Association for Information Systems
AIS Electronic Library (AISeL)

BLED 2016 Proceedings | BLED Proceedings

2016

Entropy-based approach for semi-structured processes enhancement

Julia Bilinkis (Stavenko)
National Research University Higher School of Economics (NRU HSE), Russia, ybilinkis@hse.ru

Elena Filimonova
National Research University Higher School of Economics (NRU HSE), Russia, efilimonova@hse.ru

Nikolay Kazantsev
National Research University Higher School of Economics (NRU HSE), Russia, nkazantsev@hse.ru

Anastasia Zueva
National Research University Higher School of Economics (NRU HSE), Russia, zueva_ag@mail.ru

29th Bled eConference

Digital Economy

June 19 - 22, 2016; Bled, Slovenia

Entropy-based approach for semi-structured processes enhancement

Julia Bilinkis (Stavenko)

National Research University Higher School of Economics (NRU HSE), Russia

ybilinkis@hse.ru

Elena Filimonova

National Research University Higher School of Economics (NRU HSE), Russia

efilimonova@hse.ru

Nikolay Kazantsev

National Research University Higher School of Economics (NRU HSE), Russia

nkazantsev@hse.ru

Anastasia Zueva

National Research University Higher School of Economics (NRU HSE), Russia

zueva_ag@mail.ru

Abstract

The paper analyses traditional quality control methods for business processes and their efficiency in respect to semi-structured process. We update a classical methodology to handle semi-structured processes: improve its execution and operational efficiency in a company. We propose Some new methods: under the Define step – a new method for describing business processes with the help of semantic maps; under the Measure step – a new method for the automated detection of non-random deviations (bottlenecks, errors); under the Analyse step – a new method for automated experts search to identify the root causes of business process problems based on an analysis of the information field.

Keywords: semi-structured process; business processes; flexibility; project management; system integration

1 Introduction

Business models are constantly changing. In order to react on market fluctuations and achieve higher customization of products companies seek how to make them “adjustable”. Among

them, procedures that are based on the expert assessments and big data analysis are the core of modern enterprise.

A characteristic feature of such processes is the uncertainty of the input, output and variability of process instances. That is probably the most significant extension of traditional process definition coined by Hammer, Ciampi and Davenport. A high-level description of such processes may be stable, but the detailed problems of process instances during their execution are fuzzy defined and are difficult to replicate, as they are primarily dependent on the content and user behaviour. Thus, at the modelling and process automation step the decision is to be made in advance on all participants in the process and their actions and thus to create a detailed regulation of the process so it is almost impossible to monitor the implementation on its basis. Examples are processes for the production and provision of intellectual services, which are often prevalent in such sectors as education, information technology, smart production and all industries providing intellectual services (consulting, analytics, information brokerage, marketing and banking services, etc.).

Characteristic features of such processes are listed below [1]:

1. Customization to a specific consumer (i.e. intellectual service) is not standardized: provided to one client, it cannot be provided to another customer, since it requires data re-collection, analysis and information presentation;
2. Association of consumption process with the production process through constant interaction with consumers and fast response to demand;
3. Large number of sub-processes and tasks and interdependencies between tasks;
4. Each process task depends on other tasks, which leads to a large amount of feedbacks, the availability of information on the previous and subsequent process steps;
5. The use of explicit and implicit knowledge of experts. The behaviour of process performers depends on their knowledge, which is a constantly changing mix of experiences, values and incoming information;
6. Dependence on the context. The knowledge of subject area is used to perform the process, it includes tasks, documents, experts, and other indicators. The performed process is not limited to the orchestration of Web services and the sequence of tasks, but also obtaining all relevant information about the process;
7. Focusing on the executer, collaboration and decision-making requires the development and selection of integrated solutions among the fastest possible alternatives to achieve some certain goals. Responsibility of employees increases demands on their skills and competence;
8. Distributed processes. Process participants are not only employees of the company; the successful outcome of the process is highly dependent on corporate communications due to valuable ideas coming from the external environment and outflows of ideas out of the company which have no value for it.

This paper proposes an integrated approach to improving management efficiency (reaching goal of every business process separately while minimizing costs for its achievement under the

influence of managerial decisions) based on new methods of analysis and monitoring, taking into account the specific characteristics of semi-structured processes listed above.

