strategies, IBM, white paper (2012)
20. Glancy, F.H., Yadav, S.B.: Business intelligence conceptual model. Int. J. Bus. Intell. Res. **2**(2), 48–66 (2011)
21. Gratton, S.J.: The journey to business intelligence. What does it mean? (2012). http://www.capgemini.com.technology. Accessed October 2012
22. Gurjar, Y.S., Rathore, V.S.: Cloud business intelligence—is what business need today. Int. J. Recent Technol. Eng. **1**(6), 81–86 (2013)
23. Hagerty, J.: AMR research's business intelligence/performance management maturity model, Version2 (2011). http://www.eurim.org.uk/activities/ig/voi/AMR_Researchs_Business_Intelligence.pdf. Accessed September 2011

24. Hannula, M., Pirttimaki, V.: Business intelligence empirical study on the top 50 Finnish companies. J. Am. Acad. Bus. **2**(2), 593–599 (2003)
25. Hawking, P., Foster, S., Stein, A.: The adoption and use of business intelligence solutions in Australia. Int. J. Intell. Syst. Technol. Appl. **4**(1), 327–340 (2008)
26. Herschel, R.T., Jones, N.E.: Knowledge management and business intelligence: the importance of integration. J. Knowl. Manage. **9**(4), 45–54 (2005)
27. Howson, C.: Successful Business Intelligence: Secrets to Making BI a Killer Application. McGraw-Hill, New York (2008)
28. Hribar Rajteric, I.H.: Overview of business intelligence maturity models. Management **15**(1), 47–67 (2010)
29. Inmon, W.H., Strauss, D., Neushloss, G.: DW 2.0: The Architecture for the Next Generation of Data Warehousing. Elsevier, Amsterdam (2008)
30. Isik, O., Jones, M.C., Sidorova, A.: Business intelligence (BI) success and the role of BI capabilities. Intell. Syst. Account. Finance Manag. **18**, 161–176 (2011)
31. Jordan, J., Ellen, C.: Business need, data and business intelligence. J. Digit. Asset Manage. **5**(1), 10–20 (2009)
32. Jourdan, Z., Rainer, R.K., Marschall, T.: Business intelligence: an analysis of the literature. Inf. Syst. Manage. **25**(2), 121–131 (2007)
33. Karim, A.J.: The value of competitive business intelligence system (CBIS) to stimulate competitiveness in global market. Int. J. Bus. Soc. Sci. **2**(19), 196–203 (2011)
34. Lahrmann, G., Marx, F., Winter, R., Wortmann, F.: Business intelligence maturity: development and evaluation of a theoretical model. In: Proceedings of the 44 Hawaii International Conference on System Science (2011)
35. Liautaud, B., Hammond, M.: Turning information into knowledge into profit. E-Business Intelligence. McGraw-Hill, New York (2002)
36. Lonnqvist, A., Pirttimaki, V.: The measurement of business intelligence. Bus. Intell. **23**(1), 32–40 (2006)
37. McGonagle, J.J., Vella, C.M.: Bottom Line Competitive Intelligence. Quorum Books, Westport (2002)
38. Miller, L., Schiller, D., Rhone, M.: Data warehouse maturity assessment service. TERADATA (2011).http://www.teradata.com/assets/0/206/276/3457d45f-7327-4a36-b1dc-2e5daae3d269.pdf.Accessed September 2011
39. Moss, L., Atre, S.: Business Intelligence Roadmap: The Complete Lifecycle for Decision-Support Applications. Addison-Wesley, Boston (2003)
40. Negash, S.: Business intelligence. Commun. Assoc. Inf. Syst. **13**, 177–195 (2004)
41. Negash, S., Gray, P.: Business intelligence. In: Holsapple, C.W., Burstein, F. (eds.) Decision Support Systems, pp. 175–193. Springer, Berlin (2008)
42. Nemec, R.: The application of business intelligence 3.0 concept in the management of small and medium enterprises. In: Tvrdikova, M., Minster, J., Rozenhal, P. (eds.) IT for Practice. Economicka Faculta, VSB-TU, Ostrava, p. 84–89 (2012)
43. O'Brien, J.A., Marakas, G.M.: Introduction to Information Systems, 13th edn. McGraw-Hill, New York (2007)
44. Olszak, C.M.: Tworzenie i wykorzystywanie systemów Business Intelligence na potrzeby współczesnej organizacji. Akademia Ekonomiczna w Katowicach, Katowice (2007)
45. Olszak, C.M.: Competing with business intelligence. In: Tvrdikova, M., Minster, J., Rozenhal, P. (eds.) IT for Practice, pp. 98–108. Economicka Faculta, VSB-TU, Ostrava (2012)
46. Olszak, C.M., Ziemba, E.: Business intelligence as a key to management of an enterprise. In: CohenE, Boyd E. (ed.) Proceedings of Informing Science and IT Education InSITE'2003. The Informing Science Institute, Santa Rosa (2003)
47. Olszak, C.M., Ziemba, E.: Critical success factors for implementing business intelligence systems in small and medium enterprises on the example of upper Silesia, Poland. Interdisc. J. Inf. Knowl. Manage. **7**, 129–150 (2012)

48. Olszak, C.M.: The business intelligence-based organization—new chances and possibilities. In: Ribiere, V., Worasinchai, L. (eds.) Proceedings of the International Conference on Management, Leadership and Governance. Published by Academic Conferences and Publishing International Limited Reading, Bangkok University, p. 241–249 (2013)
49. Olszak, C.M.: Assessment of business intelligence maturity in the selected organizations. In: Ganzha, M., Maciaszek, L., Paprzycki, M. (eds.) Annals for Computer Science and Information systems. Proceedings of the 2013 Federated Conference on Computer Science and Information Systems (FedCSIS) vol. 1, pp. 951–958 (2013)
50. Power, D.J.: A brief history of decision support systems (2007). http://dssresources.com/history/dsshistory.html. Accessed August 2007
51. Rayner, N.: Maturity Model of Overview for Business Intelligence and Performance Management. Gartner Inc., Research (2008)
52. Reinschmidt, J., Francoise, A.: Business Intelligence Certification Guide. IBM,International Technical Support Organization, San Jose (2000)
53. SAS: Information evaluation model (2011). http://www.sas.com/software/iem/ Accessed September 2011
54. Sauter, V.L.: Decision Support Systems for Business Intelligence. Wiley, New Jersey (2010)
55. Schick, A., Frolick, M., Ariyachandra, T.: Competing with BI and analytics at monster worldwide. In: 44th Hawaii International Conference on System Sciences, Hawaii (2011)
56. Scott, N.: The 3 ages of business intelligence: gathering, analysing and putting it to work (2013). http://excapite.blogspot-ages-of-business-ontelligence.html. Accessed January 2013
57. Tamer, Ch., Kiley, M., Ashrafi, N., Kuilbar, J.: Risk and benefits of business intelligence in the cloud. In: Northeast Decision Sciences Institute Annual Meeting Proceedings, pp. 86–95 (2013)
58. The HP Business Intelligence Maturity Model, Describing the BI Journal 2011. Hewlett-Packard. http://www.techrepublic.com/whitepapers/the-hp-business-intelligence-maturity-model-describing-the-bi-journey/1129995. Accessed September 2011
59. Turban, E., Sharada, R., Aronson, J.E., King, D.: Business Intelligence: a Managerial Approach. Pearson Prentice Hall, Boston (2008)
60. Venter, P., Tustin, D.: The availability and use of competitive and business intelligence in South African business organizations. S. Afr. Bus. Rev **13**(2), 88–115 (2009)
61. Watson, H.J.: SME Performance: Separating Myth from Reality. Edward Elgar Publishing, Cheltenham (2010)
62. Watson, H.J., Wixom, B.H.: The current state of business intelligence. IEEE Comput. **40**(9), 96–99 (2007)
63. Watson, H.J., Ariyachandra, T., Matyska, R.J.: Data warehousing stags of growth. Inf. Syst. Manage. **18**(3), 42–50 (2011)
64. Weiss, A.: A brief guide to competitive intelligence. Bus. Inf. Rev. **19**(2), 39–47 (2002)
65. Wells, D.: Business analytics—getting the point (2012). http://b-eye-network.com/view/7133. Accessed August 2012
66. White, C.: Now is the right time for real-time BI. Inf. Manage. Mag. (2004). http://www.dmreview.com. Accessed September 2004
67. Williams, S., Williams, N.: The Profit Impact of Business Intelligence. Morgan Kaufmann, San Francisco (2007)
68. Wixom, B.H., Watson, H.J.: The BI-based organization. Int. J. Bus. Intell. Res. **1**, 13–28 (2010)

Low-Frequency Signal Reconstruction and Abrupt Change Detection in Non-stationary Time Series by Enhanced Moving Trend Based Filters

Tomasz Pełech-Pilichowski and Jan T. Duda

Abstract An original approach to digital moving trend based filters (MTF) design, based on Bode plots analysis is proposed, aimed at seasonal time series decomposition and prediction [5]. A number of polynomials of different range are discussed to be used in the MTF as the LS approximation formula. The Bode plots of the MTF are shown, and the best filter is selected. Results of a seasonal time series decomposition and prediction with the best MTF is presented and compared to the classical MTF calculations (involving the linear LS approximation). The MTF enhancementsareintroduced aimed at better change detection. Efficiency of low-frequency periodic signals reconstruction and step-wise changes is illustrated.

1 Introduction

The nonstationary time series filtering with moving trends is the well-known approach to a nonparametric long term trend extraction from the series, aimed at further processing of stationary residuals and the series prediction [2, 3]. The classical moving trend filter (MTF) is based on rolling approximation of the series with the least-square (LS) linear approximation in a moving window [3]. The window width affects the extracted trend smoothness and cyclic components separation effectiveness. However typically, it is adjusted by a trial method, to reach the appropriately smooth nonparametric trend. This paper shows that much better smoothing properties and cyclic component extraction may be reached by using in MTF higher order polynomial approximations and by specification of the required filter properties in frequency domain. Hence, the MTF design is proposed by analysis of Bode plots [10] of a number of the filter variants. The MTFs designed in this way were successfully applied to analysis of hydrogeological data [7] and to financial time series

T. Pełech-Pilichowski (✉) · J.T. Duda
AGH University of Science and Technology, Al. A. Mickiewicza 30, 30-059 Krakow, Poland
e-mail: tomek@agh.edu.pl

J.T. Duda
e-mail: jdu@agh.edu.pl

M. Mach-Król et al. (eds.), *Advances in ICT for Business, Industry and Public Sector*,
Studies in Computational Intelligence 579, DOI 10.1007/978-3-319-11328-9_7

prediction [6]. Smoothing and prediction of a step change and a cyclic signal with the studied MTF was shown to illustrate their properties [5]. An Adaptive adjustments of the MTF are proposed aimed at enhancement of step-wise changes detection and the low-frequency periodic signals reconstruction [4]. In particular, the described approach allows for event/change detection of patterns based on historical data.

2 Moving Trend Based Filters—Formal Basis and Properties

Nonstationary time series $y(t)$ may be viewed as the sum of an aperiodic trend function $f(t)$, a cyclic component $C(t)$ of time period T, and a higher frequency zero-average noise $z(t)$ [1, 6, 8]:

$$y(t) = f(t) + C(t) + z(t) \tag{1}$$

The periodic component can be written in the form of the harmonic series [10]:

$$C(t) = \sum_{k=1}^{K} A_k \sin(\omega_T i_k (t - \tau_k)), \omega_T = \frac{2\pi}{T} \tag{2}$$

where $i_k, k = 1, \dots, K$ denote the set of harmonics indices of the consecutive components $k = 1, \dots K$ (e.g. $i_k = 1, 2, 10$), A_k—the amplitude of i_kth harmonic, τ_k is the delay of the kth component.

The nonparametric trend $f(t)$ may be calculated for each time step t_n by a low-pass digital filter designed in such a way to remove the ω_T and higher frequency components from the original series $y(t)$. The cyclic component $C(t)$ can be extracted from the filtering residuals by the Least Square (LS) approximation with the regression model of the form (2), and then, the regression residuals $z(t)$ may be viewed as a high-frequency stochastic process and treated with ARMA approach [2] (if its homoscedasticity can be assumed) or with GARCH models in case of its heteroscedasticity [1].

One of the techniques recommended to calculate the nonparametric trend $f(t_n)$ is a rolling approximation of the series $y(t_n)$ with the LS linear approximation in a window containing M samples, and then averaging of the approximates $y_F(i, t_n)$ obtained for each t_n [3]. It is referred to as moving trend based smoothing/filtering, which may be further used to h-samples ahead prediction of the series main component $f(t_n + h)$ by its extrapolation with a h-samples increment $\Delta_h f$ averaged with harmonic weights [3]:

$$f(t_{n+h}) = f(t_n) + \Delta_h f, \;\; \Delta_h f = \sum_{i-1}^{n-h} C_i \left(f(t_{i+h}) - f(t_i)\right) \tag{3}$$

$$C_0 = 0, \quad C_i = C_{i-1} + \frac{1}{(n-h)(n-h-i+1)}$$

Hereby we propose a generalization of the moving trend smoothing algorithm, by employing higher order approximating polynomials, with appropriately designed properties. Let us consider the polynomial of the form (4) in the time interval of M samples, with the time counted from $-M+1$ to 0:

$$y_F\,(t_i) \overset{\text{def}}{=} b_0 + b_1 t_i + b_2 t_i^2 + b_3 t_i^3 + b_4 t_i^4, \quad t_i = \{-M+1, \ldots, -1, 0\} \tag{4}$$

The approximates Y_{Fk} in a window ending at k-th sample can be calculated with the following LS formula (5):

$$Y_{Fk} = Y_k \cdot W,\ \ W \overset{\text{def}}{=} U[U^T U]^{(-1)} U^T,\ \ u_i = [1, t_i^{p1}, \ldots, t_i^{pL}] \\ i = 1, \ldots, M, t_i = -M+1, \ldots, 0 \tag{5}$$

where $Y = [y(t_1), \ldots, y(t_M)]$, $Y_F = [y_F(t_1), \ldots, y_F(t_M)]$, U—the model input matrix (M rows $u_i(t_i)$), t_{pk}—selected p_k-th order monomials, W denoted a constant $M \times M$ filtering matrix, each j-th column $w(:, j)$ of which may be viewed as a FIR (Finite Impulse Response) digital filter [2, 9] producing a value for $y_F(i, t_j)$.

The derivatives of $y_F()$ at the interval end ($t_i = 0$) can be easily shaped by fixing selected coefficients b_k at zero values, which implies different profiles of the LS approximates y_F, as shown in Fig. 1.

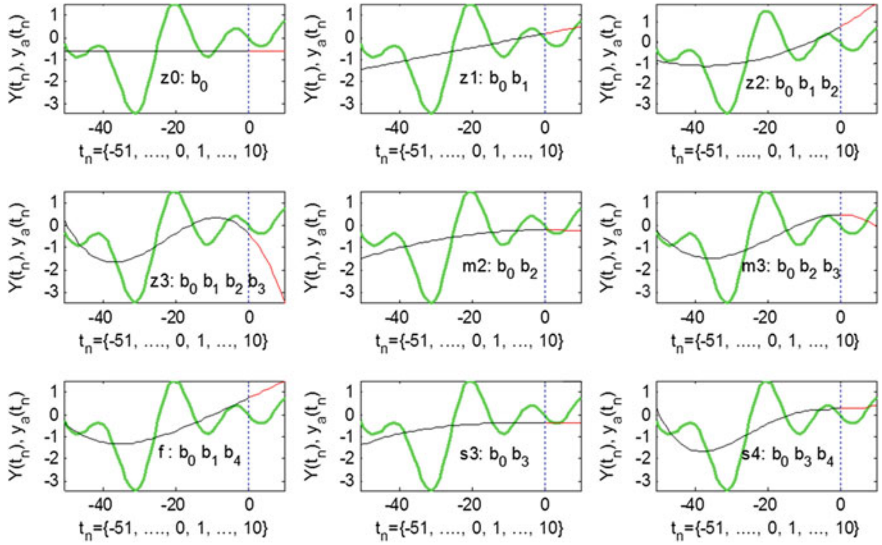


Fig. 1 Properties of the approximating polynomials expressed by Eq. 4 considered to be used in the moving trend based filters; filter codes z0, z1,…, s3, s4 used in sequel and corresponding nonzero coefficients are listed; *vertical dotted line* shows the interval end; *shadow line* $y(t_n) = \sin(0.5\omega_T t_n)\sin(\omega_T t_n) + \sin(2\omega_T t_n)\sin(3\omega_T t_n),\ \ T = M = 52$

In the moving trend algorithms the series splits into three sections. The first (starting *s*) section begins at the first (oldest) sample and finishes with the $(M-1)$th one, the filtering window width enlarges from M to $2M-2$, and the number L_n of the approximates $y_F(i, t_n)$ to be averaged increases from 1 to the $M-1$. The second (central *c*) section ranges from the *M*-th to $n-M+1$ samples, the filtering window width is $2M-1$ (constant), and the number of approximates is *M*. The third (final *f*) section contains the samples from $n-M+2$ to n (the newest one), the window width reduces from $2M-2$ to *M*, and the number of approximates $y_F(i, t_n)$ to be averaged decreases from $M-1$ to 1.

The calculations in the sections $\{s, c, f\}$ may be expressed in the FIR filtering form [1, 7]:

for $i = 1, \ldots, M-1$:

$$f(t_i) = \sum_{k=1}^{i+M-1} g_s(i,k) \cdot y(t_{(i+M-k)}) = \sum_{k=1}^{n} G_s(k,i) \cdot y(t_(n-k+1)), \tag{6}$$

$$G_s\,(k,i) \overset{\text{def}}{=} [g_s\,(i,k)\,, 0_{(n-i-M+1)}]^T,$$

for $i = M, \ldots, n-M+1$:

$$f(t_i) = \sum_{k=1}^{2M-1} g_c(k) \cdot y(t_{i+M-k}) = \sum_{k=1}^{n} G_c(k,i) \cdot y(t_{n-k+1}), \tag{7}$$

$$G_c\,(k,i) \overset{\text{def}}{=} [0_{(i-M+1)}, g_c, 0_{(n-i-M)}]^T,$$

and for the final section, $i = n-M+2, \ldots, n$:

$$f(t_i) = \sum_{k=1}^{2M-j-1} g_f(j,k) \cdot y(t_{i+M-j-k}) = \sum_{k=1}^{n} G_f(k,i) \cdot y(t_{n-k+1}), \tag{8}$$

$$G_f\,(k,i) \overset{\text{def}}{=} [0_{(i-M)}, g_f\,(j,k)]^T,$$

where g_s, g_c, g_f denote the impulse response vectors of filters in the sections ***s, c, f***, written also as the columns G_s, G_c and G_f of the unified smoothing filter matrix $G_{nx(n-1)}$, and the proper filter vector $G(k,n)$.

The impulse response coefficients $g.()$ of the moving trend filters (MTF) attributed to the sections $\{s, c, f\}$ are as follows (Eqs. 9–11):

- for the starting section *s*:

$$g_s\,(k,j) = \sum_{m=j}^{\min(M-1,k)} \frac{w(M-k+m, m-j+1)}{M-j}, j = 1, \ldots, M-1, k = 1, \ldots, 2M-2 \tag{9}$$

- for the central section c:

$$g_c\,(k,1)=\sum_{m=\max(k-M+1,1)}^{\min(M,k)}\frac{w(M-k+m,m)}{M},k=1,\ldots,2M-1 \tag{10}$$

- for the final section f:

$$g_f\,(k,j)=\sum_{m=\max(k-M+1,1)}^{\min(j,k)}\frac{w\,(M-k+m,M-j+m)}{j},j=1,\ldots,M-1,k=1,\ldots,2M-2 \tag{11}$$

The prediction formula (3) may be written in the following convolution form:

$$\begin{gathered}f(t_(n+h))=\sum_{k=1}^{n}P_h(k)\cdot y(t_{n-k+1})\\ P_h(k)\overset{\text{def}}{=}G_f(k,M-1)+\sum_{i=1}^{n-h}C_i\cdot(G(k,i+h)-G(k,i))\\ G\overset{\text{def}}{=}[G_s,G_c,G_f]\end{gathered} \tag{12}$$

Notice that P_h are strongly affected by properties of the proper (the worst) filter $G_f(k,M-1)=G(k,n)$.

By making the Fourier Transform of the filters g_s, g_c, g_f and P_h involving different approximating polynomial types $\{z0.\ldots s4\}$ with different M (see Fig. 1), one may examine their properties in frequency domain, and select a filtering variant (type, M) suitable for smoothing and/or prediction demands, usually related to ω_T viewed as the cut-off frequency of the designed low-pass filters. The Bode plots of the examined

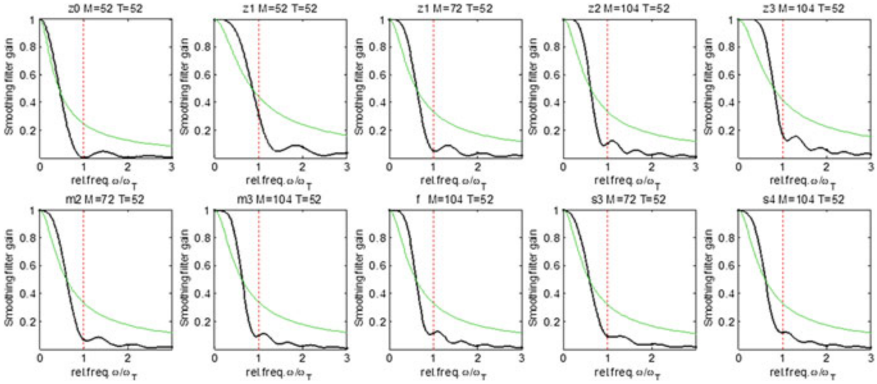


Fig. 2 Gain diagrams (vs. ω/ω_T) for the best smoothing filter (central section); *vertical point lines* show ω_T, *shadow solid lines*—gain diagram for the 1st order recursive filter of the same half-gain frequency

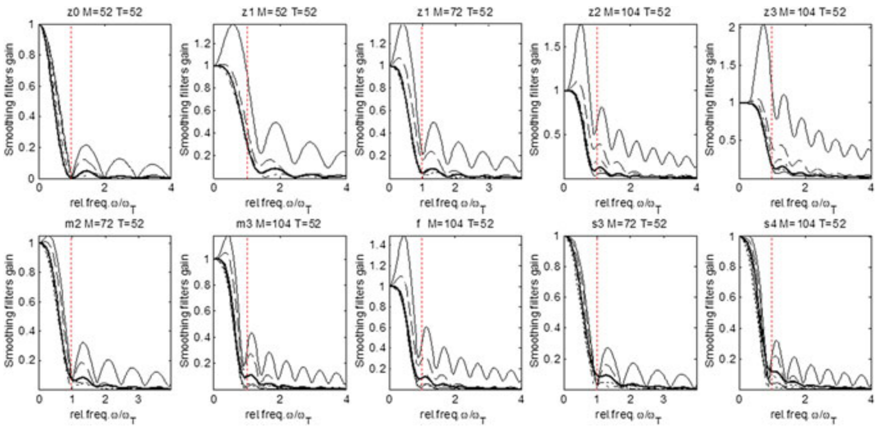


Fig. 3 Gain diagrams (vs. ω/ω_T) for the smoothing filters: central section—*bold lines* $h < M$, the final section for $h = 0$ *solid lines* (proper filter), $h = 20$ *dotted lines*, $h = 40$ *point-dotted lines*

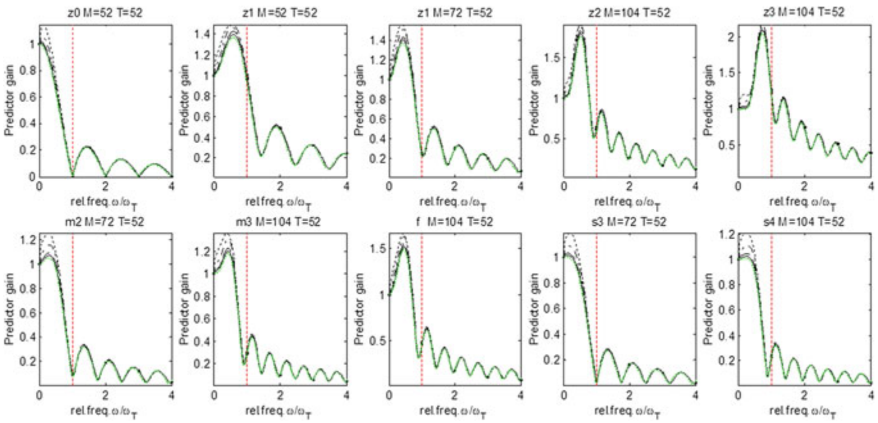


Fig. 4 Gain diagrams for the moving trend based predictors: *shadow bold line*—final filter $h = 0$; $h = 5$ *solid lines*, $h = 10$ *dotted lines*, $h = 26$ *point-dotted lines*, $h = T = 52$ *point lines*

filters are shown in Figs. 2, 3, 4, 5 and 6. We have stated that the approximation window width M affects directly g_s, g_f and P_h delays, but it is of almost no effect on a shape of all the filers gain. Hence M may be taken as the lowest value producing gains close to 1 for $\omega < \omega_T$, near zero for $\omega = \omega_T$ and close to 0 for $\omega > \omega_T$.

Figure 2 shows the central smoothing filters are much better than 1st order recursive ones.

