Building a GERT model of life cycle of educational and scientific information resources using Process Mining technology

M G Dorrer, A A Popov, E I Trishkina and N A Romanov
Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochny Av., Krasnoyarsk, Russia

E-mail: mdorrer@mail.ru

Abstract. The main priority of this article is to develop a tool for numerical forecasting of the life cycle of information resources based on actual life cycle data in the form of event logs of the resource history. To solve this problem, we used the GERT network device. To build a GERT network describes the life cycle model of an information resource, we were used methods of intellectual analysis of processes (Process Mining) implemented with the ProM Framework. In the course of the work, an analysis of the life cycle of information resources of the Scientific electronic online library was carried out. The work considers the process of extracting data used for modeling the GERT network, as well as the process of identifying the law of distribution of the probability density of the duration of the life cycle stages. The model of the GERT network is constructed and the law of distribution of the probability density of the duration of the life cycle of an information resource was described. The obtained result confirms the applicability of Process Mining technology to probabilistic analysis and forecasting of the life cycle of information resources.

1. Introduction
Currently, probabilistic modeling is actively used in various studies related to life cycle assessment. Thus, the use of the Markov chain was shown in the work on evaluating the performance and cost of the life cycle of the roofing system. The authors gave an example illustrating the development of Markov chain model, a method for identifying parameters, evaluating the Monte Carlo uncertainty, and making decisions with probabilistic information. It is shown that the uncertainty spreading in this process is not negligible and can significantly affect or even change the final decision [1].

The Monte Carlo method was used for modeling a probabilistic technical and economic model of the life cycle of four batteries, as well as for analyzing methane emissions at different stages of the well life for conventional and unconventional natural gas [2], [3].

To date, there are a large number of studies related to probabilistic models for maintaining and optimizing the life cycle characteristics of deteriorating structures, which formulate future directions in this area [4], [5], [6].

The need for a support system to achieve efficient and consistent management of life-cycle objects based on the object model was shown in the work "Facility Model for Life-Cycle Maintenance System" [7]. The effectiveness of the model is shown by the example of the object wear assessment system.
The use of probabilistic modeling for the estimation of service life is devoted to the works in which Bayesian statistical theory and a model with discrete time are applied [8], [9].

An important issue at the moment is the analysis of the life cycle of information resources, which is still insufficiently studied, in contrast to the life cycle of software and information systems. In modern information systems, there are many different types of information resources that need to be stored and processed. To improve the efficiency of using information resources (EIR) of libraries, it is necessary to assess the intensity of receipt and conduct a comparative analysis based on the data obtained, which will best combine the information resources of libraries into a single information system.

A common method for predicting life cycles is the use of Markov Chains. For example, in [10] this model was used to predict important information about the behavior of a product during its life cycle on the example of plastic products.

Process Mining technology is used to evaluate the parameters of information resource life cycles. For example, in [11] this work was done for library information resources.

However, Markov chains have some limitations in the analysis of processes. Thus, events originating from the Markov chain state can only triggered in turn and thus form a complete group, that is, the sum of the probabilities of their occurrence is equal to one. The standard situation in processes is a fork by the condition "And" (parallel) and "Or" (non-exclusive). GERT networks are becoming more and more widespread as an alternative to Markov chains. So, in [12], GERT networks used to analyze the educational process of distance education.

Thus, the task of analyzing the life cycles of information resources using stochastic GERT networks is relevant. At the same time, for the analysis of a large volume of information about the life cycles of information resources, the use of technology for intelligent analysis of business processes - Process Mining is relevant.

Thus, this paper presents a solution to the problem of building a life cycle model of information resources using the GERT network to obtain an assessment and forecast of the life cycle based on data collected using Process Mining tools.

2. Materials and methods

The GERT network is a probabilistic model of a discrete-event system that allows taking into account random deviations in the system parameters that occur when performing each of the operations [13], [14]. The operation of the business process is mapped to the branches (arcs) of the GERT network, which are characterized by additive random variables. The output characteristics of GERT networks are calculated on the basis of generating functions of moments of random variables, and the activation of each branch originating from the node is set by the probability of triggering this branch.

The GERT-network is described by a directed weight graph $G = (V, E)$,

$$G = (V, E),$$

where $V$ – is the set of nodes; $E$ – is the set of branches (arrows).

GERT network nodes are interpreted as system States, and branches (arcs) as transitions from one state to another. When describing a model of a discrete event system, such transitions are compared with the execution of generalized operations characterized by the density of the distribution of the measured value (time, expenditure of money or other resources) and the probability of execution.