The paper is structured as follows: the describe research goal, objectives and methodology. The third chapter provides insights from business on semi-structural process optimization: DMAIC model, SIPOC and TQM. In the main fourth chapter we demonstrate the proposed refinements to DMAIC model. Finally, chapter V provides conclusions.

2 Goals and Objectives

The main goal of our on-going research is the development and testing of automated methods for quality monitoring of semi-structured processes using variance analysis of semi-structured processes. Our main hypothesis in frames of current paper is that *the current level of effectiveness of semi-structured processes can be identified and improved using entropy variations of the information field as the quality indicators of business process.*

To verify this hypothesis, the following objectives were set:

1) To analyze and update current approaches for the process analysis, justification and selection of the description of semi-structured process using information field (unstructured data);
2) To determine:
    a) causes of natural variability (process output forms the distribution, stable and predictable over time),
    b) a non-random variability (disturbances in the process) by measuring entropy characteristics;
3) To compare real process problems with the problems discovered,
4) To update DMAIC methodology for semi-structured business processes.

The main method of research consists of deductive gathering of scientific information and further analysis, practical case studies. On the basis of this analysis we coin new notions. According to Hevner[1] there are two main methodologies to scrutinize the IS: the first one is Behavioral science model and the second one is Design science model. Our research bases on the latter model. This concept sees the IS scrutinizing as a science artifact, aimed to solve organizational problems of the company. The main principle of the model is enhancing the efficiency of IS within the organization via design modelling. As a modelling language we used UML. Enhanced state of the information system takes place thanks to optimization of its development and integration, staff trainings and creating new functional abilities.

March and Smith [2] suggested their own descriptive method for IS research. This method comprises two design processes and four design science artifacts. These artifacts are so-called constructs, sub-models, methods and concretizations. Constructs are playing a role of descriptive elements and shaping the common model view (i.e. verbal language). Sub-models build description of the situation out of the set of constructs. Methods contain text or logical description as well as algorithms for defining interaction between sub-models. Concretizations unite constructs, methods and sub-models into a one single model.

This paper proposes some new tools that can be used at each step of this methodology for semi-structured business processes. Our main methods are: deductive trend search in the literature review and the entropy method for determining the degree of disorder of the system (characteristics of the unstructured information field).

3 Application of DMAIC methodology on semi-structured processes

To improve the operational efficiency consultants in business often use DMAIC methodology (Define, Measure, Analyse, Improve, Control), developed around “6 sigma” concept (Fig.1).

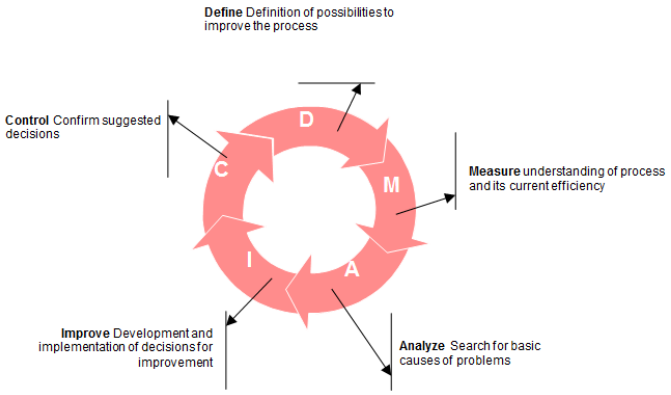

Figure 1. DMAIC methodology

DMAIC involves a gradual transition from the general understanding of the most effective solutions to the problems, with minimal cost and in the shortest time. DMAIC projects are always divided into five successive steps named after each letter and knowledge management approach for dissemination of project-related information. The cycle starts with identification requirements of customers and business, setting optimization goals, determining expert profiles for working groups what is included into a project charter.

To define the clear boundaries of the optimized process and its key participants SIPOC method (Supplier – Input – Process – Output – Customer) is often used.

S	I	P	O	C
External suppliers Domestic suppliers (employees, GSP management) Customers	Resources, documents, information, preliminary work	Description of process	Results of process	External customers Domestic customers Other interested person

VOC (Voice of Customer), technical conditions and information

Figure 2. SIPOC method: Supplier – Input – Process – Output – Customer.

The effectiveness of a structured process is easy to measure taking into account the fact that the set of works is always predetermined and structured and thus evaluating the performance of all the components, we can conclude about the effectiveness of the entire process. As

indicators for monitoring the process parameters can be selected, which are relevant objectives of each function/sub-process with a certain periodicity of collection, using which the organization receive a well-balanced system of indicators of symptoms of problems.

