ITHEA

International Journal

INFORMATION TECHNOLOGIES & KNOWLEDGE

2017 Volume 11 Number 2

International Journal
INFORMATION TECHNOLOGIES & KNOWLEDGE
Volume 11 / 2017, Number 2

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" (IJ ITK)
is official publisher of the scientific papers of the members of
the ITHEA International Scientific Society

IJ ITK rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITK are given on www.ithea.org

Responsibility for papers published in IJ ITK belongs to authors.

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Volume 11, Number 2, 2017
Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with:
Institute of Mathematics and Informatics, BAS, Bulgaria; V.M.Glushkov Institute of Cybernetics of NAS, Ukraine;
Universidad Politechnika de Madrid, Spain; Hasselt University, Belgium;
St. Petersburg Institute of Informatics, RAS, Russia; Institute for Informatics and Automation Problems, NAS of the Republic of Armenia.
Printed in Bulgaria
Publisher ITHEA®
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: info@foibg.com
Technical editor: Ina Markova
Издател: ИТЕА®, София 1000, ПК 775, България, www.ithea.org, e-mail: info@foibg.com

ISSN 1313-0455 (printed) **ISSN 1313-048X (online)**

QUALITY OF A COMPOSITE SERVICE AS A FUNCTION OF THE QUALITIES OF THE COMPRISED SUB-SERVICES A SURVEY.

Anna Otsetova, Emiliya Saranova

***Abstract**: Today more and more complex services are proposed by the companies. At the same time customer needs are changing rapidly over the time. In all service sectors the service providers used to provide services satisfying Quality of Service (QoS) requirements. In this paper we present results of conducted literal review related to composition of services, quality of composite services and the functions of the qualities of the comprised sub-services.*

***Keywords**: Service composition, sub-services, quality of services*

***ITHEA Keywords**: Introductory and survey*

Introduction

In the era of growing demand of services, service providers have started to offer different service levels in order to meet the different users' needs. Dealing with the increased changes arising from the competitive business environment service providers is attempting to survive by developing flexible composite services using various technologies and tools. They usually propose different service qualities based on user qualification and service cost.

The nature of services creates the opportunity for building composite services by combining existing elementary or complex services (referred to as sub-services) and in turn offering them as value added services.

The composite services can be presented as a result of multiple interactions of its internal components. The composite services are group of sub-services that create value by transforming inputs into more variable outputs. A composite service is defined as a coordinated and logically sequenced set of sub-services.

Information composite service aim to power the next generation data transfer so that customer are able to access and use them from anywhere in the world on demand at competitive costs depending on users' QoS requirements.

QoS has received much interest in service research because of the rapid increase of the number of services and the approximate equal qualities of the services.

Nowadays QoS has many aspects which depend on the aspect that is crucial for the user [Karawash, Mcheick, Dbouk, 2014]. Some of the QoS aspects are: response time, throughput, delay, security, service availability, reliability, timeliness, price etc. Such qualities are of interest to service providers and service consumers alike. These QoS aspects are very important, but they do have one flaw: they are exclusively inward-looking. They examine network performance, which is only one half of an increasingly important equation - and the reason that most service providers are now shifting their focus from QoS to Quality of Experience (QoE).

QoE is defined as the overall acceptability of an application or service, as perceived subjectively by the end-user. The overall acceptability may be influenced by user expectations and context [ITU-T Rec. P.10/G.100, 2006].

The agreement between the customer and the service provider is referred to as the Service Level Agreement (SLA). A Service Level Agreement is a formal agreement between two or more entities that is reached after a negotiating activity with the scope to assess service characteristics, responsibilities and priorities of every part. A SLA may include statements about performance, tariffing and billing, service delivery and compensations. Every performance reporting may include only the QoS parameters agreed in the correspondent SLA [ITU-T Rec. E860, 2002].

The QoS attributes that are frequently part of an SLA vary repeatedly and to implement the contract, these parameters need to be carefully controlled.

The quality of composite services is measured by its end-to-end quality, rather than the quality of individual service components [Yu, Lin, 2004].

QoS is increasingly significant when composing services because a degrading QoS in one of the sub-services can dangerously disturb the QoS of the complete composition.

Today's service selection solutions do not focus on QoS support from the service requester view point, but they depend on service provider interpretation. Indeed, the current form of service composition is provider driven. A consumer may interact with a composite service without knowing much about the qualities of the services that underlie it [Karawash, Mcheick, Dbouk, 2014].

For composite services, the most important QoS issue is to define the service integration model and identify the optimal service selection to meet a user's QoS requirement.

The purpose of our paper is to present results of conducted literal review related to composition of services, quality of composite services and the functions of the qualities of the comprised sub-services.

Composition of Services

In general there are two types of services – simple services and composite services. A simple service realized offered actions directly, whereas a composite service can invoke sub-services in order to meet the users` needs.

According to reference [Pichanaharee, Senivongse, 2008] service composition is a process that selects units, called sub-services, and composes them into a workflow that represents a business process. The service composition process determines how and when sub-services should run and prepares them for execution. It presents the interdependency between the sub-services and their synchronization and prioritization.

A composite service model was presented in *Figure 1*.

A composite service can be defined as a directed graph, which specifies the order in which the sub-services are executed:

$SS = \{ss_i, i = 1 \ldots n\}$ is the sub-services, where ss_i is the *i-th* sub-service, and n is the total number of sub-services;

$L = \{l_k=(ss_i, ss_j), ss_i \neq ss_j\}$ is the set of links, where the element (ss_i, ss_j) indicate that ss_i immediately precedes ss_j.

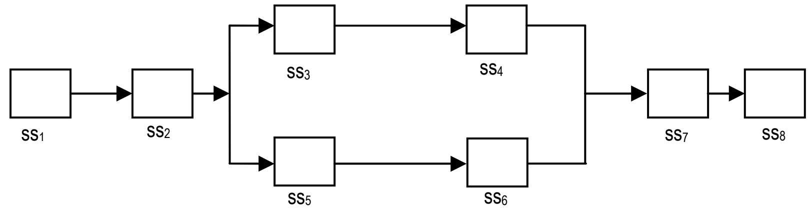


Figure 1. A composite service model with sequential and parallel sub-services

Sequential composite sub-services comprise a relation in which once a sub-service completes, another sub-service stars its execution (ss_3, ss_4).

Parallel sub-service comprises a relation in which two different services are going on in parallel (ss_3, ss_5) and synchronized at a certain point (ss_4, ss_6).

A composite service can involve loop sub-services too (*Figure 2*). A block of one or more sub-services is executed repeatedly with some probability, up to a maximum number of *k* executions.

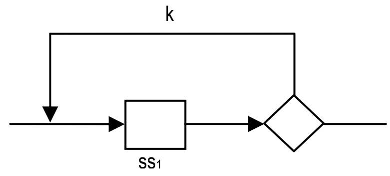


Figure 2. Loop

Another kind of services is substitute services, offered from other service provider. A substitute service is a service that performs the same or similar function as another service. If a service has more substitutes the demand for the service is more elastic. There are some conditions that increase the threat of substitutes:

- An attractive price of substitutes: The price of substitutes acts as a "ceiling" to the price of the subject service.
- Increased quality of substitutes: If the quality of a substitute is high, there is increased pressure to increase the quality of the subject service.
- Low switching costs to consumers: Switching costs to users can come in the form of monetary costs (transferring cell phone service) or lifestyle switching costs. Monetary costs effectively increase the price of the substitute services whereas lifestyle costs are more subjective and difficult to identify.

The evaluation of a service depends on its entire structure and the quality of the other sub-services invoked to compose such service.

Quality of Composite Service as a Function of the Quality of the Comprised Sub-Services

For the purpose of our study we proposed following QoS attributes:

- response time (RT);

The response time is defined as "A time period between the time when a sending of response request is triggered and the time when its response is received by the response confirmation role object [ITU-T Rec. X.748 (1999 E)].

The response time is the interval of time between the moment when a user requests the service and the moment when the user receives the response. Usually the response time includes service time (T_{sr}) and transmission time (T_{tr}). In a composite service, the transmission time is the time needed to send a user's request to the first server, to pass the result of one server to the next server in the sequential model, and to return results to the user from the last server in the server chain.

$$RT = T_{sr} + T_{tr} \quad (1)$$

- cost (C);

Cost includes service cost (C_{sr}) and transmission cost (C_{tr}).

$$C = C_{sr} + C_{tr} \quad (2)$$

- availability (A);

Availability is defined as availability of an item to be in a state to perform a required function at a given instant of time or at any instant of time within a given time interval, assuming that the external resources, if required, are provided. [ITU-T Rec. E802, 2007].

Availability is the probability that a service is available. Usually it is computed from historical data.

$$A = \frac{T_a}{T_t} \quad (3)$$

where T_a is amount of time that service is available; T_t is total time monitored [Yu, Lin, 2004].

- reliability (R).

Reliability is a probability that an item can perform a required function under stated conditions for a given time interval.

Reliability is the probability that a request is correctly handled within the expected time. It is computed from historical data.

$$R = \frac{N_s}{N} \tag{4}$$

where N_s is number of requests successfully responded and N is the number of total requests [Yu, Lin, 2004].

If the composite service involves only sub-services executed in a fixed sequential order (*Figure 3*) the outputs of some sub-services are the inputs of others and the output quality of every sub-service depend on the input quality.

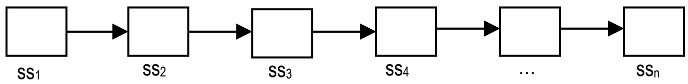


Figure 3. Sub-services executed in a fixed sequential order

In this case the QoS attributes of a composite service are as follow:

$$RT_{cs} = \sum_{i=1}^{n} RT_{SSi} \tag{5}$$

$$C_{cs} = \sum_{i=1}^{n} C_{SSi} \tag{6}$$

$$A_{cs} = \prod_{i=1}^{n} A_{SSi} \tag{7}$$

As stated above, sub-services are considered to be operating in sequential order. If failure of either of the sub-services results in failure of the combination. A composite service can be provided only if all sub-services are available. From this it follows that the combined availability is a product of the availability of all sub-services.

The implications of the above equation are that the combined availability of the composite service is always lower than the availability of its sub-services.

The reliability of the composite service follows a multiplicative rule and is the product of the reliability of every sub-service in the composition.

$$R_{cs} = \prod_{i=1}^{n} R_{SSi} \tag{8}$$

According to reference [Gupta, Londhe, Bhosale, 2011] QoS is a function of many factors having different weights ($w_1 - w_4$).

In the case in point the QoS of a composite service can be presented as shown in equation 9.

$$QoS = f\left(w_1 \times RT_{cs} + w_2 \times C_{cs} + w_3 \times A_{cs} + w_4 \times R_{cs}\right) \tag{9}$$

Weights might be calculated by simulating the end-to-end scenario and calculating the overall contribution of each parameter to QoS.

Limited work has address the more complicated cases when the composite services involve sequential and parallel sub-services but most of them are for web service composition, and for sub-services offered by different service providers.

A review of QoS-based web service selection, proposed methods and evaluation approaches are presented in *Table 1*.

Table 1. QoS-based web service selection - proposed methods and applied QoS evaluation approaches

Authors	Methods	QoS evaluation
Zeng et al., 2003, 2004 Ardagna, Pernici, 2005, 2007	Multiple Criteria Decision Making technique with Simple Additive Weighting and Linear Programming	Simple aggregation function
Liu et al., 2009	Rang based Algorithm with Simple Additive Weighting	Simple aggregation function
Doshi et al., 2004	Markov Decision Processes	Simple aggregation function
Alrifai, Risse, Nejdl, 2012	Hybrid approach	Simple aggregation function
Li, Wang, Lim, 2010	Monte Carlo method	Aggregation function with trust value and trust weight

Conclusion

One of the key issues in service composition is that of predicting and managing the QoS of composite services. QoS is dynamic and depends on many factors. QoS should be measured continuously.

It is important to estimate the QoS of a composite service based on the quality of sub- services to make sure that the composition can satisfy the expectations of end users.

The overall QoS for a composite service is computed by aggregating the QoS of its sub-services according to their structure and relation.

In order to ensure high QoS of the composite service which involves only sub-services executed in a fixed sequential order the special attention should be paid to sub-services with the highest weights. On the other hand, it is necessary to ensure the availability and reliability of each sub-service.

Our future work includes calculating QoS for composite services with more complex types of sub-services (e.g., parallel sub-services, overlapping loops) and extending the proposed method to address the problem of QoS-aware service composition.

Acknowledgement

This work is coordinated under EU COST Action IC 1304 entitled "*Autonomous Control for a Reliable Internet of Services*" and is financed by Bulgarian NSF Project DCOST 01/20.

Bibliography

[Yu, Lin, 2004] Tao Yu, Kwei-Jay Lin, Service Selection Algorithms for Web Services with End-to-end QoS Constraints, Proceedings of the IEEE International Conference on E-Commerce Technology, 0-7695-2098-7/04, pp. 1-8, 2004.

[Pichanaharee, Senivongse, 2008] Koramit Pichanaharee, Twittie Senivongse, QoS-Based Service Provision Schemes and Plan Durability in Service Composition, DAIS 2008, LNCS 5053, pp. 58–71, 2008.

[Karawash, Mcheick, Dbouk, 2014] Ahmad Karawash, Hamid Mcheick, Mohamed Dbouk, Quality-of-Service Data Warehouse for the Selection of Cloud Services: A Recent Trend, Z. Mahmood (ed.), Cloud Computing, Computer Communications and Networks, Springer International Publishing Switzerland 2014, pp. 257-276, 2014.

[Dan et al, 2004] Dan A, Davis D, Kearney R, Keller A, King A, Kuebler D, Ludwig H, Polan M, Spreitzer M, Youssef A, Web services on demand: WSLA-driven automated management. IBM Syst J 43(1):136–158, 2004.

[Gupta, Londhe, Bhosale, 2011] Priya Gupta, Priyadarshini Londhe, Arvind Bhosale, IPTV End-to-End Performance Monitoring, Advances in Computing and Communications, First International Conference, ACC 2011, Kochi, India, July 22-24, 2011, Proceedings, Part IV, 2011.

[ITU-T Rec. X.748, 1999 E] ITU-T Rec. X.748 (1999 E), Information technology – Open Systems Interconnection – Systems management: Response Time Monitoring Function, 1999.

[ITU-T Rec. E802, 2007] ITU-T Rec. E802, Framework and methodologies for the determination and application of QoS parameters, 2007.

[ITU-T Rec. E800, 2008] ITU-T Rec. E800, Definitions of terms related to quality of service, 2008.

[ITU-T Rec. E860, 2002] ITU-T Rec. E860, Framework of a service level agreement, 2002.

[ITU-T Rec. P.10/G.100, 2006] ITU-T Rec. P.10/G.100, Definition of Quality of Experience (QoE), 2006.

[Zeng et al., 2003] Zeng, L., Benatallah, B., Dumas, M., Kalagnanam, J., and Sheng, Q. Z.,. Quality driven web services composition. Proceedings of the 12th international conference on World Wide Web. 12-14 May. ACM. pp. 411-421, 2003.

[Zeng et al., 2004] Zeng, L., Benatallah, B., Ngu, A. H., Dumas, M., Kalagnanam, J., and Chang, H., QoS-aware middleware for web services composition. Software Engineering, IEEE Transactions on. 30(5), 311-327, 2004.

[Ardagna, Pernici, 2005] Ardagna, D., and Pernici, B., Global and local QoS constraints guarantee in web service selection. 2005 IEEE International Conference on Web Services (ICWS 2005). July 11-15. Orlando , Florida, USA, IEEE Computer Society. pp. 805-806, 2005.

[Ardagna, Pernici, 2007] Ardagna, D., and Pernici, B., Adaptive service composition in flexible processes. IEEE Transactions on Software Engineering. 33(6), 369-384, 2007.

[Liu et al., 2009] D. Liu, Z. Shao, C. Yu, and G. Fan.,A heuristic QoS-aware service selection approach to web service composition. In: Proceedings of the Eigth IEEE/ACIS International Conference on Computer and Information Science, Phoenix, Arizona, USA, pp. 1184-1189, 2009.

[Doshi et al., 2004] Doshi, P., Goodwin, R., Akkiraju, R., and Verma, K., Dynamic workflow composition using markov decision processes. Web Services, 2004. Proceedings. IEEE International Conference on. 6-9 July. San Diego, California, USA : IEEE. pp. 576-582, 2004.

[Alrifai, Risse, Nejdl, 2012] Alrifai, M., Risse, T., Nejdl, W. A hybrid approach for efficient Web service composition with end-to-end QoS constraints. ACM Transactions on the Web (TWEB). 6(2), Article No. 7, 2012.

[Li, Wang, Lim, 2010] Li, L., Wang, Y., Lim, E.-P., Trust-Oriented Composite Service Selection with QoS Constraints. J. UCS. 16(13), 1720-1744, 2010.

Authors' Information

Anna Otsetova – *University of Telecommunications and Post, Sofia, 1 Acad. St. Mladenov Str, Sofia 1700, Bulgaria, E-mail: a.otsetova@utp.bg*

Major Fields of Scientific Research: Quality of services, Operation management, Postal services, Customer satisfaction

Emiliya Saranova – *Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences, Acad. G. Bonchev Str., Block 8, 1113 Sofia, Bulgaria]*

University of Telecommunications and Post, Sofia, 1 Acad. St. Mladenov Str, Sofia 1700, Bulgaria, E-mail: e.saranova@utp.bg

Major Fields of Scientific Research: Information modeling, General theoretical information research, Multi-dimensional information systems

AN APPROACH FOR ARCHITECTURAL SOLUTIONS ESTIMATION

Olena Chebanyuk, Dmytro Povaliaiev

***Abstract**: Quick and effective estimation of architectural solutions that are represented as UML class diagrams allows obtaining quality information for further software development. One of the approaches to estimate class diagrams is verifying them according to SOLID design principles. Class diagrams, which follow the SOLID design principles, simplify the process of both architectural solutions and code reusing. Also such operations as class diagram refinement, merging, comparison, and analysis are performed more easily, too.*

An approach for estimation of class diagram according SOLID design principles is proposed in this paper. Entire information for class diagram is stored in XMI file that is created by software modeler environments. Estimation of class diagram correspondence to SOLID design principles is based on the automatic parsing of XMI file containing class diagram. Firstly class diagram classes and interfaces are identified from XMI file. Then relationships between them are defined. After recognizing class diagram structure queries for checking whether class diagram corresponds to every of SOLID principle are performed. Analytical foundation for forming such queries, based on predicate logic, was proposed in paper [Chebanyuk, 2016].

Then, architecture of proper software tool is described in the paper. Grounding for choosing technology for effective proceeding of XMI file content is represented. Also case study for using designed tool for class diagram estimation is represented.

***Keywords**: Abstract Syntax Tree (AST), Software Architecture, SOLID Design Principles, Class Diagram, XML Metadata Interchange (XMI), Object Constraint Language (OCL), LINQ query.*

***ITHEA Keywords**: D2 software engineering, D 2.0 Tools,*

***ACM Keywords**: Software and its engineering, Software system structure, Software system model, Model driven system engineering.*

Introduction

The more complex the software system becomes, the more complicated architecture it requires. The architecture is prominent for the software as it defines most of the technical aspects (maintainability, flexibility, extensibility etc.) that influence both the development process and the quality of the software system.