Thus, the GERT-network – is a network with sources $R$ and sinks $S$ the type of «work-on-arrow», in which each node belongs to one of six types of nodes [13], for each arrow $< i, j >$ the weight of type $[p_{ij}, F_{ij}]$ is determined, where $p_{ij}$ – conditional probability of occurrence of node $< i, j >$ with condition of activation of a node $i$, $F_{ij}$ – conditional distribution function of some random variable.

When solving a problem of calculating GERT-network’s parameters most common task is to find the first central moments of the distribution of the measured random variable of the network. Commonly for practical needs sufficient would be to find the first and second central moments of random variable - mathematical expectation and variance. More detailed research of the system can be done by creating a function of the distribution of a random variable for entire network.
Formulas and algorithms for calculation GERT-network’s parameters can be seen in works of Fillip, Garcia-Diaz [15].

To model the GERT network, you first need to define a data set for analysis, from which the necessary information will be obtained using the Process Mining technology [16]. Information was extracted and analyzed from the data set using the ProM Framework program [17]. After getting statistical data, you need to build a simulation model based on which the GERT network will be modeled. To do this, you can use the built-in Inductive visual Miner plugin. ProM uses the event log provided to it and builds its model, which allows you to see which cases were held for which events.

After building the model, it is necessary to extract duration of each event. For that a plugin «Add Time between events (Duration) as Attribute to all Events» was used. Plugin displays duration of each case as an attribute. In this case the selected time unit was in hours.

The next step is to determine the most appropriate distribution law for the edge of the GERT network. To do this, we used the MatLab program, which contains a set of tools called "Statistical toolbox" [18], necessary for determining the distribution law. Definitions of a suitable law can be divided into two stages.

The first step is to select the laws and calculate their parameters for each sample. To do this, use the MLE(x, 'distribution', 'Name') function, where x is the selection data, and Name is the distribution name. After getting the parameter data, you need to determine the value of the distribution laws for each value of each sample, using previously extracted data for the distribution laws parameters. This can be done by using cdf('Name',x,A), where name is the name of the distribution x of the sample values, A – values distribution law for a given sample.

The second stage is to select the consent criterion by which the most appropriate distribution law will be determined. In this study we used the criterion of Kolmogorov. In MatLab, this criterion can be used using the function [H, P, KSSTAT, CV] = kstest(X,cdf,alpha, tail), where x is the sample values, and cdf is a matrix with dimension m×2 (m is the number of set values of a random variable x). The first column of the CDF matrix corresponds to the specified values of the random variable X. The second column of the CDF matrix is calculated from the cumulative theoretical distribution function at points x, alpha is to set the value of the critical significance level for testing the null hypothesis, and tail is a type of alternative hypothesis. The most appropriate distribution law is determined by the parameter H and P. If the value of H is 0, then this distribution law is appropriate. If there are several such laws, then we need to consider the parameter P. The larger this parameter, the more appropriate this distribution law is. Having determined the distribution law, you can proceed to building a GERT-network model, using the previously obtained imitation model in ProM. Events should be represented as arrows and operations as nodes.

The next step is to transform the arrows in the GERT network model. Parallel arrows are transformed by formula

\[ W_{ij}(s) = W_a(s) + W_b(s), \]  

where W is W_{ij}(s)=p_{ij}M_{ij}(s), where p – probability of arrow triggering, M – distribution law. If arrows are consecutive, then they are transformed by formula

\[ W_{ik}(s) = W_{ij}(s) * W_{jk}(s). \]

3. Results
The data source is the open database of the Scientific electronic Library Online (Scientific Electronic Library Online), published in open access on the website Kaggle.com [19]. The selection was made from the "status change Log" table, which contains 23 columns. Two columns were used as the basis: Title thematic areas and Title Current Status.

Figure 1 shows that the sample contains 20 types of events and 892 events, as well as the time by which the duration of the log can be determined.
Using the ProM plugin "Inductive visual miner", based on the analysis of the event log, the structure of the process model was restored. The simulation model is shown in figure 2. In this model, ProM detected three events: current, suspended, and deceived.