Gain properties of all the smoothing filters in the central section (see Fig. 2) are similar. When assuming $M = 1.38^*T = 72$, the classical filter z1 seems to be the best due to the pass- and attenuation band properties as well as cut-off frequency gain, although z3 pass-band and attenuation of z0 look better. However a view on Figs. 3

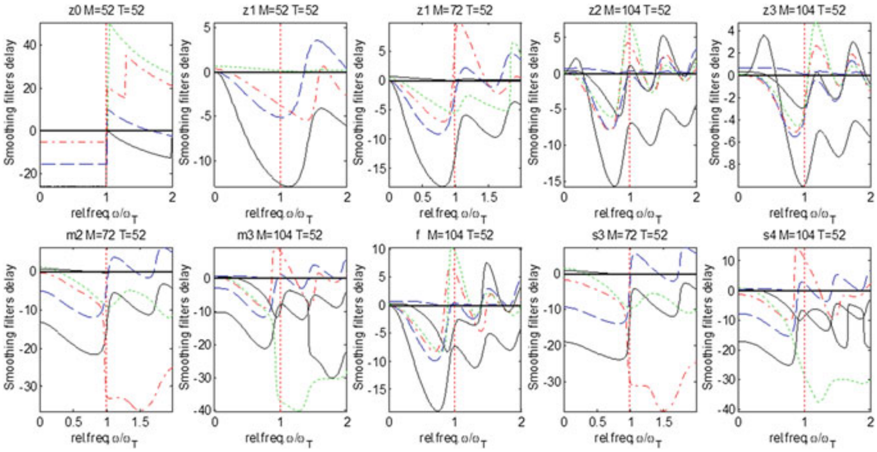


Fig. 5 Delay of the final section smoothing filters: $h = 0$ *solid lines*, $h = 20$ *dotted lines*, $h = 40$ *point-dotted lines*, $h = 60$ *point lines*, $h = 70$ *solid lines* close to the zero-delay, *bold-line* 0 delay of central section filter

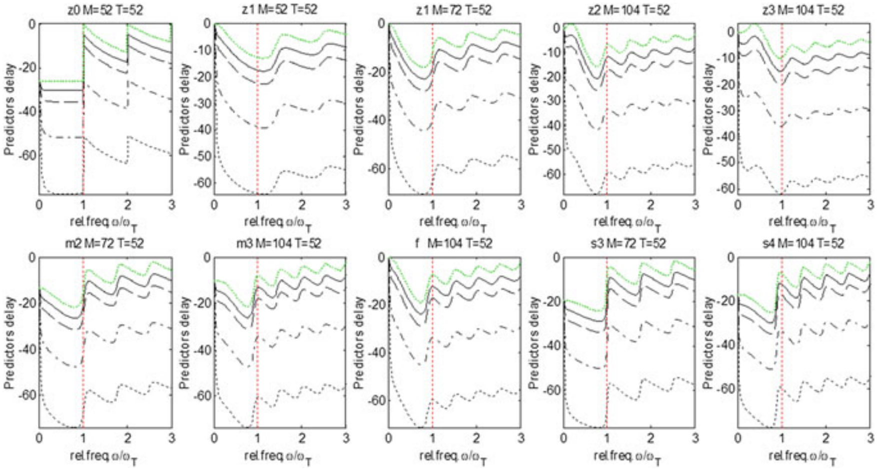


Fig. 6 Predictors delay: $h = 0$ *shadow point lines*, $h = 5$ *solid lines*, $h = 10$ *dotted lines*, $h = 26$ *point-dotted lines*, $h = T = 52$ *point lines*

and 4 gives evidence that only z0, s3 and s4 might be accepted from the perspective of final section smoothing (Fig. 3) and prediction (Fig. 4) properties. In particular, very bad pass and attenuation properties (excessive gain) of the classical filter z1 are clearly seen. Having in mind numerical problems (ill-conditioning) which can be met in s4 for larger M, one may take that the filter s3 with $M = 72(1.38T)$ is the best choice (its pass-band is noticeably better than that of z0). The same conclusion may be drawn on a basis of delay properties shown in Figs. 5 and 6. In the pass-band a

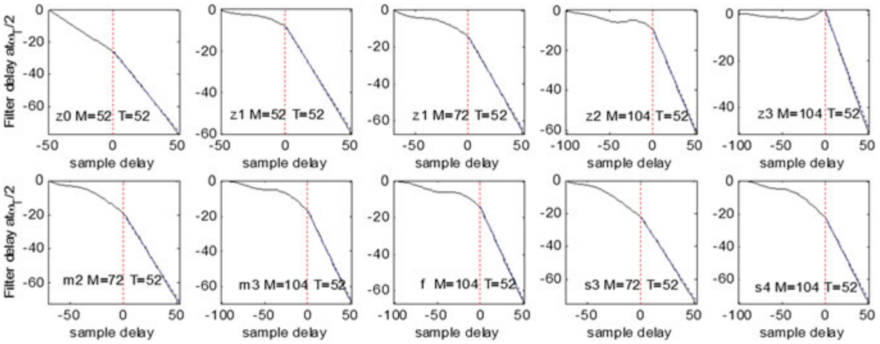


Fig. 7 Delay of smoothing filters and predictors for $\omega_T/2$ versus the sample delay

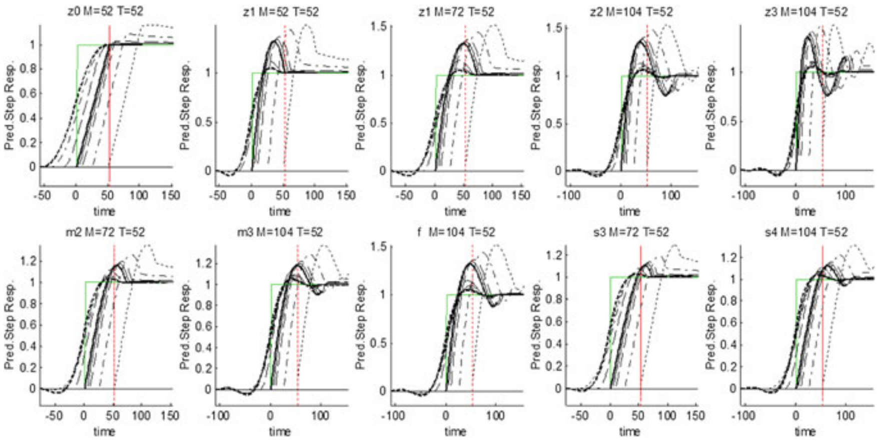


Fig. 8 Signal step-wise change smoothing and prediction: *shadow bold line*—signal; *bold dotted line*—central segment filter response ($h < -M$); *dotted point line* smoothing with $h = 40$, $h = 20$ *dotted line*; *bold line* final filter response ($h = 0$); prediction with $h = 5$ *solid lines*, $h = 10$ *dotted lines*, $h = 26$ *point-dotted lines*, $h = 52$ *point lines*

close to uniform and small delay is required (minimum delay distortion of the trend). It is satisfied only by z0 filters, but s4 delay distortion is acceptable and significantly lower than for the classical filter z1. The delay of predictors is larger than that of the final filter ($h = 0$) by prediction horizon (see Fig. 7). It means that the MTF prediction (Eq. 3) does not differ essentially from Zero Order Hold of the $f(t_n)$.

The frequency properties presented above are visible in time domain responses—see Figs. 8 and 9. Step change distortions shown in Fig. 8 are the larger, the greater irregularities of the pass-band gain and delay.

Figure 9 illustrates effects of filtering and prediction of a periodic signal (used also as the example in Fig. 1). The harmonic of $\omega = \omega_{\mathrm{T}}/2$ has to be extracted (reconstructed), but the signal contains strong components in the filters attenuation

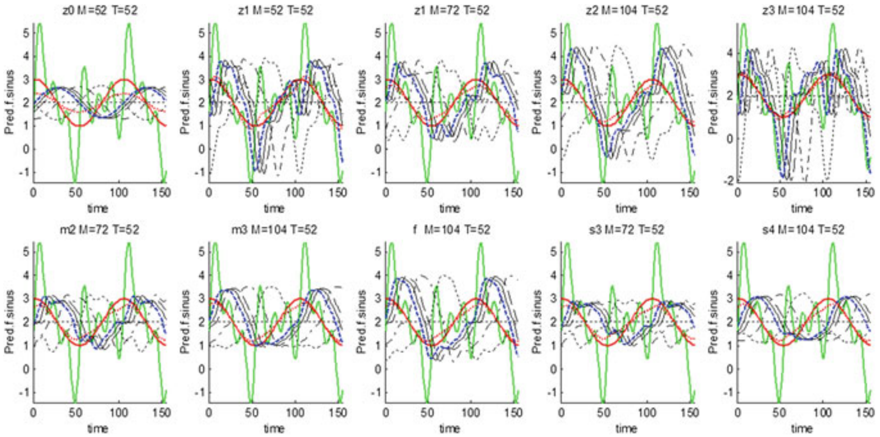


Fig. 9 Periodic signal processing with the studied filters: *shadow bold line*—the signal $y(t_n)$ (see Fig. 1); *bold line*—the main harmonic of y, $\omega = 0.5\omega_T$ (to be extracted), *bold point lines*—the main harmonic reconstruction by the central filter response ($h = M$), the main harmonic reconstruction with the final filter—*bold dotted lines*, and prediction with $h = 5$ *solid lines*, $h = 10$ *dotted lines*, $h = 26$ *point-dotted lines*, $h = 52$ *point lines*

band. Hence the attenuation gain profile is of significant effect on the extracted signal shape. The best reconstruction is reached with z0 filters. The classical (z2) filters produce highly distorted responses, both is smoothing and prediction cases, while s3 and s4 yield acceptable results. All predictions are similar in shape to the proper filter response ($h = 0$) and additionally delayed by the prediction horizon—see Fig. 7 (i.e. they do not differ noticeably from ZOH predictions).

The extracted signal distortion in the starting and final sections is significant, hence separation of the filtering residuals into periodic $C(t)$ and stochastic $z(t)$ components, by fitting the regression model (2), should be performed with the central **section** data only. Then the periodic component $C(t)$ should extrapolated on the full data interval and subtracted from the filtering residuals to get $z(t)$.

3 MTF Enhancements—Efficiency of the Low-Frequency Periodic Signals Reconstruction

The signal distortion may be significantly reduced through the signal compression by the pass-band averaged delay value. A revised package of the g_{fsj} filters of the same amplitude characteristics as g_{fi}, but of reduced delays can be obtained by changing the positions of filters g_{fi}, $I = -M + 1, \ldots, 0$, to the positions $j = M\tau_i$ (with canceling the filters previously set at this point) [4].

To avoid the compressed signal distortion, before the final filtration of seasonal time series, the periodic component from the original series in the final section has to be removed. Distortion resulting from varying delay of high-frequency random components can be reduced by the additional filtration of compressed subseries by the moving average in the window of the length $L_f = 2 \cdot d_f + 1$, where $d_f = \max\{\mathrm{int}(T/50),\ 1\}$($L_f$ value should be ranged between 3 and 11 samples).

The compression and filtering of useful signal cause loss of samples of the final section (ranged from $j_{A0} = M\eta_0 - d_f + 1$ to $j_0 = 0$ (the newest sample), $L_A = -j_{A0} + 1$ samples in total). For prediction efficiency purposes, the essential is reliable reconstruction of the loss samples, especially the last-sample-estimation.

The estimates can be produced through multi-variant fitting of suitably smoothed signal profile to the last L_A data. In the paper [4] we specified three stages:

(a) The 4th order polynomial is assumed as an approximating function $f_A(t_m)$, It has zero-derivative at the end of the fitting interval and three fixed values over the compressed signal section ($m < j_{A0}$), equal to the compressed signal values $y_{Ffs}(t_m) : f_A(t_{(-LA-k)}) = y_{Ffs}(t_{(-LA\ -\ k)}), k = \{0,\ 1,\ d_2\}$. By the samples of $k = 0$ and $k = 1$ an approximation smoothness is ensured; d_2 is fixed by trials.

(b) The conditions imposed on the approximate bind parameters a_0, a_1a_2 and a_4 of the polynomial with a_3 (the fitting parameter; Eq. 13).The model f_A is fitted with a_3, using the generalized LS method (see Eq. 14). The values for $f_A()$ over the fitting interval t_R are calculated as shown in Eq. 15. Equations (14–15) can be written as the G_{fA} matrices of digital filters. After joining the corrected filters matrices g_{fs} the full matrix G_{fs} of modified FIR filters for the final segment can be obtained.

$$a_0 = y_{Ffs}\left(t_{-L_A}\right), a_1 = a_{10} + B_1 a_3,\ \ a_2 = a_{20} + B_2 a_3,\ \ a_4 = a_{40} + B_3 a_3 \tag{13}$$

$$\begin{bmatrix} a_{10} \\ a_{20} a_{40} \end{bmatrix} = A^{-1} \begin{bmatrix} y_{Ffs}\left(t_{-L_A - d_2}\right) - y_{Ffs}\left(t_{-L_A}\right) \\ y_{Ffs}\left(t_{-L_A - d_2}\right) - y_{Ffs}\left(t_{-L_A}\right) \\ 0 \end{bmatrix} \tag{13a}$$

$$\begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} = A^{-1} \begin{bmatrix} (-d_1)^3 \\ (-d_2)^3 \\ 3L_A^2 \end{bmatrix}, A = \begin{bmatrix} -d_1 & d_1^2 & d_1^4 \\ -d_2 & d_2^2 & d_1^4 \\ 0 & 2L_A & 4L_A^3 \end{bmatrix} \tag{13b}$$

$$a_3 = [Y_R - y_{Ffs}(t_{-L_A}) - a_{10} t_R - a_{20} t_R^2 - a_{40} t_R^4] \cdot \varphi_R^T [\varphi_R \varphi_R^T]^{-1} \tag{14}$$

$$Y_R = y\left(t_{-L_A + t_R}\right), t_R = [1, \ldots, L_A], \varphi_R = B_1 t_R + B_2 t_R^2 + B_3 t_R^4 + t_R^3 \tag{14a}$$

$$\begin{aligned} f_A\left(t_{-L_A + t_R}\right) = y_{Ffs}\left(t_{-L_A}\right) + (a_{10} + a_3 B_1)\, t_R + (a_{20} + a_3 B_2)\, t_R^2 \\ + a_3 t_R^3 + (a_{40} + a_3 B_3)\, t_R^4 \end{aligned} \tag{15}$$

(c) The value for d_2 is selected by trials, to obtain the $f_A(t_m)$ approximate for $m = -L_A, \ldots, 0$ of similar shape to the signal $y_{Ffs}(t_m)$ at the beginning section

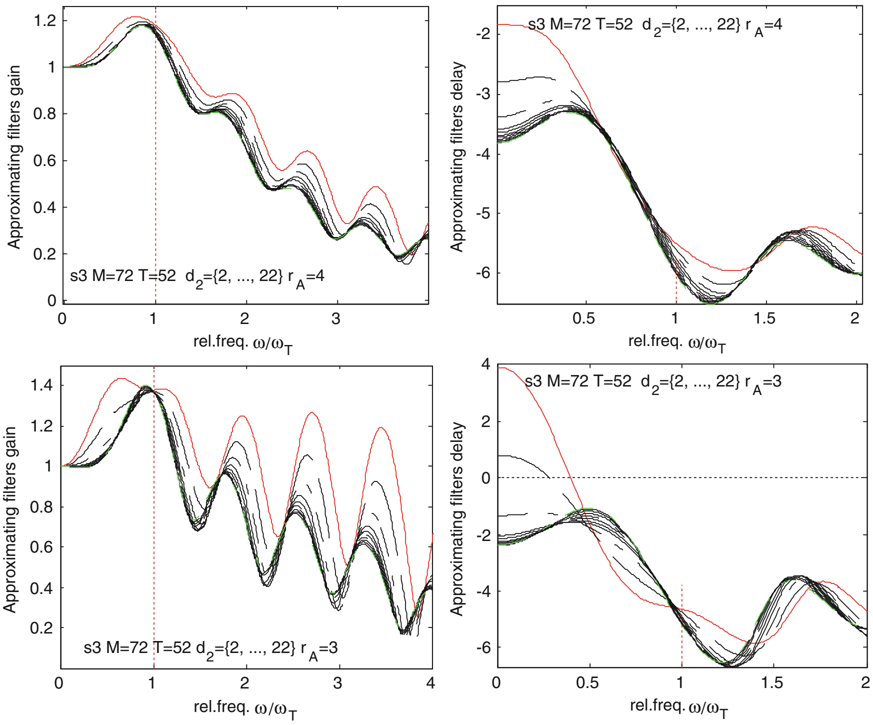


Fig. 10 Gain diagrams and delay of the final section approximate filters: 4th order ($r_A = 4$)—*upper figures* and 3rd order ($r_A = 3$)—*lower figures*, for d_2 increased from $d_2 = 2$ to $d_2 = 22$

of the final segment (for $m = -M, \ldots, j_{A0} - 1$). In this case, the most appropriate similarity measure is the difference module of the second increments mean squares of the series $f_A(t_i)$ and $y_{Ffs}(t_j)$.

As mentioned in Sect. 3, Eqs. (14–15) can be written as the f_A matrices of digital filters. After joining the corrected filters matrices g_{fs} the full matrix G_{fs} of modified FIR filters for the final segment can be obtained. Their properties are shown in Fig. 10.

The 4th order filters have much better frequency properties than 3rd order ones but they generate a low-frequency distortion and pass-band noises stronger than final section corrected filters. Thus, the signal overestimation is expected near to ω_T (c.a. 10–20 %) as slightly distorted signals (such effects are much weaker than for 3rd order filters). Delays of the 4th order filters are acceptable.

Efficiency of the described filtering method (4th order filter) in the MTF final section is illustrated in Figs. 11, 12, 13 and 14, for the periodic signal (nonparametric trend:$f(t) = \sin(2\pi/(2T_u)t) \cdot 0.85 \quad \sin(2\pi/(3T_u)t \cdot 0.25 + \sin(2\pi/(4T_u)t) \cdot 0.15$; periodic component:$C(t) = \sin(2\pi/T_u t) \quad \sin(4\pi/T_u t) + \sin(6\pi/T_u t))$.

The proposed method (especially combined with the s3 filter) can significantly improve the final signal value estimates, in the case of highly noised nonstation-

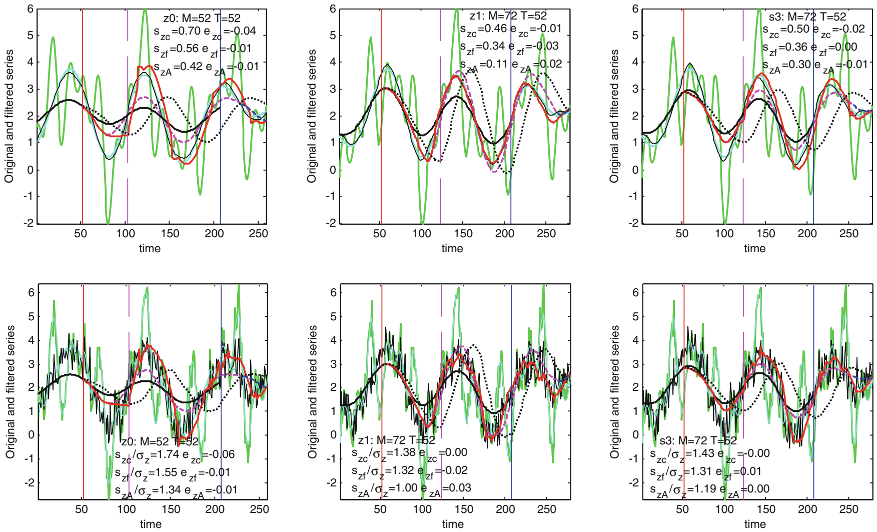


Fig. 11 Periodic time series filtering: noiseless (*upper figures*) and including random noise $\sigma_z = 0.5$ (*lower figures*). *Bold*, *solid line*—filtering output (delayed by $M - 1$ samples related to the current time/the most recent sample); *bold*, *dotted line*—output of final section filter (causal—concurrent with real time); *bold*, *dashed line*—output of the last corrected final filter (delayed by τ_0 related to the current time); *bold*, *grey line*—output of the last approximating filter (concurrent with real time); *grey line*—original/input signal; *dark line*—signal analyzed by the final filters (original one after the cyclic component extraction by regression analysis for the central section); *vertical*, *dotted lines*—the first and the last central segment end; *vertical*, *dashed line*: the first analyzed end-point of the series (current sample). s_{zc}, s_{zf} i s_{zA} denote standard deviations of filtering residuals, central filtering, corrected final one and approximated one (causal). Reconstructed signal (non-parametric trend): $f(t) = \sin(4\pi/(3T_u)t) + \sin(2\pi/(2T_u)t) \cdot 0.85 - \sin(2\pi/(3T_u)t \cdot 0.25 + \sin(2\pi/(4T_u)t) \cdot 0.15$; periodic component: $C(t) = \sin(2\pi/T_u t) - \sin(4\pi/T_u t) + \sin(6\pi/T_u t)$; random noise: $z(t_n) = \alpha z(t_n - 1) + 0.5(1 - \alpha^2)^{1/2}$, $\alpha = 0.15$ (Color figure online)

ary signal with additional harmonic of period 2*T*/3 and amplitude equal to 1 (see Figs. 11, 12, 13 and 14). This harmonic is passed weakly (referring to lower frequency components) and it significantly affects the filtering residuals in the central segment (see [4]). Thus identification of the parameters of the periodic component (A_k and τ_k) is difficult (useful signal distorted by a strong autoregressive processof standard deviation 0.5).

Figures 11, 12 and 13 illustrate final-section-filtering results in the subsequent time instants. The most important filtering results is the trend reconstruction at the current time with the approximating filter. Sensitive to disruption filtering produce better results of signal reconstruction than uncorrected final filter and corrected one (which estimates trend with a delay equal to τ_0). Figure 13 presents results of the filtering of the signal such as described in the Fig. 12 but for a longer period ($T = 261$). Figure 14 shows step-wise final-section filter responses in the moving window.

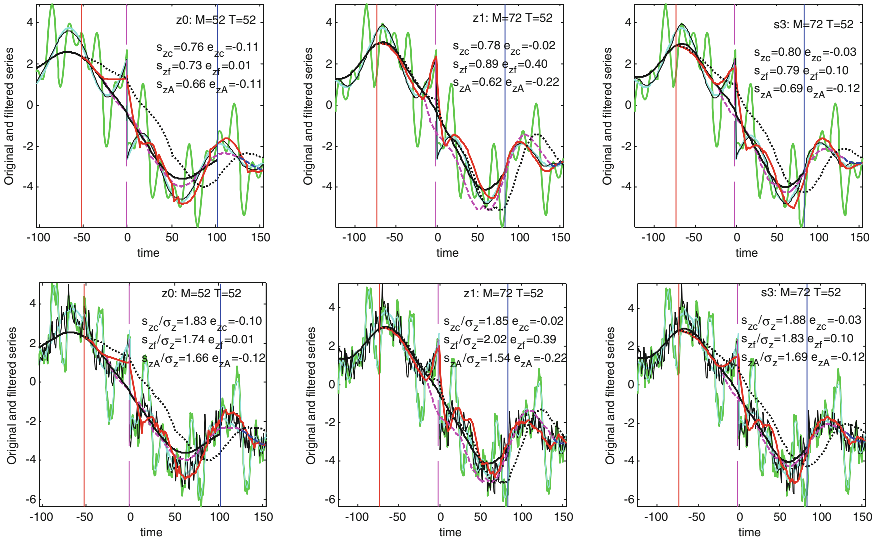


Fig. 12 Effects of periodic time series filtering (periodic signals disrupted by the step-change of amplitude equal to 4 at time point 0). Additional description—see Fig. 11

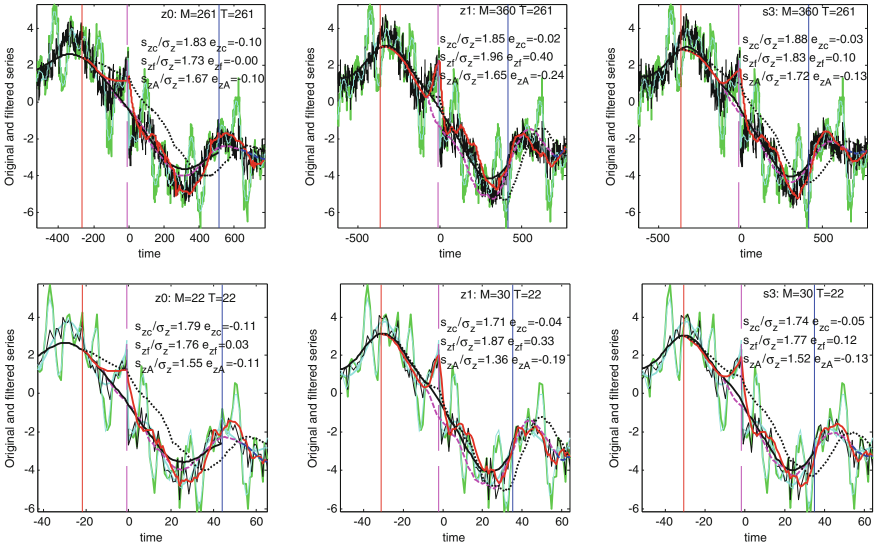


Fig. 13 Periodic time series as presented in Fig. 12 of a period $T = 261$ and $T = 22$ samples. Additional description—see Fig. 11

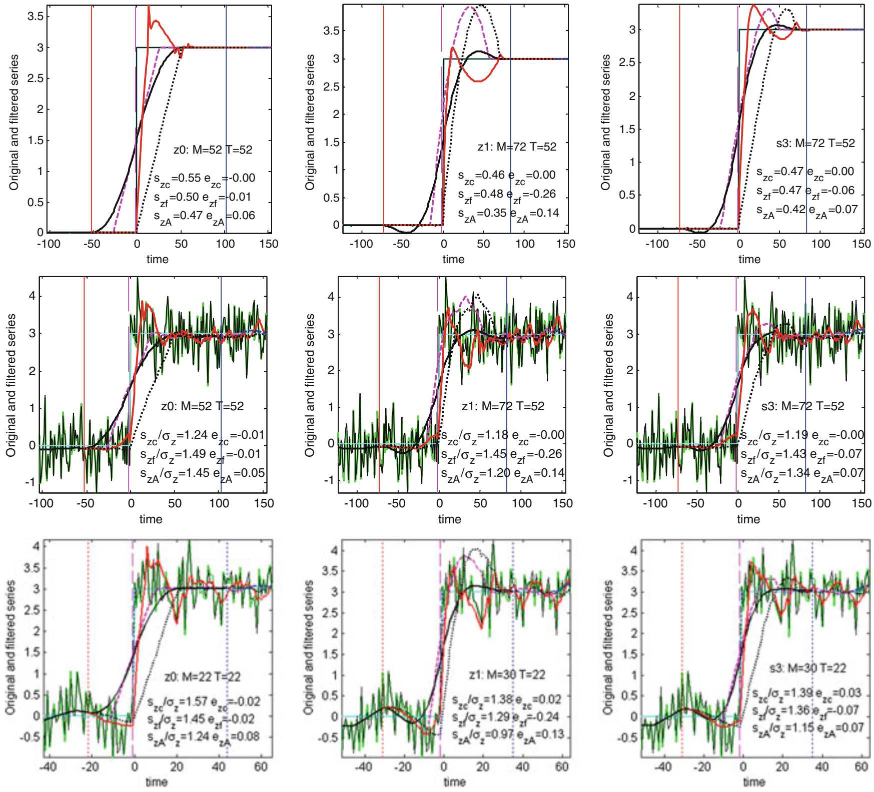


Fig. 14 Step-wise filter responses. Additional description—see Fig. 11

In all cases illustrated above, satisfactory accuracy of useful signal reconstruction in the final section is achieved, much better compared to reconstruction without correction. Especially, good results are produced by the s3 filter where approximates reach the shape closest to the reconstructed signal. It is in line with expectations based on time series analysis in frequency domain.