Another technique that postulates that the product or service of poor quality is the result of unpredictable variability of the process (or input process parameters) is Total Quality Management (TQM). According to this technique process is statistically controlled when the only source of variation is the natural cause – a variability, originating from numerous sources and inherent in the process. Natural changes behave as a system of random factors with constant parameters.

While all process instances differ, they form a certain pattern as a group; it can be described as a distribution. The reduction of this variation requires management solutions and investment capital (for example, to purchase new equipment). If this is a normal distribution, it is characterized by two parameters: the mean and standard deviation. It is impossible to measure the mean and standard deviation in practice, as this would require the measurement of all possible instances of the process. Instead they use a number of measurements taken over time by measuring the sample mean and sample variance, respectively. Until the distributions of these parameters are within predetermined limits, the process is statistically controlled and natural variations are allowable. If they come out of the specified parameters, this is due to non-random changes that are not inherited by the process. Reduction of variation requires a special analysis of its causes.

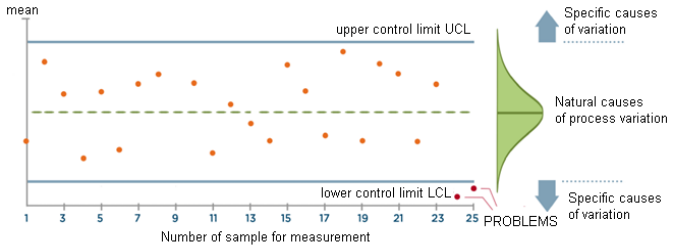

Figure 3. Natural and special causes of variation

First of all, the process must be statistically controlled; it can be done by identifying and addressing the special (non-random) causes of variability. The objective of quality control task is to supply statistical signal about the presence of non-accidental causes. Such a signal can accelerate the adoption of measures aimed at the elimination of non-random causes. This approach has proven itself to standardized processes with high volume output and low diversity.

But on the other hand, the work is not always being done consistently and with pre-defined structure, at a modern approach, it becomes clear that the process can contain both structured work and ad-hoc works, often the unique challenges. Such processes can be called nonlinear intelligent, dynamic and contextual by nature. The choice of indicators for semi-structured processes is not a trivial task, very often it's possible to use only the delayed parameters of the process result, but in case of their usage it is difficult to respond quickly at problems emerging during the process implementation. The advantage of leading indicators is that they have the

prognosis nature and allow the organization to adjust its actions quickly on the basis of comparison of the actual indicators values with planned values.

In the following sub-sections we propose to use the indicators related to the information field of the process as the elements of a subset of leading indicators for the semi-structural processes. In the following sections we describe each DMAIC methodology step with our refinements.

3.1 Define step using insights from information field

The semi structured process itself can be defined as a set of concerted efforts of the interacting participants - information and knowledge holders. We consider the activities of the company as a result of the functioning of the socio-technical system and the activities in process are aimed on searching for the required procedural decisions in the system of distributed information and knowledge. The efficiency of the entire process depends on the efficiency of search management. In terms of subject-oriented approach the subject (member of process) is the starting point to describe a situation or event. Subjects synchronize their activity through messaging to switch between their functional states. As part of the semi structured process performing, subjects generate content at a proper assessment of which it is possible to distinguish the purpose of the process and its semantic environment.

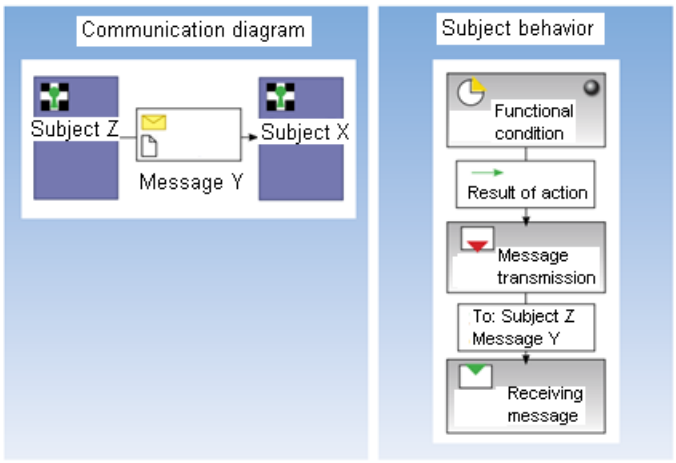

Figure 4: Presentation of the business process in terms of subject-oriented approach (Message transmission)