![Figure 1. Event Statistics.](image)

![Figure 2. Sample life cycle model based on event log.](image)

In order to determine the most appropriate distribution law, the duration of each event was obtained using the "Add Time between events (Duration) as Attribute to all Events" plugin. All extracted sample data for each event is shown in table 1.

| Event name | Lead time (hours) |
|------------|-------------------|
| Current    | 14779, 13612, 17562, 6254, 18795, 12060 |
| Suspended  | 16239, 18232, 13766, 16287, 15339, 11486, 15964 |
| Deceased   | 13445, 17685, 12467, 19457, 15634, 6140 |

Using the description of the lifecycle in the form of the ProM heuristic model (figure 2), you can start building a GERT network model. Figure 3 shows the corresponding GERT network for this model. Since States (which are analogous to processes in business process models) are mapped to GERT network arcs in the heuristic model, the current, suspended, and deceased events are represented as arcs...
in the GERT network model, and the model start, model end, and divergence operators are represented as nodes.

**Figure 3.** Sample life cycle model in the form of a GERT network.

The GERT network model contains 4 nodes and 5 arrows. Table 2 shows the designation of the nodes of the GERT network and their types.

| GERT network node | Node type |
|-------------------|-----------|
| V1                | STEOR     |
| V2                | STEOR     |
| V3                | STEOR     |
| V4                | STEOR     |
| W1                | Arrow     |
| W2                | Arrow     |
| W3                | Arrow     |
| W4                | Arrow     |
| W5                | Arrow     |

After building the model and using the obtained distribution laws on the edges and the probability of triggering each of the arrows, a table of the GERT-network parameters was compiled (table 3). Each arrow was assigned an identifier \( k \) for more convenient access to the arrow, the variable \( p \) denotes the probability of the arc being triggered, \( M(s) \) is the distribution law for a given arrow, \( W(s) \) is the function \( W_{ij}(s) = p_{ij} M_{ij}(s) \).

| Node \(<i,j>\) | k | \( p_k \) | \( M_k(s) \) | \( W_k(s) \) |
|---------------|---|----------|-------------|-------------|
| \(<V1,V2>\)   | 1 | 1        | \( \frac{1}{4083.5 \sqrt{2\pi}} e^{-\frac{(s-13843.6)^2}{2\times4086.5^2}} \) | \( \frac{1}{4083.5 \sqrt{2\pi}} e^{-\frac{(s-13843.6)^2}{2\times4086.5^2}} \) |
| \(<V2,V3>\)   | 2 | 0.8      | \( \frac{1}{1992.4 \sqrt{2\pi}} e^{-\frac{(s-15330.4)^2}{2\times1992.4^2}} \) | \( \frac{1}{1992.4 \sqrt{2\pi}} e^{-\frac{(s-15330.4)^2}{2\times1992.4^2}} \) |
| \(<V2,V3>\)   | 3 | 0.1      | \( \frac{1}{4288.5 \sqrt{2\pi}} e^{-\frac{(s-14130)^2}{2\times4288.5^2}} \) | \( \frac{1}{4288.5 \sqrt{2\pi}} e^{-\frac{(s-14130)^2}{2\times4288.5^2}} \) |
| \(<V2,V3>\)   | 4 | 0.1      | \( \frac{1}{4288.5 \sqrt{2\pi}} e^{-\frac{(s-14130)^2}{2\times4288.5^2}} \) | \( \frac{1}{4288.5 \sqrt{2\pi}} e^{-\frac{(s-14130)^2}{2\times4288.5^2}} \) |
| \(<V3,V4>\)   | 5 | 1        | \( \frac{1}{4288.5 \sqrt{2\pi}} \) | \( \frac{1}{4288.5 \sqrt{2\pi}} \) |
After carrying out the necessary transformations, which were described above and taking into account the rule for calculating GERT networks given in [15], the distribution law for the GERT network was obtained.

\[
W_{15}(s) = 0.8 \times \frac{1}{4083.5 \times \sqrt{2\pi}} e^{\frac{(s-13843.6)^2}{2 \times 4086.5^2}} + 0.1 \times \frac{1}{1992.4 \times \sqrt{2\pi}} e^{\frac{(s-15330.4)^2}{2 \times 1992.4^2}} + 0.1
\]

(4)

4. Discussion

Thus, it is shown that the GERT network allows you to predict the numerical parameters of the life cycle of information resources (for example, the duration of finding information resources at the stages of the life cycle and the duration of the life cycle in General). In contrast to traditional statistical simulation experiments, the proposed method of life cycle analysis allows you to obtain analytical expressions based on which you can directly calculate the required parameters of the life cycle. This solution has a number of advantages compared to a simulation experiment, including lower computational complexity, as well as independence of the result obtained from the number of experiments performed.