4 Conclusions

The classical moving trend smoothing algorithm (based on linear approximates) is of low efficiency, when applied to series prediction. Much better smoothing and prediction properties may be reached by employing the 3th order polynomial (s3) including only a constant and 3th order monomial (only b_0, b_3 are to be tuned by LS method). The approximation window width M may be easily adjusted by examination the Gain Plots of the moving trend based filters in frequency domain. The recommended filter s3 enables for very effective separation of the series onto low frequency ($\omega < \omega_T$) and high frequency ($\omega \geq \omega_T$) components, by taking the approximation window width $M = 1.38^*T$.

Smoothing (reconstruction of low frequency components) is the most effective (with no delay) in the central segment of the series. In the final section the low frequency signal distortion is significant, mainly due to varying delay of the consecutive final segment filters, which decreases prediction quality. The distortion produced by the recommended filter s3 is much weaker than that of the classical moving trend smoothing.

The periodic component $C(t)$ may be extracted from the filtering residuals by a regression method applied to residuals in the central section of the processed series.

The adaptive filtering approach described in the paper is based on adjusting the filter parameters referring to historical data (a specified number of last samples). Such property can be utilized for event/change detection and similarity analysis of the sample sequences, in particular patterns.

References

1. Askom, M.V., Chenouri, S., Mahmoodabadi, A.K.: ARCH and GARCH models. Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo (2001)
2. Box, G.E.P., Jenkins, G.M., Reinsnel, G.C.: Time Series Analysis, Forecasting and Control, 3rd edn. Prentice Hall, Englewood Cliffs (1994)
3. Cieślak, M. (Edt.): Economic Forecasting. Principles and practice. Polish Scientific Publishers PWN, Warsaw (2002)
4. Duda, J.T., Pelech-Pilichowski, T.: Enhancements of moving trend based filters aimed at time series prediction. Advances in systems science. Springer International Publishing, New York (2014)
5. Duda, J.T., Pelech-Pilichowski, T.: Moving trend based filters design in frequency domain. Federated Conference on Computer Science and Information Systems, 8–11 September, 2013, Kraków, Poland
6. Duda, J.T., Pelech-Pilichowski, T.: Opracowywanie prognoz sytuacji hydrogeologicznej i ostrzeżeń przed niebezpiecznymi zjawiskami zachodzącymi w strefach zasilania lub poboru wód podziemnych. Research Report, AGH University of Science and Technology, Faculty of Management, Kraków (2012)
7. Duda. J.T., Pelech-Pilichowski, T., Augustynek A.: Application of moving trend to analysis and prediction of financial time series. In: Howaniec, H., Waszkielewicz, W. (eds.) Determinants, metods and strategies of development of enterprises, ATHBielsko-Biała Publ., Poland (2013)
8. Hamilton, J.: Time Series Analysis. Princeton University Press, New Jersey (1994)
9. Oppenheim, A.V., Schafer R.W.: Digital Signal Processing. Prentice Hall (1975)
10. Otnes, R.K., Enochson, L.: Digital Time Series Analysis. Wiley, New York (1972)

The Assessment of the EPQ Parameter for Detecting H-Index Manipulation and the Analysis of Scientific Publications

Rafał Rumin and Piotr Potiopa

Abstract The work presents the analysis of mechanisms for determining the susceptibility of parametric indices (such as the h-index) of evaluation of scientific articles published on the modification of parameters not resulting from essential value of the research work. Currently, most methods for verifying the article is focused on the selection of works potentially strongly influence the international position of a journal. To this end, editorial offices wide use of parametric methods of assessment. In addition, the work attempts to identify the used criterion functions, namely the assessment parameters and guidance, the risks associated with using this type of method to change the popular parametric indexes for authors and journals. These parameters are divided into categories and offered their initial verification based on statistical analysis of already published articles in various journals. Each parameter has attributed weight function, which allows to define its impact on the total evaluation of an article, and also adaptation of formula to any academic journal. Weight functions will be determined with the usage of neural networks or genetic algorithms, aiming to their individual adaptation to particular journal.

1 Introduction

Relentless pursuit of scientific journals to obtain the greatest possible number of points in the created rankings enhances continuous improvement of parametric algorithms to verify the quality of the article and the assessment of its author (Philadelphia List, Impact Factor, quoting indicators etc.) [1–3].

All the time created a new methods of evaluation of journals and modifications of existing criteria result in a situation that merits evaluation of the article can be replaced by a parametric assessment forced by the publisher [4–12].

R. Rumin · P. Potiopa (✉)
AGH University of Science and Technology, Al. Mickiewicza 30, 30-059 Kraków, Poland
e-mail: ppotiopa@zarz.agh.edu.pl

R. Rumin
e-mail: rumin@agh.edu.pl

M. Mach-Król et al. (eds.), *Advances in ICT for Business, Industry and Public Sector*,
Studies in Computational Intelligence 579, DOI 10.1007/978-3-319-11328-9_8

The result is a situation that good, exploratory research publications may be assessed or unfairly withdrawn from the publications on the ground that have been poorly prepared for parametric criteria. Thus, the following aspects parameterization are to determine the influence of subjective factors in the evaluation of scientific articles in specific journals. Furthermore, there were presented series of factors which, if they are taken into consideration during writing of scientific articles, they have a chance to increase probability of obtaining positive review and in effect the acceptance of publication in renowned journals. In the further process of research works, realization of automatic information system is planned, which role will be connected with the verification of the working version of an article, before sending it to the journal and the definition of the probability of obtaining high parametric evaluation. Described parametric evaluation will determine the coefficient EPQ—Estimated Paper Quality. This co-efficiency will be helpful for scientists who concentrate mainly on essential, and less over the editorial part of their scientific article. The low value of EPQ should induce the author to analyze and supplement his publication before sending article to editorial office of the previously chosen journals.

2 Mathematical Models for Authors Evaluation

Authors of scientific articles are subject to verification by placing in the ranking reflects their contribution to the development of the field of scientific work. One of main parameters applied in relation to authors of publication is proposed in year 2005 the Hirsch index (h-index) [13]. As easily can be envisaged, such evaluation can be sensitive on manipulation on the side of several cooperating with each other authors, who mutually will quote their works (apart from their essential contribution into researches). The parametric evaluation of publication issues from the category of scientific research. Scientific researches require financing, and one of the popular sources of learning financing are exploratory grants. To obtain financing it is expected that the scientist will carry out planned investigations and their effect will have visible influence on a given exploratory field. How such influence is measured?

Most legible measures are publications and their quotations. For this reason, scientists who have a suitably high Hirsch index, are treated as trustworthy to commit to them public money on carried researches. Legible dependencies appear between the financing of research, with the quantity of publication, and with their quotations which put each other greater chance for future financial resources.

2.1 Mathematical Models for Journals Evaluation

The high quality journal tries to be visible for society of scientists. To found a difference between quality of journals, the special parametric factors was proposed.

Below are some of them [14–20]: Source-Normalized Impact per Paper (SNIP), Relative Citation Rates (RCR), SCImago Journal Rank (SJR), Journal to Field Impact Score (JFIS), Article Influence (AI), and the most popular—Impact Factor (IF) (1).

"IF" counts all citations from particular calendar year, and it divides them by the amount of "cited" publications from last two years (C).

$$IF = \frac{B}{C} \tag{1}$$

Other indicators also reflect the parametric quality rating of journal, but they are not so popular. Each of them characterizes different factors which influence final evaluation of journals. Different journal evaluation criteria cause the inhomogeneity in resultant rankings. Furthermore, algorithms of evaluation are subjected to continuous changes aiming to the most reliable definition of publications quality. For this reason, the aim of publishing companies, instead of valuing scientific publications, having less 'popular' character (though substantially equally good if not much better), could be the wish of achievement of the highest parametric coefficients evaluating the other of their publications.

It can be accepted that, as the evaluation of the given journal is higher in the ranking, an article published in it has the chance to obtain greater range, and consequently receive greater quantity of quotations. It seems that there exists the conformity of business among a journal and an author of the article, however this concerns only the wishes of obtaining the maximum quotations quantity at other publishers through the large number of scientists.

Willing to check our chances for the publication in the given journal, we often set incorrect question—*Will this journal publish my article?*

To show existing dependences and conflicts of interests between an author and an editor, one ought to set himself the question:

How will my article help the journal to obtain better position in the ranking (more points in the parametric evaluation of journals)?

The answer depends on many factors, which can subjectively influence the evaluation of an article, apart from its essential value. Figure 1 shows a general scheme of the relations between a publishing house and an author.

2.2 Mathematical Models for Article Evaluation

Each article has its own essential value which cannot be measured automatically. The article evaluation is limited to a group of parameters defining its quality from the interest of a journal point of view. Unfortunately this can cause a conflict of interests between publishing houses and authors [21].

During the evaluation of an article, the essential value can be estimated by additional parameters: the range of carried out researches, description of theoretical

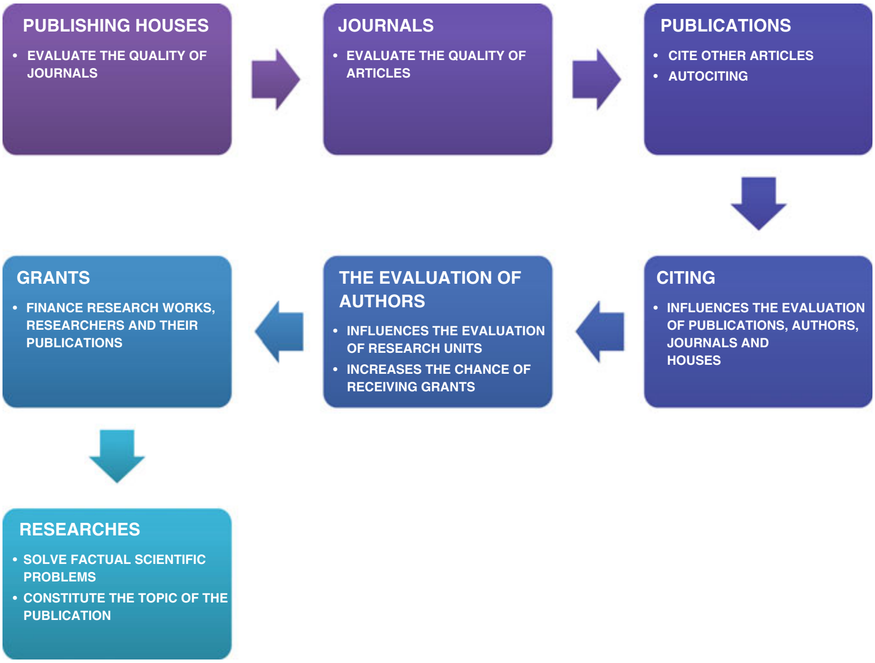

PUBLISHING HOUSES
• EVALUATE THE QUALITY OF JOURNALS
JOURNALS
• EVALUATE THE QUALITY OF ARTICLES
PUBLICATIONS
• CITE OTHER ARTICLES
• AUTOCITING
CITING
• INFLUENCES THE EVALUATION OF PUBLICATIONS, AUTHORS, JOURNALS AND HOUSES
THE EVALUATION OF AUTHORS
• INFLUENCES THE EVALUATION OF RESEARCH UNITS
• INCREASES THE CHANCE OF RECEIVING GRANTS
GRANTS
• FINANCE RESEARCH WORKS, RESEARCHERS AND THEIR PUBLICATIONS
RESEARCHES
• SOLVE FACTUAL SCIENTIFIC PROBLEMS
• CONSTITUTE THE TOPIC OF THE PUBLICATION

Fig. 1 The scheme of the relations between a publishing house and an author

models, simulation models, experiment. If it has only theory it can be classified lower than articles containing simulation or experiment.

Articles containing the experimental verification of carried researches will be evaluated as the best ones. Separately, articles containing rich and complex reviews of the literature from the given field can display significantly high classification, because this type of articles are quoted often many times. This results from the specific approach of scientists to carried out researches and wishes of using elaborated earlier literature review, which often requires a lot of time and belongs to “less attractive” researches.

Thereby, at the evaluation of articles’ value, nobody can foresee how often he will be quoted in the future. To put it simply, it can be assumed, that at the initial phase of article analysis, each has the evaluation for the essential value on the same level. Since the quantity of elaborated article future quotations cannot be influenced, it can be influenced who the author quotes in his own publication. This way the quantity of “gained” quotations from the journal’s point of view, can be controlled. The issue here is the period of time in which journals are subjected to evaluation in rankings. For the calculation of Impact Factor, last 2 years are taken into consideration which means that the auto quotation of other articles which appeared in the same publishing house within a period of last 2 years have a positive influence on IF indicator increase. Therefore, the publishing house will be willingly promoting articles which already

Table 1 Defining parameters for the calculation of the EPQ indicator—basic parameters

P_i	Meaning of value substituted to P_i	Range	Formula on Pi
P_1	H—authors' Hirsch index	H=[0:inf]	$P_1 = (1 - \frac{1}{1-H}) * w_1$
P_2	I—the quantity of authors' indexed publications	I=[0:inf]	$P_2 = (1 - \frac{1}{1-I}) * w_2$
P_3	C—quantity of authors' indexed quotations	C=[0:inf]	$P_3 = (1 - \frac{1}{1-C}) * w_3$
P_4	S—degree/ the scientific title of the author (none/engineer/MSc/the doctor/assistant professor/professor)	S=[0:5]	$P_4 = (1 - \frac{1}{1+20*S}) * w_4$

show quotations from their own journal, is a method to obtain higher place in the ranking. However, if there exists a group of journals given by the common institution, then cross quotations of other journals belonging to the same publisher constitute also an added value. Here arises a threat regarding the reliability of one published articles, because one can apply the mechanism which would permit ranking speculations between journals. Following the paragraphs of this article, they contain the case study describing such situations.

3 The New Indicator for the Parametric Evaluation of an Article—EPQ

All articles can be parameterized by the Estimated Paper Quality coefficient (EPQ). This model can indicate many factors which participate in the evaluation of given article. It can be presented as weighted mean of individual parameters, with suitably assorted weights functions. The value of parameters is standardized so that it contains itself in the range from 0 to 1. This type of method descends from Churchman and Ackoff (1954) researches, under the name Simple Additive Weighting (SAW) [22, 23]. SAW is one of most popular solutions in Multi-Attribute Decision Making type (MADM) problems of which undoubtedly is the problem described in the work. Elaborated process of EPQ calculation is similar to the above methods, however differences in designating of individual parameters appear. Differences are caused by different ways of P_i parameters of values determination.

$$EPQ = \frac{1}{n} \sum_{i=1}^{n} P_i * w_i \tag{2}$$

where P_i is appropriate parameter of evaluation with following index n appointed, and w_i is the weight for a given parameter. Below, in the table (cf. Tables 1, 2 and 3) the list of parameters together with their asserted values and ranges is presented. All parameters P_i are situated in the same range: $P_i \in [0, 1]$.

Table 2 Defining parameters for calculation of the EPQ Indicator—content rating of an article

P_i	Meaning of value substituted to P_i	Range	Formula on Pi

P_5	The calculated Gaussian distribution basing on the quantity of all quotations contained by an author in the article, where:
	d—the height of the Gaussian curve top,
	x—quantity of all quotations contained by an author in the article,
	σ—standard deviation of Gaussian distribution,
	μ—expected value, equal average quantity of quotations devolving on one article in the given journal,
	a—quotations devolving on one article (k) in the given journal
		d = [0:1]
		x = [0:inf]
		σ = [0:inf]
		μ = [0:inf]
		a = [0:inf]
		k = [0:inf]
			$P_5 = \left(d * e^{\frac{-(x-\mu)^2}{2\sigma^2}}\right) * w_5$
			$\sigma = \sqrt{\frac{1}{k-1}\sum_{k}^{i=1}(x_i - \mu)^2}$
			$\mu = \frac{1}{k}\sum_{i=1}^{k} a_i$

P_6	A—the quantity of quotations coming from archival numbers of the same journal to which the publication is submitted	A=[0:inf]	$P_6 = \left(1 - \frac{1}{1-A}\right) * w_6$
P_7	B—the quantity of quotations coming from archival numbers of remaining journals belonging to the same publishing house to which publication is submitted	B=[0:inf]	$P_7 = \left(1 - \frac{1}{1-B}\right) * w_7$

P_8	The indicator of the publication originality.
	O—the quantity of similar articles earlier published by the author.
	D—the sum of "duplicates", measured by the coefficient of similarity of genuine text and small pictures between previous articles of the author, and with his current publication
		O=[0:inf]
		D=[0:inf]
			$P_8 = \left(1 - \frac{1}{1+O}\right) * w_8$
			$O = \sum_{i=1}^{n} D_i$

P_9	R$\mathrm{R_d}$—the quantity of cited publications of the current editor of journal to which publication is submitted	$\mathrm{R_d}$=[0:inf]	$P_9 = \left(1 - \frac{1}{1+\mathrm{R}_d}\right) * w_9$
P_{10}	$\mathrm{R_c}$—the quantity of cited publications of current reviewer of journal to which publication is submitted	$\mathrm{R_c}$=[0:inf]	$P_{10} = \left(1 - \frac{1}{1+\mathrm{R}_c}\right) * w_{10}$

3.1 The Methodology of Calculation the EPQ

Calculation of the EPQ coefficient is based on a lot of other indicators described bellow. Particularly essential from the usage of EPQ indicator point of view, is the

Table 3 Defining parameters for calculation of the EPQ indicator—other parameters

P_i	Meaning of value substituted to P_i	Range	Formula on Pi
P_{11}	J—the quantity of authors' publications quoted by a current editor or reviewer of the journal to which publication is submitted	[0:inf]	$P_{11} = (1 - \frac{1}{1+J}) * w_{11}$
P_{12}	K—the quantity of authors' common publication articles and a current editor or reviewer of a journal to which publication is submitted	[0:inf]	$P_{12} = (1 - \frac{1}{1+K}) * w_{12}$
P_{13}	Z—the quantity of elements from the range of carried out researches (the form of survey): review, theory, model, simulation, experiment, lack/other	[0:5]	$P_{13} = (1 - \frac{1}{1+20+Z}) * w_{13}$

possibility of weights definition w_i in way compatible to parametric evaluations applied by the given journal. The large number of academic journals cause different approach to the parametric evaluation of accepted articles to editorial office and the review of article. Basing on the data from previous years, considering all publications printed within the framework of one publishing-title, we are able to determine weights of individual parameters individually for the given journal.

For that purpose we will use neural networks with the feedback which will learn to recognize the influence of the given parameter on the positive acceptance of article to the publication. In case of the analysis, already printed publications, we will subordinate the quantity of published articles from the value of individual parameters. The more articles will have e.g. the high parameter P6, the greater influence on the printing of publication has the quantity of archival articles quotations laded from the same journal.

Initial values of the weight parameter w_1 amount to 1. Due to the implication of matching algorithm, the weights shall be modified in the 0 to 1 bracket through artificial neural networks. The aim of the modification is the selection of appropriate levels of weights to a given journal.

3.2 *The Example of the EPQ Calculation*

The definition of the exact value EPQ does not decide about "the success" and the publication of the given magazine article. This will permit however finishing up and improving of the editorial part which could not take into the above-mentioned factors influencing decision of editors and reviewers. System elaborated in such a way, using the IT network will permit quick definition of the article modification. Outwardly, basing on obtained result EPQ it will enable to propose the alternative academic

Table 4 The results from this case are as follows

No	Parameter P_i	Initial data	P_i results
1	P_1	$H = 5$	0,833
2	P_2	$I = 100$	0,990
3	P_3	$C = 500$	0,998
4	P_4	$S = 5$	0,990
5	P_5	$d = 1; x2 = 90; a1 = 80; a2 = 100; a3 = 120$	0,882
6	P_6	$A = 2$	0,667
7	P_7	$B = 2$	0,667
8	P_8	$D = 0; O = 0$	1
9	P_9	$Rd = 0$	0
10	P_{10}	$Rc = 2$	0,667
11	P_{11}	$J = 0$	0
12	P_{12}	$K = 0$	0
13	P_{13}	$Z = 3$	0,984
Average			0,667
EPQ			0,67

journal which parameters answer to the result. The value EPQ was calculated basing on the example of the publication based on the *Matlab* software.

Below, the calculated value of EPQ was presented for a model publication.

(a) Data concerning the author: The current Hirsch index of the author amounted to 5, for 100 indexed publications and 500 of all his connective cited publications. The author obtains the academic title of professor.

(b) Data concerning a publication: The article contains 90 of citations, from which 2 citations come from archival journal, to which the publication is being composed. In total, there are 2 citations from other archival issues of journals belonging to the same publishing house, to which the publication is being composed. The publication demonstrates original quality, since there have not been any of its duplicate samples and publications of similar content of the same author. In the publication there are no citations from the works of the members of editorial board, however there have been 2. citations of works developed by reviewers. The publication contains at least 3. elements of scientific publication (e.g. review, theory, model).

(c) Data concerning the journal: The statistical average number of citations in a single article published in this journal amounts to 100. Editors and reviewers have not cited any other works of the author, they have not had any mutual articles with the author.

The range of the selected parameters together with separate results of the calculations are depicted above (Table 4).

4 SEO, Hirsch Index and Impact Factor

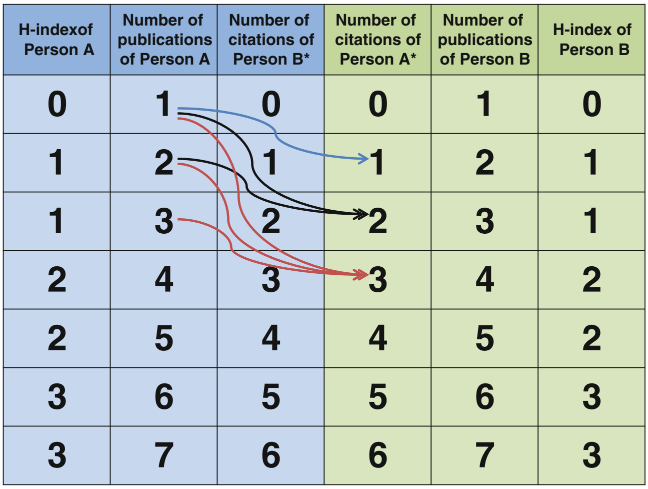
4.1 The Similarity of the Hirsch Index and Impact Factor to Page Rank, and Threats Resulting from Black Hat SEO Methods

The growth of the Hirsch index and IF strictly depends on the quantity of the given author's publication quotations. This model can be compared to the published ranking of websites (PR—Page Rank) used in Google search engine [24–26]. The similarity refers to the quantity of quotations which correspond to quantities of returnable links indicating given page of data sources.

There are known general methods of influencing the algorithm of search engine in this way, so that the indicated page will be higher in the Search Engine Results Page ranking (SERP). These methods are divided on the so-called white and black. White Hat SEO—means the positioning of the website in compliance with official guidelines of search engines, which should result in better page adaptation to Web-crawler's and engines of search engines requirements. Good preparation of the website facilitates, quick indexing of it in the search engine base of data, however increasing number of valuable references to page (gained naturally and resulting from its popularity and uniqueness) permits its positioning and obtaining of high place in the SERP ranking. As valuable references are acknowledged, links from pages about high PR are often visited by users (e.g. thematic, community websites). There also exists Black Hat SEO which is characterized by the use of all possible gaps in the search engine, for the purpose of raising the ranking of given website. Such effects are achieved through the manipulation with the quantity of returnable links and their "artificial" addition through generating large quantity of pages with links. So many of manipulation methods constitute the necessity of continuous algorithms change of search and qualitative selection of websites.

From obvious reasons, exact parameters of the algorithm are not revealed for the purpose of their protection before the manipulation. There can be only estimated general dependencies and on their base there can be created algorithms improving the position of website in ranking of searches. Methods of rankings creating e.g. PR and IF, and also H-index cause the risk of appearing methods taken from SEO, which in the artificial way will manipulate results of the above-mentioned rankings. Probably there is no possibility of obtaining 100 % reliable and objective ranking not burdened with the above risk.

From this reason, the essential evaluation of publication can be shaken, in the interest of the parametric evaluation. This can cause the reverse to intended effect i.e. these rankings will promote less ambitious scientific discoveries, but artificially will overvalue indexes across the elaboration of their manipulation method. The case study is presented below, which in the mental experiment, could result with "artificial" increasing of IF for the journal or with "artificial" increasing of the H-index for given scientist.



H-indexof Person A	Number of publications of Person A	Number of citations of Person B*	Number of citations of Person A*	Number of publications of Person B	H-index of Person B
0	1	0	0	1	0
1	2	1	1	2	1
1	3	2	2	3	1
2	4	3	3	4	2
2	5	4	4	5	2
3	6	5	5	6	3
3	7	6	6	7	3

Fig. 2 Mutual citations of 2 authors

4.2 Threats Resulting from the Usage of Artificial Methods of Increasing the indexes

Among academic researchers, there is no unambiguous method of scientific achievements' evaluation, which would credibly and objectively determine the value of a research work. There are numerous publications that describe threats connected with manipulation of indexes [27, 28]. A short analysis of the case that highlights the susceptibility of the used indexes for artificial manipulation is shown below. It is a similar situation that the search engine Google encountered. It is subjected to continuous attempts of manipulation of websites' rankings that are displayed in the first 10 results of a search. The search engine algorithm was evolving, taking into account increasingly different parameters so as to extract artificial positioning. If the algorithms of calculating the indexes are not modified, the intentional manipulation to increase indexes is very likely to occur, which is shown below.