The most widespread approach of software development lifecycle management nowadays is Agile. Every operation in software development lifecycle management is performed by means of software model processing. Software models are represented as UML diagrams. They play both cognitive and communicative roles in collaboration between stakeholders and between stakeholders and customers [Chebanyuk and Markov, 2015 ICP].

As the customer may change software requirements, consequently all software development artifacts can be modified. Updated software models contain actual information about requirements, design, test cases, and other software artifacts.

Software models, reflecting architectural solutions, are modified too. They serve for many tasks, for example simplify a constant revision of the architecture of software being developed.

An important step in the architecture solutions designing is their verification. Unfortunately, performing of this step requires many facts that are complex to be formalized. In some companies this step is usually missed, therefore leading to mistakes in design and higher overall project cost. Lack of software tools, which allow to perform class diagram verification in automatic mode, become a motivation for authors fo provide this research.

This paper continues investigation, started in paper [Chebanyuk, 2016]. It was grounded using predicate logic for class diagram processing. Then corresponded predicate expressions for estimation class diagrams for accordance to SOLID design principles were proposed. Also criteria for estimation of analysis results were formulated.

Contribution of this paper is grounding choosing of technology to process class diagrams in automated mode and formulating LINQ queries for XMI file parsing that stores class diagram. Description of an approach for class diagram processing, which represent SOLID recondition algorithm and software architecture designing, is also proposed in this paper.

Related papers

Proper approach to software design process is based over understanding of some fundamental principles and guidelines. One can see a variety of articles describing benefits of sticking to SOLID design principles and drawbacks of neglecting them, for example [Ocampo J, 2009], [Lakhal F., 2012], [Chebanyuk, 2016]. These papers help to compose approaches for estimation of architectural solutions characteristics in accordance to SOLID design principles.

Author of book [Ocampo, 2009] investigates SOLID principles in details and proposes practical recommendations for verification class diagram and designing code. Many advices for flexible designing from the reusability point of view are listed in this book.

Question of architectural solution estimation are important for processes where it is necessary to obtain stable model of problem domain. Consider a paper, which touches question of estimating structural characteristics of software models, represented as UML diagrams.

Article [Sanchez et al 2015] considers aspect for designing stable component diagrams for adaptive systems. Authors introduce quality attributes of these systems as a their non-functional properties. Respectively, such properties are classes attributes and methods. Authors also touch a problem of reusing domain entities, transferring attributes of system, developed earlier to new ones.

Authors propose a framework for the specification, measurement and optimization of quality attribute properties expressed on top of feature models. Then it was shown how these properties can be specified by means of feature attributes and evaluated with quality metrics in the context of feature models. Structure of diagram, reflecting feature models modified by means of aggregate functions over the features is changed in the run time. For instance, if properties of component or its structure are changed, the system structure should be changed at run time and the system takes to reconfigure itself. Authors propose configuration selection algorithms for enrich obtained model by necessary amount of qualities to design quality model.

Consider researches of software tools designing for optimizing architectural solutions.

Other paper that manage quality of software models by means of analyzing interconnection between domain entities is devoted to modifying of problem domain profiles when borders of application domain are changed [Lakhal, 2012]. Consequently the structures of profiles are changed too.

The models have then to be fitted to the new profiles version. Implementing designed tool authors propose to refuse from manual adaptation of software models, using the combination of Eclipse plugins to compare two software models. Then defined differences was a cause to make some changing in profiles and finally to adapt the models to the new version of the UML profile [EMF, 2011].

Paper [Sielis et al, 2017] describes the design, development and evaluation of a software prototype, named ArchReco, an educational tool that employs two types of Context-aware Recommendations of Design Patterns, to support users (students or professionals) who want to improve their design skills when it comes to training for High Level Software models. The tool’s underlying algorithms take advantage of Semantic Web technologies, and the usage of Content based analysis for the computation of non-personalized recommendations for Design Patterns. The recommendations’ objective is to support users in functions such as finding the most suitable Design Pattern to use according to the working context, learn the meaning, objectives and usages of each Design Pattern. Authors also design an environment for visualization of design of high level software diagrammatic models that obtained after recommendations gathering and processing. Design patterns are chosen according recommendations. These recommendations are proceeded involving semantic web technologies.

Nowadays there are lots of tools that allow us to automate many processes within software development life cycle, but they still pay not enough attention to keeping track of correspondence to SOLID principles. In fact there are no tools that can grant SOLID consistency.

Using an advanced modeling environments, as IBM Rational Software Architect [IBM, 2015] or Eclipse plugins modeling software, for example Papyrus [Eclipse, 2015], class diagram may include constraints to precise requirements of application domain.

The most widespread Object Constraint Language (OCL) [OCL, 2014] performs check of class diagram components for accordance requirements of application domain. Theoretically, OCL may be used for checking whether class diagram corresponds to SOLID design principles. But such operation should be performed manually. For example let us consider the procedure of checking whether class conforms to single responsibility design principle by method, proposed in [Chebanyuk, 2016]. Idea to do this, proposed by authors is to define the number of public methods in class. Using OCL it is necessary to type:

context class_name:OCL expression

for every class in class diagram. To observe class diagram for checking such feature of all its classes will take less time in comparison with composing OCL expression for every class.

The most convenient way for class diagram estimation will be creating of software tool or plugin for development environment to perform this task.

Investigation of class diagram storing format

Many modeling environments store information about UML diagrams in formats representing software model structure as a hierarchical tree. The key advantage of such modeling environments is using Xml Metadata Interchange (XMI) standard [XMI, 2015]. This allows to store relations between classes, and, this way, it allows user to design constraints to apply on these relations. However, there are few environments that support XMI. These are IBM Rational Software Architect, Papyrus, and UML Designer [Eclipse, 2015]. Other environments use their own not unified format for storing UML diagram data (JSON schemas, custom XML tag systems) [ASTM™, 2011].

XMI representation corresponds to theoretical approach Abstract Syntax Tree (AST). An AST is a formal representation of the syntactical structure of software that is more amenable to formal analysis techniques than is the concrete or surface syntax of software. Construction of ASTs typically involves the use of parsing technologies. AST model structures permit the expression of compositional relationships to other language constructs and provide a means of expressing a set of direct and derived properties associated with each such language construct [ASTM™, 2011].

The data structures from which the abstract syntax trees are composed provide an exhaustive collection of formal compositional elements for a language. These language model elements (or constructs) are generally defined in a type (or class) hierarchy. There are many ways to define these ASTs. The AST may be derived from an analytical process that can be applied to the surface syntax of the software asset or may be captured through a process that involves the application of rewrite rules to other data structures. For instance, a common, or language-neutral, AST model might be generated by the application of rewrite rules that generate a language specific AST model of some application, or a generic AST model might be generated directly from a UML class diagram or action diagram by means of a series of refinement rules. An AST may be an invertible representation. In other words, it may be possible to traverse the AST and reconstruct the “surface syntax” of the legacy system or reconstitute it in textual form from the abstract structures. An AST may be augmented; it may be analyzed and updated using additional structures that describe other properties about the software. Common analyses that augment an AST with additional properties include constraint analysis, data-flow analysis, control flow analysis, axiomatic analysis, and denotational analysis. ASTs are generally augmented with additional analyses layers, such as type analysis, control-flow analysis, or data-flow analysis (to support code optimization). Augmentation may also support capture of software engineering metrics and documentation. Having a standard metamodel to represent ASTs will facilitate interchange at a foundational level for all architecture-driven modernization work. Hence, the AST provides an appropriate formalism for the derivation of properties required for detailed knowledge discovery. Formally, in computer science, an AST is a finite, labeled, directed tree, where the internal nodes are labeled by operators, and the leaf nodes represent the operands of the node operators. Thus, the leaves have nullary operators, i.e., variables or constants. In computing, it is used in a parser as an intermediate between a parse tree and a data structure, the latter which is often used as a compiler or interpreter's internal representation of a computer program while it is being optimized and from which code generation is performed. The range of all possible such structures is described by the abstract syntax. An AST differs from a parse tree by omitting nodes and edges for syntax rules that do not affect the semantics of the program. The classic example of such an omission is grouping parentheses, since in an AST the grouping of operands is explicit in the tree structure. In contrast, an Abstract Semantic Graph (ASG) is a data structure used in representing or deriving the semantics of an expression a formal language (for example, a programming language). The scope of article includes only XMI-stored diagrams, considering that it is the most expressive and informative of the others [ASTM™, 2011], [Newcomb, 2005].

Abstract syntax tree helps to design XMI schemas. XMI schemas have hierarchical structure and tree serve to represent hierarchical structure of class diagram. Every XMI schema consists of the following declarations [XMI, 2015]:

1. An XML version processing instruction;
2. An optional encoding declaration that specifies the character set, which follows the ISO-10646 (also called extended Unicode) standard;
3. Any other valid XML processing instructions;
4. A schema XML element;
5. An import XML element for the XMI namespace;
6. Declarations for a specific model. Every XMI document consists of the declarations, unless the XMI is embedded in another XML document;
7. An XML version processing instruction.

Conclusion from the review

Review of software engineering standards, related to representing of information about software models shows the next:

- general approach to store software models is based on hierarchical representation of text information;
- rules for composing of this representation are based on abstract syntax tree;
- the format to represent hierarchical structure in readable manner is XML, namely the XMI;
- to choose software tool for effective (quick and strick) processing of such an information it is necessary to investigate text query based techniques.

Task

Task is to propose a technique and a software tool for estimation class diagram for accordance them to SOLID design principles. In order to accomplish this task it is necessary to do the following:

- investigate the format of class diagram storing (see previous chapters);
- ground choice of software techniques for class diagram verification;
- propose techniques for class diagram verification, that is based on analytical approach, proposed in paper [Chebanyuk, 2016];
- represent an algorithm for software tool working;
- describe a software architecture components.
- represent case study;
- verify case study results by means of analytical apparatus, proposed in [Chebanyuk, 2016].

Grounding of the choice of development environment and tools for XMI files processing

Concretely, each modeling environment tool stores and manages the models with its own internal format or its own “dialect,” even if a standard format is adopted. To improve interoperability between modeling tools, standardization bodies have defined specific model interchange languages.

The best-known model interchange language is XMI (XML Metadata Interchange), a standard adopted by OMG for serializing and exchanging UML and MOF models. Unfortunately, even if this has been devised as an interchange format, different tools still adopt it with different flavors.

Therefore, to provide a framework for operations on XMI-compliant diagrams, the information should be processed by the Model-to-Text transformation [XMI, 2015].

There are several ways to perform this operation and each of them provides a specific tool for extraction of information from the text. The most widely used is a regular expression (regex) engine that provides a special notation to complete this task (e.g. .”*name="\K.*?(?=".*)” to extract the name of the class). However, regular expressions are suitable for processing of natural texts and other non-structured information, but perform much worse when the task is to extract information from the text with strict and specific structure like the XML is. Other downsides include low level of extensibility and readability and therefore increased cost of maintenance.

The other possible solution is a Language INtegrated Queries (LINQ) featured in .NET platform:

class.Attributes().FirstOrDefault(f => f.Name.LocalName.ToLower() == “name”).

LINQ extracts the name of the class. Such queries allow easy manipulation of XML documents via elements of functional programming like Lambda expressions. By using this approach, one can benefit from easy-to-use and logical syntax of predicates.

Therefore, the LINQ was chosen for proceeding class diagram. It is a more robust technology that decreases the time the development cycle takes. This decision defines the overall technological stack of the project - the cross-platform and open source .NET Core framework.

Being based on a cross-platform technology, application of plug-in for class diagram verification may be used in many different ways and embedded into various different applications and systems. Therefore the core functionality should be implemented as a portable library in a .NET Standard format. It provides a full-featured transformation and analysis Application Program Interface (API) for any types of applications: desktop, web, command-line.

Because of increasing role of web technologies in the modern software development industry, the reference system is implemented both in the form of REST-compliant web application and command-line tool. Web application allows it for user to easily analyze diagrams and manage the output of results. On the other hand, command line tool provides a simple and straightforward interface suitable for both desktop and server environment.

Designing of patterns for identification of class diagram components from XMI file

XMI stores the models in a tree-like structure where the root element is “XMI” and its descendants store information about UML entities such as classes, interfaces, and relations between them. According to XMI standard each entity should have a unique string ID that allow it to be referenced by the other entities. Self-sufficient such as class, interface, or relation are represented in the form of “packagedElement” XML elements with the corresponding “type” attribute. Their properties such as name shown or visibility level are specified with additional attributes. The embedded entities such as operations and attributes are represented as child elements of corresponding class and interface tags as “ownedOperation” and “ownedAttribute” respectively.

The generalization is represented in the form of attribute of the class that is a derived one (the same scheme is applied to interface realization). Such links are marked as “generalization” and “interfaceRealization” tags. Generalization stores a string ID of the parent class in its single “general” attribute, while the interface realizaition has three of them: (1) the “supplier” with identifier of the interface being implemented, (2) the “client” with ID of the class that implements it and (3) the “contract” with ID of the contract specified by the interface (suitable for contract programming, stores the same ID as “supplier” by default).

Table 1. Patterns and LINQ queries for XMI file processing

Aim of LINQ request	Fragment of XMI file	LINQ queries to parse XMI tags
Extract all classes	<packagedElement xmi:type="uml:Class" xmi:id="_qu0M8ASgEeeImrsnVV7-jw" name="Class2">	var class = diagramDoc.Descendants().Where(w => w.Name.LocalName.ToLower() == "packagedelement" && w.Attributes().Any(a => a.Name.LocalName.ToLower() == "type" && a.Value == "uml:Class"));
Get all operations of the particular class	<ownedOperation xmi:id="_HUTN4A-dEeeRAcEtZoXAZQ" name="Create"/>	var generalization = child.Descendants().FirstOrDefault(w => w.Name.LocalName.ToLower() == "ownedoperation");
Get all attributes of the particular class	<ownedAttribute xmi:id="_Ja4tMAZYEeeIgspGntkIFA" name="attribute1" visibility="private"> </ownedAttribute>	var generalization = child.Descendants().FirstOrDefault(w => w.Name.LocalName.ToLower() == "ownedattribute");

Aim of LINQ request	**Fragment of XMI file**	**LINQ queries to parse XMI tags**
Extract information about interface realizations of the particular class	<interfaceRealization xmi:id="_2WPxUA-cEeeRAcEtZoXAZQ" supplier="_u_OgcA-cEeeRAcEtZoXAZQ" client="_qu0bMA-cEeeRAcEtZoXAZQ" contract="_u_OgcA-cEeeRAcEtZoXAZQ"/>	var implementations = child.Descendants().Where(w => w.Name.LocalName.ToLower() == "interfacerealization");
Get information about attributes of the particular class representing ends of Associations' relation	<ownedAttribute xmi:id="_WaWnBA-dEeeRAcEtZoXAZQ" name="testClass2" visibility="private" type="_PE6DcA-dEeeRAcEtZoXAZQ" association="_WaWnAA-dEeeRAcEtZoXAZQ"> </packagedElement>	var associationAttributes= xmlClass.Descendants().Where(w => w.Name.LocalName.ToLower() == "ownedattribute" && w.Attributes().Any(a => a.Name.LocalName.ToLower() == "association")&& ! w.Attributes().Any(a => a.Name.LocalName.ToLower() == "aggregation"));

Aim of LINQ request	**Fragment of XMI file**	**LINQ queries to parse XMI tags**
Get information about attributes of the particular class representing ends of Aggregations' relation	<ownedAttribute xmi:id="_WaWnBA-dEeeRAcEtZoXAZQ" name="testClass3" visibility="private" type="_PE6DcA-dEeeRAcEtZoXAZQ" association="_WaWnAA-dEeeRAcEtZoXAZQ" aggregation="composite"> </packagedElement>	var aggregationAttributes= xmlClass.Descendants().Where(w => w.Name.LocalName.ToLower() == "ownedattribute" && w.Attributes().Any(a => a.Name.LocalName.ToLower() == "association")&& w.Attributes().Any(a => a.Name.LocalName.ToLower() == "aggregation"));
Extract all interfaces	<packagedElement xmi:type="uml:Interface" xmi:id="_u_OgcA-cEeeRAcEtZoXAZQ" name="TestInterface"/>	var interface = diagramDoc.Descendants().Where(w => w.Name.LocalName.ToLower() == "packagedelement" && w.Attributes().Any(a => a.Name.LocalName.ToLower() == "type" && a.Value == "uml:Interface"));
Extract all Association and Aggregation relations on class diagram	<packagedElement xmi:type="uml:Association" xmi:id="_WaWnAA-dEeeRAcEtZoXAZQ" memberEnd="_WaWnAQ-dEeeRAcEtZoXAZQ _WaWnBA-dEeeRAcEtZoXAZQ"/>	var xmlAssociations = diagramDoc.Descendants().Where(w => w.Name.LocalName.ToLower() == "packagedelement" && w.Attributes().Any(a => a.Name.LocalName.ToLower() == "type" && a.Value == "uml:Association"));

Description of class diagram estimation algorithm

The defined task consists of two parts: the class diagram importing and further verification. Steps for class diagram refinement algorithm are represented below:

1. Obtaining XMI representation of class diagram. Performing this step model to text transformation is done. This action is executed by software modeling environment.
2. Defining classes of class diagram.
3. Defining interfaces of class diagram.
4. Defining relations between classes and interfaces.
5. Checking class diagram to accordance to SOLID design principles.
6. Generating a report of class diagram accordance to SOLID design principles
7. Modification of class diagram by software architector in modeling environment (optional).

Points 2-6 are performed by designed software tool by means of LINQ queries proposed in previous points.

The general sequence diagram of the model import and verification is represented at the Figure 1.

To perform the model verification of SOLID principles compliance, it should be first transformed from the XMI-compliant text representation to the corresponding data structure. Parsing of XMI file, containing XMI class diagram representation is performed in several passes. At the first pass, a class diagram entities (classes, interfaces, use cases and actors) are extracted along with their attributes (e.g. operations and properties for class diagrams). Those attributes include class diagram components that are used in XMI scheme as ends of relations (e.g. inheritance or association) between corresponding entities. On the next pass, the information about the relations is extracted and the edges of the graph are established. The direction of edges is determined by the direction of the original relations on which they are built upon.

The developed application represents the class diagram (software model) as a directed graph, where the entities (classes and interfaces) are graph vertexes and nodes and relations between them (association, aggregation, composition, and inheritance) are the graph edges.