The prospects for the development of this work are the following:

- Using the analytical expressions for the GERT-network, it is possible to produce a numerical solution to the problem of predicting the parameters of the duration of the life cycle – in the form of graphs of the probability density distribution, and important in practice, values of Central moments – mathematical expectation and variance of the duration of the life cycle.
- Analytical expression for the distributions of numeric parameters will allow to formulate and solve analytically the optimization problem for the target parameters of life cycles by varying the parameters of the individual steps, the probability of operation of the branches life cycle and structure of the life cycle in General.
- The forecast of probabilistic parameters of the life cycle of information resources will solve the problem of optimizing the storage infrastructure of analyzed information resources.

5. Conclusions

The analysis determined the total duration of the event log, as well as the duration of each event contained in the sample. Using a simulation model, we were able to construct a probabilistic model of the GERT network. Based on the data analysis, the law of distribution of values for the GERT network was determined.

Thus, the proposed approach to solving the analysis and forecasting of information resource life cycles using GERT networks has shown its efficiency. In this article, this work is based on data from the library of scientific publications, but it can be assumed that the life cycle of any information resource can be modeled in this way, for which it is possible to get a log of life cycle events.

References

[1] Zhang Y and Augenbroe G 2005 Uncertainty Analysis in Using Markov Chain Model to Predict Roof Life Cycle Performance 10DBMC Int. Conf. On Durability of Building Materials and Components (Lyon: Georgia Institute of Technology) 17-20
[2] Battke B, Schmidt T S, Grosspietsch D and Hoffmann V H 2013 A review and probabilistic model of lifecycle costs of stationary batteries in multiple applications Renewable and Sustainable Energy Reviews 25 (Zurich: Elsevier) 240-50
[3] Shahriar A, Sadiq R and Tesfamariam S 2014 Life cycle greenhouse gas footprint of shale gas: a probabilistic approach Stochastic Environmental Research and Risk Assessment 28(8) 2185-204
[4] Van Noortwijk J M and Frangopol D M 2004 Two probabilistic life-cycle maintenance models for deteriorating civil infrastructures Probabilistic Eng. Mech. 19(4) 345-59
[5] Frangopol D M, Kallen M -J and van Noortwijk J M 2004 Probabilistic models for life-cycle performance of deteriorating structures: review and future directions Prog. Struct. Eng. Mater 6(4) 197-212
[6] Swei O, Gregory J and Kirchain R 2015 Probabilistic life-cycle cost analysis of pavements: Drivers of variation and implications of context Transp. Res. Rec. 2523 47-55
[7] Takata S, Hiraoka H, Asama H, Yamaoka N and Saito D 1995 Facility Model for Life-Cycle Maintenance System CIRP Ann. - Manuf. Technol. 44(1) 117-21
[8] Val D V, Stewart M G and Melchers R E 2000 Life-Cycle Performance of RC Bridges: Probabilistic Approach Comput. Civ. Infrastruct. Eng. 15(1) 14-25
[9] Fischer S A 1973 Life Cycle Model of Life Insurance Purchases Int. Econ. Rev. 14(1) 132
[10] Afrinaldi F 2020 Exploring product lifecycle using Markov chain Procedia Manuf. 43 391-8
[11] Dorrer M. G, Popov A. A and Bartuzanova A N 2020 Analysis of the life cycle of a library information resource using Process Mining technology J. Phys. Conf. Ser. 1582 012024
[12] Dorrer M and Dorrer A 2020 Forecasting e-Learning Processes Using GERT Models and Process Mining Tools Smart Innovation, Systems and Technologies 172 857-66
[13] Pritsker A A B 1966 GERT: Graphical Evaluation and Review Technique (Santa Monica: RAND Corporation)
[14] Neumann K 1990 Stochastic Project Networks 344 (Berlin, Heidelberg: Springer Berlin Heidelberg)
[15] Phillips D T and Garcia-Diaz A 1981 Fundamentals of network analysis Networks 12(2) 209-10
[16] van der Aalst W M P et al. 2012 Process Mining Manifesto 169-94
[17] van Dongen B F, de Medeiros A K A, Verbeek H M W, Weijters A J M M and van der Aalst W M P 2005 The ProM Framework: A New Era in Process Mining Tool Support 444-54
[18] Statistics and Machine Learning Toolbox Retrieved from: https://www.mathworks.com/products/statistics.html/
[19] Publication and usage reports Retrieved from: https://www.kaggle.com/scieloorg/publishing-and-usage-reports-1998-201710-br/