4.3 Hirsh Index Manipulation

The Hirsch index has lots of advantages, however it is also subjected to the risk of being manipulated. Suppose there are at least 2 researchers working in similar field of study, they may cite each other's work (Fig. 2), only to increase each other's Hirsch index.

Of course the publication that contains several or a dozen of citations of the same author may focus the attention of the reviewer as far as the legitimacy of citing is

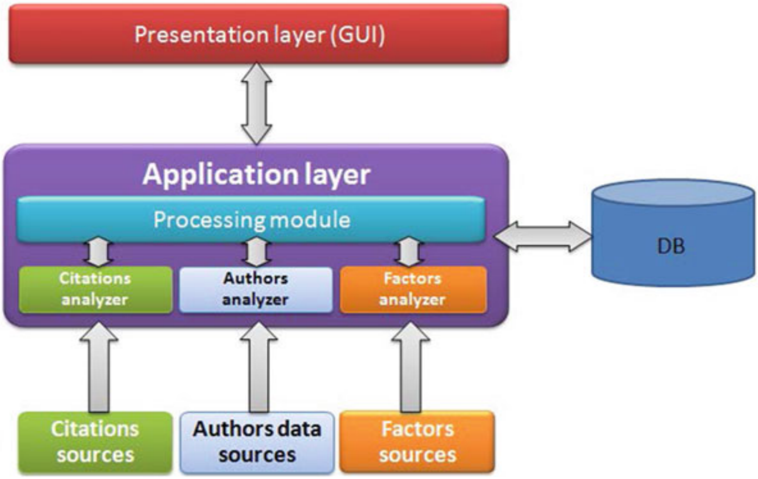


Fig. 3 The architecture of a proposed IT system to determine EPQ indicator

concerned. In such a case there exists the possibility of developing the model for a specified group of authors, where division of citations between each other will be evenly determined so as not to undermine the legitimacy. Furthermore, it is worth to add that the publication drawn up by an author A that contains more or less 5 citations of an author B is fully sufficient to increase the Hirsch index from 0 to 3 level. In the cooperation of at least 6 authors, the increase of H index to the level 5 for each of the author will demand only the cooperation in the scope of citing of 5 publications (6 counting from the first one, which has not been cited before by nobody).

*—for each next publication one person has to cite other person's all publications (the number of citations per new publication).

5 Application Realizing EPQ Designation

Determination of the individual parameters can be achieved through the use of available databases of scientific publications and names of authors like: SCOPUS, WEB OF KNOWLEDGE, Google Scholar and other smaller databases. The system collects this data, it will turn collecting the following information: author, citations, journals, publications, and then assessed the parametric analyzed publication. Based on this evaluation, it can propose suggestions, ref. introductions of changes in the article or present proposal of the alternative journal to which the parametric evaluation was better. The system architecture may be built based on the client-server methodology which is presented on Fig. 3.

5.1 The Architectural Schema

On the after-mentioned Fig. 3 the general architectural schema of the system is presented.

In presented architecture system we distinguish:

1. *Presentation layer*—a layer of the application responsible for the presentation of results and communication with user, receiving data from user (proposed article, survey for the author)
2. *Application layer*—layer responsible for the resumption of data and processing of results which consists of:
   - *citations' analyzer* (module processing the quotation categorizing and counting quotations of authors works.)
   - *authors' data analyzer* (module processing data of authors (also reviewers and editors), checking relations of author with journals across quotations as well as categorizing his achievement)
   - *authors' analyzer* (module being supposed for the task to process available data sources information of used in the algorithm coefficients for journals and authors)
3. *DB*—a layer of database recording source data and results of calculations with application layer, permitting caching of data sources in the situation when data don't need to be refreshed at every operation of weight-coefficients calculation weight-coefficients.
4. *Sources*—a layer of gaining data from chosen sources dividing into sources of quotations gaining ("citations sources"), given authors ("authors the date sources") and coefficients used in the algorithm of EPQ count ("factors sources").

The system architecture in case of further development can be calibrated because the module of processing may receive partial results of calculations (weights of component parameters) from individual modules which can be found on separate instances of servers. Each module of gaining data can have the separate database in which it will store the received results of the data sources indexing. In case of presentation layer, the system can communicate with software of the thin client type in case of approach users (authors of articles) and with the software of the fat client type in case of the administrator who can control work of the processing module (settings control).

6 Conclusion

Determining the actual accuracy of parametric evaluation for scientific journal article by calculating the proposed EPQ parameter is possible only after verification on figures. Methodology is based on foundations that a substantially good article can be evaluated wrongly due to remaining factors on which reviewers and editors of journals pay attention to. The elaborated system, proper verifying of targets and correction of an article before its delivery to publishing houses will permit to carry out essential research on equally high level and to regard subjective 'expectations' from the side of publishing house in relation to an author. So an improved article has greater chance for printing in a renowned journal, which can positively rebound on future publications of many authors.

On the other hand, it will be possible to verify retrogradely articles that have too high parametric evaluation to indicate potential authors or journals, which were subjected to artificial mechanisms that are used to increase index parameters. Through this process it will be possible definition of what elements have been streamlined parametric evaluation of the use of illegal solutions, which include the use of observed SEO methods.

References

1. Egghe, L., Rousseau, R.: Introduction to Informetrics. Elsevier, Amsterdam (1990)
2. Moed, H.F., Vriens, M.: Possible inaccuracies occurring in citation analysis. J. Inf. Sci. **15**, 95–107 (1989)
3. Todorov, R.: Journal citation measures: A concise review. J. Inf. Sci. **I4**, 47–65 (1988)
4. Zhou, D., Orshanskiy, S.A., Zha, H., Giles, C.L.: Co-Ranking Authors and Documents in a Heterogeneous Network, In: International Conference on Data Mining (2007)
5. Chen, P., Xie, H., Maslov, S., Redner, S.: Finding scientific gems with Google. J. informetr. **1**, 8 (2007)
6. Garfield, E.: Citation analysis as a tool in journal evaluation. Science **178**(60), 471–479 (1972)
7. Liu, X., Bollen, J., Nelson, M.L., Van de Sompel, H.: Coauthorship networks in the digital library research community (2005). arXiv:cs/0502056
8. Bianchini, M., Gori, M., Scarselli, F.: Inside pagerank. ACM Trans. Internet Tech. **5**(1), 92–128 (2005)
9. De Moya, F.: The SJR indicator: A new indicator of journals' scientific prestige (2009). arXiv:0912.4141
10. Garfield, E.: The history and meaning of the journal impact factor. JAMA **295**, 90–3 (2006)
11. Amin, M., Mabe, M.: Impact factors: Use and Abuse. Perspectives in Publishing, vol. 1, pp. 1–6 (2000)
12. Huang, M.-H., Cathy Lin, W.-Y.: The influence of journal self-citations on journal impact factor and immediacy index. Online Information Review, **36**5, 639–654 (2012)
13. Hirsch, J.E.: An index to quantify an individual's scientific research output. In: Proceedings of the National Academy Science (PNAS), **102**(46) (2005)
14. SNIP and SJR at Journal Metrics. www.journalmetrics.com
15. SCImago Journal Rank (SJR). http://www.scimagojr.com/
16. Source - Normalized Impact per Paper (SNIP). www.journalindicators.com
17. Impact Factor (IF). http://thomsonreuters.com/products_services/science/free/essays/impact_factor
18. h-index. http://help.scopus.com/robo/projects/schelp/h_hirschgraph.htm
19. Article Influence (AI). www.eigenfactor.org
20. Relative Citation Rates (RCR)/Journal to Field Impact Score (JFIS)
21. Wenneras, C., Wold, A.: Nepotism and sexism in peer-review. Nature **387**(6631), 341–343 (1997)

22. Churchman, C.W., Ackoff, R.L.: An approximate measure of value. J. Operat. Res. Soc. Am. **2**(2), 172–187 (1954)
23. Widayanti, D., Oka, S., Arya, S.: Analysis and implementation fuzzy multi-attribute decision making saw method for selection of high achieving students in faculty level. IJCSI International Journal Computer Science Issues **10**(1), 2 (2013). ISSN 1694–0784
24. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: WWW7 Proceedings of the Seventh International Conference on World Wide Web 7, pp. 107–117. Elsevier Science Publishers B.V. (1998)
25. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Technical Report, Stanford Digital Library Technologies Project (1998)
26. Maslov, S., Redner, S.: Promise and pitfalls of extending Google's PageRank algorithm to citation networks. J. Neurosci. **28**, 11103 (2008)
27. Christoph, B., Kokkelmans, S.: Detecting h-index manipulation through self-citation analysis. Scientometrics **87**(1), 85–98 (2011)
28. Bihui, J., et al.: The R-and AR-indices: Complementing the h-index. Chin. Sci. Bulletin **52**(6), 855–863 (2007)

Fuzzy Multi-attribute Evaluation of Investments

Bogdan Rębiasz, Bartłomiej Gaweł and Iwona Skalna

Abstract Most companies have a large number of projects that they would like to accomplish for various reasons. However, financial and material limitations cause that only some of the investments can be undertaken, which rises the problem of selecting the portfolio of the most effective investment projects. Selecting a portfolio from available project proposals is crucial for the success of each company. This paper proposes a practical framework for modelling projects portfolio selection problem with fuzzy parameters resulting from uncertainty associated with decision makers' judgment. A fuzzy multi-attribute decision-making approach is adopted. A two-step evaluation model that combines fuzzy AHP and fuzzy TOPSIS methods is used to rank potential projects. The proposed approach is illustrated by an empirical study of a real case from steel industry involving fifteen criteria and ten projects. The case study shows the effectiveness and feasibility of the proposed evaluation procedure.

1 Introduction

Decisions on investment projects have a direct impact on a company's success. The quality of these decisions has long-term influence upon the development and company's market position. However, financial and material limitations cause that only some of the investments can be undertaken, which rises the problem of selecting the portfolio of the most effective investment projects. This problem is particularly difficult to solve, because of the ubiquitous uncertainty associated with any business activity. Moreover, it can be observed that all branches of economy and industry are affected by the increase in the variability of products prices, growing pressure to reduce production costs, increase in competitions etc. This causes that the

B. Rębiasz · B. Gaweł (✉) · I. Skalna
AGH University of Science and Technology, Kraków, Poland
e-mail: bgawel@zarz.agh.edu.pl

B. Rębiasz
e-mail: brebiasz@zarz.agh.edu.pl

I. Skalna
e-mail: skalna@agh.edu.pl

M. Mach-Król et al. (eds.), *Advances in ICT for Business, Industry and Public Sector*,
Studies in Computational Intelligence 579, DOI 10.1007/978-3-319-11328-9_9

project portfolios selection (PPS) becomes more and more complex decision task, which motivates managers to utilise modern techniques and tools to optimise capital allocation. Methods for effective project portfolio selection gained popularity in early 90s and since then they are of great interest for business activity practitioners. They are constantly developed and improved by researchers in the field of project managements.

Nowadays, the company's investment decision process should take into account not only financial but also environmental, market, technological and human resources aspects. For this reason, the decision process can be viewed as a multi-attribute evaluation of a project portfolio selection. Obviously, that the more parameters of an investment project that are subjected to the analysis, the more reliable the selection process is. Unfortunately, wide spectrum of criteria makes the analysis more costly. Moreover, the excessive number of criteria may hamper the interpretation of the results. At present, there are a lot of methods that can be applied to solve the PPS problem, including Economic Analysis, Decision Theory, Optimisation and Multi-criteria methodologies. In order to deal with both financial and non-financial project attributes, the multi-criteria decision making (MCDM) analysis is a preferred approach.

The goal of the multi-criteria decision making (MCDM) analysis is to "provide a set of attributes aggregation methodologies that enable the development of models considering the decision makers' (DMs') preferential system and judgement policy" [1]. Intuition and simple rules do not suffice to reach this goal. Every company should develop its own MCDM based framework for solving complex decision problems. In order to select an adequate framework, various factors should be considered including: the nature of the decision problems, types of choices that have to be made, measurement scales, dependency among attributes, type of uncertainty, expectation of decision makers, quality and quantity of available data and judgements [2]. All of this causes that nearly every company and every industry has its own unique MCDM framework. In general, MCDM methods may be divided into two groups: multi-objective decision making (MODM) and multi-attribute decision making (MADM). MADM have been used to solve problems with discrete decision choices and a predetermined or limited number of alternative choices. A comparative study on various MCDM methods is presented in [3, 4].

In this paper, an MADM approach to project portfolio selection (PPS), which is one of the links in project portfolio management chain, is applied. The classical approach is expanded to deal with uncertainty expressed in the form of fuzzy numbers. Modelling of uncertainty is of great importance in modern knowledge-based organisation. It enables to describe errors resulting from the conversion of real life problems into knowledge representations. There is a range amount of scientific publications which develop very sophisticated methods for describing uncertainty. Meanwhile, according to the survey of Hubbard [5], modern enterprises still asses and mitigate risk using old fashioned methods which have not evolved much for several decades. This paper attempts to fill out the gap between the theory and practise. A practical framework to deal with this problem is developed. The reminder of this paper is organised as follows: Sect. 2 briefly describes problems with modelling uncertainty in PPS. It also provides explanation of benefits in using fuzzy approach to model uncertainty in practise. Sect. 3 presents methodological framework of proposed methods: fuzzy numbers and intervals, fuzzy AHP and fuzzy TOPSIS are described in details. A numerical example is shown in Sect. 5. The paper ends with conclude remarks.

2 Risk and Uncertainty in Project Portfolio Selection

Nowadays, methods of risk assessment constitute a fundamental tool for supporting enterprise decision-making process. Parameters of economic calculus are usually burdened with uncertainty. For many years, the only tool which allowed to express uncertainty in mathematical language was probability calculus. However, in numerous decision-making situations, the nature of uncertainty of economic model parameters does not satisfy the assumptions of the probability theory. This is when uncertainty stems mainly from insufficient information about model parameters or when it has an epistemological nature. In practise, it is often not possible to determine probability distribution or perform probability calculus, especially when there is no sufficient volume of data to use statistical tests. On the other hand, assumption of "no data available at all" is not true as well. In general, there is always some information available. Most often, in the form of experts' estimates of unknown values.

There is no universally accepted definition of business risk and uncertainty, but in the PPS context they may be understood as potential problems with availability and certainty of information, and also imprecise choices. To deal with uncertainty in PPS, it must be first noted that PPS usually consists of two stages. In the first stage, projects are selected on the basis of the threshold criteria which are determined by decision-makers. In order that a project can be passed to the next stage, it must strictly fulfill these criteris. Typical threshold criteria include financial criteria, e.g., $\text{NPV} > 0$, $\text{IRR} >$ threshold IRR etc. The selected projects are input for MADM method. An MADM method usually firstly calculate the weights of criteria, and then determines the ranking of potential projects. Each part of MADM method is associated with different type of uncertainty.

The main source of uncertainty in determination of criteria weights is imprecision of expert judgements. Due to cognitive biases, decisions may be deviated from a standard of rationality or good judgement. To take into account these systematic errors, fuzzy numbers are used instead of crisp numbers. Unfortunately, practical usage of fuzzy numbers as criteria ratios faces many problems. The most important ones are the following: Whether to use crisp numbers or fuzzy numbers to describe uncertainty? What type of fuzzy sets should be used—fuzzy numbers or fuzzy intervals? How to transform linguistic description into fuzzy scale? What kind of representation should it be used to compare choices—fuzzy or crisp form obtained by defuzzification?

There is no agreement between scientists on how to answer the above questions. Later in this paper a fuzzy AHP method is used to obtain fuzzy criteria weights.

Some researchers believe that classical Saaty's AHP method has some weaknesses which are connected with uncertainty. In [6] authors points out that mapping experts judgement to crisp numbers and cognitive biases generates uncertainty which is not taken into account by the AHP method and may have huge impact on AHP result. To deal with this problem, some researches fuzzified AHP (e.g., [7] has considered trapezoidal fuzzy intervals for comparison ratios in AHP and [8] has proposed approach for triangular case). However, Saaty points out that pairwise comparison matrix is already fuzzy because it lies in the methods assumption. He believes that AHP is already a fuzzy process because most criteria for ranking are in fact very specific fuzzy numbers representation and there is no theoretical proof that fuzzifying the comparison data leads to better results. Therefore, it cannot be proved that fuzzifying AHP is a confident idea [9]. He also points out that fuzzification of the process does not give much different results. This work is usually cited as a major point of criticism against fuzziness in AHP methods. It is important to note that the main criticism is based on assumption that comparison ratios are based on expert consensus. In practise, comparison ratios are usually averages of expert ratios. As long as there are no agreement between experts, everyone of them interprets linguistic variables in different ways. In this situation, translation of expert verbal possibilities into numbers should better be better used a fuzzy then crisp numbers.

The second aspect of representing model uncertainty in terms of fuzzy numbers concerns the problem of which type of fuzzy numbers should be used. Generally speaking, there are two approach to fuzzification of comparison ratios—fuzzy numbers [10, 11] and fuzzy intervals [12]. The empirical study shows [13] that membership functions of numerical equivalents of linguistic terms are similar to fuzzy numbers, which are not distributed equidistantly along the possibility scale and which vary considerably in symmetry and vagueness. Table 1 in Sect. 5 provides summary of translation developed based on this study.

Usually, after the first stage, the calculated criteria weights are defuzzified. In the proposed approach, fuzzy weights are passed to the second stage. This guaranties that uncertain judgements of decision-makers are taken into account also during the second phase of PPS.

The second phase of PPS determines the ranking of potential projects based on the weights obtained in the first stage. In this phase, uncertainty concerns attributes of alternatives. The attributes are divided into two groups: objective (numerical) and subjective (linguistic). A majority of authors argue that only subjective criteria should be described in terms of fuzzy numbers. In the proposed approach it is assumed that quantification of financial attributes of investment projects should be modelled as mixture of possibility and probability distribution. In [14, 15] Rębiasz presents methods for selection of efficient portfolios in a situation where objective parameters in the calculation of effectiveness are expressed in form of interactive fuzzy numbers and probability distribution. In this paper, a new hybrid fuzzy multiple criteria approach for project selection is proposed. It includes financial, social, and environmental effects of an investment, strategic alliance, organisational readiness, and risk of investment.

3 Methodology

The proposed methodology of selecting an efficient portfolio of investment projects consists of the following steps. First, multiple criteria that are considered in the decision-making process for the decision-makers are identified. Then, criteria weights are calculated according to the fuzzy AHP methodology. After constructing the relationship of a criteria decision matrix, the fuzzy TOPSIS approach is used to achieve the final ranking results.

3.1 Fuzzy Numbers

From the mathematical point of view, fuzzy sets generalise classical set theory by replacing the binary membership function with a function of continuum grades having values from the interval [0, 1].

Definition 1 A fuzzy set $\tilde{A} \subseteq X$ is characterised by a membership function

$$\mu_{\tilde{A}} : X \ni x \to \mu_{\tilde{A}}(x) \in [0, 1].$$

The function value $\mu_{\tilde{A}}(x)$ is called the grade of membership of x in $\tilde{A}$.

Definition 2 Given a fuzzy number $\tilde{A}$ in X and a real number $\alpha \in [0, 1]$, then the α-cut or α-level of $\tilde{A}$, denoted by $\tilde{A}^\alpha$ is the crisp set

$$\tilde{A}^\alpha = \{x \in X \mid \mu_{\tilde{A}}(x) \geqslant \alpha\}. \tag{1}$$

Definition 3 A fuzzy set $\tilde{A} \subseteq \mathbb{R}$ is called a fuzzy number if the following conditions hold:

1. $\exists!\, x_0 \in \mathbb{R}$ such that $\mu_{\tilde{A}}(x_0) = 1$ (*normality*),
2. $\mu_{\tilde{A}}(\lambda x + (1-\lambda)y) \geqslant \min\{\mu_{\tilde{A}}(x), \mu_{\tilde{A}}(y)\}, \forall x, y \in \mathbb{R},\ \forall \lambda \in [0, 1]$ (*convexity*),
3. $\forall x_0 \in \mathbb{R}, \forall \varepsilon > 0$ there exists a neighbourhood $V(x_0)$, such that $\forall x \in V(x_0)\ \mu_{\tilde{A}}(x_0) \leqslant \mu_{\tilde{A}}(x) + \varepsilon$ (*upper semi-continuity*),
4. the support $supp(\tilde{A}) = cl\{x \in \mathbb{R} \mid \mu_{\tilde{A}}(x) > 0\}$ is bounded (*compactness*).

The set of all fuzzy numbers will be denoted by $F(\mathbb{R})$.

The most commonly used fuzzy numbers are trapezoidal and triangular fuzzy numbers.

Definition 4 A trapezoidal fuzzy number (fuzzy interval) $\tilde{A} \in F(\mathbb{R})$ is represented by a 4-tuple $\tilde{A} = (a_1, a_2, a_3, a_4)$ and has the following membership function:

$$\mu_{\tilde{A}}(x) = \begin{cases} 0 & x < a, \\ \frac{x-a_1}{a_2-a} & a_1 \leqslant x \leqslant a_2, \\ 1 & a_2 < x < a_3, \\ \frac{a_4-x}{a_4-a_3} & a_3 \leqslant x \leqslant a_4, \\ 0 & x > a_4. \end{cases} \tag{2}$$

Basic arithmetic operations on trapezoidal fuzzy numbers can be defined using α-cuts or, as given below, using the 4-tuple representation. Given two independent trapezoidal fuzzy numbers $\tilde{A} = (a_1, a_2, a_3, a_4)$ and $\tilde{B} = (b_1, b_2, b_3, b_4)$, their addition and subtraction are defined as:

$$\tilde{A} + \tilde{B} = (a_1 + b_1, a_2 + b_2, a_3 + b_3, a_4 + b_4), \tag{3}$$

$$\tilde{A} - \tilde{B} = (a_1 - b_4, a_2 - b_3, a_3 - b_2, a_4 - b_1), \tag{4}$$

Multiplication and division on trapezoidal fuzzy numbers can be defined in different ways, and the result is not necessarily a trapezoidal fuzzy number. The following formulae [16, 17] ensure that property:

$$\begin{aligned} \tilde{A} \cdot \tilde{B} &= (c_1, c_2, c_3, c_4), \text{ where} && (5) \\ c_1 &= \min\{a_1 b_1, a_1 b_4, a_4 b_1, a_4 b_4\}, \\ c_2 &= \min\{a_2 b_2, a_2 b_3, a_3 b_2, a_3 b_3\}, \\ c_3 &= \max\{a_2 b_2, a_2 b_3, a_3 b_2, a_3 b_3\}, \\ c_4 &= \max\{a_1 b_1, a_1 b_4, a_4 b_1, a_4 b_4\}, \\ \tilde{A}/\tilde{B} &= (d_1, d_2, d_3.d_4), \text{ where} && (6) \\ d_1 &= \min\{a_1/b_1, a_1/b_4, a_4/b_1, a_4/b_4\}, \\ d_2 &= \min\{a_2/b_2, a_2/b_3, a_3/b_2, a_3/b_3\}, \\ d_3 &= \max\{a_2/b_2, a_2/b_3, a_3/b_2, a_3/b_3\}, \\ d_4 &= \max\{a_1/b_1, a_1/b_4, a_4/b_1, a_4/b_4\}. \end{aligned}$$

In the case of division, it is assumed that $0 \notin [b_1, b_4]$.

A trapezoidal fuzzy number $\tilde{A} = (a_1, a_2, a_3, a_4)$ with $a_2 = a_3$ is called a triangular fuzzy number.

3.2 Fuzzy AHP

The *Analytic Hierarchy Process* (AHP) method developed by Saaty [18] is a powerful and flexible MCDM tool for complex problems where both qualitative and quantitative aspects need to be considered [19]. The AHP integrates different measures into a single overall score for ranking alternative decisions. It is based on pairwise comparison judgements. By reducing complex decisions to a series of simple comparisons and rankings, then synthesising the results, the AHP not only helps the analysts to arrive at the best decision, but also provides a clear rationale for the choices made [8].

The main steps of AHP are the following [8]:

1. Define decision criteria in the form of a hierarchy of objectives.
2. Build judgement matrices by pairwise comparisons.
3. Calculate a priority vector in order to weight the elements of the matrix.
4. Calculate global priorities by aggregating all local priorities using a simple weighted sum.
5. Use the eigenvalue in order to assess the strength of the consistency ratio of the comparative matrix and determine whether to accept the information. If the comparison matrices are not consistent, the elements in the matrices should be adjusted and a consistency test should be carried out until they are consistent.

The fuzzy AHP [8] is the fuzzy extension of AHP to deal with the fuzziness of the data involved in the decision making process. Fuzzy AHP enables decision makers to specify preferences in the form of natural language expressions about the importance of each performance attribute.

3.3 Fuzzy TOPSIS

Technique for Order Preference by Similarity Ideal Solution (TOPSIS) is another popular approach to MCDM. It was proposed by Hwang and Yoon [20] and is widely used in the literature [21]. The main idea behind the method is that the best alternative should have the shortest distance from the (positive) ideal solution and the farthest distance from the negative ideal solution.

The TOPSIS process is carried out as follows:

1. Create an evaluation matrix E consisting of m alternatives (rows) and n criteria (columns), with the intersection of each alternative and criteria given as e_{ij},
2. Create a normalised matrix $R = (r_{ij})_{m\times n}$ using the following normalisation:

$$r_{ij} = \frac{e_{ij}}{\sqrt{\sum_{k=1}^{m} e_{kj}^2}}, i = 1, 2, \ldots, m, \;\; j = 1, 2, \ldots, n,$$

   The normalisation is performed in order to eliminate anomalies with different measurement units and scales.

3. Calculate the weighted normalised decision matrix

$$T = (t_{ij})_{m\times n} = (w_j r_{ij})_{m\times n}, i = 1, 2, \ldots, m$$

   where $w_j = W_j / \sum_{j=1}^{n} W_j$, $j = 1, 2, \ldots, n$ so that $\sum_{j=1}^{n} w_j = 1$, and W_j is the original weight given to the indicator v_j, $j = 1, 2, \ldots, n$.