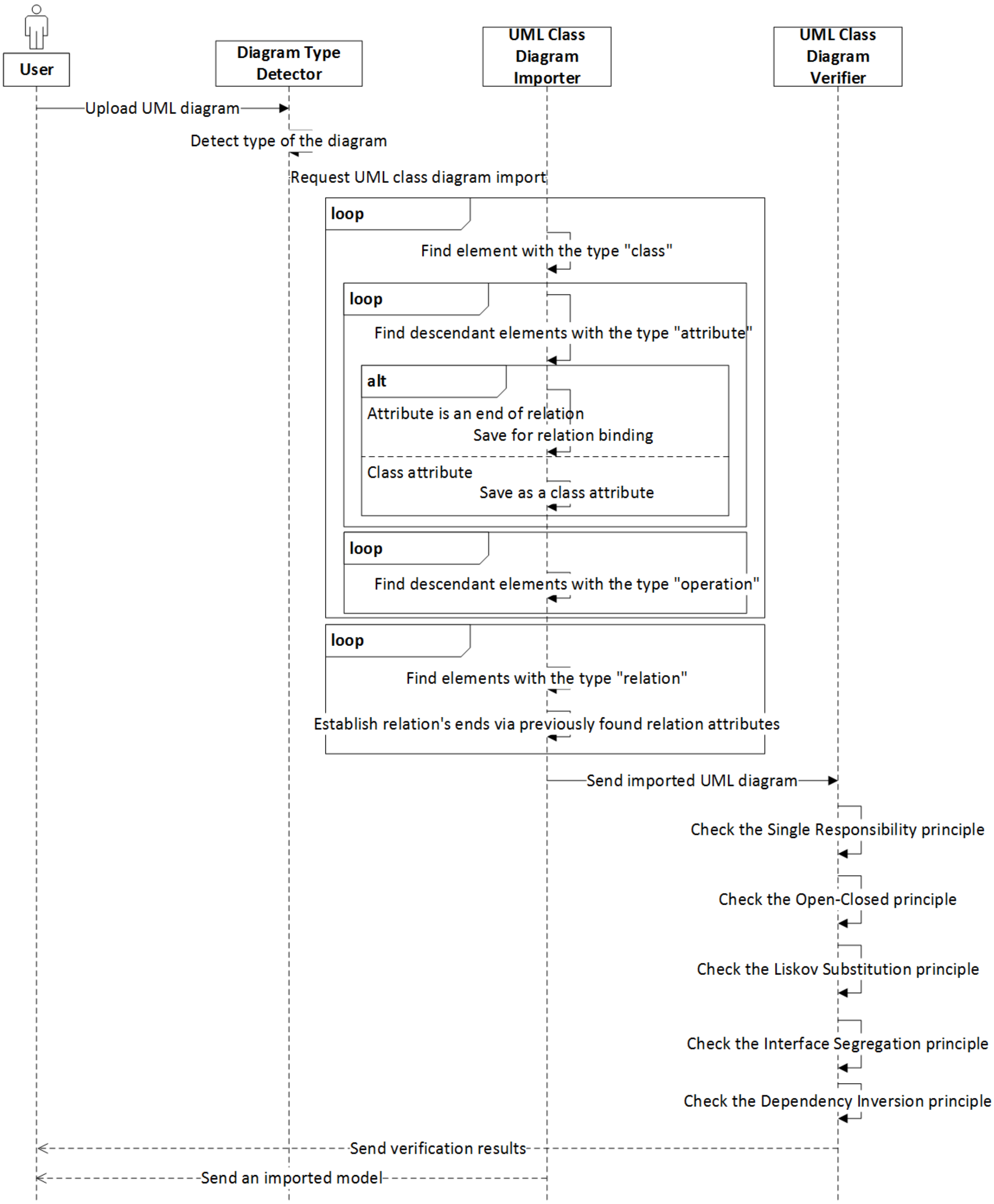


Figure 1. Sequence diagram of class diagram verification approach in accordance to SOLID design principles

The second task of designed approach is the software model verification to accordance to SOLID design principles. The verification module uses LINQ-to-Objects mechanism to provide the metrics of compliance using the syntax consistent with one used for extraction of information from the text representation of model.

Those LINQ queries verify all the SOLID design principles, namely:

- Single Responsibility;
- Open-Closed;
- Liskov Substitution;
- Interface Segregation;
- Dependency Inversion Principles.

Single Responsibility design principle is applied analyzing every class diagram entity separately. Other SOLID design principles are analyzed considering some of the interconnections between class diagram entities [Chebanyuk, 2016].

Software Product Architecture Description

The architecture of the software product is represented on Figure 2. The aim of designed software tool is to verify class diagram for accordance to SOLID design principles. Software architecture consists from several packages.

- “Model package” is used to prepare class diagram for further analysis;
- “Provider package” is used to extract the information from the XMI files and create the representation of the diagram on their basis;
- “Verification package” is used to provide verification tasks of model compliance to the SOLID design principles.

“Model package” contains classes for storing information about software model. These classes are UmlDiagram, UmlRelation, UmlStereotype. Also this package containes enumerations (UmlDiagramType, UmlRelationType and UmlVisibility). A list of classes that provide functionality of specific UML diagrams can be placed in separate sub-packages, namely UmlClass, UmlAttribute, UmlInterface, UmlOperation classes with IUmlClassMember interface for Class diagrams. They act as a software model entities and store properties like name of the entity, level of visibility etc. The collections of elements like members of particular class or classes of the diagram are represented as a dictionary with XMI entity ID acting as a key to facilitate quicker access.

“Provider package” contains classes for class diagram import. It has importer XmiClassDiagramImporter which is responsible to handling import class diagram into the project. “Verification package” containes a single ModelVerifier class that provides methods for verification of the whole diagram or particular class. LINQ queries (Table 1) contained here allow flexible verification and provide results for frontend parts of the application.

To address the problem of verification, the product contains five methods of ModelVerifier class for SOLID compliance check, each addressing their own software design principle:

- IsSingleResponsibilitySatisfied checks the number of public operations of a particular UML class or interface. To mark the whole diagram as compliant, all classes and interfaces must be compliant;
- IsOpenClosedSatisfied checks if the number of classes participating in class hierarchy is big enough;
- IsLiskovSubstitutionSatisfied checks that there are no operations defined in the base classes or interfaces and not implemented on lower level of class hierarchy. It is the most complex one, utilizing algorithms such as cyclical graph traversal to perform verification;
- IsInterfaceSegregationSatisfied works specifically for the interfaces and verifies if the number of operations is not too big. Same as Single Responsibility, this principle must implemented by all interfaces to make the diagram compliant;
- IsDependencyInversionSatisfied enumerates all the classes on the lowest level of hierarchy and checks that there are no association or composition relations with them;

The results of verification of UML Class diagrams are combined with the lists of recognized entities and relations between them to facilitate the detection of non-well-formed entities and their correction.

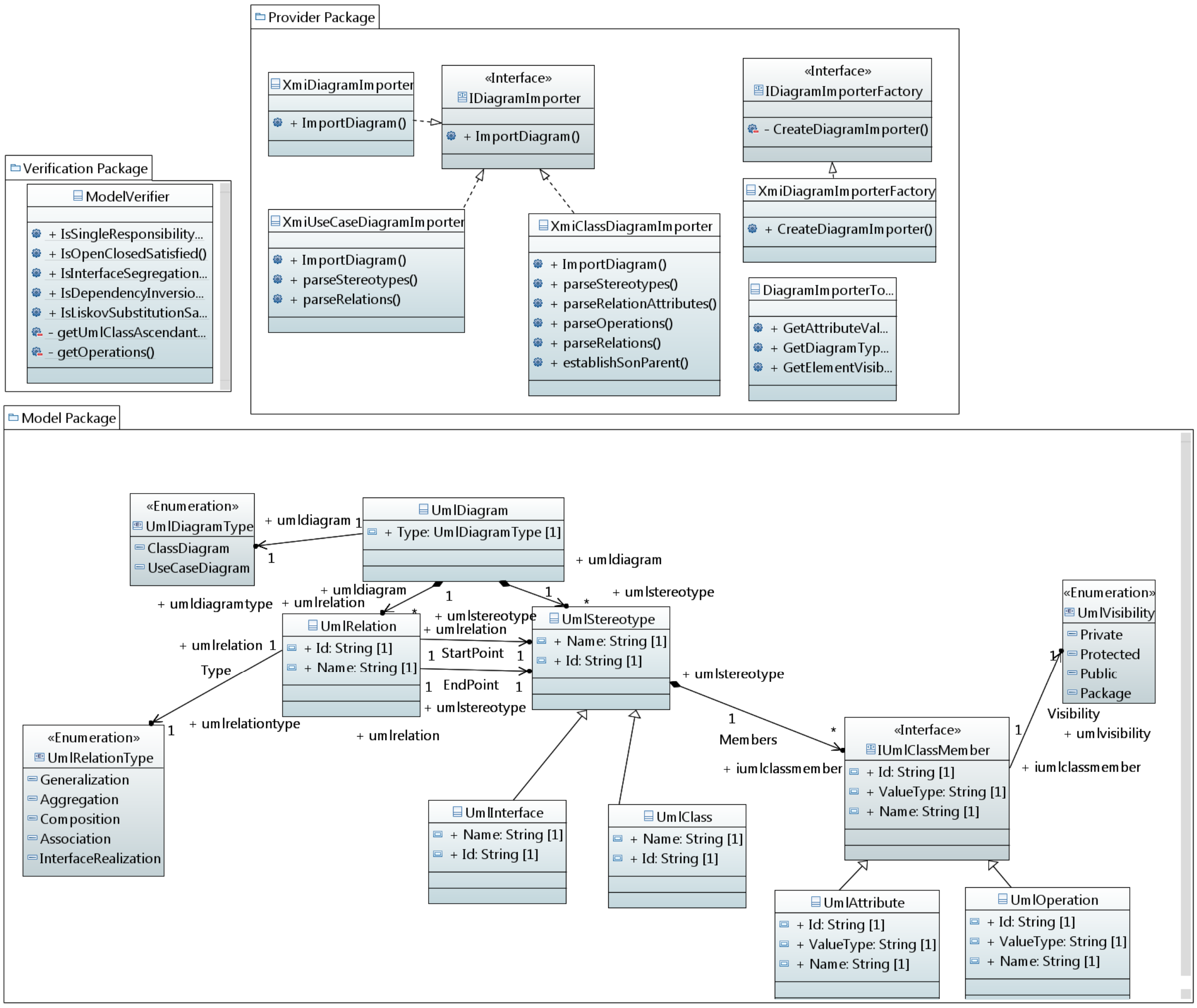


Figure 2 Architecture of software product for class diagram refinement

Case Study

To perform the case study, an example Class Diagram was created (Figure 3).

It contains several classes:

— Class1 contains a big amount of operations and several attributes;

— Class2 is a descendant of Class1 with one own attribute and is linked via aggregation with the Class3;

— Class 3 implements interfaces Interface1 and Interface2 each containing one operation;

— Class4 is connected with the Class3 via association.

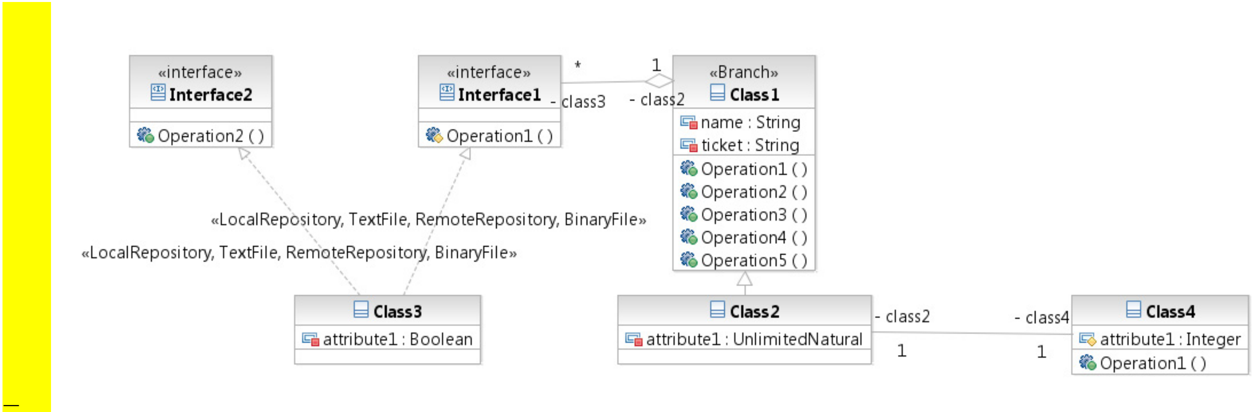


Figure 3 Example of Class Diagram

After selecting the correct type of diagram (XMI produced by IBM Rational Software Architect) and uploading the file (figure 4), the application begins the process of diagram verification.

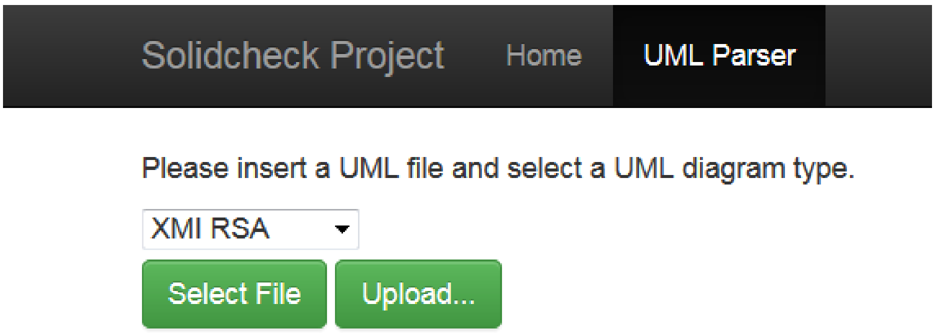


Figure 4 Diagram Upload Page

The result of class diagram estimation is represented on figures 5, 6 and 7.

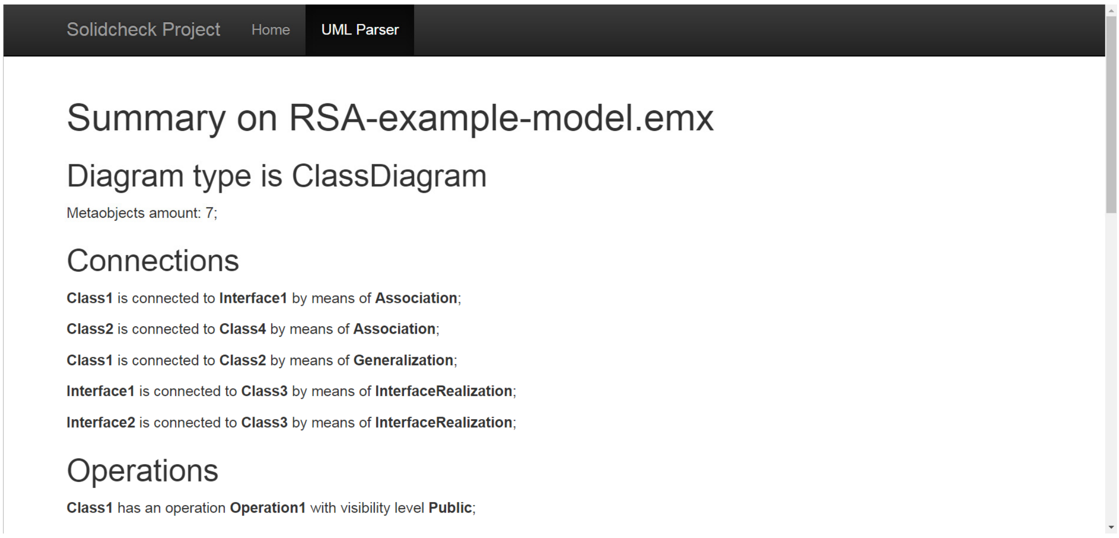


Figure 5 Results of Diagram Verification (Part 1)

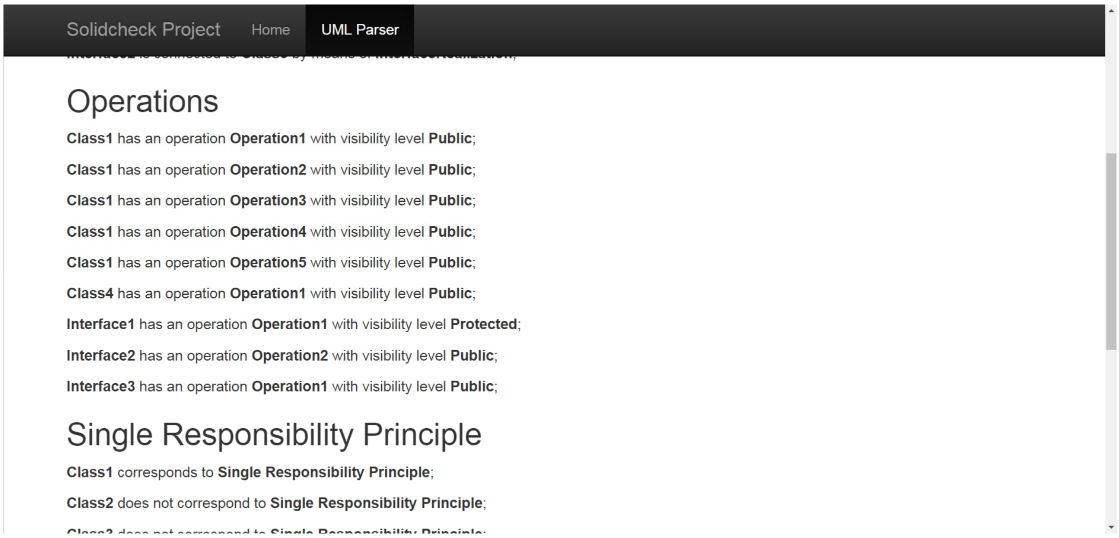


Figure 6 Results of Diagram Verification (Part 2)

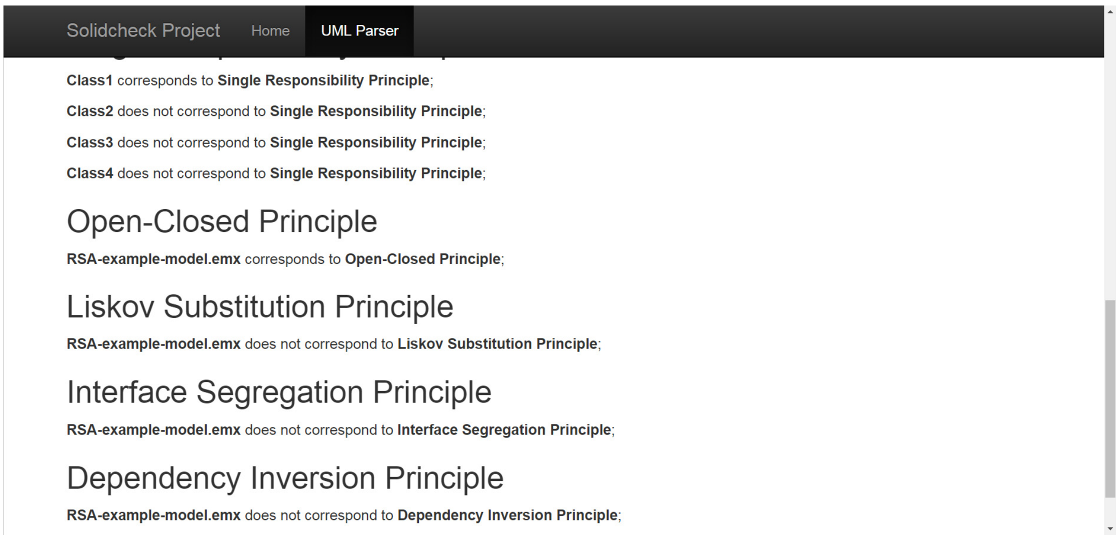


Figure 7 Results of Diagram Verification (Part 3)

Textual representation:

Summary on RSA-example-model.emx

Diagram type is ClassDiagram

Metaobjects amount: 7;

Connections

Class1 is connected to Interface1 by means of Association;

Class2 is connected to Class4 by means of Association;

Class1 is connected to Class2 by means of Generalization;

Interface1 is connected to Class3 by means of InterfaceRealization;

Interface2 is connected to Class3 by means of InterfaceRealization;

Operations

Class1 has an operation Operation1 with visibility level Public;

Class1 has an operation Operation2 with visibility level Public;

Class1 has an operation Operation3 with visibility level Public;

Class1 has an operation Operation4 with visibility level Public;

Class1 has an operation Operation5 with visibility level Public;

Class4 has an operation Operation1 with visibility level Public;

Interface1 has an operation Operation1 with visibility level Protected;

Interface2 has an operation Operation2 with visibility level Public;

Interface3 has an operation Operation1 with visibility level Public;

Single Responsibility Principle

Class1 corresponds to Single Responsibility Principle;

Class2 does not correspond to Single Responsibility Principle;

Class3 does not correspond to Single Responsibility Principle;

Class4 does not correspond to Single Responsibility Principle;

Open-Closed Principle

RSA-example-model.emx corresponds to Open-Closed Principle;

Liskov Substitution Principle

RSA-example-model.emx does not correspond to Liskov Substitution Principle;

Interface Segregation Principle

RSA-example-model.emx does not correspond to Interface Segregation Principle;

Dependency Inversion Principle

RSA-example-model.emx does not correspond to Dependency Inversion Principle;

As it can be seen, most of the classes does not correspond to the Single Responsibility principle because of number of their operations being too low to successfully perform even a single function. No other SOLID design principles except Open-Closed principle are satisfied.