The idea of presenting processes with the help of unstructured information has been described by several authors. The formation of automation system model of an enterprise as a multilayer taxonomy has been made in work [3], where the company is regarded as a "*scale significant collection of various information entities*" [3] which can be classified by created taxonomy. In the work [3] structural units of the enterprise are root taxa, keywords are the end ones and define a business process – operations. Ontology definition and formal models of business processes of telecommunication enterprise are described in detail in the paper [4]; where the taxonomy of business processes was described based on the reference ontology, then the ontology was created based on real business processes of the organization, after their

comparison and analysis. Both ontologies have been formed on the basis of a common glossary of terms.

That's why within the scope of the study the process would be described through:

• *Information field*, which is a set of messages which are exchanged between participants during process execution to achieve a certain objective;

• *Subjects* (participants of the project) - implicit expert knowledge carriers. The information field is given by the characterizing topics of the text information, which in turn are defined by terms.

Topic is not just a set of terms (keywords), it is not random, but stable characteristics of a set of semantically related terms, characterizing the process. This was confirmed in study [5].

3.2 Refinement of the second step (Measure) using performance of information field

At Measure step, the measurement is performed, it is gathered data on the problem (Y) and possible root causes (Xn): Y = f (x1, x2, …,xk), where Y is the result of a process, Xn – inputs and internal factors of the process.

For structured process at this step based on a generalized map of SIPOC/VSM a detailed process map is created, it includes marked process steps, tasks, operations, and decision-making moments in chronological order. The purpose for the semi-structured process is clearer than the way of its achieving, therefore the choice of the way will be determined in the course of achieving goal based on the information and would not be planned in advance, so it cannot be fixed in a low-level regulations or process model.

Defining of semi-structured processes can be impractical because of the measurement of the same parameter for different instances of the same business process, different values will be received faster due to the following reasons:

• Business process is characterized by complex logical and temporal structure, so a copy of it can be developed in different ways depending on numerous conditions;

• Objects incoming to the business process for different instances can have different values of the same parameters and that significantly affects the development of the whole business process;

• Subject actions of the business process can be changed under the influence of external and internal environment;

• Some activities within the business process related to decision-making can have an informal unregulated character.

• The above-described control maps method can be used only for structured processes. Control maps cannot be applied, if the inputs are not homogeneous, the process is not regular, the output parameters are unique. I.e. for semi-structured processes there must be developed a new approach which enables to:

• Define the non-random variation causes of the semi-structured process in order to respond to the situation when process parameters go beyond the scope and thus to identify problems before client meets them.

• Reduce the variability of the process, to improve the product or process input parameters, thus reducing the likelihood of unpredictable variations impact the process and lead to problems.

If we define a business process in the form of its information field; its content is designed to implement a specific goal in case of non-random changes terms describing these changes should appear in the information field. The indicators of the information field will be used as leading indicators, which are determined by detecting the abnormal variation in the references frequency to relevant terms by process participants. The paper shows that the meaningful terms should have strongly unequal distribution of the relative frequency of use among the employees and commonly used ones show approximately the same relative frequency of use.

To solve the task of the significant terms extraction from the commonly used ones, it was calculated the statistics of the *relative usage frequency* of t_i term for all texts written by a particular employee p_j:

$$TF(t_i, p_j) = \frac{m(t_i, p_j)}{\sum_k m_k}, \quad (1)$$

where$m(t_i, p_j)$is the number of uses t_i term by p_j person, and the denominator is the total number of occurrences of all terms by p_j person.

As a result, for each term, you can make a sample of the relative usage frequency of this term by the authors$TF(t_i, p_j)$, where n is the number of terms in the information field of the organization, N is number of authors:

$$\begin{matrix} TF(t_1, p_1) & \dots & TF(t_1, p_N) \\ \dots & \dots & \dots \\ TF(t_n, p_1) & \dots & TF(t_n, p_N) \end{matrix} \quad (2)$$