4. Determine the worst alternative (A_w) and the best alternative (A_b):

$$\begin{aligned} A_w &= \{\langle \max(t_{ij} \mid i = 1, 2, \ldots, m) \mid j \in J_- \rangle, \langle \min(t_{ij} \mid i = 1, 2, \ldots, m) \mid j \in J_+ \rangle\} \\ &\equiv \{t_{wj} \mid j = 1, 2, \ldots, n\}, \end{aligned} \tag{7}$$

$$\begin{aligned} A_w &= \{\langle \min(t_{ij} \mid i = 1, 2, \ldots, m) \mid j \in J_- \rangle, \langle \max(t_{ij} \mid i = 1, 2, \ldots, m) \mid j \in J_+ \rangle\} \\ &\equiv \{t_{bj} \mid j = 1, 2, \ldots, n\}, \end{aligned} \tag{8}$$

where

$$J_+ = \{j \mid 1 \leqslant j \leqslant n, j \text{ associated with the criteria having a positive impact}\},$$

and

$$J_- = \{j \mid 1 \leqslant j \leqslant n, j \text{ associated with the criteria having a negative impact}\},$$

5. Calculate the Euclidean distance between the target alternative i and the worst condition A_w:

$$d_{iw} = \sqrt{\sum_{j=1}^{n} (t_{ij} - t_{wj})^2}, i = 1, 2, \ldots, m,$$

and the Euclidean distance between the alternative i and the best condition A_b

$$d_{ib} = \sqrt{\sum_{j=1}^{n} (t_{ij} - t_{bj})^2}, i = 1, 2, \ldots, m.$$

6. Calculate the similarity to the worst condition:
$s_{iw} = d_{ib}/(d_{iw} + d_{ib}), 0 \leq s_{iw} \leq 1, i = 1, 2, \ldots, m.$
$s_{iw} = 1$ if and only if the alternative solution has the worst condition; and
$s_{iw} = 0$ if and only if the alternative solution has the best condition.
7. Rank the alternatives according to $s_{iw} (i = 1, 2, \ldots, m)$.

The TOPSIS method has also been extended to deal with fuzzy numbers. In order that it can be used to deal with fuzzy MCDM problems, several approaches have been proposed. The simplest one is to change fuzzy MCDM into a crisp one by using defuzzification. This approach, however, can lead to the loss of information. Another approach is to define a crisp Euclidean distance between fuzzy numbers. For example, Chen [22] defines the distance between two triangular fuzzy numbers $\tilde{A} = (a_1, a_2, a_3)$ and $\tilde{B} = (b_1, b_2, b_3)$ as

$$d(\tilde{A}, \tilde{B}) = \sqrt{\frac{\sum_{i=1}^{3} (a_i - b_i)^2}{3}}.$$

An approach based on α-cuts can also be found in the literature [23].

4 Proposed Framework for Project Portfolio Selection

Based on methodology described in Sect. 3, a new approach to project portfolio selection problem is proposed. It consists of five stages.

4.1 Identification of Available Investment Projects and Criteria

In this stage a committee of decision-makers who come from different managerial levels of the company is formed. They identify m potential investment projects $A_1,\ldots,$ A_m and n criteria $C_1, \ldots, C_n$ that the projects must fulfil. To properly assess a project, many factors should be considered. There is a wide area of publications concerning choosing project investment criteria. McKown and Mohamed [24] presents multi-criteria project selection where uncertainty of profitability parameters is described by fuzzy numbers. They point out that the selection of investment projects should consist of two kinds of parameters—financial (e.g., net present value (NPV), internal-rate-of-return (IRR) and pay-back period) and non-financial (e.g., social, environmental, strategic an organisational).

In this paper, a method that allows to aggregate financial and non-financial indicators is proposed. First, the criteria are divided into two groups—objective and subjective. Objective criteria are described by fuzzy numbers which usually result from fuzzy modelling or aggregation of historical data. Subjective criteria are qualitative criteria with values that are specified by decision makers in the form of linguistic variables. In the approach presented here, linguistic variables are transformed into fuzzy numbers (triangular or trapezoidal). It simplifies further ranking of the projects.

4.2 Calculation of Synthetic Importance Weights

Obviously, the problem of calculating the importance weights of the criteria is a typical multi-variable and multi-objective optimisation problem. To calculate importance weights of the criteria we use fuzzy AHP. AHP has been widely used and successfully applied to many practical decision-making problems. The fuzzy AHP method is used here because decision makers have different understanding of meaning of linguistic variable. To make a pairwise comparison, linguistic scale which are shown on Table 1 is developed. Decision-makers use this scale to give their judgement indirectly using pairwise comparison. Then, every judgement is converted into triangular fuzzy number through a designed rating scale. To obtain pairwise comparison matrix, average value of decision-makers judgement is used. Finally, synthetic importance weight are obtained by fuzzy AHP algorithm. The final scores of criteria are also represented by fuzzy numbers.

Table 1 Comparison of relative importance of criteria for fuzzy AHP (linguistic terms and associated fuzzy numbers)

Linguistic terms	Crisp intensity of importance (classical approach)	Fuzzy intensity of importance
Equally important	1	(1, 1, 1)
Moderately more important	3	(1, 3, 5)
Strongly more important	5	(2, 5, 6)
Very strongly more important	7	(6, 7, 8)
Extremely more important	9	(8, 9, 9)

4.3 Development of Performance Ratings for Projects

In this stage, the performance ratings of objective and subjective parameters are calculated. All ratings of all subjective parameters for alternative projects are obtained using preference ratings. Each expert evaluates ratings of subjective criteria in terms of linguistics variables. Then, all linguistic variables are converted into fuzzy numbers. Subsequently, arithmetic mean is used to calculate the rating of each criteria. The objective variables are also described by fuzzy numbers.

At the end of this stage the threshold selection is made. Only those projects which have passed the threshold selection are taken into account in the next stage of the PPS.

4.4 Calculating Hierarchy of Projects Using Fuzzy TOPSIS

In this stage, the hierarchy of projects is established. Both subjective and objective parameters are combined using fuzzy weights obtained from the fuzzy AHP method. First, objective and subjective parameters are normalised. Normalisation is a common modification process that involves the division of each value by the largest value, resulting in the range is between 0 and 1. Then, the overall ranking of projects is calculated. The ranking allows a decision-maker to select the most appropriate investment option.

5 Numerical Example

The proposed approach was applied for PPS in steel industry. Fig. 1 presents the hierarchy of criteria. There are five criteria $C1, \ldots, C5$—financial, market, technology and environment, staff and compliance with the company's strategic objective. Each of them is divided into subcriteria. The following objective subcriteria were used:

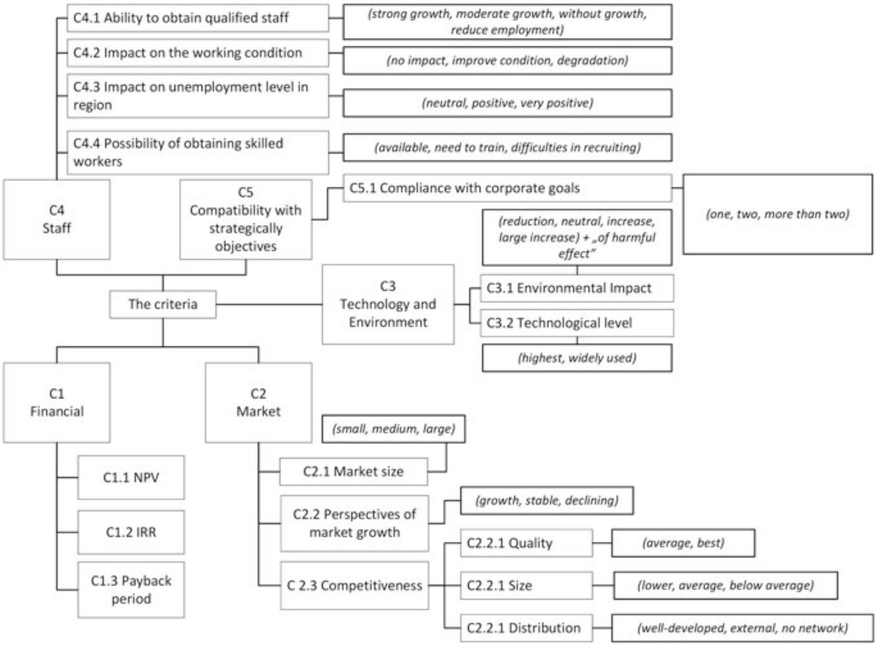

C4.1 Ability to obtain qualified staff
(strong growth, moderate growth, without growth, reduce employment)
C4.2 Impact on the working condition
(no impact, improve condition, degradation)
C4.3 Impact on unemployment level in region
(neutral, positive, very positive)
C4.4 Possibility of obtaining skilled workers
(available, need to train, difficulties in recruiting)
C4 Staff
C5 Compatibility with strategically objectives
C5.1 Compliance with corporate goals
(one, two, more than two)
(reduction, neutral, increase, large increase) + „of harmful effect"
The criteria
C3 Technology and Environment
C3.1 Environmental Impact
C3.2 Technological level
(highest, widely used)
C1 Financial
C2 Market
(small, medium, large)
C1.1 NPV
C2.1 Market size
C2.2 Perspectives of market growth
(growth, stable, declining)
C1.2 IRR
C2.2.1 Quality
(average, best)
C1.3 Payback period
C 2.3 Competitiveness
C2.2.1 Size
(lower, average, below average)
C2.2.1 Distribution
(well-developed, external, no network)

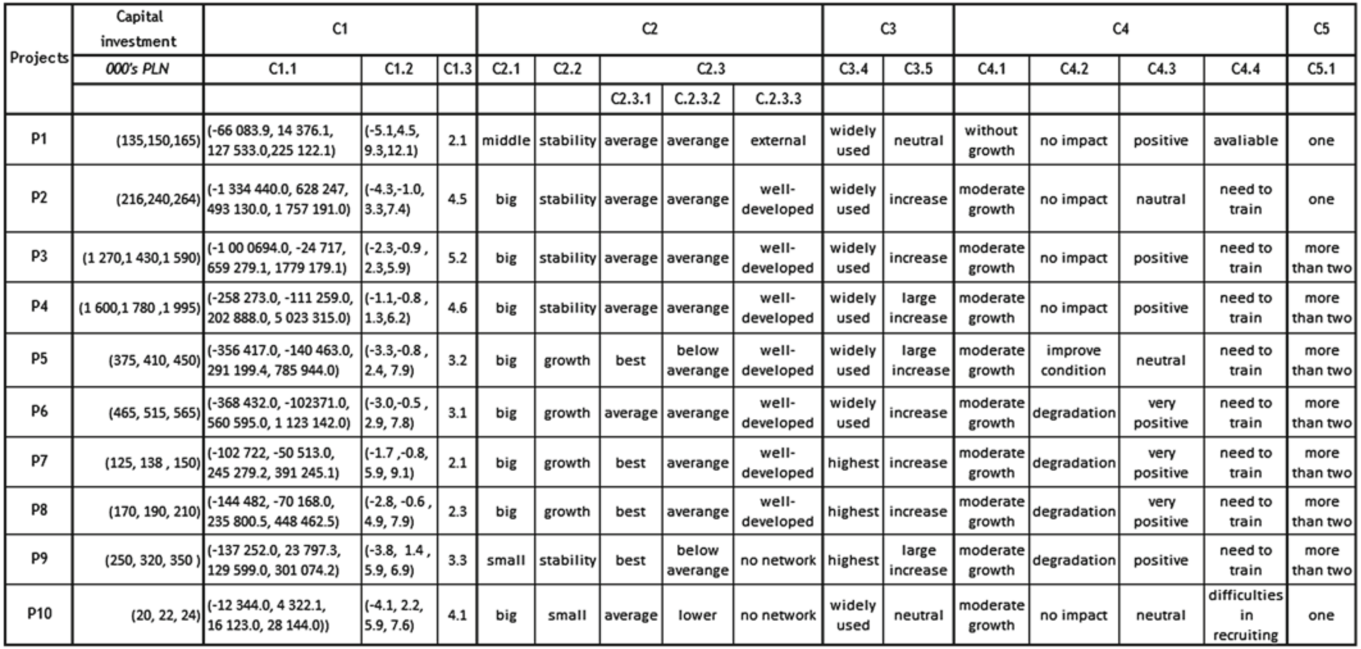
Fig. 1 The hierarchy of project requirements

1. **NPV**—represents difference between two cash flows: inflows and outflows. It compares the present value of money today to the present value of money in future, taking inflation and returns into account.
2. **IRR**—discount rate at which the net present value of the investment costs equals the net present value of the benefits.
3. **Payback period**—represents the amount of time that it takes for a Capital Budgeting project to recover its initial cost.

The rest of subcriteria is subjective. There is also the third level of subcriteria for the C2 criteria. They are called attributes.

To calculate weights of criteria, a team of decision makers make pair-wise comparison of criteria. The results of this comparison are presented in Tables 2, 3 and 4. Then, using the fuzzy AHP global priorities are obtained. They are shown in Table 5. The priorities are presented in terms of fuzzy numbers. It can be seen that the higher hierarchy of the criteria is the wider fuzzy number are. For example, fuzzy weight C2.3.1 range between 0 to nearly 0.3. This illustrates the well-known phenomena of accumulation of uncertainty. That is why in next step the consistency degree should be used (e.g., fuzzy preference programming).

In the next step, evaluation matrix is created. Matrix consists of 15 criteria and 10 projects ($P1, \ldots, P10$) (Fig. 2). The alternatives were as follow:



Projects	Capital investment	C1			C2					C3		C4				C5
	000's PLN	C1.1	C1.2	C1.3	C2.1	C2.2	C2.3			C3.4	C3.5	C4.1	C4.2	C4.3	C4.4	C5.1
							C2.3.1	C.2.3.2	C.2.3.3							
P1	(135,150,165)	(-66 083.9, 14 376.1, 127 533.0,225 122.1)	(-5.1,4.5, 9.3,12.1)	2.1	middle	stability	average	averange	external	widely used	neutral	without growth	no impact	positive	avaliable	one
P2	(216,240,264)	(-1 334 440.0, 628 247, 493 130.0, 1 757 191.0)	(-4.3,-1.0, 3.3,7.4)	4.5	big	stability	average	averange	well-developed	widely used	increase	moderate growth	no impact	nautral	need to train	one
P3	(1 270,1 430,1 590)	(-1 00 0694.0, -24 717, 659 279.1, 1779 179.1)	(-2.3,-0.9 , 2.3,5.9)	5.2	big	stability	average	averange	well-developed	widely used	increase	moderate growth	no impact	positive	need to train	more than two
P4	(1 600,1 780 ,1 995)	(-258 273.0, -111 259.0, 202 888.0, 5 023 315.0)	(-1.1,-0.8 , 1.3,6.2)	4.6	big	stability	average	averange	well-developed	widely used	large increase	moderate growth	no impact	positive	need to train	more than two
P5	(375, 410, 450)	(-356 417.0, -140 463.0, 291 199.4, 785 944.0)	(-3.3,-0.8 , 2.4, 7.9)	3.2	big	growth	best	below averange	well-developed	widely used	large increase	moderate growth	improve condition	neutral	need to train	more than two
P6	(465, 515, 565)	(-368 432.0, -102371.0, 560 595.0, 1 123 142.0)	(-3.0,-0.5 , 2.9, 7.8)	3.1	big	growth	average	averange	well-developed	widely used	increase	moderate growth	degradation	very positive	need to train	more than two
P7	(125, 138 , 150)	(-102 722, -50 513.0, 245 279.2, 391 245.1)	(-1.7 ,-0.8, 5.9, 9.1)	2.1	big	growth	best	averange	well-developed	highest	increase	moderate growth	degradation	very positive	need to train	more than two
P8	(170, 190, 210)	(-144 482, -70 168.0, 235 800.5, 448 462.5)	(-2.8, -0.6 , 4.9, 7.9)	2.3	big	growth	best	averange	well-developed	highest	increase	moderate growth	degradation	very positive	need to train	more than two
P9	(250, 320, 350)	(-137 252.0, 23 797.3, 129 599.0, 301 074.2)	(-3.8, 1.4 , 5.9, 6.9)	3.3	small	stability	best	below averange	no network	highest	large increase	moderate growth	degradation	positive	need to train	more than two
P10	(20, 22, 24)	(-12 344.0, 4 322.1, 16 123.0, 28 144.0))	(-4.1, 2.2, 5.9, 7.6)	4.1	big	small	average	lower	no network	widely used	neutral	moderate growth	no impact	neutral	difficulties in recruiting	one

Fig. 2 Description of the projects alternatives

Table 2 Pairwise comparison matrices

	Criteria				
	C1	C2	C3	C4	C5
C1	(1,1,1)	(2,5,6)	(8,9,9)	(8,9,9)	(8,9,9)
C2	(0.17,0.2,0.5)	(1,1,1)	(1,3,5)	(1,3,5)	(1,3,5)
C3	(0.11,0.11,0.13)	(0.2,0.34,1)	(1,1,1)	(1,1,1)	(1,1,1)
C4	(0.11,0.11,0.13)	(0.2,0.34,1)	(0.2,1,1)	(1,1,1)	(1,1,1)
C5	(0.11,0.11,0.13)	(0.2,0.33,1)	(0.2,1,1)	(0.2,1,1)	(1,1,1)

Table 3 Pairwise comparison matrices—subcriteria

	Sub-criteria			
	C1.1	C1.2	C1.3	
C1.1	(1,1,1)	(1,1,1)	(6,7,8)	
C1.2	(1,1,1)	(1,1,1)	(6,7,8)	
C1.3	(0.13,0.14,0.17)	(0.12,0.14,0.17)	(1,1,1)	
	C2.1	C2.2	C2.3	
C2.1	(1,1,1)	(0.12,0.14,0.17)	(1,3,5)	
C2.2	(6,7,8)	(1,1,1)	(0.16,0.2,0.5)	
C2.3	(0.2,0.33,1)	(2,5,6)	(1,1,1)	
	C3.1	C3.2		
C3.1	(1,1,1)	(6,7,8)		
C3.2	(0.13,0.14,0.17)	(1,1,1)		
	C4.1	C4.2	C4.3	C4.4
C4.1	(1,1,1)	(0.17,0.2,0.5)	(0.2,0.33,1)	(1,3,5)
C4.2	(2,5,6)	(1,1,1)	(1,3,5)	(8,9,9)
C4.3	(1,3,5)	(0.2,0.333,1)	(1,1,1)	(8,9,9)
C4.4	(0.2,0.33,1)	(0.11,0.11,0.13)	(0.11,0.11,0.13)	(1,1,1)

Table 4 Pairwise comparison matrices—attributes

	Attributes		
	C2.3.1	C2.3.2	C2.3.3
C2.3.1	(1,1,1)	(1,3,5)	(1,1,1)
C2.3.2	(0.2,0.333,1)	(1,1,1)	(1,3,5)
C2.3.3	(1,1,1)	(0.2,0.333,1)	(1,1,1)

- $P1$—Modernisation of the heavy section mill,
- $P2$—Increase of the capacity of the hot rolling mill,
- $P3$—Construction of the cold rolling mill with the capacity of 1,000 thousand t/year,

Table 5 Importance weights of individual requirements

	Weight		Weight
C1	(0.429,0.633,0.917)	C3.1	(0.039,0.056,0.124)
C2	(0.073,0.175,0.372)	C3.2	(0.006,0.008,0.018)
C3	(0.051,0.064,0.122)	C4.1	(0.003,0.007,0.039)
C4	(0.051,0.064,0.122)	C4.2	(0.013,0.036,0.145)
C5	(0.051,0.064,0.122)	C4.3	(0.008,0.019,0.085)
C1.1	(0.181,0.292,0.471)	C4.4	(0.001,0.003,0.013)
C1.2	(0.181,0.292,0.471)	C5.1	(0.051,0.064,0.122)
C1.3	(0.025,0.042,0.072)	C2.3.1	(0.002,0.027,0.313)
C2.1	(0.006,0.05,0.307)	C2.3.2	(0.001,0.02,0.264)
C2.2	(0.018,0.061,0.204)	C2.3.3	(0.002,0.015,0.127)
C2.3	(0.011,0.063,0.31)		

- $P4$—Construction of the cold rolling mill with the capacity of 1,500 thousand t/year,
- $P5$—Construction of the hot dip galvanising line with the capacity: 300 thousand t/year,
- $P6$—Construction of the hot dip galvanising line with the capacity: 400 thousand t/year,
- $P7$—Construction of the organic coating line with the capacity of 200 thousand t/year,
- $P8$—Construction of the organic coating line with the capacity of 300 thousand t/year,
- $P9$—Construction of the tinning plant with the capacity of 100 thousand t/year,
- $P10$—Construction of the wire drawing plant.

The objective criteria are characterised by fuzzy intervals and are taken fuzzy modelling. The level of subjective criteria are specified by experts. The subjective criteria are translated into triangular fuzzy numbers using procedure as follows:

In the presented example, there are two kinds of subjective attributes—some of them describe patterns, and some of them judgements. Market size criterion C2.1 and prospects for market growth criterion C2.2 belong to first group. They describe the belief of decision maker that market for project i will behave in accordance with some pattern. For example, pattern *stable* means the dynamic of the market growth which may be described by the fuzzy number $(-1.02, 0, 1.02)$.

The second group that is subjective criteria represents judgements of experts. Therefore, they are treated as ordinal fuzzy variables. Since all of subjective criteria are ordinal (variable with order), thus fuzzy ordinal rank transformation is used. After translation of linguistic variables—the fuzzy TOPSIS is applied. The obtained final ranking of projects is presented in Table 6.

Table 6 Final ranking of projects

Project	Rank
P9	0.7154
P10	0.7101
P1	0.7095
P4	0.7011
P3	0.6817
P7	0.6789
P8	0.6782
P5	0.6770
P6	0.6714
P2	0.6615

6 Conclusions

The evaluation and selection of industrial projects is one of the most important aspects of PPS. This paper proposed a combined fuzzy MADM approach based on fuzzy AHP and fuzzy TOPSIS techniques. A real world case study from steel industry was presented to explain approach. The paper introduced fuzzy decision making concept, when some data is burden with uncertainty. It is argued that if a fuzzy MADM problem is defuzzified into crisp one to early, then the advantage of modeling uncertainty becomes negligible. The rational approach is to defuzzify imprecise values at the very end of methods. Based on this argument, we perform deffuzzification at the very end of MADM method during calculate weight of criteria.

More research is needed to examine projects interaction and dependency. Further research is also required with respect to the subjective criteria of project selection. The problem of quantifying the qualitative factors remains a difficult and sometimes controversial tasks.

References

1. Doumpos, M., Zopounidis, C.: Multiattributes Decision Aid Classification Methods. Kluwer Academic Publishers, Boston (2002)
2. Vincke, P.: Multiattributes decision aid. Wiley, New York (1992)
3. Archer, N.P., Ghasemzadeh, F.: An Integrated framework for project portfolio selection. Int. J. Proj. Manag. **7**(4), 207–216 (1999)
4. Figueira, J., Greco, S., Ehrgott, S.: Multiple Attributes Decision Analysis: State Of The Art Surveys. Springer, New York (2005)
5. Hubbard, D.W.: The Failure of Risk Management: Why It's Broken and How to Fix It. Wiley, New Jersey (2009)
6. Yang, ChCh., Chen, BSh: Key quality performance evaluation using Fuzzy AHP. J. Chin. Inst. Ind. Eng. **21**(6), 543–550 (2004)
7. Buckley, J.J.: Fuzzy hierarchical analysis. Fuzzy sets and syst. **17**(3), 233–247 (1985)

8. Chang, D.-Y.: Applications of the extent analysis method on fuzzy AHP. Eur. J. Op. Res. **95**, 649–655 (1996)
9. Saaty, T.L., Tran, L.T.: On the invalidity of fuzzifying numerical judgments in the analytic hierarchy process. Math. and Comput. Model. **46**(7), 962–975 (2007)
10. Mahmoodzadeh, S., Shahrabi, J.: Project selection by using fuzzy AHP and TOPSIS technique. International Journal of Humanities and Social Sciences **1**(3):135–140 (2007) Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.4107&rep=rep1&type=pdf
11. Wang, J., Fan, K., Wang, W.: Integration of fuzzy AHP and FPP with TOPSIS methodology for aeroengine health assessment. Expert Syst. with Appl. **37**(12), 8516–8526 (2010)
12. Liu, P., Su, Y.: The extended topsis based on trapezoid fuzzy linguistic variables. J. Converg. Inf. Technol. **5**(4), 38–53 (2010)
13. Bocklisch, F., Bocklisch, S.F., Krems, J.F. (n.d.): How to Translate Words into Numbers? Fuzzy Approach for the Numerical Translation of Verbal Probabilities. Lecture Notes in Computer Science **6178**, 614–623 (2010)
14. Rębiasz, B.: Fuzziness and randomness in investment project risk appraisal. Comput. Opera. Res. **34**(1), 199–210 (2007)
15. Rębiasz, B.: Selection of efficient portfolios-probabilistic and fuzzy approach, comparative study. Computers Industrial Engineering, Perganon (2013)
16. Anile, A.M., Deodato, S., Privitera, G.: Implementing fuzzy arithmetic. Fuzzy Sets and Syst. **72**(2), 239–250 (1995)
17. Kuchta, D.: Soft mathematics in management. Use of interval and fuzzy numbers in managerial accounting (monograph in Polish). Oficyna Wydawnicza Politechniki Wroclawskiej, Wroclaw (2001)
18. Saaty, T.L.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)
19. Bevilacqua, M., Braglia, M.: The analytic hierarchy process applied to maintenance strategy selection. Reliab. Eng. Syst. Saf. **70**(1), 71–83 (2000)
20. Hwang, C.L., Yoon, K.: Multiple Attribute Decision Making: Methods and Applications. Springer, New York (1981)
21. Boran, F.E., Gen, S., Kurt, M., Akay, D.: A multi-criteria intuitionistic fuzzy group decision making for supplier selection with TOPSIS method. Expert Syst. with Appl. **36**(8), 11363–11368 (2009)
22. Chen, C.T.: Extension of the TOPSIS for group decision-making under fuzzy environment. Fuzzy Sets and Syst. **114**, 1–9 (2000)
23. Wang, Y.-M., Elhag, T.M.S.: Fuzzy TOPSIS method based on alpha level sets with an application to bridge risk assessment. Expert Syst. with Appl. **31**, 309–319 (2006)
24. Mahomed, S., McKown, A.K.: Modelling project investment decisions under uncertainty using possibility theory. Int. J. Proj. Manag. **19**, 231–241 (2001)

Theory of Digital Data Transformation in the ICT

Lubomyr B. Petryshyn

Abstract Functional analysis is crucial for digital data processing. There are a lot of bases and systems of functions used for the decomposition of signals. This article presents the results of the research that identified the mathematical interdependence between bases of functions and coding systems. These results allowed systematizing and analytically describing the procedure for transforming the form of information in information systems.