Verifying case study results by means of analytical apparatus

Prove that considered class diagram corresponds to SOLID design principles.

Consider C as a set of classes in class diagram.

$$C=\{class1, class2, class3, class4\} \quad (1)$$

1. Single responsibility design principle

Number of public methods of class $c \in C$ is denoted as follows: $n(B_c^{public})$ where - B_c^{public} is a set of public methods of class $c \in C$ [Chebanyuk, 2016]. Consider that for performing one task classes should have no more than four public methods, namely $c \in C, n(B_c^{public}) \in \{3,4,5,6,7,8,9\}$,

$$class1 \in C, n(B_{class1}^{public}) = 5, class2 \in C, n(B_{class2}^{public}) = 0,$$
$$class3 \in C, n(B_{class3}^{public}) = 0, class4 \in C, n(B_{class4}^{public}) = 1 \quad (2)$$

Consider predicate expression for checking single responsibility design principle.

$$P(c, n(B_c^{public})) == c\ has\ n(B_c^{public})\ public\ methods$$

$$P(class1, n(B_{class1}^{public})) = class1\ has\ 5\ public\ methods = true,$$
$$P(class2, n(B_{class2}^{public})) = class2\ has\ 0\ public\ methods = false,$$
$$P(class3, n(B_{class3}^{public})) = class3\ has\ 0\ public\ methods = false,$$
$$P(class4, n(B_{class4}^{public})) = class4\ has\ 1\ public\ methods = false. \quad (3)$$

Expressions (3) mach results of software tool according single responsibility design principle. (Figure 5).

2. Open-Close SOLID design principle

Class diagram, represented in Figure 8 has structural characteristics that match to design pattern “Abstract Factory”. Consider OCL Constraints of this pattern [ICER, 2015]. Class diagram is also taken from [ICER, 2015].

Correspondence of abstract factory participants to elements of class diagram presented at the Figure 8 is described in Table 2.

Table 2 Correspondence of abstract factory participants to elements of class diagram

classes of class diagram on figure3	abstract factory participants (figure 8)
Class1	AWindows
Class2	motifWindow
Class3	PMwidgetFactory
Interface1	WidgetFactory

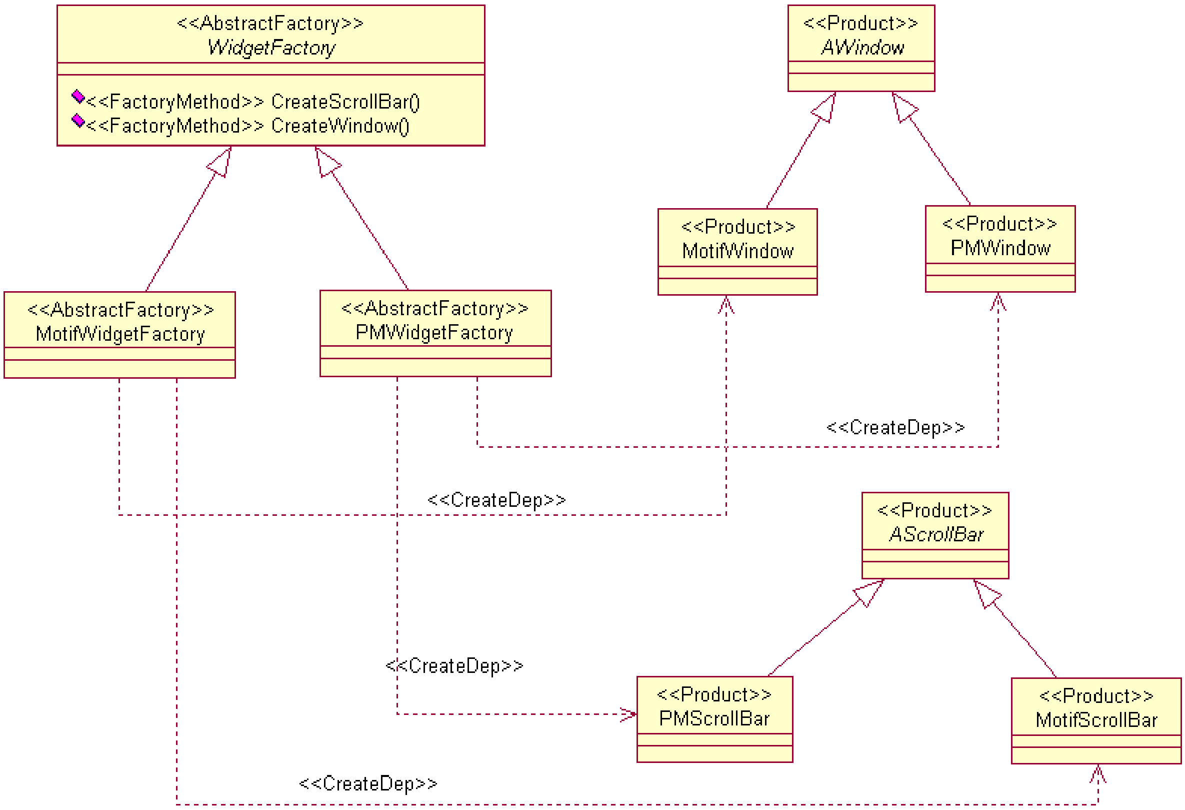


Figure 8. Abstract factory class diagram [ICER, 2015]

Rewrite OCL constrains from [ICER, 2015] to class diagram, represented on Figure 8:

context class1
inv: self.isRoot implies (not self.isLeaf and self.isAbstract)

inv: self.isAbstract implies (self.specialization->size() > 0 and self.specialization->forAll (c | c.child.oclIsKindOf(class1)))

inv: not self.isAbstract implies self.createDep->size() = 1

context WidgetFactory
inv: self.isRoot implies (not self.isLeaf and self.isAbstract)

inv: self.isAbstract implies (self.specialization->size() >= 1 and self.specialization->forAll (c | c.child.oclIsKindOf (WodgetFactory)))

inv: not self.isRoot implies (self.createDep->size() > 0)

Authors of [ECIR, 2015] formulate necessary structural characteristics of Abstract Factory design pattern. Represent characteristics, that match to classes in Figure 8, i.e. each abstract factory class hierarchy has an abstract base class, which needs to have at least two sub-classes (to support multiple families of products, one family per abstract factory).

In paper [Chebanyuk, 2016] set of operation *OPER* for interconnection between classes is defined.

$$OPER = \{inh, ass, aggr, comp\} \tag{4}$$

where: *inh* - inheritance, *ass* - association, *aggr* – aggregation, *comp* - composition.

Refer to the term “functionality” from paper [Chebanyuk, 2016]. Consider $c, c' \in C$. General idea of spreading class functionality $F(c)$, when c and c' are connected by means of $oper \in OPER$, is denoted as follows:

$$F(c)^{oper} = F(c) \cup F(c') \tag{5}$$

Sign ╰ depicts that functionality of class *c* ($F(c)$) is extended with functionality of class c' ($F(c')$).

The same is true when functionality of $c \in C$ is spreading by inheritance or including reference to $i \in I$. It is denoted as follows:

$$F(c)^{oper} = F(c) \cup i \tag{6}$$

At the figure 3 *class2* inherits *class1.* Denote it by the following:

$$\begin{aligned} &\exists class1 \in C, \exists class2 \in C, \\ &F(class2)^{inh} = F(class2) \cup F(class1) \\ &P(F(class2)^{inh}) = true \end{aligned} \tag{7}$$

The root of each product tree must be abstract.

At diagram (Figure 3) *class3* inherits interfaces. Thus, the functionality of this class $F(class3)$ is represented by the following:

$$\begin{aligned}
&\exists class3 \in C, I = \{\text{int}\,erface1\}, \\
&F(class3)^{inh} = F(class3) \cup \text{int}\,erface1 \\
&P(F(class3)^{inh}) = true
\end{aligned} \tag{8}$$

Expressions 7 and 8 show that to make some changes to class diagram it is not necessary to modify code inside classes. As class3 and class2 depend upon abstractions one can modify class diagram not performing changes to existing methods of the classes. The same conclusion was donw by the software tool (figure 5). Thus class diagram correspond to Open-Close Design principle.

3. Interface Segregation

In order to follow interface segregation principle it is necessary to define the next:

- number of methods in interfaces, that are parent for classes should be less than five;
- class, that inherits interfaces should not contain empty overridden methods

$$\begin{aligned}
&\exists c \in C, \exists i \in I, \\
&F(c)^{inh} = F(c) \cup i \\
&B_c^{public} = \{\beta_c^{public} \mid \beta_c^{public} \neq \varnothing\}
\end{aligned} \tag{9}$$

Consider interface1 and interface2 from class diagram. They are following to the first condition, namely, every interface contains one public method.

Consider class3, inheriting interface1 and interface2. The number of public methods in it is zero. Also there are no other methods in this class. The second condition for interface segregation principle is not performed.

$$\exists class3 \in C, \exists interface1 \in I, \quad (10)$$
$$F(c)^{inh} = F(c) \cup interface1$$
$$B_{class_3}^{public} = \{\beta_{class_3}^{public} \mid \beta_{class_3}^{public} = \varnothing\}$$

Expression (10) contradicts to expression (9). That why, class diagram does not follow interface segregation design principle.

4. Liskov Substitution design principle.

According to [Chebanyuk, 2016] represent the necessary conditions for Liskov Substitution SOLID design principle.

a) $\exists c \in C$ has a reference to another class diagram class $c^{|} \in C$. These classes are not connected by inheritance relationship. Then:

$$\exists c, c^{|} \in C \quad F(c)^{acc} = F(c) \cup F(c^{|})$$
$$P(F(c)^{acc}) = F(c)^{acc} = true \quad (11)$$

b) $c^{|} \in C$ has at least two inherited classes. Denote a set of such inheritors as $Cl\,(c^{|})$ .Then:

$$\exists c^{|} \in C, \exists c^{|}_{1} \in C, \quad F(c^{|}_{1})^{inh} = F(c^{|}_{1}) \cup F(c^{|})$$
$$P(F(c^{|}_{1})^{inh}) = F(c^{|}_{1})^{inh} = true \quad (12)$$
$$Cl(c^{|}_{1}) = \{c^{|}_{1} \mid P(F(c^{|}_{1})^{inh}) = true, c_{1}, c^{|}_{1} \in C\} \mid Cl(c^{|}) \mid > 2$$

Expression (12) should be applied for other classes' inheritors in $c^{|} \in C$ hierarchy, namely.

c) $c_1^{|} \in Cl(c^{|})$ has non empty overridden methods. Denote a set of these methods as

$$B_{c_1}^{override} \subset B_{c_1}^{public} \quad (13)$$

Then:

$$\beta_{c_1}^{override} \neq \varnothing, \quad \beta_{c_1}^{override} \in B_{c_1}^{override} \tag{14}$$

Expression (14) also should be applied for other classes' inheritors in $c^{|} \in C$ hierarchy.

Necessary condition for corresponding to Liskov substitution design principle is that class diagram should contain inheritance. UML diagram contains inheritance relationship between class1 and class2. Expression (12) shows this. But $|\, Cl(class1^{|}) \,| = 1$. That is why class diagram on figure 3 does not follow to Liskov Substitution design principle.

5. Dependency Inversion design principle

Represent the last SOLID design principle: Dependency Inversion design principle. Dependency Inversion Design Principle requires only one structural characteristic: functionality of $c \in C$ should be extended by means of one of the three variants, namely: $c^a \in C$, $c^{|} \in C$ (classes that are related on a top of hierarchy) or $i \in I$.

$$F(c)^{acc} = \begin{cases} F(c) \cup F(c^a) \\ F(c) \cup i \\ F(c) \cup F(c^{|}) \end{cases} \tag{15}$$

Consider class3 and class2

$$\begin{gathered} F(class3)^{inh} = F(class3) \cup \mathrm{int}erface1 \cup \mathrm{int}erface2 \\ F(class2)^{inh} = F(class2) \cup F(class1) \end{gathered} \tag{16}$$

$$F(class4)^{ass} = F(class4) \cup F(class3) \quad (17)$$

Class diagram (figure 3) does not follow to Dependency inversion design principle. Functionality of class4 is spread by class3 (17) that is not spread by means of abstractions.

Conclusion

An approach for class diagram estimation is proposed in this paper.

To refine class diagram it is necessary to check whether it corresponds to SOLID design principles. It is done in automatic way by means of proposed software tool. Description of this tool architecture is represented in the paper.

Contribution of this paper is a representation of approach and techniques for defining constituents of class diagram. Having information about class diagram content stakeholders may solve many important tasks in software development life cycle process more effectively. “Provider package” extracts initial information about class diagram. This information is used for all operations of software model processing, namely model comparison, merging, refinements, refactoring, and versioning.

Let's describe advantage of the proposed approach in respect to procedure of class diagram estimation by means of IBM RSA.

Using of IBM RSA, for estimation of class diagram, tools, supporting OCL, are involved. To estimate class diagram in accordance to SOLID design principles, OCL expressions should correspond with predicates for verifying class diagram, proposed in [Chebanyuk, 2016]. But in OCL expression context should be mentioned. There are no tools, allowing changing context automatically. That why software architect spends the same time to rewrite contexts for all OCL expressions of class diagrams or to analyze them mentally.

The tool, represented in this paper, allows to estimate automatically class diagrams according to SOLID design principles and to reuse them many times when changes to software are performed.

Further Research

It is planned to design an approach and software tool for Model-to-Model transformation, using collaboration of existing tools and environments, proposed in paper [Chebanyuk, 2017], namely IBM Rational Software Architect 7.5.5.2, Medini QVT, and Eclipce plug-ins. It is planned to obtain information about class diagram components using LINQ queries (Table 1). Then modify QVT scripts for supporting Many-to-Many transformation. Source and target software models are visualized in IBM Rational Software Architect.

Acknowledgement

Authors thank to Marco Brambilla for discussing ideas of paper [Chebanyuk, 2016] at the conference MODELSWARD 2016.

This discussion was a starting point to start research represented in this article.

Bibliography

[ASTM™, 2011] Architecture-Driven Modernization™ (ADM™): Abstract Syntax Tree Metamodel™ (ASTM™) http://www.omg.org/spec/ASTM/1.0

[Chebanyuk, 2016] Chebanyuk E. and Markov K. An Approach to Class Diagrams Verification According to SOLID Design Principles.In Proceedings of the 4th International Conference on Model-Driven Engineering and Software Development - Volume 1: MODELSWARD (2016), ISBN 978-989-758-168-7, pages 435-441. DOI: 10.5220/0005830104350441 access mode

http://www.scitepress.org/DigitalLibrary/PublicationsDetail.aspx?ID=HASwCJGMcXc=&t=1

[Chebanyuk, 2017] Elena Chebanyuk and Kyryl Shestakov An Approach for Design of Architectural Solutions Based on Software Model-To-Model Transformation International Journal "Information Theories and Applications", ISSN 1310-0513 Vol. 24, Number 1, 2017, pages 60-84

[Eclipce, 2016] Eclipce desctip IDE https://eclipse.org/ide/

[EMF, 2011] EMF Compare, 2011: Available from World Web: http://wiki.eclipse.org/index.php/EMF_Compare

[IBM, 2015] IBM, Rational software architect designer http://www-03.ibm.com/software/products/ru/ratsadesigner

[ICER, 2015] Western Michigan University project https://www.cs.wmich.edu/~OODA/patterns/AbsFactory.html

[Lakhal, 2012] Lakhal F., Dubois H. and Rieu D. (2012). P²E: A Tool for the Evolution Management of UML Profiles.In Proceedings of the 7th International Conference on Software Paradigm Trends -

Volume 1: ICSOFT, ISBN 978-989-8565-19-8, pages 211-217. DOI: 10.5220/0004081002110217 access mode http://www.scitepress.org/DigitalLibrary/PublicationsDetail.aspx?ID=jbI8fcFzMFo=&t=1

[Newcomb, 2005] Philip Newcomb. Chief Executive Officer Abstract Syntax Tree Metamodel Standard ASTM Tutorial 1.0 access mode: http://www.omg.org/news/meetings/workshops/ADM_2005_Proceedings_FINAL/T-3_Newcomb.pdf

[Ocampo, 2009] Joe Ocampo, Jason Meridth, Chad Myers, Sean Chambers, Ray Houston, Jimmy Bogard, Gabriel Schenker, and Derick Bailey, Pablo's solid software development, 2009 e-book access mode : https://lostechies.com/wp-content/uploads/2011/03/pablos_solid_ebook.pdf

[OCL, 2014] Object Constraint Language Version 2.4 OMG standard http://www.omg.org/spec/OCL/2.4/PDF

[Papyrus, 2012] Papyrus, 2012. Available from World Web: www.papyrusuml.org. Unified Modeling Language (UML), 2011. Available from World Web: http://www.omg.org/spec/UML/2.3/

[Sanches, 2015] Sanchez, L.E., Diaz-Pace, J.A., Zunino, A. et al. An approach based on feature models and quality criteria for adapting component-based systems Journal Software Engineering Research and Development (2015) 3: 10. doi:10.1186/s40411-015-0022-1

[Sielis, 2015] Sielis, G.A., Tzanavari, A. & Papadopoulos, G.A. ArchReco: a software tool to assist software design based on context aware recommendations of design patterns Journal Software Engineering Research and Development (2015) 2. doi:10.1186/s40411-017-0036-y

[XMI, 2015] XML Metadata Interchange access mode http://www.omg.org/spec/XMI/

Authors' Information

Olena Chebanyuk – *Software Engineering Department, National Aviation University, Kyiv, Ukraine,*

Major Fields of Scientific Research: *Model-Driven Architecture, Model-Driven Development, Software architecture, Mobile development, Software development,*

e-mail: chebanyuk.elena@ithea.org

Dmytro Povaliaiev – *Software Engineering student, National Aviation University, Kyiv, Ukraine,*

Major Fields of Scientific Research: *Software Architecture, Software Development, Model-Driven Architecture, Model-Driven Development,*

e-mail: dmytro.povaliaiev@gmail.com

THE COMPARATIVE ANALYSIS THE ALGORITHM OF EXTRACTION BOUNDARIES OF OBJECTS IN THE SPACE IMAGES

Rasul Shamsiev Zairovich

Abstract:*The results of the comparative analysis of algorithms of extraction borders of objects using various approaches to search and localization of borders are given in article. For comparison are used criteria of an assessment of parameters which received results by testing result. Is offered the algorithm of a choice the best solution of a subject task to extraction the borders of objects and control of its parameters. Results of researches are presented in the form of the table of set of estimates of the generalized indicator of quality.*

Keywords: *matrix, filter, algorithm, space image, the pixel, contour, boundary, gradient.*

Introduction

Interpretation when deciphering space images depends on the quality of detection of objects by segments. Segmentation separates the object from the background for further image processing to identify its containing information. When segmenting, the selection of contours is the main component of the process of identifying the detected objects. Consequently, the efficiency of solving many problems in the processing of space images depends on the quality of the extracted boundaries using segmentation algorithms, which are based on spectral discontinuities. The process of delineation depends on the quality of the accuracy of detecting the brightness discontinuities in pixels. The boundaries of objects consist of groups of pixels and they change from sudden changes in breaks. When solving problems to identify the boundaries of objects in images, there are many different algorithms, but the space images contains more information than the ordinary image, that is, the difference in the number of pixels is several tens of thousands of times. To determine the appropriate algorithm for delineating boundaries on a space photograph, it is necessary to take into account multiple brightness differences in the contour structure. Each algorithm is distinguished by the quality of the selection of boundaries of a certain type. Very important information is the quality of the output information obtained by preliminary, thematic processing and classification. Therefore, the aim of the study is to compare the resulting contours of space images, analyze and evaluate the quality of the algorithms.