Several metrics can be offered to determine the significance of the term, based on the idea of nonlinear probability distribution. The simplest metric for the degree of changes in the values of the sample is determined by calculating their variances$D(t_i)$,that is, the variance of the relative frequency of use of the$t_i$term and standard deviationσ_i:

$$D(t_i) = \frac{(TF(t_i,p_1)-M)^2+(TF(t_i,p_2)-M))^2+\dots+(TF(t_i,p_N)-M))^2}{N-1} = \frac{1}{N-1}\sum_{k=1}^{N}(TF(t_i, p_N) - M))^2, \quad (3)$$

where M is an estimation of the expectation (the sample mean), the relative frequency of use of the term, is calculated as follows:

$$M(t_i) = \frac{TF(t_i,p_1)+TF(t_i,p_2)+\dots+TF(t_i,p_N)}{N} = \frac{1}{N}\sum_{k=1}^{N} TF(t_i, p_N) \quad (4)$$

The standard deviation of the relative usage frequency of the term is the square root of the variance:

$$\sigma(t_i) = \sqrt{D(t_i)} \tag{5}$$

Then the significance of the term for a particular author exceeds the expectation $M(t_i)$,measured in standard deviations $\sigma(t_i)$:

$$Impact(t_i\,; p_j) = \frac{(TF(t_i,p_j) - M(t_i))}{\sigma(t_i)} \tag{6}$$

According to the equation (8), it turns out that the negative values of $Impact(t_i\,; p_j)$ will have the authors which have rarely used the term, i.e. not experts on this term.

To normalize the weight of $Impact(t_i\,; p_j)$ calculation of arctangent function is used. The higher is the value of arctangent, the more significant is this word for the author:

$$\text{FinalImpact}(t_i\,; p_j) = \text{arctg}\,(\text{Impact}(t_i\,; p_j)) \tag{7}$$

i.e. innovations are determined by identifying abnormal deviation of the term usage frequency.

Identification of the business process characteristics, which affect its quality indicators can be combined into a single function. All the considered characteristics are the variables in the objective function of the business process. Since the function is directional and has more than two parameters, we consider in this case the multi-criteria optimization, and an appropriate vector, which can be used to control the quality of the test process.

Some tasks can be solved based on analysis of information criterion, for instance, determining whether a business process corresponds to the reference one (officially required), revealing hidden company's processes, identifying trends and dependencies in the life cycle of business process. All of the tasks of identifying and applying the results in practice lead to the optimization of business processes and activities of the company as a whole.

3.3 The third step of the project optimization: search for experts, analysis of the root causes of problems

The root causes of process problems are determined and confirmed during the Analysis step. To do this, one should provide a list of reasons (critical) and how they are determined by an expert report. Then experts using tools such as the Ishikawa diagram and 5 Why perform a causal analysis to determine the root causes and make the prioritization of the most important reasons. To do this, one should identify the human expert for each deviation, a person who chooses and makes decisions based on information provided by the decision support system.

The amount of effort that are to be made in order to find an expert in the organization depends on several factors such as the size of the organization, the level of automation, the power of social relations within the organization, etc. In case if there is no automated expert search engine a task of finding the right person may take considerable efforts.

To identify experts within the process you need to:

1. the concept of topic, i.e. the set of semantically related terms
2. define the most relevant topic for each author.
3. match the topic with search query to select relevant authors (experts) not by words but by semantic units.

It is necessary to group all the important terms so that for any term to choose the relevant topic (group of terms) on the basis of the above tasks. For this reason, we used model in which each term is associated with many others. This model can be represented as a graph, its nodes are the terms and subgraphs are the topics, the adjacent topics have common terms (graph nodes).

Clustering is not the most effective way to solve this problem its result is a partition of the terms into clusters (one term cannot not be included in more than one cluster). A more efficient way is to calculate the semantic proximity of terms in which a list of related terms (TOP 50) is selected for each term. Different techniques are used to calculate the semantic similarity between the terms: PMI1, LSI2, LDA3, etc. [6]. PMI method is computationally less expensive, and thus works quite predictably, in addition it shows the greatest consistency in terms of experts [7].

This method involves determining the PMI-related terms based on their co-occurrence in each text. Input information is a set of unstructured text of each author, which they exchanged at work, this is a collection of D-documents.

The term means the word w, extracted from text ∈D (data) and having a high importance for informative text. Topic T is a set of semantically related key phrases, T = {w1, …, wN }. In the best case, all the terms which form the topic T refer the same category which is sufficiently narrow.