1 Introduction

Modern information and communication systems process and manage large volumes of digital data. The efficiency of such systems is determined by methods used for transforming the form and digital processing of information. Since nowadays there are a lot of mathematical methods of coding and data processing, the choice of methods for coding and processing data adapted to the nature of information sources is an important direction of research in the theory of information technology.

Lately, complex information systems that include systems function of creating, forming, converting, transmitting, processing and documenting digital data [1, 2] become widespread. When building an information system, the problem of selecting an appropriate coding system is considered at four levels:

1. theory of information sources and data formation;
2. the method for conversion of the form of information;
3. transmission and receiving of digital data;
4. processing, displaying and storage of information.

L.B. Petryshyn (✉)
AGH University of Science and Technology, 30 Mickewicha Al., 30-059 Cracow, Poland
e-mail: L.B.Petryshyn@gmail.com

L.B. Petryshyn
Precarpathian National University, 57 Shevchenko Str., Ivano-Frankivsk 76-025, Ukraine

M. Mach-Król et al. (eds.), *Advances in ICT for Business, Industry and Public Sector*,
Studies in Computational Intelligence 579, DOI 10.1007/978-3-319-11328-9_10

In order to solve these problems of information forms transformation and digital data processing this paper analyzes and defines systems of the basic functions and identified analytical relationship between them. This allowed analyzes functions systems for creating of code and coding systems. On the basis of the defined functions systems a transition is realized to appropriate codes and coding systems. The analyzed code systems are used to solve such problems as creating, forming, converting, processing, compressing and storing information [2, 3].

The present research focuses on methods of analysis, synthesis and design of the coding methods, which are applied in discrete devices for transformation of the information form and process digital data and are based on the theory of orthogonal series. The main purpose of this work is the development of a unified approach to solving the number-theoretic transformation and digital data coding. It is based on the representation of functions systems of the logic algebra by means of orthogonal or linearly independent series. This approach makes it easier to obtain solutions of several important problems from the theory of designing digital information systems [4–8]. In order to do this systems of basic functions for unitary and Libaw-Craig-Lippel codes have been defined and analytical interdependence between them has been described. The presented results are obtained on the basis of the cross-cutting analysis of the main number-theoretic bases [1, 9].

2 Number-Theoretic Foundations of Bases and Coding Systems

The methodology of constructing mutually simple orthogonal functions is presented in detail using mathematical apparatus [5, 6, 8]. In order to determine the number-theoretic foundations for the basis of the most common code systems, their corresponding basis functions are investigated and the functional interdependence between different bases is established. This allows attaining high technical and economical characteristics in the hardware implementation of converters of the form of information. Below, the most widely used systems of functions, codes and coding systems will be analyzed.

In the theory and technique of digital signal processing, linear-convergent functions are widely applied. Analog signals are represented using non-harmonic basis of Legendre, harmonic basis is represented by Fourier series (with the sin- and cos-components) and harmonic-frequency basis [1, 5, 9, 10].

Creating, converting, transmitting and processing of digital signals are strictly connected with the use of discrete number-theoretical bases. Thus, it is important to analyze the effectiveness of their application and investigate properties of codes and code systems which use them.

The problem of analyzing the effectiveness of the relationship between various discrete bases is repeatedly considered by scientists in the field of digital signal processing and data conversion [1–5, 9]. Important results were obtained among others by Gold and Rader [11] for the discrete-harmonic basis, by Huang [12] for Hadamard basis and by Walsh [13] for multiplicative basis. However, the theoretical generalization of full relationships between different systems of functions was not given. The least is known about the Galois basis and coding systems that use it. The relationships between different bases and corresponding coding systems was not established as well, which prevented the full implementation and analytical transition between bases.

In this paper, the innovative work relies on the analysis of systems of discrete functions. Mathematical foundations, analytical dependencies, and definitions of key properties are determined. Based on them the classification of basis, the corresponding code systems and basis matrixes of number-theoretic transformations is made.

Systems of discrete linear-convergent functions are divided into three groups: discrete-harmonic, discrete-irregular and combined functions. The analysis of the number-theoretic basis starts from the first group as such systems of functions play a key role.

3 Discrete-Harmonic Systems of Functions

The following conditions will help formalize the transition from the real functions represented with weight $\rho(t)$ to the binary logic suitable for the synthesis of digital computing devices [14]:

- intervals of certainty of the basis functions [0, *T*], where $T = 2\pi$, in some specified cases [–*T*/2, *T*/2] may be considered;
- linear independence between the functions

$$\mathrm{B}(\tau) = \{\beta_0(\tau), \beta_1(\tau), \ldots, \beta_\nu(\tau), \ldots\};$$

- for non-normalized functions (e.g., Haar functions)

$$(h_i, h_j) = \sum_0^T h_i(t) \cdot h_j(t) \cdot \rho(t)\Delta t = \begin{cases} 0 & \text{for } n \neq m \\ 1 & \text{for } n = m \end{cases},$$

where $\rho(t)$ is the value of the weighting coefficient;

- for graphics, matrix and baseline presentation the following sign-rounding procedure will be realized

$$\operatorname{sign} \rho(t) = \begin{cases} 1 & \text{for } \ x(t) > 0 \\ -1 & \text{for } \ x(t) < 0 \\ 0 & \text{for } \ x(t) = 0 \end{cases}$$

this will result in functions in the normalized form.

Table 1 16-bit unitary code field example

N_i	u_{16}	u_{15}	u_{14}	u_{13}	u_{12}	u_{11}	u_{10}	u_9	u_8	u_7	u_6	u_5	u_4	u_3	u_2	u_1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
4	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
5	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
7	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
8	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
9	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
10	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
11	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
13	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
14	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Unitary codes, also called thermometric, are used as primary codes in converters of the form of information, particularly in parallel ADC architectures and unitary processors [1, 2]. In order to mathematically represent the unitary code, the unitary functions of the following form are described analytically:

$$\mathrm{Uni}(n, \theta, i) = \mathrm{sign}[\sin(2^n \pi\theta + 2^n(\pi/4)i)],$$

where $i = 0, 1, 2, \ldots, n$ is the number of functions, θ is a normalized time parameter ($\theta = t/T$, where t is the current time value), $n = \log_2 N$ is the order of the functions system of number-theoretic transformations and N is a module of integer basis values.

The functions of lower order n are obtained as a result of pair-wise multiplication of functions of higher n+1 orders according to the following equation

$$\mathrm{Uni}(n, \theta, i) = \mathrm{Uni}(n+1, \theta, i)\,\mathrm{Uni}(n+1, \theta, i + 2^{n-1}).$$

The graphic representation of the first n sets of unitary functions of the third order are shown in Fig. 1.

The main disadvantage of the unitary functions systems is their non-orthogonality, which indirectly causes non-compact packing of unitary code elements, which are presented in Table 1.

Unitary code is characterized by the code element capacity P of the code field:

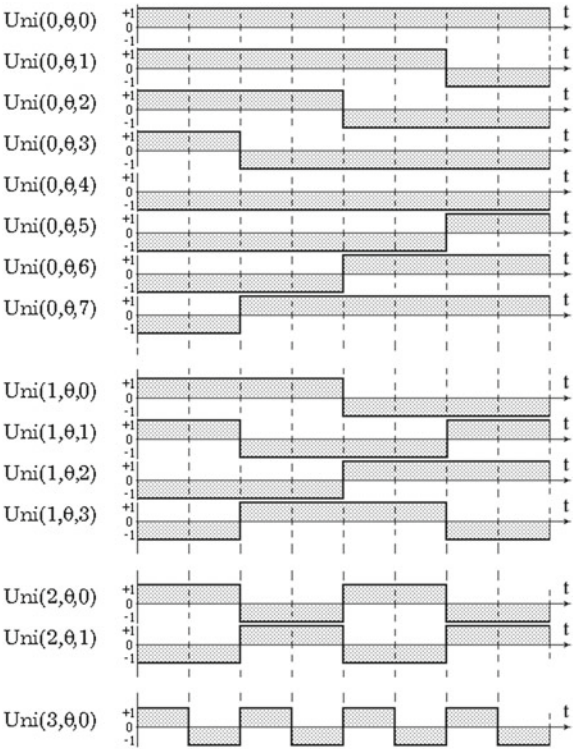


Fig. 1 Unitary functions of the third order

$$P = Nm,$$

where m is the number of bits of the codeword, N-modulo of coding system.

Because for the unitary coding method $N = m$, (from Fig. 1 and Table 1), then

$$P = N^2.$$

The unitary coding method requires the use of a data bus, whose number of bits m is equal to modulo N. This leads to a considerable redundancy of information streams, especially in remote data transmission. However, the practical implementation of unitary codes in the data converting devices is an expensive realization and has low reliability of technical implementation. There is a need for the reduction of the bits number of format data word $m = N$ to the lowest possible level according to Shannon relation $m = \log_2 N$. At the same time, there are examples of application of unitary codes in digital unitary correlation processors.

In the literature there are no reports on transformation between unitary functions system and other ones. This complicates the solution of the problem of data

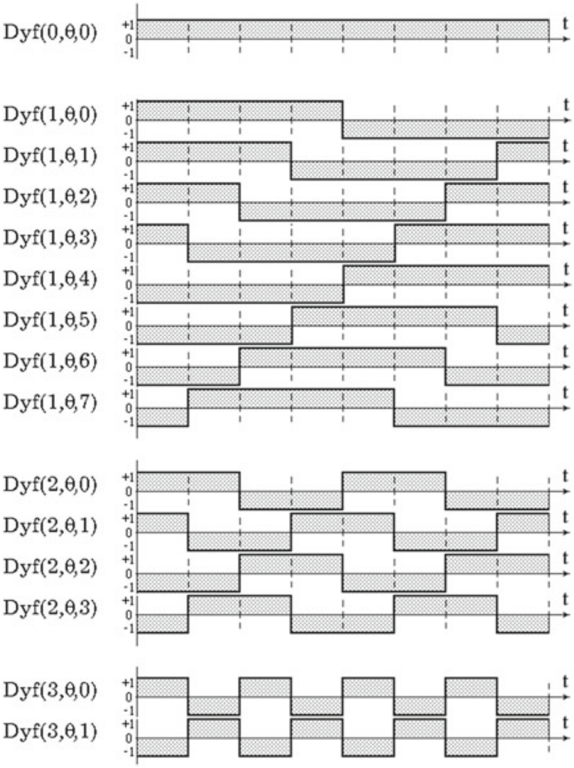


Fig. 2 System of discrete-phase functions

conversion between different codes and coding systems. To solve the transformation problem, the author introduces the system of discrete-phase functions, which is the basis for the creation of Libaw-Craig [15] and Lippel codes [16]. The system of discrete-phase functions is derived from the system of unitary functions by extending the definition of the period T of unitary functions twice.

Formally, the system of unitary functions of order n–1 defines its period 2π as a half of period π of the system of discrete-phase functions of order n. Thus, the even-numbered functions of n order from the system of discrete-phase functions are a direct reflection of functions of n–1 order from the system of unitary functions which can be expressed as follows

$$\mathrm{DF}(n, \theta, 2i) = \mathrm{Uni}(n-1, \theta, i).$$

The analysis of the graphical representation of functions system (Fig. 1) makes it obvious that the set of unitary functions of the 2nd order is a set of discrete-phase functions of the 2nd order in period $0 \div \pi$ (Fig. 2). The presentation of the system of discrete-phase functions in the dimension of higher order needs to increase the accuracy of transformation. This is achieved by introducing additional intermediate functions with the phase shifted by the half of the period of the functions of the highest order.

Discrete-phase functions of order n are described by the following generalized analytical expression

$$\mathrm{DF}(n, \theta, i) = \mathrm{sign}[\sin(2^n \pi\theta + 2^{n-1}(\pi/4)i)].$$

The system of discrete-phase functions in the period $T{=}2\pi$ reflects the phase shift $\Delta\theta$ of sign-procedures over the function $\sin 2\pi$ where

$$\Delta\theta = T/N = 2\pi/N = 2\pi/2^n \ = \pi/2^{n-1}.$$

The graphic representation of the first three sets of discrete-phase functions of order 1, 2, 3 is shown in Fig. 2. It should be noted that the functions of lower orders are the result of the pair-wise multiplication of the selected functions of higher orders

$$\mathrm{DF}(n, \theta, i) \ = \ \mathrm{DF}(n+1, \theta, i)\ \mathrm{DF}(n+1, \theta, i+2^{n-1}).$$

The proposed system of discrete phase functions is the basis for creating Libaw-Craig-Lippel codes for which the bits number of data format is equal $m = N/2$. Table 2 shows an example of forming the code field for Libaw-Craig-Lippel codes.

Since the Libaw-Craig-Lippel encoding method requires a data bus format, whose number of bits m is twice reduced respectively to modulo $N(M = N/2)$, then code element capacity P of the code field is equal:

$$P = Nm = NN/2 = N^2/2$$

Libaw-Craig-Lippel codes are practically applied in the positioning devices of linear and angular movements.

On the other hand, discrete-phase functions are the basis for the generation of the systems of Rademacher, Gray and Walsh functions [13, 17, 18]. To show this, it is necessary to make several discrete trigonometric transformations. Since the complete set of discrete-phase functions contains all the components of the phase shift of the function of the form $\mathrm{sign}[\sin(2\pi + \Delta\varphi)]$, then with phase shift $\Delta\varphi = \pi/2$ the function is transformed into function of the type of $\mathrm{sign}[\cos 2\pi]$.

A set of sin- and cos-components of the basis of discrete-phase functions (Fig. 3) is represented as:

$$\mathrm{DF}(0, \theta, 3) = \begin{cases} \mathrm{sign}\,[\sin\, 2\pi\theta] \\ \mathrm{sign}\,[\cos\, 2\pi\theta] \end{cases}$$

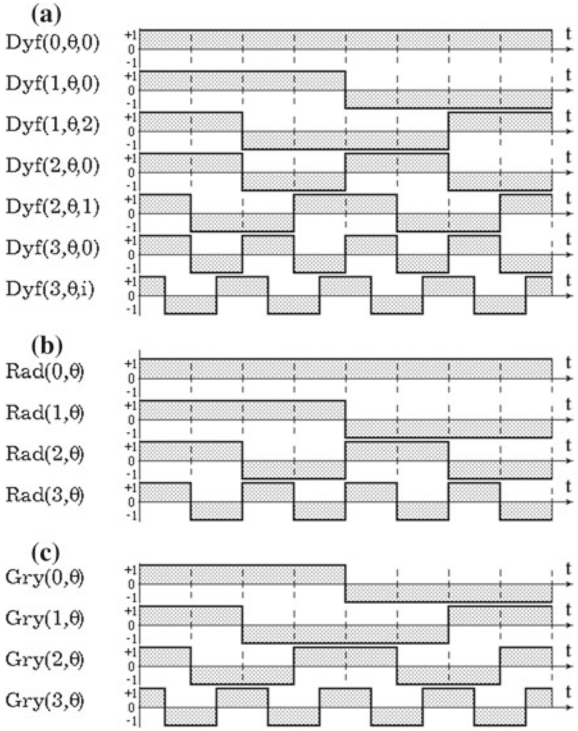


Fig. 3 Discrete-phase functions (**a**), sin- (**b**) and cos-components (**c**)

$$\begin{aligned}
\mathrm{DF}\,(2,\theta,3) &= \ \mathrm{sign}\,[\cos 2\pi\theta] \\
\mathrm{DF}\,(0,\theta,2) &= \ \mathrm{sign}\,[\sin 4\pi\theta] \\
\mathrm{DF}\,(1,\theta,2) &= \ \mathrm{sign}\,[\cos 4\pi\theta] \\
\mathrm{DF}(0,\theta,1) \ &= \ \mathrm{sign}\,[\sin 8\pi\theta]\,.
\end{aligned}$$

The sin-components from each of the systems of *n*-order functions form the basis of Rademacher, and cos-components the Gray basis:

$$\begin{aligned}
\mathrm{Rad}(n,\theta) &= \mathrm{DF}(n,\theta,0) = \mathrm{sign}\left[\sin 2^n\pi\theta\right], \\
\mathrm{Gry}(n,\theta) &= \mathrm{DF}(n,\theta,3) = \mathrm{sign}\left[\cos 2^n\pi\theta\right].
\end{aligned}$$

For example if n = 3, then

$$\begin{aligned}
\mathrm{Rad}(0,\theta) &= \mathrm{sign}[\sin\pi\theta] = \mathrm{DF}(0,\theta,0), \\
\mathrm{Gry}(0,\theta) &= \mathrm{sign}[\cos\pi\theta] = \mathrm{DF}(0,\theta,0), \\
\mathrm{Rad}(1,\theta) &= \mathrm{sign}[\sin 2\pi\theta] = \mathrm{DF}(1,\theta,0),
\end{aligned}$$

Table 2 16-bit Libaw-Craig-Lippel codes field example

N_i	d_8	d_7	d_6	d_5	d_4	d_3	d_2	d_1
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	1	1
3	0	0	0	0	0	1	1	1
4	0	0	0	0	1	1	1	1
5	0	0	0	1	1	1	1	1
6	0	0	1	1	1	1	1	1
7	0	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	0
10	1	1	1	1	1	1	0	0
11	1	1	1	1	1	0	0	0
12	1	1	1	1	0	0	0	0
13	1	1	1	0	0	0	0	0
14	1	1	0	0	0	0	0	0
15	1	0	0	0	0	0	0	0

$$\begin{aligned}
\mathrm{Gry}(1,\theta) &= \mathrm{sign}[\cos 2\pi\theta] = \mathrm{DF}(1,\theta,2),\\
\mathrm{Rad}(2,\theta) &= \mathrm{sign}[\sin 4\pi\theta] = \mathrm{DF}(2,\theta,0),\\
\mathrm{Gry}(2,\theta) &= \mathrm{sign}[\cos 4\pi\theta] = \mathrm{DF}(2,\theta,1),\\
\mathrm{Rad}(3,\theta) &= \mathrm{sign}[\sin 8\pi\theta] = \mathrm{DF}(3,\theta,0)\\
\mathrm{Gry}(3,\theta) &= \mathrm{sign}[\cos 8\pi\theta] = \mathrm{DF}(3,\theta,3).
\end{aligned}$$

The graphic representation of the first four Rademacher and Gray functions is shown in Fig. 3.

The set of Rademacher functions is the basis for the binary number system (Table 3a) and nowadays is used in most computer technologies and control systems [4, 5, 17].

The system of Gray functions, which is the basis for building the Gray code (Table 3b), code scales and converters of the information form, thanks to its unique properties has also a wide range of applications [17, 19].

Methods of binary and Gray coding used data bus format, whose number of bits is reduced to Shannon relation $m = \log_2 N$.

Binary and Gray codes are characterized by the following code element capacity P of the code field:

$$P = Nm = N\log_2 N.$$

Table 3 Examples of the binary (a) and Gray (b) codes with 4-bits data base format

N_i	2^0	2^1	2^2	2^3	N_i	g^0	g^1	g^2	g^3
0	0	0	0	0	0	1	1	1	0
1	1	0	0	0	1	0	1	1	0
2	0	1	0	0	2	0	0	1	0
3	1	1	0	0	3	1	0	1	0
4	0	0	1	0	4	1	0	0	0
5	1	0	1	0	5	0	0	0	0
6	0	1	1	0	6	0	1	0	0
7	1	1	1	0	7	1	1	0	0
8	0	0	0	1	8	1	1	0	1
9	1	0	0	1	9	0	1	0	1
10	0	1	0	1	10	0	0	0	1
11	1	1	0	1	11	1	0	0	1
12	0	0	1	1	12	1	0	1	1
13	1	0	1	1	13	0	0	1	1
14	0	1	1	1	14	0	1	1	1
15	1	1	1	1	15	1	1	1	1

The main disadvantage of the Rademacher-Gray basis is the low convergence of numerical series, which does not allow to effectively perform the information transformation and processing of high-frequency signals [17].

Rademacher and Gray bases are mutually complementary, since the Rademacher basis contains only the set odd functions ($\mathrm{f}(t) = -\mathrm{f}(-t)$), and the set of Gray functions contains only even functions ($\mathrm{f}(t) = \mathrm{f}(-t)$). Using Rademacher-Gray basis, which is shown in Fig. 3 in the form of extraction of discrete-harmonic sin- and cos-components, improves the presentation of the odd and even analyzed signals by discrete functions.

4 Systems of Discrete-Harmonic Piecewise-Defined Functions

A system of harmonic piecewise-defined functions is a subclass of discrete-harmonic basis. The basis of positional coding is a theoretical-numerical basis of the Haar functions, which is described by the following equation [2, 6, 10, 20]:

$$\mathrm{Har}\,(n, \theta, i) = \mathrm{sign}\left[\sin i 2^n \pi \theta\right].$$

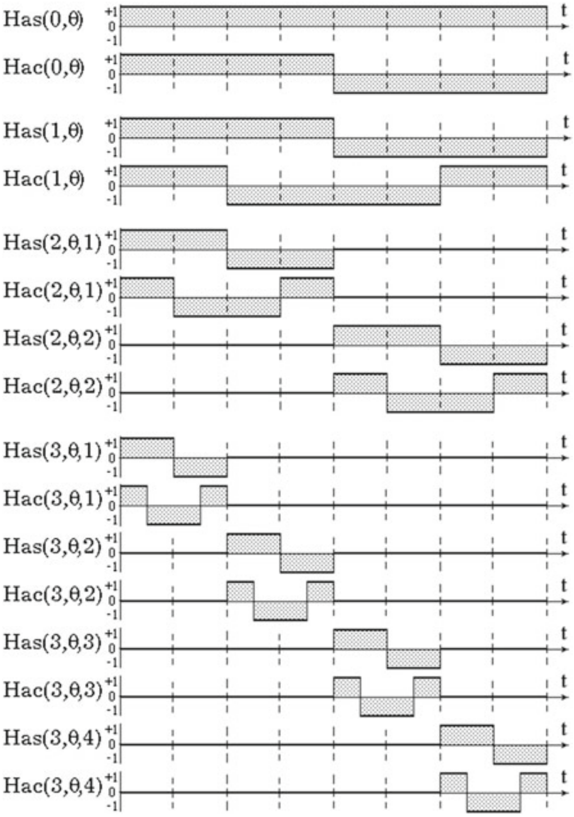


Fig. 4 System of sin- and cos-components of the Haar functions

It should be noted that the incompleteness of the aforementioned basis greatly limits the volume of its practical application and possibilities of mathematical mutual conversion to other systems of functions. The limitations of this basis are caused by the fact that it includes only sin-components (Has), which are described by odd functions $\mathrm{f}(-t) = -\mathrm{f}(-t)$. In order to extend the functionality of the Haar basis, it is proposed here to introduce the basis of cos-components (Hac) of Haar functions, which will allow to represent even functions $\mathrm{f}(-t) = \mathrm{f}(t)$. A complete Haar basis is described by the following sin-and cos-components set:

$$\mathrm{Har}\,(n, \theta, i) = \begin{cases} \mathrm{Has}(n, \theta, i) = \mathrm{sign}[\sin i 2^n \pi \theta] \\ \mathrm{Hac}(n, \theta, i) = \mathrm{sign}[\cos i 2^n \pi \theta] \end{cases},$$

where $i = 1, 2, \ldots, 2^n$.

Graphic representation of the complete set of Haar basis functions is shown in Fig. 4. It is obvious that the Haar basis is generated by the Rademacher-Gray basis by

Table 4 Example of a 16-bit position code

N_i	h_{16}	h_{15}	h_{14}	h_{13}	h_{12}	h_{11}	h_{10}	h_9	h_8	h_7	h_6	h_5	h_4	h_3	h_2	h_1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

masking the selected sin- and cos-components on the period of 2π of basis functions according to the order n to determine the required period $T = 2^{\pi+1}\pi$:

$$\mathrm{Har}\,(n, \theta, i) = \begin{cases} \mathrm{Has}(n, \theta, i) = \mathrm{Rad}(n, \theta) \cdot 2\pi i \\ \mathrm{Hac}(n, \theta, i) = \mathrm{Gry}(n, \theta) \cdot 2\pi i \end{cases}.$$

The Haar basis functions and positional codes that are generated by them have been widely used: in converters of the form of information, e.g., as intermediate analog-to-digital converters in the displacement sensors (linear and angular), to initialize the dispersed elements of computing systems, in particular of memory cells, etc. Table 4 shows an example of a 16-bit position code.

Since for positional code $n = N$ (Table 4), the code elements field capacity P is equal to a unitary code elements field capacity:

$$P = Nm = N^2.$$

5 Efficiency of Digital Data Presentation

The above facts can be summarized that each method has a certain value for the efficiency of digital data presentation and conversation. Table 5 shows the analytical

Table 5 Example of a 16-bit position code

Code and code systems	Data base ormat (bit)	Code elements field capacity (bit)
Unitary; Positional	$m = N$	$P = N^2 = m^2$
Libaw-Craig-Lippel	$m = N/2$	$P = NN/2 = N^2/2$
Binary; gray	$m = log_2N$	$P = Nlog_2N$

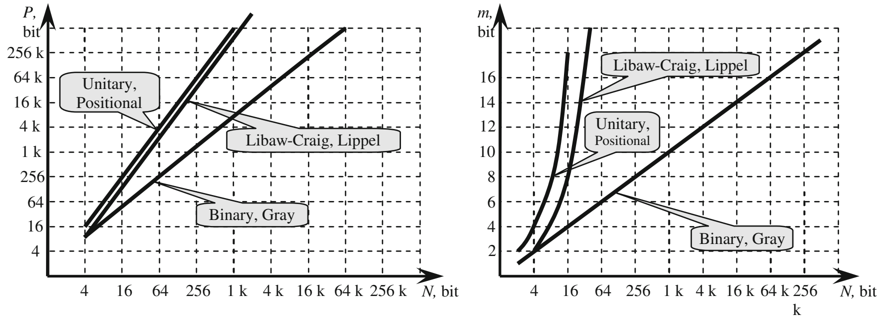


Fig. 5 Efficiency diagrams of digital coding and data transformation methods

dependencies for each of the above methods of information forms transformation on which a graph of (Fig. 5) is built.