Preliminary processing of space images Landsat 7

For more accurate detection of the boundaries of objects, it is required to pre-process the space image.The methods and algorithms used in known software, in some cases, do not allow to obtain qualitative results in image processing and decoding. When solving problems of qualitative interpretation of images through software, there are problems of improving the efficiency of image processing methods and algorithms, such as image filtering, improving visual quality, improving the resolution of objects, combining multispectral images, classification, etc.

Each software has its own thematic functions and drawbacks and when using different modules of different soft, taking into account their quality of solving the tasks assigned, the simultaneous application of several soft becomes profitable and effective.

For example, after the computational experiments using the Landsat 7 space image, the following processing results were obtained on the ESRI, ERDAS Imagine, ENVI, PCI Geomatics software products (principal of component analyse), as shown in Figure1.

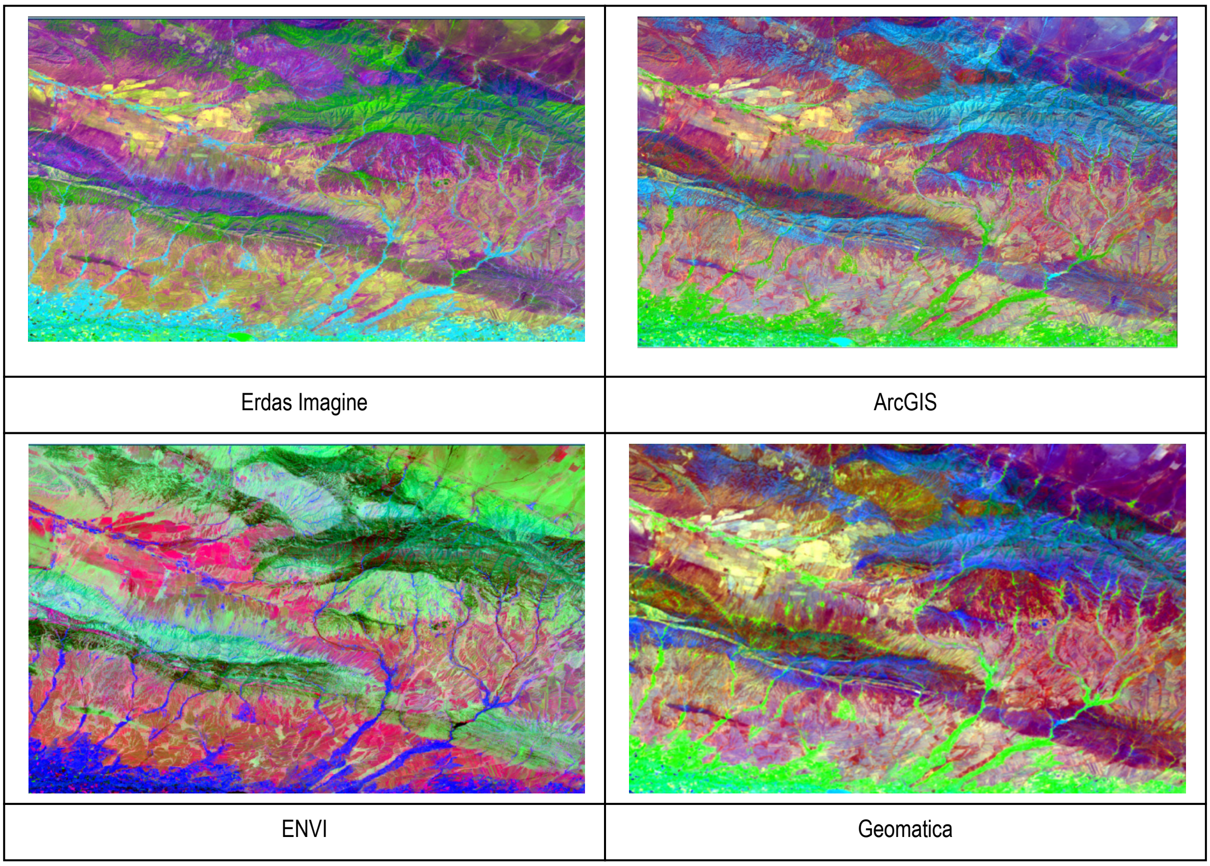


Figure1.Processed space image with using the PCA method in different GIS programs.

When assessing the results obtained, space images are distinguished by color gamut, brightness, clarity of the boundary line and the shape of the objects. When compared with the original image and the geological map, the image that was processed through ENVI and Geomatica looked more informative. By the speed of processing, Geomatica is ahead of many other programs, the boundaries of objects are slightly blurred. By the combination of spectral channels, the ENVI software delivers images of more bright, clear boundaries of structures, more revealing vegetation, geological and technogenic objects. Erdas Imagine and ArcGIS showed an identical result for the rest of the images, except for the combination of spectral channels, which is inferior to ArcGIS. In ArcGIS software, developers paid more attention to processing speed by reducing the number of channels of the image, which became more problematic for visual interpretation. The high speed of Geomatic depends on the output file. The size of the "pix" format memory is smaller than other formats, this format has its drawbacks, many programs do not read them, saving processing time is lost when converting to other formats. But sometimes the information obtained is more important than the speed and depends on the availability of improved processing algorithms that are not available in other programs. According to the experimental calculations, Table №1 was compiled in which the entire list of activities, the characteristics of the snapshot, the results of the time and the properties of the identified objects were included. Of all the programs, the ENVI program's PCA module in terms of informative parameters outruns the modules of other programs, except processing speed and output file size, these disadvantages can be eliminated by preselecting a combination of channels.

Table 1.Experimental calculation for different software products.

List	ArcGIS 9.3	Erdas Imagine 9.2	ENVI 4.7	PCI Geomatics 2002
Types of algorithms	PCA	PCA	PCA	PCA
Space image	Landsat7	Landsat 7	Landsat 7	Landsat 7
Type of the space image	spectral	spectral	spectral	spectral

Number of spectral channels	7	7	7	7
Solvability of the space image	x(0,00029763759),y (0,00029763759)	x (0,00029763759),y (0,00029763759)	x (0,00029763759),y (0,00029763759)	x (0,00029763759),y (0,00029763759)
Scale study	1: 200 000	1: 200 000	1: 200 000	1: 200 000
Study of area	Samarkand region	Samarkand region	Samarkand region	Samarkand region
Processing speed with RAM 4 GB	1 min 40 sec + 1 min selection of parameters	2 min 15 sec + 1 min selection of parameters	3 min + 1 min selection of parameters	1 min + 1 min selection of parameters
Processing more than three spectral channels	no	yes	yes	yes
Ignoring zero values	no	yes	yes	no
The combination of spectral channels	no	yes	yes	yes
Outputfileformat	GRID Stack 7.x	Imagine	HDR	PIX
Pixeltype	unsigned integer 16	unsigned integer 8	Float 32	signed integer 8
Image type	continuous	continuous	continuous	continuous
Solvability output image	x(0,00025705621),y (0,00025705621)	x (0,00025705621),y (0,00025705621)	x (0,00025705621),y (0,00025705621)	x (0,00025705621), y (0,00025705621)

Qty. sp. channels of the output file	3	7	7	7
Volume of input file	67,45 mb	67,45 mb	67,45 mb	67,45 mb
Volume of output file	57,81 mb	75,28 mb	269,81mb	67,49 mb
Format integration	yes	yes	yes	no
Qty. geol. objects	35	60	70	52
Qty. landscape objects (rivers, flora)	1 (river)	2 (rivers, flora)	2 (rivers, flora)	2 (rivers, flora)
Qty. technogenic objects	1(roads)	3 (roads, houses, arable land)	4 (roads, houses, arable lands, power lines)	2 (roads, houses)
Clarity of object boundaries	low	middle	high	middle
Blurring objects	high	middle	low	middle-high
Color difference objects	low	middle	high	middle
Visual informative	low	middle	middle-high	middle

Total processing quality	low	middle	middle-high	low-middle

In this regard, taking into account the properties of different software tools, an adaptive segmentation is carried out, which consists in multi-stage processing of the space image. This method of segmentation, in contrast to the standard used in ArcGIS, takes into account the boundaries and integrity of objects and allows you to significantly reduce the number of cases where the boundaries of blocks divide the image visible in the image into separate parts, in some scattered. The processing of a space image using this method made it possible to identify geological objects that were hidden in previous images. The use of this approach can be applied to other space images of the type Landsat TM. If images are needed at different times, spectral transformations are required.

The obtained results and their comparison with cartographic data on the example of the Samarkand region for assessing the accuracy of segmentation are shown in Table №2.

As you can see in the table №2, the image of the past segmentation has a large amount of informative content than the previous one. Further, on the above segmented images, it is possible to reveal the linear boundaries of natural objects.

Methods for selecting the boundaries of objects in a space images

Recently, known algorithms are often used, such as operators Roberts, Prewitt, Kirsch and Sobel. These operators are based on the basic properties of the luminance signal as a discontinuity. The search for breaks is performed by a method of processing a snapshot with the help of sliding windows, called filters, cores, etc. Filters are a square matrix consisting of a group of pixels of the original picture. A square matrix includes elements called coefficients (Figure 2). The change in coefficients for local transformations is called filtering.

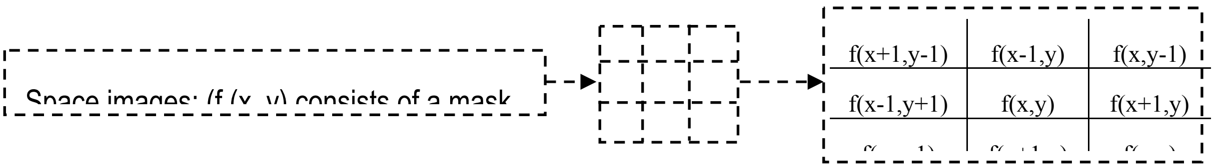
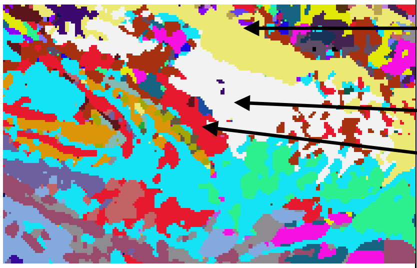
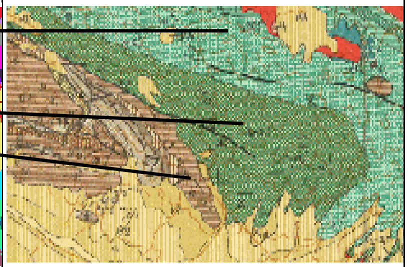
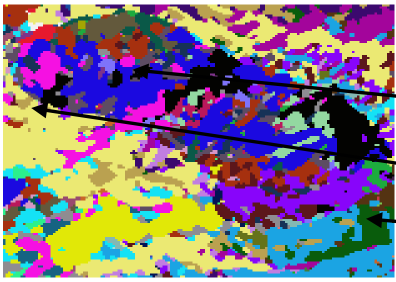
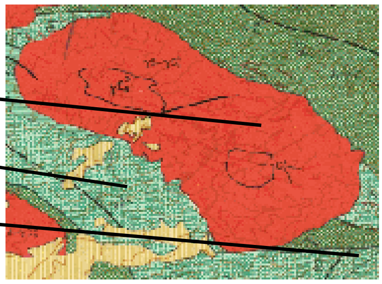
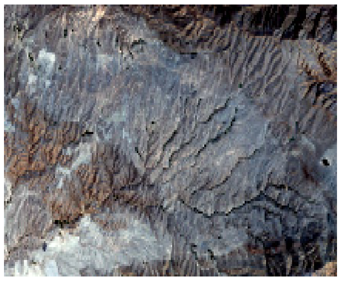
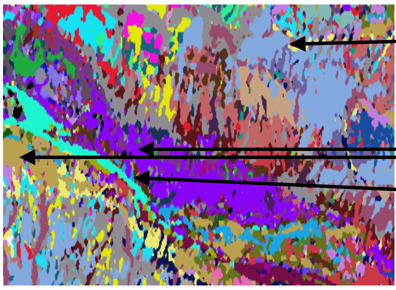
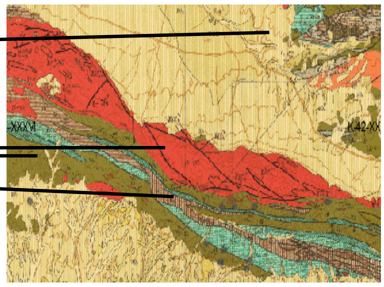
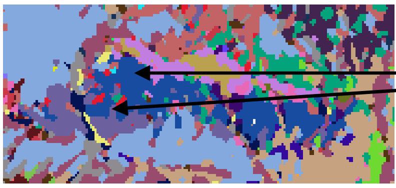
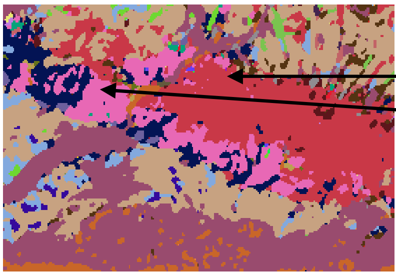
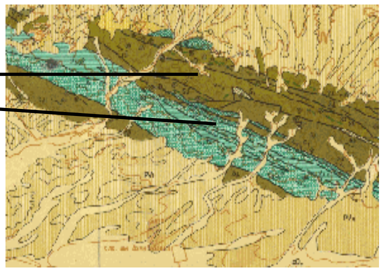
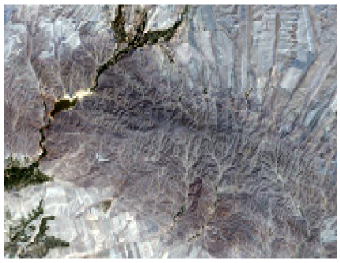


Figure 2.Structure of the matrix filter of the space image.

Table 2. Segmentation the space images

№	Segmented space images	Geological map	Previous
1			
2			
3			
4			

5			

The existing algorithms can also be added to other frequently used operators Contour (Suzuki), SUSAN and Canny. All operators are very acceptable from the point of view of what kinds of images they are applying, however, for each of them, the ambiguity of marking points in real situations is characteristic due to the need for different choices (choice of brightness coincidence thresholds, choice of digital masks, etc.).

The task is to find the optimal method in space images, which contain information on the structure and boundaries of natural objects. To select the optimal algorithm for processing a space image, it is necessary to analyze all of these algorithms and give an estimate of the performance of each. We give a brief description of the properties of the algorithms for delineating space images, and then highlight the advantages and disadvantages of each after testing.

Algorithms Sobel is used to calculate the approximation of the gradient of the brightness of the image. They calculate the gradient of the brightness of the image at each point.

Algorithm Canny. This algorithm is the most popular in using for delineating boundaries. The algorithm itself is executed in several steps:

1. Elimination of noise and unnecessary details;
2. Counting the gradient of the image;
3. Reduction of edge thickness (edge thinning);
4. Linking individual edges to the edge (edge linking).

The SUSAN algorithm is used to calculate the edges and angles of objects in images. The processing speed of this algorithm is higher than the rest of the operators due to the lack of application of the convolution process.

The Contour (Suzuki) algorithm allocates closed contours. Counting the image gradient and reducing the thickness of the edges (edge thinning);

The LoG (Laplacian of Gausian) algorithm smooths the image and calculates the Laplace function, which leads to the formation of double contours. The definition of loops reduces to finding zeros at the intersection of double boundaries.

The Roberts algorithm is easy to implement and has a high speed, it quickly calculates a two-dimensional spatial measurement, but is highly sensitive to interference.

Combined algorithm of extraction boundaries

The studying the properties of the above algorithms, there was possible to develop and test a combined algorithm using the example of the segmented space image Landsat TM (resolution 30 meters), below is a brief description of this algorithm and the block diagram in Figure 3.

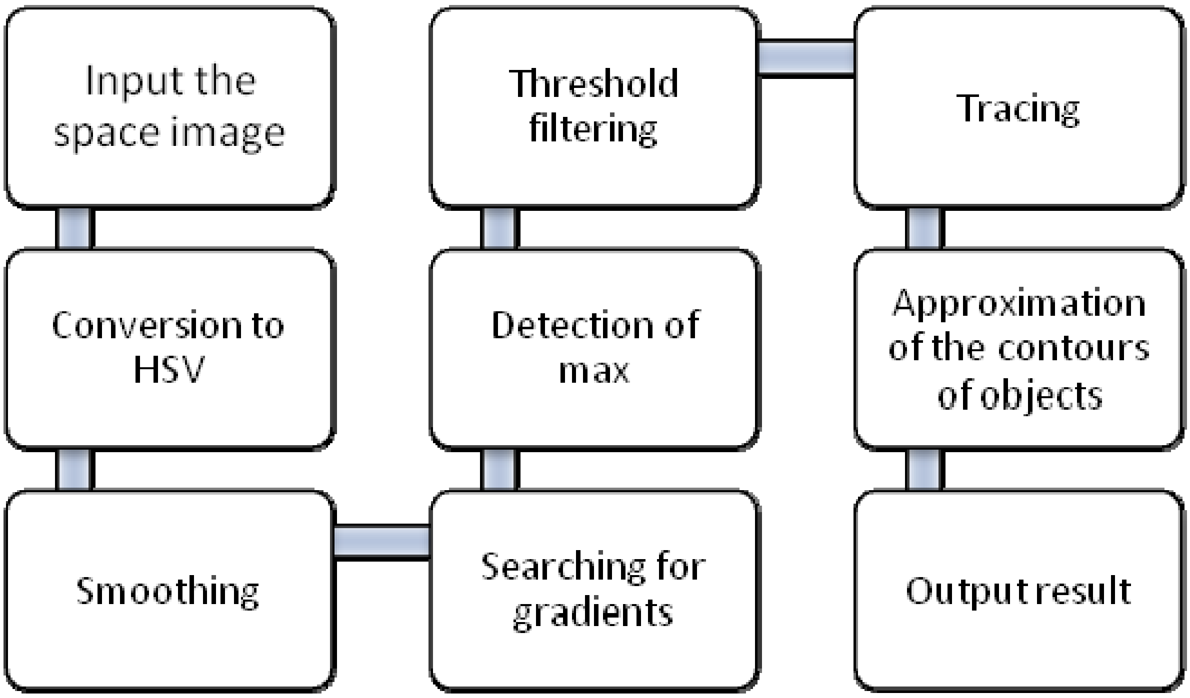


Figure 3.Block diagram.