Statistical methods are based on the calculation of performance based on the co-occurrence of words. At the same time all the words are treated as points in N-dimensional space, and the problem of determining the semantic distance is reduced to two basic steps:

1) Set the coordinates of points in space.
2) Calculate the distance between the points. The last step requires the selection of suitable metric. In order to reduce the dimension of task being solved by clustering the N-dimensional space in which the presentation and clustering points is performed is built based on a set of T key phrases, selected in this step as follows:
   a) Set T is regarded as fixed dictionary at this step Vc, |Vc|=N.
   b) All key phrases v ∈ Vc are numbered.

1 Probabilistic Latent Semantic Indexing

2 Latent Semantic Analysis

3 latent Dirichlet allocation

c) Now, each word or phrase w ∈ V can be represented in the N-dimensional vector values of co-occurrence (fw) = (f1…fn), where fi indicates how often a word w occurs in conjunction with the word vi.

Thus, each key phrase appears as a point in N-dimensional space. The values expressing the degree of semantic proximity of sentences with phrases from Vc are to be taken as the coordinates of points fi having such characteristic: the bigger is the semantic proximity of the two considered words or phrases the bigger is the value. This PMI value for v and w phrases has such form:

$$\mathrm{PMI}(\mathrm{v},\mathrm{w}) = \log\frac{\mathrm{p}(\mathrm{v},\mathrm{w})}{\mathrm{p}(\mathrm{w})\mathrm{p}(v)} \quad (8)$$

Where p(w,v) is the frequency of co-occurrences of terms w and v, p (w) – frequency of the term w in the texts, p (v) –frequency of the term v in the texts. Words w and v are considered to be met jointly when they met at a distance not of less than N words. If two words are statistically independent, their PMI is equal to “0”.

4 Conclusions and constraints

All mentioned above refinements are sufficient from our experience *for handling semi-structured processes in modern business models.* The proposed changes need to be verified not only on set of industries but also in different cultural environments. Those are our intentions for continuation of this on-going research. The latter works would be aimed on fostering the fourth step and fifth step (controlling) solutions for improving the process on the basis of the identified factors. During fifth step the developed solutions are extended and fixed.

References

- **Journal Articles**

von Alan, R. H., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS quarterly, 28(1), 75-105.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. Decision support systems, 15(4), 251-266.

Konev, K.A. (2012): "Conceptual model of enterprise automation in the aviation industry based on multilayer taxonomy", UGATU.

Chistiv D.A., Kamaev V.A., Naboka M.V. "Ontological re-engineering of business processes of service provider", VolGTU.

David Newman, Sarvnaz Karimi, Lawrence Cavedon: "External Evaluation of Topic Modes", NICTA and The University of Melbourne Parkville, Victoria 3010, Australia.

Anton Korshunov, Andrey Gomzin. Topic modelling in natural language texts, ISP RAS, Moscow, Russia [Electronic resource].

http://www.ispras.ru/proceedings/docs/2012/23/isp_23_2012_215.pdf

J. Hockenmaier. Introduction to Natural Language Processing. Lectures at University of Illinois at Urbana-Champaign [Electronic resource].–Mode of access: http://www.cs.uiuc.edu/class/fa08/cs498jh/

Дулесов А.С., & Хрусталев В.И. (2012): Определение энтропии как меры информации при сопоставлении прогнозных и фактических показателей предприятия, Современные проблемы науки и образования, ISSN 2070-7428.

Зеленков Ю.А. (2013): «Об измерении эффективности бизнес-процессов и поддерживающих их информационных систем», Управление большими системами, ОАО «Научное Объединение «Сатурн», Рыбинск

Jae-Yoon Jung: «Measuring Entropy in Business Process Models», The 3rd International Conference on Innovative Computing Information and Control, 2008 г.

Olga Streibel: "Mining Trends in Texts on the Web", Networked Information Systems, Free University Berlin, K¨onigin-Luise-Str.24-26 , 14195 Berlin, Germany.

Лопатин В.А. (2008) Система управления бизнес-процессами // Управление в кредитной организации. № 6.

Кини Р.Л., Райфа Х. (1981) Принятие решений при многих критериях: предпочтения и замещения. — М: Радиоисвязь.

Gromoff A., Kazantsev, N., Kozhevnikov, D., Ponfilenok, M. and Stavenko, Y. (2012). Newer Approach to Create Flexible Business Architecture of Modern Enterprise. Global Journal of Flexible Systems Management. 13(4), Springer-Verlag, 207-215