The results allow to quantify the assessment of the efficiency of the considered coding methods. In the presented class of the parallel coding methods the binary and Gray coding techniques have minimal redundancy, which led to their widespread use in ICT.

These results have an intermediate stage of research on the coding systems effectiveness, which continues in the study of the combined information form transformation methods on the basis of Walsh functions and Galois recursive sequence.

6 Conclusion

This paper analyzes systems of discrete-harmonic functions, identifies analytical relationship between them and determines their corresponding coding systems. Systems of basic functions for unitary and Libaw-Craig-Lippel codes are defined; and analytical interdependence between them is described. Functional relationships between the discrete-phase functions and bases of Rademacher and Gray functions are defined. The results of the research on determining the mathematical interdependence between the function systems allow to systematize and analytically describe the procedure of the information form transformation in information systems. The efficiency of the analyzed digital data coding methods is assessed.

The perspectives of further research were defined; they will focus on the transition to a class of bases of discrete-irregular Walsh and Galois systems functions, which possess the best technical and economic characteristics. The use of these functions systems determines the direction of development of information form transformation and digital signal processing methods.

References

1. Varichenko, L.V., Labunec, V.G., Rakov, M.A.: Abstraktnye algebraicheskie sistemy i cifrovaya obrabotka signalov, -Kiev, Naukova Dumka, p. 248, (in Russian) (1986)
2. Zalmanzon, L.A.: Preobrazovaniya Fourier'a, Walsh'a, Haar'a i ih primenenie v upravlenii, svyazi i drugih oblastyah, -M., Nauka, p. 496, (in Russian) (1989)
3. Golubov, B.I., Efimov, A.V., Skvorcov, V.A.: Ryady i preobrazovaniya Walsh'a: Teoriya i primeneniya, Nauka, p. 343, (in Russian) (1987)
4. Proakis, J.G.: Digital Signal Processing: Principles, Algorithms, and Applications, p. 1156. Pearson Education, Upper Saddle River (2007)
5. Smith, S.W.: Digital Signal Processing: A Practical Guide for Engineers and Scientist, p. 650. Newnes, Boston (2003)
6. Haar, A.: Zur theorie der ortogonalen funktionsysteme. Math. Ann. **69**, 331–371 (1910)
7. Haar, A.: Zur theorie der ortogonalen funktionsysteme. Math. Ann. **71**, 38–53 (1912)
8. Paley, R.E.A.C.: A Remarkable Series of Ortogonal Funktions. Proc. London Math. Soc. **2**(34), 241–279 (1932)
9. Blahut, R.E.: Fast Algorithms for Digital Signal Processing, p. 441. Addison-Wesley Pub. Co., New York (1985)
10. Gonorovskii, I.S.: Radiotehnicheskie cepi i signaly, -M., Radio i svyaz, p. 512, (in Russian) (1986)
11. Gold, B., Rader, C.M.: Digital Processing of Signals, p. 269. McGraw-Hill, New York (1969)
12. Huang, T., et al.: Fast Algorithms For Digital Image Processing, p. 224. Radio i Svyaz, Moscow (1984)
13. Walsh, J.L.: A closed set of orthogonal functions. Amer. J. Math. **45**, 5–24 (1923)
14. Karpovskii, M.G., Moskalev, E.S.: Spektral'nye metody analiza i sinteza diskretnyh ustroistv. -L., Energiya, p. 144, (in Russian) (1973)
15. Libaw, W.H., Craig, L.J.: A photoelectric decimal-coded shaft digitizer, Trans. IRE on Electronic Computers, v. EC-2, 3 (1953)
16. Lippe1, B.: A decimal code for analog-to-digital conversion, Trans. IRE on Electronic Computers, v. EC-4, 4 (1954)
17. Rademacher, H.: Einige Satze von allgemeine Ortogonalfunktionen. Math. Annalen **87**, 122–138 (1922)
18. Petryshyn, L.B.: Teoretychni osnovy peretvorennya formy ta cyfrovoi obrobky informacii v bazysi Galois'a. -Kyiv, IZiMN MOU, p. 237, (in Ukrainian) (1997)
19. Gray, F.: Pulse code communication, March 17, 1953 (filed Nov. 1947). U.S. Patent 2,632,058 (1953)
20. Walsh, J.L.: A property of Haar's system of ortogonal functions. Math. Ann. **90**, 38–45 (1923)

The Opportunities and Challenges Connected with Implementation of the Big Data Concept

Janusz Wielki

Abstract This paper is devoted to the analysis of the Big Data phenomenon and the opportunities and challenges connected with it. It is composed of seven parts. In the first, a general overview of the situation related to the transformation of the economy from the industrial into the post-industrial one is given. In this context, the growing role of data and information as well as the rapid increase in the new socio-economical realities and the notion of Big Data are discussed. In the next section, the notion of Big Data is defined and the main sources of growth of data are characterized. In the following part of the paper the most significant opportunities and possibilities linked with Big Data are presented and discussed. The next part is devoted to the characterization of tools, techniques and the most useful data in the context of Big Data initiatives. In the following part of the paper the success factors of Big Data initiatives are analyzed. The penultimate part of the paper is focused on the analysis of the most important problems and challenges connected with Big Data. In the final part of the paper, the most significant conclusions and suggestions are offered.

1 Introduction

The Big Data phenomenon is attracting an increasing amount of attention of managers of contemporary organizations, as confirmed by the results of various pieces of research. According to the results of a survey of Fortune 1,000 C-level executives and executives responsible for Big Data conducted in 2013 by NewVantage Partners, 91 % of those surveyed stated that that their organization had a Big Data initiative in progress or planned [1]. Also the results of the global survey conducted in the same year by the EMC Corporation in fifty countries show that there is a significant

J. Wielki (✉)
Faculty of Economics and Management, Opole University of Technology,
Ul. Bielska 32/7, 45-401 Opole, Poland
e-mail: janusz@wielki.pl

M. Mach-Król et al. (eds.), *Advances in ICT for Business, Industry and Public Sector*,
Studies in Computational Intelligence 579, DOI 10.1007/978-3-319-11328-9_11

level of interest from organizations in the utilization of solutions belonging to this IT trend [2].

This fact is also confirmed by data concerning the value of the Big Data technology and services market. According to the research firm IDC, this market will grow at a rate of 27 % (year on year) and in 2017 it will reach 32.4 billion USD. It means that this market is expected to grow at a rate that is six time stronger than the whole IT market [3].

The fact that organizations are very interested in solutions relating to Big Data is also confirmed by the expectations of their managers concerning the level of investment for this purpose. According to the above mentioned results of the research conducted by NewVantage Partners, 68 % of those surveyed expected that their organizations would spend in 2013 over 1 million USD for these types of investments, with the percentage rising to 88 % by 2016 [1].

In the context of the above mentioned facts, there are two important questions that arise. Namely, what new possibilities does the Big Data phenomenon provide and what kind of opportunities and benefits can it bring organizations?

2 The Notion of Big Data and the Most Significant Sources Behind the Growth Of Data

Contemporary organizations are dealing with quickly and ever increasing amounts of data that are being generated in the socio-economic space. According to the estimations of Siegel, 2.5 quintillion bytes of data are added every day [4]. And it is emphasized that this amount doubles about every 40 months [5]. This is the result of the rapidly growing quantity of data that is generated not only by the organizations themselves but also in the organizations' business environments by both their stakeholders and other entities operating there.

But contrary to its name, the Big Data phenomenon does not solely relate to the issues connected with large amounts of data. The character of the data that organizations are dealing with has radically changed as well [6]. The data sets are becoming unstructured to a considerable degree, more and more granular, increasingly diverse, iterative, fast-changing, and available in real-time.

If the definition of the term Big Data is considered according to the Leadership Council for Information Advantage, this term is not precise "(…) it's a characterization of the never-ending accumulation of all kinds of data, most of it unstructured. It describes data sets that are growing exponentially and that are too large, too raw or too unstructured for analysis using relational database techniques" [7]. On the other hand, according to the NewVantage Partners "Big Data as a term used to describe data sets so large, so complex or that require such rapid processing (sometimes called the Volume/Variety/Velocity problem), that they become difficult or impossible to work with using standard database management or analytical tools" [8].

Generally, there have been four significant trends that have caused a considerable increase in data generation. They are [9]:

1. The growth in traditional transactional databases.
2. The increase of multimedia content.
3. The growth of the 'Internet of Things'.
4. The growing popularity of social media.

The growth in traditional transactional databases is chiefly connected with the fact that organizations are collecting data with greater granularity and frequency. This is due to reasons such as the increasing level of competition, increasing turbulence in the business environment and the growing expectations of customers. All of these factors necessitate organizations to react rapidly and with maximum flexibility to the changes taking place and adjust to them. In order to be able to do this, they are forced to conduct more and more detailed analysis concerning marketplaces, competition and the behavior of consumers.

The second trend is connected with the rapid increase in the use of multimedia in the industries of the contemporary economy. For example, in the health care sector over 95 % of the clinical data generated is in video format. More generally, multimedia data already accounts for over half of Internet backbone traffic [9].

The third trend which has caused a growth in the amount of data being generated is the development of the phenomenon called "The Internet of Things", where the number of physical objects or devices that communicate with each other without any human interference is increasing at a fast pace. As they are equipped with various sensors or actuators, they collect and send huge amounts of data [10, 11]. By 2015, the amount of data generated from the 'Internet of Things' will grow exponentially as the number of connected nodes deployed in the world is expected to grow at a rate of over 30 % per year [9]. It is worth mentioning that in this context more and more often the broader term is being used. It is the 'Internet of Everything', understood as "the networked connection of people, process, data, and things" [12].

Social media is the fourth extremely significant source of the increase of data. In the case of Facebook, in 2013 over 1.2 billion of its users generated huge amounts of data every day. They uploaded an average of 350 million of photos, sent 10 billion messages, and shared 4.75 billion items [13]. Twitter users sent 500 million Tweets per day, while in the case of YouTube 100 h of video were uploaded every minute [14, 15]. In addition, smart phones are playing an increasingly important role in social networks. Although the penetration of social networks is increasing for both PCs and smartphones, it is significantly higher for smartphones. If frequent users are considered in the case of PC's it is 11 % p.a. while in the case of smartphones it is 28 % p.a [9]. This has caused a rapid increase in mobile data traffic which doubled between the third quarter of 2011 and the third quarter of 2012. It is predicted that mobile data traffic will grow twelve fold by 2018 [16].

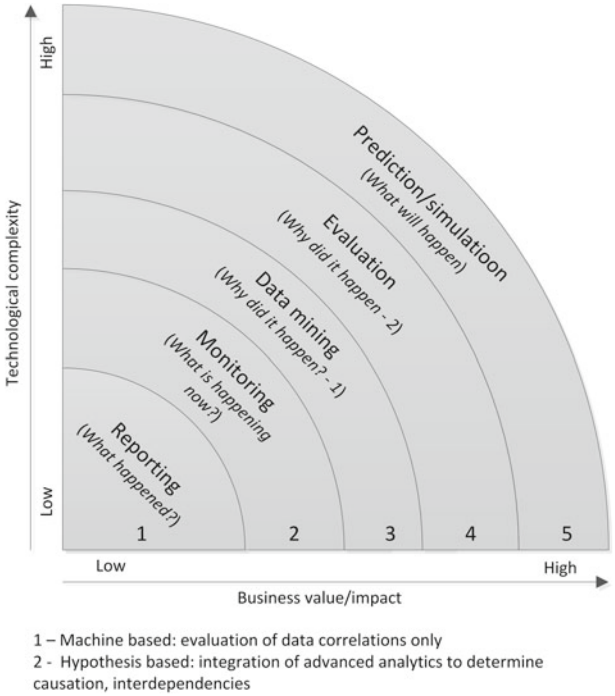


Fig. 1 Expending of the analytical possibilities connected with development of the Big Data systems (*Source* [18])

3 Opportunities and Benefits Connected with Big Data Utilization

The development of the Big Data phenomenon and its associated tools and techniques, is not something which has been separated from the wider processes which have been taking place in organizations over recent years. In fact, it is becoming more and more common in organizations concerned with the field of analytics and has significantly expanded the possibilities available within the scope of business intelligence (BI) tools [9, 17] (it especially relates to the areas 4 and 5—see Fig. 1).

Given their role in providing organizations with numerous possibilities and opportunities in the sphere of analytics, business intelligence systems are well suited for aggregating and analyzing structured data [19]. But there are, however, some types of analyses that BI can not handle. These mainly relate to situations where data sets become increasingly diverse, more granular, real-time and iterative.

Such types of unstructured, high volume, fast-changing data, pose problems when trying to apply traditional approaches based on relational database models. As a result, it has become apparent that there is growing demand for a new class of technologies and analytical methods [7].

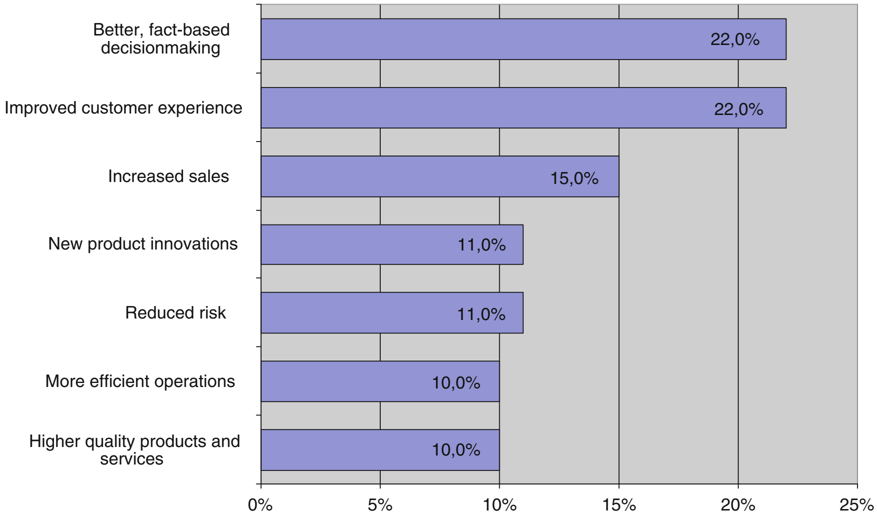


Fig. 2 Tangible benefits expected to be achieved through Big Data initiatives (*Source* [20])

According to the McKinsey Global Institute, there are five key ways in which the Big Data phenomenon creates value for organizations. They are [9]:

1. Creating transparency by integrating data and making it more easily accessible to all relevant stakeholders.
2. Enabling experimentation to discover needs, expose variability, and improve performance.
3. Segmenting populations in order to customize actions.
4. Replacing or supporting human decision making with automated algorithms.
5. Innovating new business models, products and services.

According to the results of a survey conducted in summer 2012 by NewVantage Partners, among C-level executives and function heads from many of America's leading companies, there are seven basic groups of benefits connected with Big Data initiatives [20] (see—Fig. 2).

Better, fact-based decision making and an improved customer experience are the most important of these benefits, coupled with the overall message that the expectation is to make better decisions faster. Organizations use Big Data platforms to give them answers to important questions in seconds rather than months. Thus, the key value of Big Data is to accelerate the time-to-answer period, allowing an increase in the pace of decision-making at both the operational and tactical levels [19–21].

Also the results of other surveys confirm that the expectations of executives connected with Big Data mainly relate to the issues connected with the improvement of the quality of their decision making and that they are counting on the possibility of faster, fact-based decision making [1, 2]. According to the results of the survey conducted by IDC, making better decisions and faster, are the two key goals in the context of the programs connected with Big Data utilization [3].

As it was mentioned earlier, an extremely important new element, in the context of decision-making, connected with the Big Data phenomenon, is the possibility for constant business experimentation to guide decisions and test new products, business models, and customer-oriented innovations. Such an approach even allows, in some cases, for decision making in real-time. There are many examples of companies using this in practice. For example multifunctional teams in Capital One perform over 65,000 tests each year. They experiment with combinations of market segments and new products [22]. On the other hand, eBay systematically improves its Internet-based products based on huge sets of behavioral data generated by users utilizing them at scale [23]. The online grocer FreshDirect is another example. It adjusts, on a daily basis or even more frequently, prices and promotions based on on-line data feeds [6]. On the other hand Netflix, the provider of on-demand Internet streaming media, captures quick feedback to learn what particular programs have the greatest audience, using this knowledge to adjust its offer [24].

As for Tesco this company gathers transaction data on its millions of customers through a loyalty card program and uses it to analyze new business opportunities. For example how to create the most effective promotions for specific customer segments and inform them about decisions concerning pricing, promotions, and shelf allocation [22]. Walmart is another example. This company created the Big Data platform (The Online Marketing Platform) which is used, among other things, to run many parallel experiments to test new data models [25]. Also such dot-com giants as Amazon, eBay and Google have been using testing in order to drive their performance [22].

All of these examples highlight that Big Data is turning the process of decision-making inside out. It means that instead of starting with a question or hypothesis, organizations use Big Data techniques and technologies to see what patterns they can find. When the discovered patterns provide them with a business opportunity or reveal a potential threat, organizations are then able to determine how to react [19].

There are many other examples of utilization of Big Data concerning various aspects of companies functioning. Ford intensively and successfully utilizes the Big Data phenomenon in its business activity and support of the decision-making processes. Thanks to founding, in various units, analytics centers of excellence, the company significantly improved its financial results. One of the basic fields of the utilization of analytics and the Big Data phenomenon is focused on customer preferences and making the best choices concerning the next new models of cars and the technologies used in them. In this context the company, among other issues, monitors and analyses social media streams and the signals that emerge in them. Another very important field where Ford uses business analytics and Big Data is the prediction of vehicle sales [26].

As in the case of Ford, LinkedIn, the biggest business networking Website, utilizes its own analytical group (Data Science) to develop new product features and functions. Some of them have had a significant impact on growth and persistence of the customer [6].

The utilization of Big Data in competitive intelligence systems has also become a very significant sphere in the support of decision-making. It relates to such issues as the usage of various tools for "catching" and analyzing signals (mentioned in the context of Ford), sometimes every week, generated in social media. This type of real-time information can be crucial in the context of preempting the actions of competitors or adjustments of strategy [27]. Such signals can be also used in the context of making decisions such as when to launch new products or services or help solving customers' problems. TomTom, a manufacturer of automotive navigation systems, is an example of a company functioning in this way. It "catches" and mines signals coming from the on-line driving community stream for ideas on design and product features and to quickly troubleshoot new offerings [28].

As far as marketing activities are concerned, apart from the above mentioned competitive intelligence, there are many fields where the Big Data phenomenon has been utilized. One of the most important of these fields is targeted advertising i.e. reaching a consumer with precise, individualized offers prepared on the basis of knowledge of him or her, generated based on the analysis of various data sets, such as behavioral ones. The Target Corporation, which is one of the biggest American retail companies, is an example of an organization which is very successful in this field. The company very strongly developed its usage of business analytics which had a significant impact on its revenues. It became famous for the fact that their business analytics were able to build a model that, based on the shopping habits of their clients, allowed the company to predict, with a high degree of probability, which of their customers was pregnant and her delivery date. Linking this information with other data (client ID, name, credit card number or e-mail address) enables the company to reach expectant mothers with precise offers for baby items [29, 30].

Activities from the field of targeted advertising are strictly connected with another quickly developing trend whereby marketers use the tools associated with the Big Data phenomenon: marketing automation. This refers to dynamic and automated actions, based on the collection and analysis of any available data and information about individual customers and their behavior which allows an organization to predict the best offer for the moment of interaction [31]. Audioteka is an example of an organization which implements such kind of solutions. The company is the biggest Internet-based service in Poland, dealing with the sale and distribution of audiobooks. Its utilization of Big Data allows it to create dynamic segments of clients, providing it with the possibility of the automation of additional sale and promotion actions aimed at particular segments of customers. It is possible, as a result of combining and analyzing data concerning the behavior of consumers visiting the Audioteka web site with data about the purchase of audiobooks [32]. Google also broadly uses business analytics of huge, unstructured data sets for the automation and individualization of advertising offers. In the case of Gmail, the company uses the automated scanning and processing of sent or received email content [33, 34].

Human resources is a further example of a functional field where the Big Data phenomenon is being utilized, and in this context the term "people analytics" has emerged. This means the fast-growing practice of using large amounts of various data sets generated by employees and business analytics to quantify those already employed and to assist in the selection of new ones [35, 36]. Google is one of the leaders in this field and has completely rebuilt its HR system as a result [35, 37, 38].

Generally, the results of the research conducted in February 2012 among 607 executives from around the world by the Economist Intelligence Unit confirm the value of Big Data utilization by companies. The surveyed executives claim that Big Data initiatives have improved the performance of their organizations over the past three years by around 26 %. Simultaneously they expect that such initiatives will improve performance by an average of 41 % over the next three years [19]. In addition, it is worth noticing that according to the results of the research of Brynjolfsson et al. firms where decision making is based on data and business analytics have 5–6 % higher output and productivity. Decision making based on data and business analytics also impacts on other performance measures such as asset utilization, equity return and market value [39].

As in the case of BI initiatives [40] Big Data systems have been used for two purposes—human decision support and decision automation. According to the results of the above mentioned research conducted by the Economist Intelligence Unit, Big Data is used, on average, for decision support 58 % of the time and for decision automation around 29 % of the time, based on the level of risk connected with the decision [19].

4 Techniques, Tools and the Most Useful Data in the Context of Big Data Initiatives

The effective implementation of Big Data initiatives requires an undertaking of appropriate organizational actions, including ensuring organizations are provided with all the necessary resources to enable analysis of the ever-growing data sets to which they have access. In this context, the application of proper techniques and technologies is one of the key issues. In practice, organizations use many various techniques and technologies to aggregate, manipulate, analyze, and visualize Big Data. They come from various fields such as statistics, computer science, applied mathematics, and economics. Some of them have been developed intentionally and some of them have been adapted for this purpose. Examples of techniques utilized for the analysis of Big Data are: A/B testing, data fusion and data integration, data mining, machine learning, predictive modeling, sentiment analysis, spatial analysis, simulation or time series analysis.

Examples of technologies used to aggregate, manipulate, manage, and analyze of Big Data are: Big Table, Cassandra, Google File System, Hadoop, Hbase, MapReduce, stream processing, visualization (tag cloud, clustergram, history flow, spatial information flow) [9].

Increasingly, there are a number of new analytical toolkits for the analysis of Big Data. Examples of such solutions are [27]:

- Alterian, TweetReach (network intelligence tools for real-time analysis of the reactions and responses to changes of industry players),
- NM Incite, Social Mention, SocMetrics, Traackr, Tweepi (sentiment analysis tools for estimating the buzz around a product or service, influencer intelligence tools for identifying key influencers and targeting for marketing or insights),
- Attensity, Autonomy (live testing tools for getting direct feedback from users on new products or ideas, data mining tools for text-analytics to estimate market size).

In addition, a very important element of Big Data initiatives is properly trained people. In this context, a specific type of worker is indicated, known as data scientists, who are properly trained to work with Big Data. In practice, it means that they should be people who know how to discover the answers to an organization's key questions from huge collections of unstructured data (it is estimated that more than 80 % of the world's data is unstructured or semi structured [17]). These people should be a hybrid of analyst, data hacker, communicator and trusted advisor [41]. In addition to analytical abilities and substantial and creative IT skills, they should be close to the products and processes inside the organization [42]. As the acquisition of in-depth domain knowledge from data scientists typically takes years [43], most organizations build platforms to close the gap between the people who make decisions and data scientists, such as that created by Walmart—the Social Genome Platform. It facilitates cooperation among buyers, merchandisers, product managers and other people who have worked in retail for years and data scientists [44].

In addition to proper techniques, tools and people, the basic resource required for Big Data initiatives is appropriate data. As was mentioned earlier, a lot of data from various sources is currently flowing into contemporary organizations but not all Big Data sets are equally valuable [18, 45] (see Fig. 3). Business activity data such as sales, purchases, costs etc. is definitely the most important source of data. Office documentation is the second key source of data, closely followed by social media. In certain sectors such as healthcare, pharmaceutical, and biotechnology, data sets from social media are more important that those from office documentation [19]. Generally in the context of Big Data initiatives, the data used should be "smart data". That is, it should be *accurate* (it must be precise enough), *actionable* (it must drive an immediate scalable action) and *agile* (it must be flexible) [46].

5 Success Factors of Big Data Initiatives

Through an analysis of implemented Big Data initiatives, various success factors can be determined, each with their own set of recommendations. Marchand and Peppard have identified five important guidelines for the success of a Big Data project. They include [47]:

1. Placing people at the heart of the Big Data initiative.
2. Emphasizing information utilization as the way to unlock value from information technology.

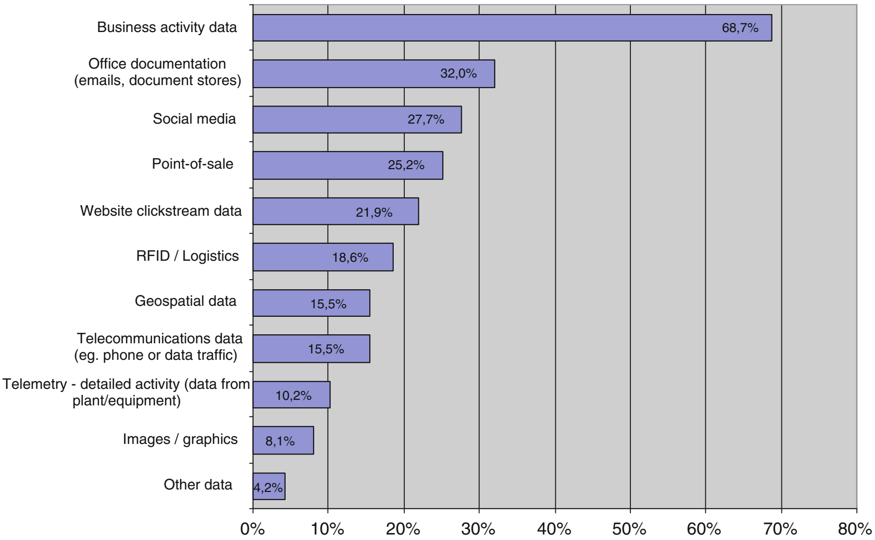


Fig. 3 The Big Data sets which add the most value to organizations (*Source* [19])

3. Equipping IT project teams with cognitive and behavioral scientists.
4. Focusing on learning.
5. Worrying more about solving business problems than about deploying technology.