Input f (x, y), segmented satellite imagery Landsat TM (1).

$$f(x,y) \rightarrow HSV \rightarrow f1(x,y) \quad (1)$$

Conversion an RGB space image matrix to an HSV image matrix (2):

$$\begin{aligned}&HSV(vM) = rsubM(rM,0,rows(rM)-1,0,cols(rM)/3-1) + gsubM(gM,0,\\&rows(gM)-1,cols(gM)/3,cols(gM)2/3-1) + bsubM(bM,0,\\&rows(bM)-1,cols(bM)2/3,cols(bM)-1)\end{aligned} \quad (2)$$

Smoothing: Remove noise using the LoG filter. In this process, the differentiation of the smoothing filter g is carried out by a two - dimensional function, after the convolution process (3).

$$g(x, y) = 1/(2\pi * \sigma^2) * \exp(-x^2 + y^2/(2\sigma^2)) \quad (3)$$

To optimize noise suppression, you need to select the best filter parameters. The degree of blurring is determined by the parameter.

Search for gradients (4). To search, we apply the Sobel filter, this filter consists of two 3x3 matrices, the second matrix differs from the first angle of 90 degrees. The filter calculates the approximate values of the brightness gradient of the image and the resulting result is either a gradient of the brightness gradient or its norm.

$$G = \sqrt{G_x^2} + \sqrt{G_y^2} \quad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} * A \quad and \quad G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * A \quad (4)$$

The process proceeds to double threshold filtering to determine whether or not the boundary is at a given point in the snapshot. The first filtering threshold calculates pixel values above the upper limit if it takes the maximum value (the boundary is assumed to be reliable), the second threshold, if lower - the pixel is suppressed, the points with a value falling in the range between the thresholds take a fixed average value. In this process, the operator Canny is involved, the filter of which can be well approximated to the first derivative of the Gaussian (5).

σ = 1:

$$B_x = \frac{1}{159}\begin{bmatrix} 2 & 4 & 5 & 4 & 2 \\ 4 & 9 & 12 & 9 & 4 \\ 5 & 12 & 15 & 12 & 5 \\ 4 & 9 & 12 & 9 & 5 \\ 2 & 4 & 5 & 4 & 2 \end{bmatrix} * A \quad \Theta = \arctan(\frac{G_y}{G_x}) \quad (5)$$

Where the angle of the direction of the gradient vector is rounded and can take such values 45 -135, for the current process the value to be received is 45.

Tracing selects groups of pixels that have received an intermediate value at the previous stage, and assigning them to the border or suppressing them. The pixel is added to the group if it touches it in one of 8 directions.

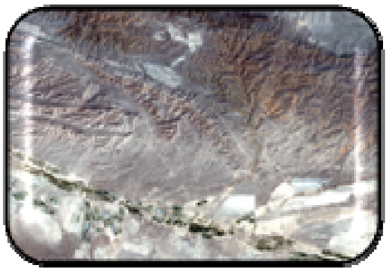
Approximation. The space image contains an extremely large number of objects, since each pixel of the image is considered when the boundaries are selected, we get a lot of points of contact, so for further actions the problem of approximating the set of boundary points of the object is solved.

The result of computational experiments using space imageLandsat 7

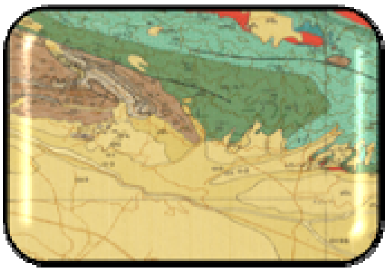
For each of the above algorithms, segmented space images of Landsat 7 were processed (table 3).

When analyzing the results obtained (Table 3), the following patterns were revealed: the algorithms Kirsch, Roberts and Prewitt give the same results. The rest of the average parameters almost identically detected the contours of objects. By the results of the Contour (Suzuki) algorithm, an important parameter is the selection of the contours of objects by classes, that is, by the color characteristics. Closer to the high parameters of the results, we can conclude that the algorithms Canny and Contour (Suzuki) have coped much better. When processing a test snapshot, better results can be observed after the proposed algorithm works, although with the other methods of preliminary and thematic processing the best method may be the other. The contours highlighted by the method proposed in this paper are lines with a thickness of 1 to 2 pixels, the solvability of each pixel depends on the type of space photograph. This algorithm solves the problem associated with the thickness of contour lines, preserving the distribution of color intensity by class of objects. The application of the algorithm with the properties of different algorithms shows high efficiency and improves the quality of processing (Table 4). At present, in many technical literature, methods for detecting object boundaries are well described, but the laborious task still remains in GIS technology for certain sectors of the national economy. So, in the direction of geology, the application of many algorithms to the processing of space images shows shortcomings that have not been identified in other types of images and require improvement or the creation of more efficient methods of processing.

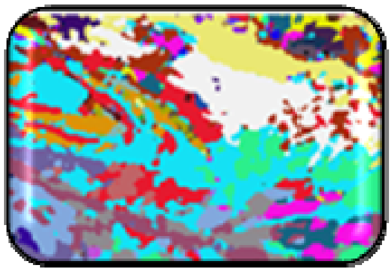
Table3.The revealed linear boundaries of geological structures



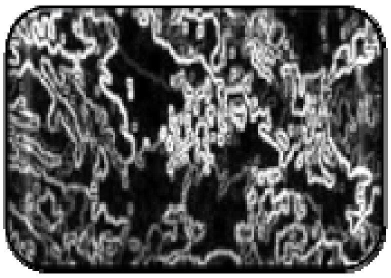
Initial space image Landsat TM



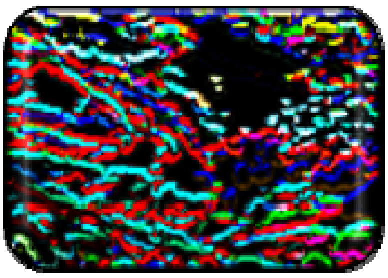
Fragment of the geological map



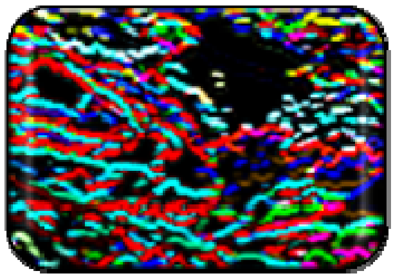
Segmented space image



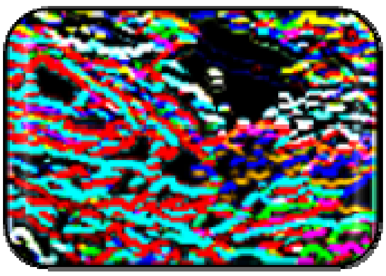
Sobel



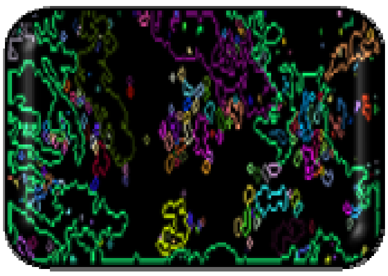
Prewitt



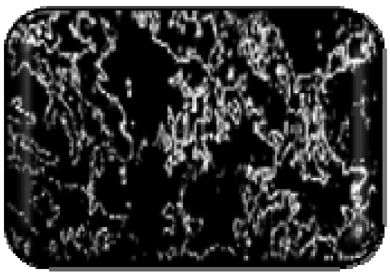
Roberts



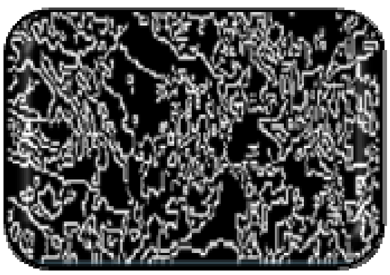
Kirsch



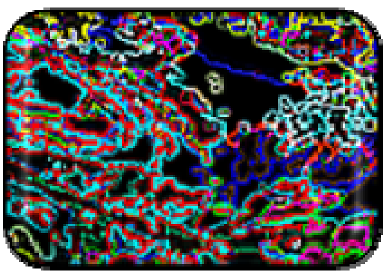
Contour(Suzuki)



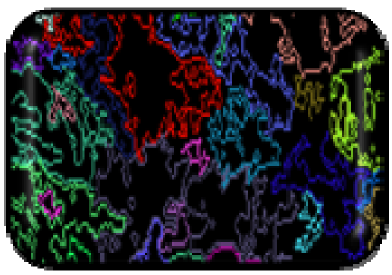
SUSAN



Canny



LoG



Suggested

Table 4. The quality processing

		Preservation of object classes	Line sharpness	Visibility of contours by objects	Line thickness	Processing time	Intensity of noise	+/-
№	Algorithm	a	b	c	d	e	f	
1	Sobel	- (0%)	-(30%)	+(75%)	-(25%)	+(100%)	-(45%)	2/4
2	Prewitt	- (0%)	-(15%)	-(15%)	-(30%)	+(100%)	-(35%)	1/5
3	Roberts	-(0%)	-(15%)	-(15%)	-(35%)	+(100%)	-(25%)	1/5
4	Kirsch	-(0%)	-(10%)	+(10%)	-(25%)	-(45%)	-(45%)	1/5
5	Contour(Suzuki)	+(65%)	+(65%)	+(60%)	+(60%)	-(45%)	+(55%)	5/1
6	SUSAN	-(0%)	+(55%)	-(35%)	+(50%)	+(65%)	-(45%)	3/3
7	Canny	-(0%)	+(60%)	+(75%)	+(60%)	+(50%)	+(65%)	5/1
8	LoG	-(0%)	+(50%)	+(50%)	-(25%)	-(35%)	+(50%)	3/3
9	Suggested	+(65%)	+(75%)	+(80%)	+(75%)	+(65%)	+(70%)	6/0

Conclusions

Very important is the factor of the accuracy of the coordinates of the areas of different natural structures for geological prospecting or other ground operations, since the cartographic data are taken from ground operations, there is hardly any technology for calculating the boundaries of huge areas and recognizing their structure. To determine some of the rocks that can be distinguished from space images, laboratory studies and decent financing are required. The development of methods and technology for the processing of space images can significantly change the conduct of various types of ground-based research, thereby saving themselves and simultaneously obtaining valuable results.

Bibliography

[Canny J. 1986] A Computational Approach to Edge Detection. IEEE Transactions on pattern analysis and machine intelligence. No. 6 1986, p 679.

[Shamsiev R. Z., 2016] Analysis and evaluation of software modules in the solution of task of processing the space images.Tashkent. Scientific - technical and information-analytical journal TUIT 2016, №2 (38), pp.54-56.

[Shamsiev R.Z., 2016] Recognition the structural differences of space images. Tashkent. Scientific - technical and information-analytical journal TUIT 2016, №3 (39), pp. 3-5.

[Shamsiev R.Z., 2016] Algorithm for detection of ore field with the use of spectral values of a space image. Tashkent. Scientific - technical -analytical journal TSTU 2016, №1 (94), pp.210-212.

[Shamsiev R.Z., 2016] Algorithm for detecting the boundaries of geological objects on space image. Tashkent. Science and practice integration as the mechanism of effective development of geological industry of the republic of Uzbekistan. Tashkent. 2016, p.366.

[Shamsiev R.Z., 2016] Algorithms for detecting minerals from the spectral characteristics of the Aster space image. Science and practice integration as the mechanism of effective development of geological industry of the republic of Uzbekistan .Tashkent.2016, p.369.

[Esri, 2017] Documentation. Principal Components Analysis (PCA). http://desktop.arcgis.com/ru/documentation/.

Author's Information

Rasul Shamsiev Zairovich *- scientific researcher. The Center for Software Development and hardware-software complexes at the Tashkent University of Information Technologies; 100084, Uzbekistan, Tashkent, Kichik halka str., 2; e-mail: rasulshamsiev@rambler.ru*

Major Fields of Scientific Research: Remote sensing and GIS technologies.

KNOWLEDGE ACQUISITION SYSTEM TRAINER1 (SIMULATOR FOR TEACHING FINANCIAL ANALYSIS) AS INEVITABLE COMPONENT OF DISTANCE LEARNING

Tea Munjishvili

Abstract: *In article the existing systems of collection of knowledge are analyzed, need of computer simulators of creation for collection of knowledge in economy is checked, necessary conditions for their functional capabilities and use are defined, the initial computer simular for the collection of knowledge created by us is figured. In article the existing systems of training of the financial analysis are analyzed. One of ways in improving of quality of e-learning of objects of an economic profile - development and applications during training of simulator. In article the technology of preparation and training of the financial analysis with use of an assessment bankruptcy of the enterprise by means of the logician - probability and the Z-score of models, the developed author of software package TRAINER1 the exercise machine. The TRAINER1 used for the multi-scenario analysis and generalization of results to predict assumptions of possible errors and to create recommendations of their correction.*

Keywords: *Knowledge acquisition, Simulator, Financial analysis, Semantic analysis, Debriefing. FINSIM1.1*

ITHEA Keywords: *Please use keywords from* http://idr.ithea.org/tiki-browse_categories.php .

Introduction

Achieved results in the field of informatization makes development of distance learning possible. Distance learning can be considered as form of teaching by correspondence approved in practice during decades, which is by correspondence in form but is relevant to form of attended teaching by its sense. Components of distance teaching of subject are as follows: textbook created by using hypertext technology, video lectures, special programmed environment, which provides: 1. Acquisition of knowledge in training regime; 2. Model typical situations and acquire decision making skills; 3. Demonstrate and assess acquired knowledge.

Problems of knowledge acquisition, demonstration and assessment have always been, is and always will be topical issue. Process of searching is infinite in time and space. Three approaches are singled out: traditional (informal), using computer systems (formal) and combination of these two.

Traditional method is effective one that has been experienced during centuries and is very unlikely to be changed ever. In the last three decades intense research has been taking place and computer programs of different kinds and functional possibilities have been created. These programs are oriented towards knowledge acquisition, demonstration and assessment.

In the field of knowledge acquisition transferring knowledge using video lectures, multimedia electronic books and printed publication are frequently used today. In addition, online consultations with the teachers is also widely accepted.

Many scientific publications are written regarding improving teaching process using electronic systems of knowledge acquisition. One of the directions is drawing out computer simulators.

In the exploitation of complex technical systems and training of management specialists different kinds of simulators have been used for a long time. In the article {Трухин 2008} teaching simulators of different aims existing in Russian Federation are analyzed. In the recent years simulators have been widely used in medical institutions in the process of teaching medical personnel and improving their qualification [Свистунов 2011]. Working out of simulators of different type and aim - started from physical ended by electric - is taking place in today's world. Mainly three kinds of simulators are prevalent: physical, electronic and combination of two.

Experience in drawing out and use of electronic simulators in the training of specialists in the area of management includes: Railway tickets cashier, managing and exploiting heating systems in train coach and etc. [3].

Training of specialist in any field including economist consists of two main components: theoretical knowledge and practical skills. Object of our research is acquisition of knowledge oriented towards searching for a method of elaboration mechanism of practical skills. Acquiring skills for operating aircraft, ships, boats and similar apparatus is impossible without physical simulators.

Object of the learning of economic processes are informational flows depicting movement and modification of material currents. Because of this only computer simulator can be used in training of economists. Their role is particularly important in the distance learning. Methodological foundations of realization of simulator are economic-mathematic models and computer systems. Teaching practical skills using simulator is achieved by solving situational problems.

Scientific research works done for elaboration of simulators for subjects of economic profile represented in scientific papers and dissertations are very interesting [Ельцин 2013, Рышкевич 2013].

Analysis of informational materials shows that there is no common understanding between the concepts of electronic simulator and simulator. Part of the researchers understand electronic simulator as a unity of textbook made using hypertext technology and informational-programed complex. Using electronic textbook introduction to and acquisition of the issue is done, while programed complex provides fixation of acquired issues. In this process student solves the problems in an interactive regime and answers theoretical questions. In order to answer correctly to these questions student uses reminder represented by text, graph, and video. Process of the teaching and exam for certificate is made online.

Hence, for training high quality specialists of economic profiles and permanent qualification improvement of management specialist's two interlinked problems have to be solved: first – drawing out universal program package of electronic simulator, second – filling knowledge base of simulator by knowledge base of problematic section.

First problem consists of two interlinked problems: first – elaboration of electronic textbook using hypertext technology and second – elaboration of programmed environment for problem solving online in the interactive regime. In a special literature theoretical and practical issues of creating electronic textbooks using hypertext technologies are widely discussed. Simple programs – editors are available in the environment of which any teacher or professor working in the field of economics can create electronic textbook after some training. Today, many electronic textbooks are created. Any textbook is presented with text, graph, video. It serves as a basis for knowledge acquisition. After introducing materials in the textbook it is necessary to solve practical problems, analyse theoretical issues and etc. This goal can be achieved by program – simulator.

Different kinds of simulators have been used in exploitation of complex technical systems and training of specialists for a long time. In article /1/ simulators existing in the field of business and used on the levels of teaching and functioning are given: AdSim Advertising, Baton Simulations, Blue Ocean Strategy Simulation (BOSS), Capsim Foundation and etc.

In article /2/ simulators of teaching purpose existing in Russian Federation are analyzed. In the recent years simulators have been widely used in medical institutions in the process of teaching medical personnel and improving qualification /3/. This is not surprising, because any mistake made by technical system and medical personnel brings immediate results. Moreover, mistake of technical system – made by ship or aircraft crew brings their life under risk.

Because of the fact that result of the mistake caused by low qualification of an economist will show up after Δt period, it remains unnoticed for the decision makers. Result of the mistake can be fatal for the society.

Mistakes made by economists on a different level of management are the results of impunity syndrome. Economists' inability to predict the possible result of their decisions comes from the low level of their training. Processes of transferring, acquiring, revealing of knowledge must be changed as a matter of principle.

Our focus is made on searching for the methods, ways and means of reduction or in ideal case – liquidation of errors made by economists that are caused by low qualification.

At the moment simulators of different kinds and purpose are being drowned out in developed countries – started from the physical simulators ended by electronic ones. It won't be mistake if we say that there are mainly three kinds of simulators being constructed: physical, electronic and combination of two

For the training of economists we can create and use only electronic simulators.

Modern information tehnologies allow us to look differently to the teaching process and cunduct it in an unusual way.

Analysis of informational materials shows that there is no common understanding between the concepts of electronic simulator and simulator. Part of the researchers understand electronic simulator as unity of textbook made using hypertext technology and informational-programed complex. Using electronic textbook introduction to and acquisition of the issue is done, while programed complex provides fixation of acquired issues. In this process student solves the problems in an interactive regime and answers theoretical questions.

Under the simulation modeling informational-programed environment is considered in which specific type of situation is being modeled.

Aim of the elaboration and use of simulators is:

First – Enabling for understanding management of the situations arising on a real object, decision making and analysis of expected results of decision in the teaching process.

Second – Enabling modeling existing situation with real data and looking at expected results before making decision in the practical activities.

It is worth to be mentioned that like electronic diagnostics it is impossible to achieve adequate modeling of actual situation regardless the perfection of the situational model. Because of this, final word is on decision maker and he/she makes decision under self-responsibility.

Topics discussed in the specific discipline include discussion and acquisition of typical finite situations for those simulation models should be drowned out.

Second problem consists of two interlinked problems: first – elaboration of electronic textbook using hypertext technology and second – formation of system knowledge base.

Simulators can be used invariably or with the little modification in the following areas with the common legal space:

1. In the higher education institutions in different countries, including Georgia:
   - For the training of specialists;
   - For the retraining of specialists in the field of management: economists, financiers, marketing specialists and others.
2. In the process of enterprise management.

Preconditions for drawing out simulators for economic processes exist – these are theoretical and practical experience of elaboration and use of intellectual systems of decision making in different problematic areas (medicine, technical systems and etc.). Like economic processes in medical diagnostic systems here we deal with the information processes. Because of this while drawing out simulators of economic processes modeling mechanisms for knowledge demonstration and decision making can be taken into account. Clearly, like medical diagnostic systems at first classification of economic simulators using certain criteria must be done. This criteria include: problematic areas (entrepreneurial, macroeconomic processes), kinds of problematic areas, situation types and etc.

Problems of elaboration and use of simulators depicting economic processes for training of economists stand in the center of our attention. Form of the realization of economic simulator is computer programmed complex. Its aim is to ensure that in the teaching process student acquires skills for situation analysis, economic decision making and assess results of the decision for the different values of real data.

Use of simulators in the teaching process changes methods and approach to the transfer of knowledge. The role of the teacher is particularly increasing, which should be able to:

- Comprehend the working principles of the simulator;
- Understand the realization algorithm of the problematic situation that is simulated;
- Interpret the results of the simulation;
- Correct of the simulation results taking into account the factors considered by the simulator;
- Approve/reject simulation results

On the practical lessons student is required to explain simulation results. Simulation realization and analysis of the results depend on the knowledge about problematic area. Because of this, introduction to and acquisition of issues regarding the simulation are integral parts of the simulator.

Components of study using the simulators are as follows: computer turned in a global network, program simulator, electronic textbook realized by hypertext technology linked to simulation organized on a server. Integral part of the any kind of studying process including the studying process using simulators is video clip representing the process of studying. After classes the video is placed on the internet and student gets it.

Our modest experience in drawing out economic simulators enables us to form necessary requirements for simulators.

Simulator is used to:

- Model problem depicting actual economic process using certain algorithm;
- Solve linear and non-linear optimization problems;
- Form multi-optional plans;
- Comply and select options;
- Visualize of simulation results in a form of diagrams, tables, videos or their combination;
- Model N independent variable in the problem depicting model by $\forall n_i$ variable and represent the results in the form of diagram;
- Model by any of n_i and n_{i+1} variable from the set N simultaneously and represent the results in the form of table;
- Model using number of variables simultaneously for the different values of variables;
- Represent multi-optional results of modeling;
- Explain simulation results.

Using program package it can be possible to:

- Introduce to the essence of the problem, algorithm;
- Realize simulation in many natural language.

Work with the simulator must be allowed for every interested person without any kind of registration and passwords.

Simulator should enable us to make as situation modeling as well as simulation of solving problem using known algorithm and research the algorithm to solve this problem. Clear example of it is solving the problem in Excel with its financial functions. Observing the process of problem solving in Excel using its functions we can see how the result is being changed in response to the changes made in the values of arguments. Arguments themselves reflect occurrence of the certain fact such as taking credit, investing, amortization and etc.

We see conducting lectures and practical classes using simulators in such way: /Fig. 1, 2/

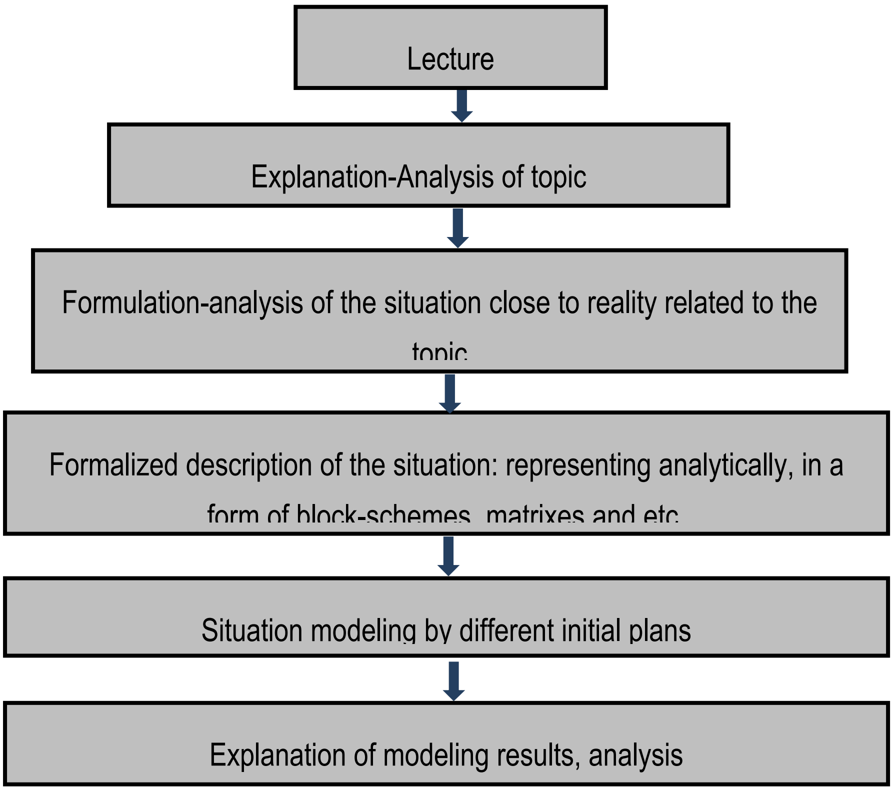


Figure 1. Block-scheme for giving lecture using simulators

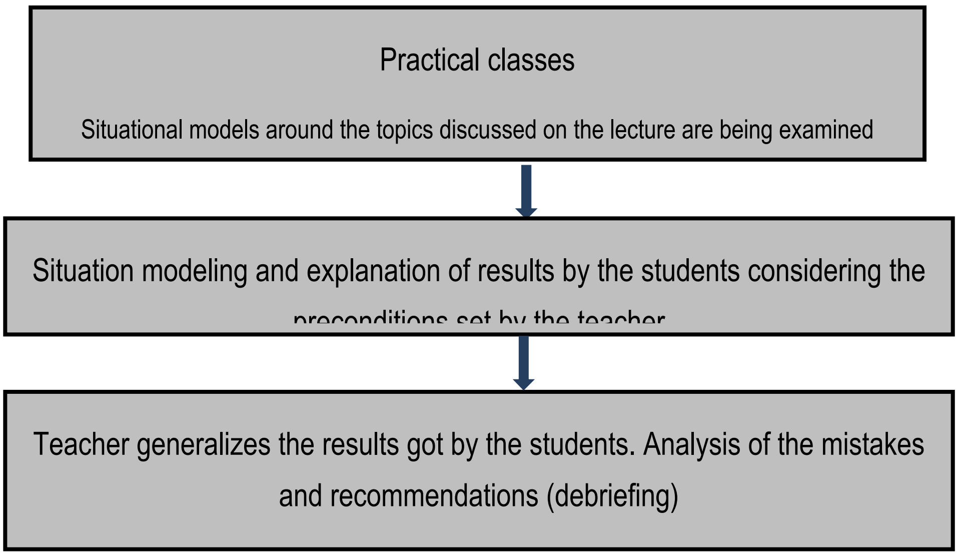


Figure. 2. Block-scheme of conducting practical classes using the simulators

Use of the simulator in the teaching process consists of the stages of planning and functioning.

On a **planning stage** from the set of the simulators teacher selects simulator(s) depicting the given topic;

Functioning stage /Fig. 3/

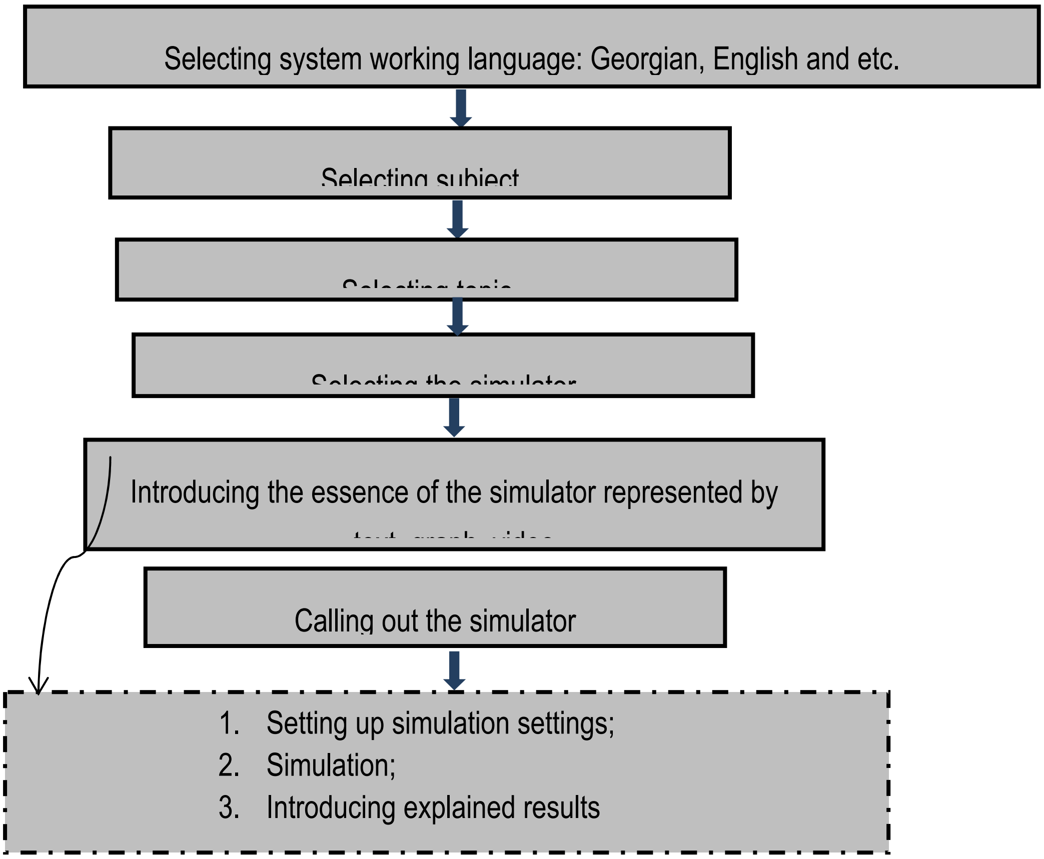


Figure 3. Stage of functioning in conducting teaching with the simulator

In the program package FINSIM1.1 created by us simulators are realized that are used in: first, modeling evaluation of financial stability of enterprise by Altman, Springate, Fulmer, Brzezinski, DuPont models. With these models financial stability of enterprise is evaluated using one, two and many (all the variables included in the model) variables; second – simulating 15 functions of Excel by one or two variables.

Algorithmic-programmed environment of the simulator is universal, but its informational inside part depends on the teacher of subject, its approach to the transfer of knowledge and etc. Because of this, drawing out typical simulators in the teaching process as a standard is unacceptable. Simulators by subjects can have recommendation-auxiliary role.

Currently, analysis of the simulators used in trainings of the specialists in different fields and our years-old experience in computer systems for demonstrating and assessing knowledge and practical exploitation allows us to form necessary requirements for the computer simulators oriented towards economic profile subjects:

3. Existence of theoretical and practical issues;
4. Formation of knowledge base according to the theoretical and practical issues of the subjects, subject topics and sub-topics;
5. Formation of standard tasks for training;
6. Explanation of any task in the exercise demonstrating specific method, algorithm and solution;
7. Formation of base of advices and recommendations for avoiding repetition of the errors made during the training;
8. Introducing specific results of the training, advices for correction of the errors and showing recommendations;
9. Making analysis of the errors made during the training in the certain period of time;
10. Changing trajectory of training by student based on the answers.

Formation of the knowledge base of the simulator.

Basis of the training is formation of knowledge base by the teacher according to the subjects, subject topics and sub-topics. Knowledge base general tests have two types – closed and open.

Closed tests are ones for which on every posed question maximum seven answers comes out of which three are correct. Answering to the posed question is brought down to selection of the right answers out of maximum seven options.

Open test is a test in which the answer of the student must be written in the form of numerical value, sentence or expression.

Based on our experience it is better to have 70% open tests in the test base and remaining to be closed tests, which can preferably be distributed according to the number of right answers in the following way:

Table 1.

Name	Number (%)
Tests with one right answer	25%
Tests with two right answers	35%
Tests with three right answers	40%

In the closed tests assessment mustn't be partial, because selecting two or three right questions out of seven is quite hard and requires thorough knowledge of the issue. Open tests are tasks. Possible answers to the tasks aren't limited. Teacher assesses solution result according to his/her attitude and makes ranging of total points based on the answers. Formation of the test base according to the sub-topics occurs in a way that closed and open tests describe typical situations.

For the closed tests firstly, correct answers with detailed proofing are provided by the teacher and secondly, in case of wrong answers advices and recommendations are provided.

For the open test (task) the teacher indicates: firstly, solution for the specific task, secondly, method for solution such task, algorithm, mathematic model, area of use, solution technique and technology, programmed means such as Excel, MATLAB used in solution and etc.; thirdly, the reasons for the wrong answers and the ways and means to avoid them.

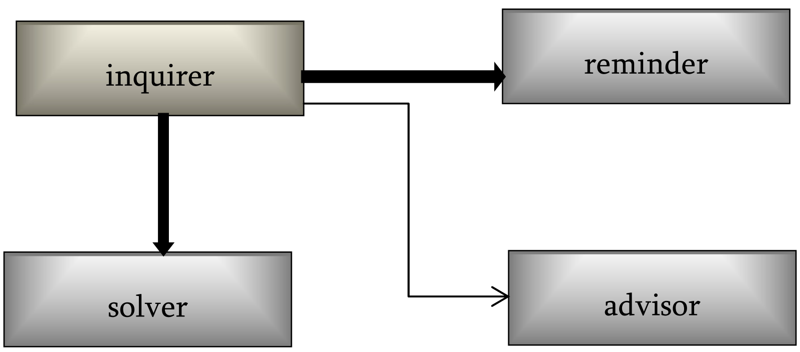


Fig. 4. Essence structure of the tests

Hence, basis of representing any test is information given in a following structure /pic. 1/

“Inquirer” consists of two sections: question (task) text and the possible answers.

“Solver” consists of question, task solution algorithm, model, solution description formulated in specific “inquirer”.

In the “reminder” solution method, experience of solution of analogous task, expected results in case of wrong answer and etc. are given.

In the “advisor” negative results are discussed. Reasons causing them and the ways of avoiding negative results are shown here.

In order to study causes for negative results and factors affecting them there is tight connection between the “reminder” and the simulator of specific economic processes given in the “inquirer”. Using it the process of the modeling by one, two and multiple variables, formation of multi-optional plans and etc. are made [Tea Munjishvili, 2017].

Any functional block given on the pic. 1 can be shown by the text, graph, video or their combination.

Planning the training.

Planning the training - selection of the topics, sub-topics, tests, time of the training – is made by the teacher within the simulators familiar to us and individual potential of the student isn’t taken into account. In the ideal case student should have possibility to select topics for training from the test base and sub-topics, their number and training time. In practice a compromise option can be found, particularly:

- Teacher forms standard exercises according to the sub-topics showing all the tests regarding this sub-topic;
- Certain points are assigned to the answers for any closed and open test;
- Partiality of assessment for any test and necessity for semantic analysis are showed for any test.

Before a start of the training 1) student selects topics and sub-topics, 2) defines number of the tests for the training according to the sub-topic and 3) sets total time for the training.

Teaching the subject using the simulator is done in the following sequence /Fig. 5/:

In the given scheme of knowledge acquisition we pay attention to the discussion of the training results – debriefing stage (debriefing is an English word meaning discussion of the results of the work done). In some way this is actually done in many cases by the lecturers, but has no clearly shaped form. Errors made by the students while working with the simulator are shown on the computer display of the lecturer in the form of report. Lecturer analyses it and discusses the results with the students after the training – debriefing.

The process of debriefing is informal. In the article A. A. Svistunov [Свистунов 2013] describes the tasks and aim of debriefing in details. It pays particular attention to the role of the lecturer as a leader in the process of the debriefing. He notes that the leader is obliged to create emotionally compatible environment. It is unacceptable to concentrate attention on the mistake done by the student, on the contrary it should be represented as group mistake and allow the members of the groups to express their thoughts. In such discussion ideas, problem solving ways and methods are generated among the group members.