Based on their experiences gained from cooperation with companies from data rich industries, Barton and Court, on the other hand, came to the conclusion that full exploitation of data and analytics requires three capabilities [48]:

1. Choosing the right data.
   In this context two aspects are important:
   - creative sourcing of internal and external data,
   - upgrading IT architecture and infrastructure for easy data merging.
2. Emphasizing information utilization as the way to unlock value from information technology.
   In this context two aspects are important:
   - focusing on the biggest drivers of performance,
   - building models that balance complexity with ease of use.
3. Equipping IT project teams with cognitive and behavioral scientists.
   In this context two aspects are important:
   - creating simple, understandable tools for people on the front lines,
   - updating business processes and developing capabilities to enable tools utilization.

According to Biesdorf et al. any successful Big Data plan should focus on three core elements. They include: choosing the right data (both internal and external ones) and integrating them, selecting advanced analytical models and tools. There is one element which is considered as a critical enabler of the whole Big Data initiative. This is the organizational capabilities necessary to exploit the potential connected with Big Data [49].

According to Barth et al. organizations that benefit from Big Data base their activities on three fundamental issues [42]:

1. Paying attention to data flow as opposed to stocks.
2. Relying on data scientists and product and process developers rather than data analysts.
3. Moving analytics away from the IT function, into core business, with operational and production functions.

6 Challenges and Implications Connected with Big Data Utilization

As in the case of other IT-related initiatives, Big Data also has its own set of problems and challenges. They include difficulties of a technical, organizational and legal nature. Those in the first category are connected with such issues as the selection of the most useful techniques, technologies and tools or the creation of an efficiently functioning system for feeding the most relevant data (internal and external) to support the organizational goals.

In the case of organizational challenges, these are connected with such issues as the selection of the proper people (as it was mentioned earlier the term "data scientists" is most commonly used), updating business processes and developing capabilities to enable Big Data tools utilization or changes in organizational culture. In the context of this last challenge the key issue is to make decisions as data-driven as possible, instead of basing them on hunches and instinct [5].

If the techno-organizational challenges are considered there are some obstacles indicated by the executives as the most important impediments to the effective utilization of Big Data for decision-making (see Fig. 4).

"Organizational silos" were the most significant barrier, which result from the fact that data connected with particular organizational functions (i.e. sales, distribution, accounts receivable etc.) are collected in "function silos" rather than pooled for the benefit of the entire company. The second, although no less important, issue is the lack of appropriately skilled people (data scientists) prepared to analyze data. The third aspect is the excessively long time it takes organizations to analyze huge data sets. As was mentioned earlier, organizations expect to be able to analyze and act on data in real time. The fourth barrier is the difficulties concerned with the analysis of ever increasing amounts of unstructured data. Finally, the inability of senior management

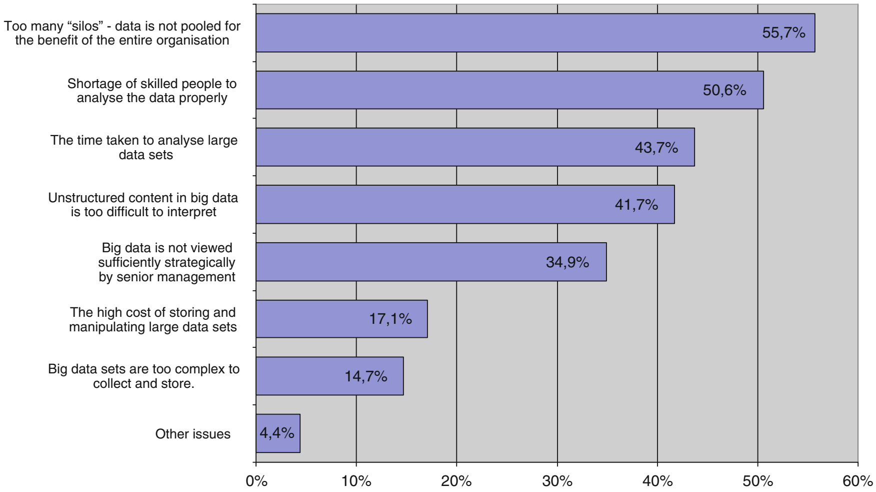


Fig. 4 The biggest impediments to using Big Data for effective decision-making (*Source* [19])

to view Big Data in a sufficiently strategic way is the fifth key impediment [19]. Similar results were also obtained in other surveys (see [50]).

As it was mentioned earlier there are also challenges of legal nature. They relate to such issues as copyright, database rights, confidentiality, trademarks, contract law, competition law [51]. An additional legal challenge worth mentioning is the risk associated with the utilization of Big Data to increase the automation of decision-making.

There is one more important danger which is underlined in the context of Big Data. It is connected with the fact that Big Data might not be providing the whole picture for a particular situation. There are several reasons for this i.e. biases in data collection, exclusions or gaps in data signals or the constant need for context in conclusions [52].

At the same time, existing pre-Big Data challenges and threats are still developing, such as the problem of securing collected data and information [53]. These issues chiefly relate to the problem of how to protect competitively sensitive data and data that should be kept private by organizations (e.g. various types of consumer data) [9]. As a result, the problems connected with the broadly defined security of the IT infrastructure of organizations and protection against various attacks becomes an even more important issue than previously [54]. The increasing dependence, as a result of the Big Data phenomenon, of organizations on the efficient and reliable functioning of their IT infrastructure, means that securing it has become even more important.

Apart from the above mentioned challenges of a technical, organizational and legal nature, the development of the Big Data phenomenon and its utilization is connected with numerous socio-ethical disputes. It is import for companies to find and maintain a balance between the benefits and tremendous opportunities connected with Big Data and the potential risks and challenges. This relates to both established challenges which have been deepening and to emerging new ones. Organizations have to remember that technologies connected with Big Data, as with any other technology, are ethically neutral, but their utilization is not [43].

According to Davis and Patterson, there are four elements which organizations should address in the context of the challenges connected with Big Data ethics. They include [43]:

- identity,
- privacy,
- ownership,
- reputation.

If the first issue is considered, it is argued that the identity of people is multifaceted (according to Poole, identity is "prismatic"), which means that it is hard to be aggregated for a consumption by a single organization. Big Data technologies change this situation. They provide an organization with the ability to summarize, aggregate or correlate various aspects of the identity of every consumer without his or her participation or agreement and to undertake appropriate or inappropriate action [43].

The above mentioned activities of Target are an example of the ethical challenges related to this issue. Based on the analysis of the shopping habits of their clients, the company is able to predict, with high probability (about 90 %), which of their clients is pregnant and the delivery date. Combining this information with other information, the company can reach expectant mothers with precise offers (in this case coupons for baby items). But due to a situation where Target knew about a teenage girl's pregnancy before her father, made the company aware that in spite of benefits of business analytics usage, there were also significant risks connected with it. They relate to questions around how Target was able to identify its pregnant customers and the problem of potential accusations from clients that the company was spying on them [29, 30]. This last issue can be important because of the fact that companies who are providers of Big Data analytics solutions for physical storefronts (such as RetailNext) collect shopper data from a variety of sources such as POS systems, RFID tags, and also surveillance video [55].

In the context of identity there are some issues which are extremely important. They include: personally identifying information, the anonymization of data sets, and reidentification or cracking back [43]. These issues have emerged in the context of activities of companies such as Facebook, MySpace and several other social-networking sites and were identified a few years ago. Namely, they have been sending data to advertising companies that could be used to find consumers' names and other personal details [54]. These issues are also very important in the context of the business activities of enterprises called "data brokers". These companies are collecting, analyzing and packaging the most sensitive personal information, i.e. that which is individually identified, and selling it to each other or to advertisers [56].

The second challenge in the context of Big Data ethics is privacy. This aspect is repeatedly a matter of interest and discussions in the context of the utilization of information technology and the range of challenges connected with this issue have significantly increased with the entrance of the Internet into the contemporary socio-economic reality. The issues connected with privacy are extremely important because of the fact that personal data has become a "new asset" in the contemporary economic reality, and as the above mentioned "data brokers" treat them as a "commodity" [56, 57].

There are numerous questions and issues that arise in this context, but the most important ones relate to the problem of whether individuals should have the ability to control data about them and to what extent [43]. These issues are extremely important in the context of the activities of firms utilizing business analytics such as Google, Facebook, but also for "data brokers" (e.g. Axciom).

In the case of Google, it relates to activities such as the above mentioned automated scanning and processing of e-mail content for emails that are sent or received by Gmail users. Particular reservations are connected with the fact that the company also applies the scanning to e-mails sent to Gmail users from other e-mailing systems by people who haven't accepted its privacy policy [33, 34]. Another example connected with Google relates to its bypassing of the privacy settings of millions of people using Safari Web browser on their iPhones and computers. This allowed Google to track the Web-browsing habits of users who had knowingly blocked this kind of monitoring [58].

In the case of Facebook, the company is currently being sued over claims that it mines users' private messages. Users believed that they had a private and secure mechanism for communication with their acquaintances and friends and they felt the company had breached that princple [59].

As far as Axciom, the largest "data broker", is concerned, this company recently launched a Website which gives consumers some control over the data collected about them. Namely, they can review, edit and restrict the scope of distribution of their personal data collected by the company [56, 60].

The third challenge in the context of Big Data ethics is ownership. This issue is connected with answers to such questions as [43]:

- who owns data?
- can rights to data can be transferred?
- what are the obligations of companies who use data?

The issues connected with ownership were previously a subject of challenges several years ago in the context of early dot-com retailers (see the case of HealthCentral .com—[61]), and their significance grows in the Big Data era. The issues are connected with such situations as the above mentioned case of Facebook who mined private messages of its users to advertisers or Barclays who started selling information on their 13 million clients' spending habits to other companies. In this last case, the bank has stated that location data derived from any mobile device used by their customers can be collected as well, enabling the bank to pinpoint where in the world a customer is at any particular moment in time) [59, 62]. The issues relating to ownership are particularly important from the point of view of such sensitive sectors as health care, which is strongly involved in Big Data utilization [9, 18]. They are also significant in the context of the activities of such companies as the above mentioned "data brokers" [56].

Reputation is the fourth challenge connected with Big Data ethics. In this context it is necessary to note that issues related to this aspect are on one hand connected with companies implementing programs relating to the utilization of the Big Data phenomenon, while on the other hand with the reputation of their clients. The above mentioned activities of Google (users of Gmail or Safari Web browser) or Facebook (mining users' private messages) influence their reputation and image in a decidedly negative way. In the case of banks (Barclays) reputation is a key issue, so an announcement about the selling of information about clients' spending habits to other companies will not reflect positively on them. As far as Target is concerned, in its case the duality of influence of Big Data phenomenon can be clearly noticed. On the one hand suspicions of spying on clients can lead to a public-relations disaster, and consequently influence the financial results of the company. On the other hand, there are issues connected with the impact on the reputation of its clients. In this case the teenager who was a recipient for advertisements for baby items from Target, before she had made her situation widely known.

In the context of reputation it is worth mentioning that there is one more significant challenge connected with the utilization of data in Big Data systems. This aspect relates to the issue of how to determine what data is trustworthy. It is an important challenge because of the fact that the amount of data and the ways organizations can interact with them is increasing exponentially [43].

7 Conclusions

The rapidly increasing amounts and diverse sources of data and information which organizations have at their disposal are connected with a growing number of possibilities which offer rapidly developing analytical tools that are having an increasingly significant impact on their functioning and competitive advantage. Because of this, an increasing number of organizations have been implementing Big Data initiatives in order to utilize opportunities connected with business analytics usage. They include such issues as enabling experimentation to discover needs, expose variability and improve performance, better segmenting populations in order to customize actions, replacing or supporting human decision making with automated algorithms or innovating new business models, products and services.

But if initiatives aimed at the practical usage of Big Data sets are to be successful at giving an organization a competitive advantage and be of value, it is not enough to just collect and own the appropriate data sets. In fact, this is only the starting point of every Big Data initiative. Further essential elements are suitable analytical models, tools, skilled people, and organizational capabilities. Lack of all of these necessary components can lead to a situation whereby instead of expected benefits there is only disappointment and a belief that Big Data initiatives are only the next wave in a long line of management fads.

In the context of such initiatives, a number of challenges, typical for IT-related projects, emerge (technical, organizational, legal etc). But in the case of the projects connected with Big Data, one sphere is at the forefront. This is the sphere of ethical challenges. It is connected with the fact that Big Data creates, compared to other IT-related initiatives, a much broader set of ethical challenges and concerns. They especially relate to a broadly understood privacy, which is the issue which is more and more commonly publicly noticed and discussed [56].

So if organizations want to be able to utilize emerging opportunities connected with the Big Data phenomenon, these issues must be especially carefully addressed by them. It is particularly important that every company utilizing Big Data implements internal practices which will reinforce proper data management [Brown, 2014], including establishing ethical decision points which should assure the existence of a proper relationship between the values held by organizations and aligning those values with the actions undertaken by them [43].

Generally, although Big Data solutions have a huge potential for both commercial organizations and governments, there is uncertainty concerning the speed with which they can be utilized in a secure and useful way [63].

References

1. NewVantage Partners. New vantage big data executive survey 2013: Business adoption backs up the big data buzz. http://newvantage.com/wp-content/uploads/2013/09/NVP-Big-Data-Press-Release-090913.pdf (2013). Accessed 21 Oct 2013
2. EMC Corporation. Even though 79% think big data will improve decision making, one-third have no big data plans. http://www.emc.com/about/news/press/2013/20131212-01.htm (2013). Accessed 15 Dec 2013
3. Bednarz, A.: How today's enterprises use Big Data. Network World. http://www.networkworld.com/news/2014/020514-big-data-278471.html (2014). Accessed 05 Feb 2014
4. Hayashi, A.: Thriving in a big data world. Sloan Management Review. http://sloanreview.mit.edu/article/thriving-in-a-big-data-world (2012). Accessed 20 Dec 2012
5. Brynjolfsson, E., McAfee, A.: Big data: The management revolution. Harv. Bus. Rev. **90**, 60–68 (2012)
6. Davenport, T., Kim, J.: Keeping Up with the Quants. Harvard Business School Press, Boston (2013)
7. Leadership Council for Information Advantage. Big data: Big Opportunities to create business value. http://poland.emc.com/microsites/cio/articles/big-data-big-opportunities/LCIA-Bigdata-Opportunities-Value.pdf (2011). Accessed 23 Mar 2012
8. NewVantage Partners. Big data executive survey: Themes & trends. http://newvantage.com/wp-content/uploads/2012/12/NVP-Big-Data-Survey-themes-trends.pdf (2012). Accessed 30 Dec 2012
9. McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/mgi/publications/big_data/pdfs/MGI_big_data_full_report.pdf (2011). Accessed 21 Jun 2011
10. Chui, M., et al.: The internet of things. McKinsey Quarterly. https://www.mckinseyquarterly.com/article_print.aspx?L2=4&L3=116&ar=2538 (2010). Accessed 17 Apr 2010
11. Commission of the European communities. Internet of things-An action plan for Europe. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0278:FIN:EN:PDF (2009). Accessed 19 Mar 2013

12. Bradley, J., et al.: Internet of everything: A 4.6 trillion public-sector opportunity. http://internetofeverything.cisco.com/sites/default/files/docs/en/ioe_publicsector_vas_white%20paper_121913final.pdf (2013). Accessed 12 Jan 2014
13. Smith, C.: By the numbers 75 amazing facebook statistics. http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/ (2014). Accessed 16 Jan 2014
14. Twitter. About. https://about.twitter.com/company/ (2014). Accessed 16 Feb 2014
15. YouTube. Statistics. http://www.youtube.com/yt/press/statistics.html (2014). Accessed 17 Feb 2014
16. Ericsson. Ericsson Mobility Report. http://hugin.info/1061/R/1659597/537300.pdf (2012). Accessed 03 Dec 2012
17. Lampitt, A.: Hadoop: Analysis at massive scale. InfoWorld. http://resources.idgenterprise.com/original/AST-0084522_IW_Big_Data_rerun_1_all_sm.pdf (2013). Accessed 04 Apr 2013
18. Groves, P., et al.: The_big_data_revolution_in_healthcare: Accelerating value and innovation. McKinsey Quarterly. http://www.mckinsey.com/insights/health_systems/ /media/7764A72F70184C8EA88D 805092D72D58.ashx (2013). Accessed 06 Feb 2013
19. Capgemini. The deciding factor: Big data & decision making. http://www.capgemini.com/insights-and-resources/by-publication/the-deciding-factor-big-data-decision-making/?d=6C800B16-E3AB-BC55-00F4-5411F5DC6A8C (2012). Accessed 09 Mar 2012
20. NewVantage Partners. Big data executive survey: Creating a big data environment to accelerate business value. http://newvantage.com/wp-content/uploads/2012/12/NVP-Big-Data-Survey-Accelerate-Business-Value.pdf (2012). Accessed 30 Dec 2012
21. Olavsrud, T.: How to use big data to make faster and better business decisions. Computerworld. http://www.computerworld.com/s/article/print/9235604/How_to_Use_Big_Data_to_Make_Faster_and_Better_Business_Decisions?taxonomyNa (2013). Accessed 11 Jan 2013
22. Bughin, J., et al.: Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. McKinsey Quarterly. https://www.mckinseyquarterly.com/article_print.aspx?L2=20&L3=75&ar=2647 (2010). Accessed 01 Sep 2010
23. Ferguson, R., et al.: From value to vision: Reimaging the possible with data analytics. Research Report. Sloan Management Review. http://bf6837e6b8471f0219c29a1c5d9301ad213e0e491b46094b7800.r64.cf2.rackcdn.com/MITSMR-SAS-Data-AnalyticsReport-2013.pdf (2013). Accessed 01 Sep 2013
24. Rosenzweig, P.: The benefits-and limits-of decision models. McKinsey Quarterly. http://www.mckinsey.com/insights/strategy/the_benefits_and_limits_of_decision_models (2014)
25. Walmartlabs. Big data platform and demand generation. http://www.walmartlabs.com/platform/ (2013). Accessed 30 Apr 2013
26. King, J.: How analytics helped Ford turn its fortunes. Computerworld. http://www.computerworld.com/s/article/9244363/How_analytics_helped_Ford_turn_its_fortunes (2013). Accessed 02 Dec 2013
27. Harrysson, M., et al.: How 'social intelligence' can guide decisions. McKinsey Quarterly. https://www.mckinseyquarterly.com/article_print.aspx?L2=21&L3=37&ar=3031 (2012). Accessed 03 Dec 2012
28. Harrysson, M., et al.: The strength of 'weak signals'. McKinsey Quarterly, http://www.mckinsey.com/Insights/High_Tech_Telecoms_Internet/The_strength_of_weak_signals (2014). Accessed 06 March 2014
29. Duhigg, C., How companies learn your secrets. The New York Times. http://www.nytimes.com/2012/02/19/magazine/shoppinghabits.html?pagewanted=all_r=0 (2012). Accessed 20 Feb 2012
30. Hill, K.: How target figured out a teen girl was pregnant before her father did. Forbes. http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/ (2012). Accessed 20 Feb 2012
31. Marketo. The definitive guide to marketing automation. http://www.marketo.com/_assets/uploads/The-Definitive-Guide-to-Marketing-Automation.pdf (2013). Accessed 29 Dec 2013
32. Kpka, M.: Audioteka: Marketing Automation wybrany do obsługi tzw. Big Data. http://www.internetstandard.pl/news/388687/Audioteka. Marketing. Automation. wybrany.do.obslugi.tzw.Big.Data.html (2013). Accessed 27 Feb 2013

33. Kirk, J.: Google’s Gmail scanning unclear to users, judge finds. Computerworld. http://www.computerworld.com/s/article/9242748/Google_s_Gmail_scanning_unclear_to_users_judge_finds (2013). Accessed 27 Sep 2013
34. Rushe, D.: Google: don’t expect privacy when sending to Gmail. The Guardian. http://www.theguardian.com/technology/2013/aug/14/google-gmail-users-privacy-email-lawsuit/ (2013). Accessed 15 August 2013
35. Cutter, A.: People analytics: Big data hits hiring. ClickZ. http://www.clickz.com/print_article/clickz/column/2325600/-people-analytics-big-data-hits-hiring (2014). Accessed 20 Jan 2014
36. Peck, D.: They’re watching you at work. The Atlantic. http://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/ (2013). Accessed 20 Nov 2013
37. Garvin, D.: How Google sold its engineers on management. Harv. Bus. Rev. **91**, 74–82 (2013)
38. Sullivan, J.: How Google is using people analytics to completely reinvent HR. TLNT. http://www.tlnt.com/2013/02/26/how-google-is-using-people-analytics-to-completely-reinvent-hr/ (2013). Accessed 26 Feb 2013
39. Brynjolfsson, E., et al.: Strength in numbers: How does data-driven decisionmaking affect firm performance?. http://ssrn.com/abstract=1819486 (2011). Accessed 03 Feb 2012
40. Davenport, T., Harris, J.: Competing on Analytics. Harvard Business School Press, Boston (2007)
41. Davenport, T., Patil, D.: Data scientist: The sexiest job of the 21st century. Harv. Bus. Rev. **90**, 70–76 (2012)
42. Barth, P., et al.: How big data is different. Sloan Management Review. Fall, pp. 21–24 (2012).
43. Davis, K., Patterson, D.: Ethics of Big Data. O’Reilly Media, Sebastopol (2012)
44. Ferguson, R.: It’s all about the platform: What walmart and google have in common. Sloan Management Review. http://sloanreview.mit.edu/article/its-all-about-the-platform-what-walmart-and-google-have-in-common/ (2012). Accessed 05 Dec 2012
45. eMarketer. What do marketers want from big data?. http://www.emarketer.com/Articles/Print.aspx?R=1009798 (2013). Accessed 10 Apr 2013
46. Ferguson, R.: Better decisions with smarter data. Sloan Management Review. http://sloanreview.mit.edu/article/better-decisions-with-smarter-data/ (2014). Accessed 21 Feb 2014
47. Marchand, D., Peppard, J.: Why IT fumbles analytics. Harv. Bus. Rev. pp. 104–113 (2013).
48. Barton, D., Court, D.: Making advanced analytics work for you. Harv. Bus. Rev. **90**, 78–83 (2012)
49. Biesdorf, S., et al.: Big data: What’s your plan? McKinsey Quarterly. http://www.mckinsey.com/insights/business_technology/big_data_whats_your_plan?p=1 (2013). Accessed 01 Apr 2013
50. SAS. Data visualization: Making big data approachable and valuable. http://www.findwhitepapers.com/force-download.php?id=25943 (2012). Accessed 01 Sep 2012
51. Kemp Little LLP. Big Data - Legal Rights and Obligations. http://www.kemplittle.com/Publications/WhitePapers/Big%20Data%20-%20Legal%20Rights%20and%20Obligations%202013.pdf (2013). Accessed 07 Apr 2013
52. Ferguson, R.: Competitive advantage with data? Maybe ... Maybe Not”. Sloan Management Review. http://sloanreview.mit.edu/article/competitive-advantage-with-data-maybe-maybe-not/?utm_source=facebook&utm_medium=social&utm_campaig (2013). Accessed 26 Mar 2013
53. Security for Business Innovation Council. Information security shake-up. http://www.emc.com/collateral/industry-overview/h11391-rpt-information-security-shake-up.pdf (2012). Accessed 05 Jan 2013
54. Wielki, J.: Modele wpływu przestrzeni elektronicznej na organizacje gospodarcze Wydawnictwo Uniwersytetu Ekonomicznego, Wrocław (2012).
55. Parasuraman, K.: How big data and analytics are transforming in-store experience for retailers. ClickZ. http://www.clickz.com/clickz/column/2260001/how-big-data-and-analytics-are-transforming-instore-experience-for-retailers (2013). Accessed 09 Apr 2013
56. Kroft, S.: The Data Brokers: Selling your personal information. CBSNews. http://www.cbsnews.com/news/the-data-brokers-selling-your-personal-information/ (2014). Accessed 09 Mar 2014

57. World Economic Forum. Rethinking personal data: strengthening trust. http://www3.weforum.org/docs/WEF_IT_RethinkingPersonalData_Report_2012.pdf (2012). Accessed 29 Dec 2012
58. D Angwin, J., Valentino-Devries, J.: Google's iPhone tracking. The Wall Street Journal (2012). Accessed 18 Feb 2012.
59. Rushton, K., Trotman, A.: Facebook mined private messages to advertisers, lawsuit claims. The Telegraph. http://www.telegraph.co.uk/technology/facebook/10548196/Facebook-mined-private-messages-to-advertisers-lawsuit-claims.htm (2014). Accessed 02 Jan 2014
60. Brown, B., et al.: Views from the front lines of the data-analytics revolution. McKinsey Quarterly. http://www.mckinsey.com/Insights/Business_Technology/Views_from_the_front_lines_of_the_data_analytics_revolution?cid=other-eml-alt-mkq-m (2014). Accesses 14 Jan 2014
61. Wielki, J.: Consumers in the marketspace - the ethical aspects of electronic commerce. In: Bynum, T., et al. (eds.) Proceedings of the Fifth International Conference on the Social and Ethical Impacts of Information and Communication Technologies, vol. 2, pp. 259–268. Wydawnictwo Mikom, Gdańsk (2001)
62. Jones, R.: Barclays to sell customer data. The Guardian. http://www.theguardian.com/business/2013/jun/24/barclays-bank-sell-customer-data (2013). Accessed 25 June 2013
63. National Intelligence Council. Global trends 2030: Alternative worlds. http://globaltrends2030.files.wordpress.com/2012/11/global-trends-2030-november2012.pdf (2012). Accessed 03 Jan 2013