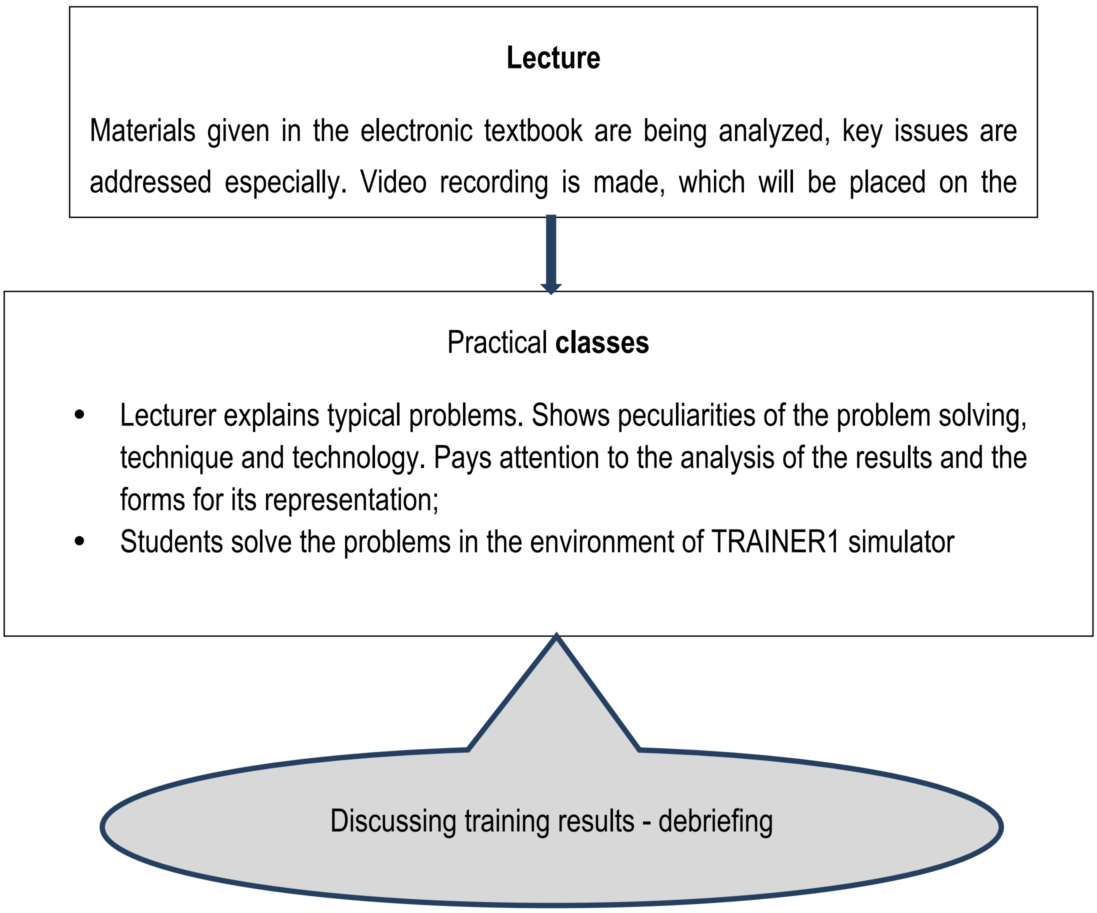


Figure 5. Teaching the subject using the simulator

Requirements set for the leaders in [Свистунов 2011] aren't new. They are formed fully and in details in the work regarding the method of managing intellectual processes, psycho-heuristic programming method [В.В.Чавчанидзе 1974].

Thoughts and proposals expressed by the students in the process of group discussion are taken into account by the lecturer while working on preparing materials and making corrections for the simulator.

Exercise created by the student can be represented before the start of the training graphically in a form of looped finite oriented G graph, where $\forall v_i, \mathrm{v}_i \in V$. From the set of the tests indicated in the exercise P is the selected test by probability. $\forall v_i, \mathrm{v}_i \in V$ is linked to the neighboring knot with one rib $e_i \in E$. $\forall v_i, \mathrm{v}_i \in V$ of G graph has loop (cycle). Number of the cycle depends on a student. Realization of the loop emerges in the system when the test is wrongly answered. Student can call out the "advisor" N times.

We have created electronic textbooks in economic subjects that have been functioning since 2013. These are: operational management, strategic management (author: Prof. Badri Ramishvili), financial functions in Excel (author: Prof. T. Minjishvili) and zero version of simulator [Tea Munjishvili 2013,....., Badri Ramishvili 2013]. Training is done using program package "Cyber1" [Tea Munjishvili 2014]. Considering negative and positive sides in the process of "Cyber1" exploitation, program-simulator TRAINER1 is elaborated in which abovementioned requirements are realized [Tea Munjishvili 2017].

Following are ensured using our computer simulator TRAINER1:

1. Formation of standard exercise by the teacher taking into account leading teachers in subjects, topics, sub-topics;
2. Formation of individual exercise by the student before the start of the exercise;
3. Possibility of elaboration of closed and open tests;
4. Existence of maximum three right answers out of seven options in the closed tests;
5. Existence of any number of the answers in open tests;
6. In the closed tests fixing the answers only after marking needed number of right answers and giving answer fixing command;
7. Using words, numbers, sentences or their combination, also abbreviations in the open tests;
8. Understanding indicative sentence used in the answers in case of breaking synchronization and inserting words;
9. Writing words used in the indicative sentences and general answers in any case, using wrong versions of these words;

10. Identification of actually written answer to the problems depicting certain subjects, topics, sub-topics, sub-chapters to the standard value of this answer. Breaking synchronization, writing words in any case, omitting or inserting words aren't allowed here;
11. Posing the problem using text, graph, video or their combination;
12. Existence of support on subject, test using text, graph video or their combination;
13. Existence of advices and recommendations for the mistakes in closed tests;
14. Existence of advices and recommendations for the mistakes in any answer in open tests;
15. Fixing respective points $q_i \in Q$ to the $\forall n_i, n_i \in N$ answer of the problem. Point can be whole or decimal positive number;
16. Receiving different kinds of diagnostic messages during the training process. For example, omitting words, numbers and etc., while using unknown words;
17. Ordering tests in a probability sense at the beginning of the testing process and selection of tests in a probability sense in the process of training;
18. Getting different analytical information after the training, particularly: report on the training process, advices and recommendations for avoiding errors made at specific training (debriefing), statistics of mistakes made in certain period of time and their representation by the form of diagram;
19. Generalization of the errors made by the students, group discussion, formulation and realization of proposals and recommendations.

Necessary conditions for using computer systems:

1. Proper work of technical system: computers, computative network, server, program means. This condition is an axiom. It doesn't require noting;
2. Original discussion of topics, sub-topics according to the syllabus and formation of relevant tests by the lecturers;
3. Creation of comprehensive support according to the typical test, represented in the form of text, graph, video information;
4. Periodically renewing information in the support base, tests, adding new tests, modifying existing ones.

Metods: In TRAINER1 engineering approach to the semantic analysis of sentence is realized, i.e. method of “productive grammar”, that bears certain universalism and is insured from the lack of table compatibility principle. Essence of this method, algorithms and programmed realization are discussed [Tea Munjishvili 2014; Tea Munjishvili 2015]. We describe algorithm of semantic analysis used in TRAINER1 briefly.

Indicative sentences are used in the possible answers of the open tests.

$$G = \bigcup g_j, \text{ j} = 1....\text{n}.$$

Elements of the G set g_i, $\text{i} = 1.....\text{m}$, are answers to open tests.

Formulating following conditions for G set:

1. G is preliminary known finite set. Elements of the set are natural language sentences or word orders;
2. Indicate sentence with conditional number i by g_i and the word in this sentence with conditional number λ by $g_{i,\lambda}$. Then used $g_{i,\lambda}$ words constitute L dictionary. Contents of G depend on demonstration of knowledge by simulator and assessment subjects and sentences used in open answers of these subject tests;
3. Any two elements of G differs by one word at least;

$$\forall(g_i, g_{i_0} \in G) \Rightarrow (g_{i_0} \setminus (g_{i_0} \cap g_i) \neq \{\} \wedge g_i \setminus (g_{i_0} \cap g_i) \neq \{\})$$

4. Any pair can comprise words with similar essence.
5. In TRAINER1 any answer of specific test, phrase $g_i \in G$ is unequivocally matched by certain command ψ_i, or $g_i \to \psi_i$.

ψ_i can be phrase, word, unity of symbols and etc. Hence, elements of G set $g_i \in G$ are represented on G^* set.

$$\psi_i \quad \in G^*, f : G \to G^* \quad \text{i.e.} \quad f(g_i) = \psi_i.$$

Desired results will be achieved if $\forall g_i \in G$, then g_i can be considered as production. g_i words $a_k \in L$ are considered as conditions, while ψ_i respective to g_i - as actions. In this sense “semantic analysis" of any indicative sentence is brought down to finding production system, organization of dictionary and relevant productions of incoming facts.

In the systems of knowledge demonstration and assessment, including TRAINER1, specific production (in our case – sentence) is selected, because we discuss specific answer with known value. It is necessary to define its relevance with existing one according to the incoming facts (words). Hence:

We have: not empty set of predicates (natural language words or word orders used in the open tests of the given subject) $L = \{a_k\} \neq \otimes,\ k = 1...n$, not empty set of activities $G^* = \{\psi_i\}, \mathrm{i} = 1..\mathrm{m},$not empty set of productions (indicative sentences) $G = \{g_i\}, \mathrm{i} = 1..\mathrm{m},$not empty set of informativeness. Elements of this set are scalar functions defined on G and they measure certain parameters of predicate by certain scale. Any criterion of informativeness has certain weight h_α so that

$$h_\alpha \in Q_+, \quad \sum \mathrm{h}_\alpha = 1.$$

Our tasks:

1. Reveal set of criteria of informativeness and ranging them according to their values;
2. Evaluation of each predicate from $a_k \in L$ dictionary by the given informativeness coefficient while organizing dictionary;
3. Understanding of the sentence given in the answer while entering n number of predicats in the system:
   - While changing word order in g_i;
   - While omitting and inserting words in any part of the sentence g_i.

Basis of the organization of knowledge base are answers to the open tests – G set of the sentences (productions). We can consider G as a text consisting with semantically unrelated indicative sentences, while in the production system – choosing production as searched form.

We have answers of two types – R and RT. In the first type of problems in R or RT type answers where the words and their synonyms can't be used in different case, only the words used in the answers of the tests or their incorrect versions will be selected and directed to the entrance of the system. Morphological basis, wrong versions and synonyms aren't indicated in the dictionary. In this case described algorithm for sentence identification is used only with the difference that at the entrance of the system in $L_w\{a_1, a_2,...a_s\}$ phrase the order of the words must be strictly protected and $L_w^{'} \equiv L_w$ .In the second type of problems while searching for relevant information for R or RT type answers written by the student synchronization can be broken. Words can be used in any case and conjugation, wrong versions, and excess-any number words that exist or not in the dictionary can exist.

Despite its simple character proposed way of understanding indicative sentence can be considered unnecessary because, in many cases students try to write a sentence with the sequence familiar to them. Words are used in many cases and the words are omitted and inserted frequently. Writing erroneous versions of the words is rare.

Identification of open test answer – sentence is made in a following way: /Fig. 6/

1. Suppose natural language words $a_0 \in L$, $L_w\{a_1, a_2,...a_s\}$ have come into the system. a_0 word can be wrong version of the word written in the answer, word written in any case, synonym or abbreviation.

↓

2. a_0 Word is being searched in L_1 dictionary. If one of the words from $g_i \subset G$ phrase regardless of its position coincides with a_0, or its base regardless of its position coincides with a_0, word is identified, otherwise next step is made.

↓

3. Finding synonyms. a_0 Word is being found in L_1 dictionary. If morphological basis of a_0 is found, then identification of the respective $g_i \subset G$ word is made by using L_1 dictionary.

↓

4. a_0 Word is being searched in L_3 dictionary. If the words respective to abbreviation in the dictionary coincide with the words $a_k \in g_i$ used in the answers in sequence then abbreviation is identified.

↓

5. If all words in open test answer $g_{i,\lambda} \subset G$ regardless their position coincides to $L_w\{a_1, a_2,...a_s\}$ set or $L_w^{'} \in L_w$ sub-set, then the searched phrase is identified, otherwise the message regarding specification of the answer shows up.

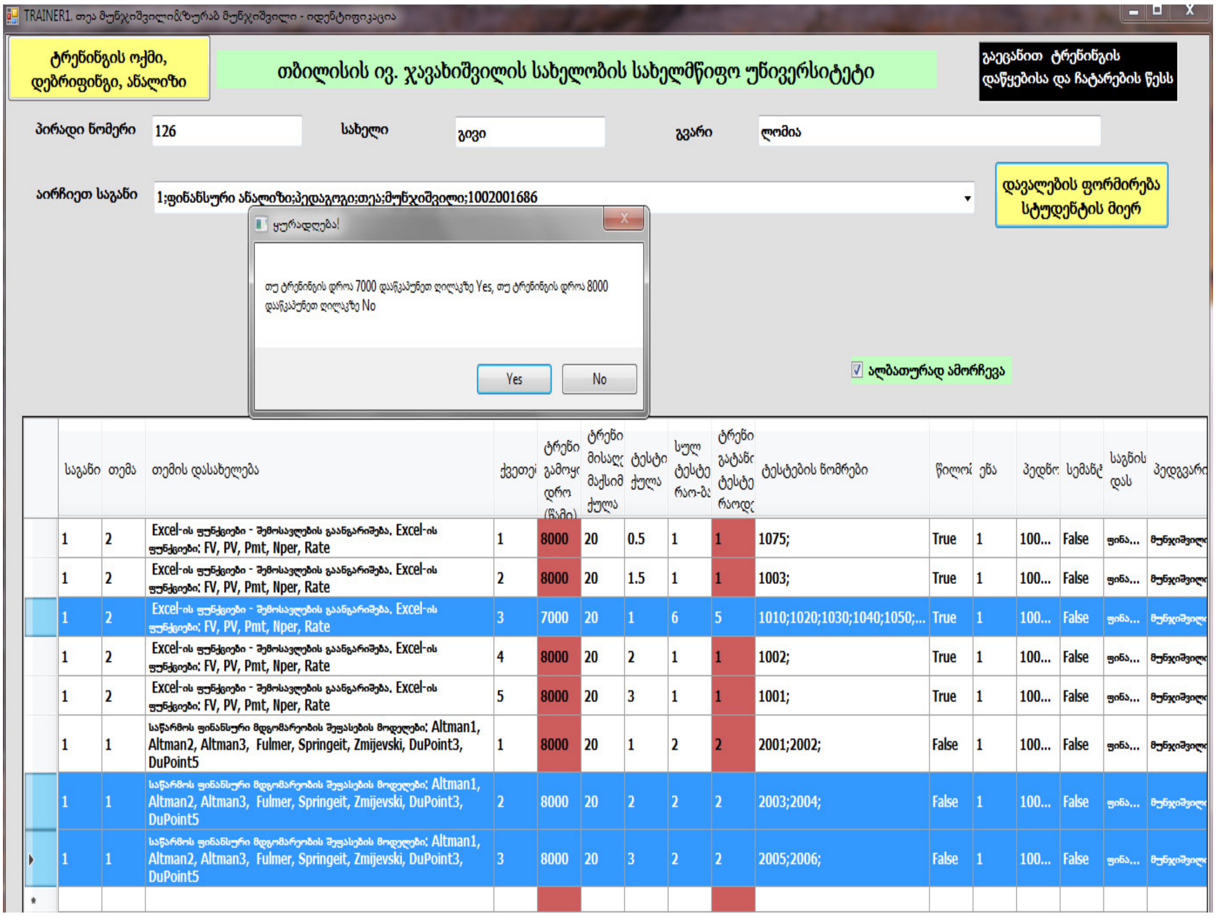
Figure 3. Stages of semantic analysis of the sentence in TRAINER1

TRAINER1 program is written in VB.NET 2010, database is organized on SQL Server 2008. System is multilingual. Information on the same subjects can be represented in several languages.

Implementation of results

Down we discuss the process of starting and running of training with TRAINER1:

Any person interested in can conduct training. Person is identified by ID number or any number consisted of 16 digits. After identification and subject selection window will appear with the list of standard exercises from selected topic written by the teacher according to the sub-topics. Student can change number of the tests for the training and training time in any standard exercise. Before selection of sub-topic student can view at the tests in the exercise. /pic 1/

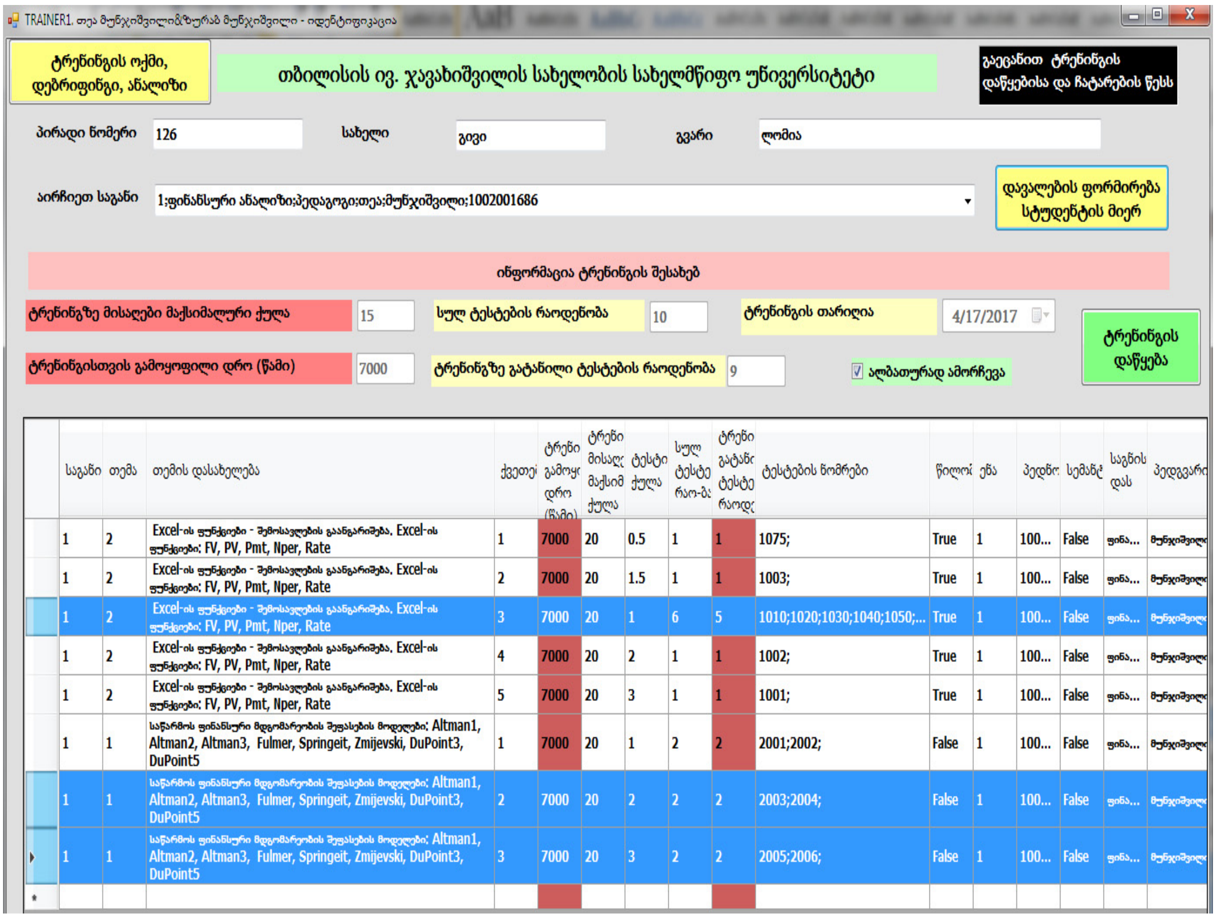


Pic. 1. Formation of individual task of training by the student on the basis of standard exercise

Following situation is depicted on the picture:

- Table pillars: number of the tests for training, training time. Values that can be changed are colored;
- Number of tests for the training on third sub-topic of second topic is changed (5 is written instead of 4);
- Total time for the training is changed (7000 seconds is written instead of 8000 seconds);
- Third, seventh and eighth lines in the table are marked to bring on training, i.e. second and third sub-topics from topic 1 and third sub-topic from topic 2;
- After pressing on a button “formation of exercise by student” message of choosing training time will appear.

Window appearing after pressing on a button “formation of exercise by student” is following: /pic.2/

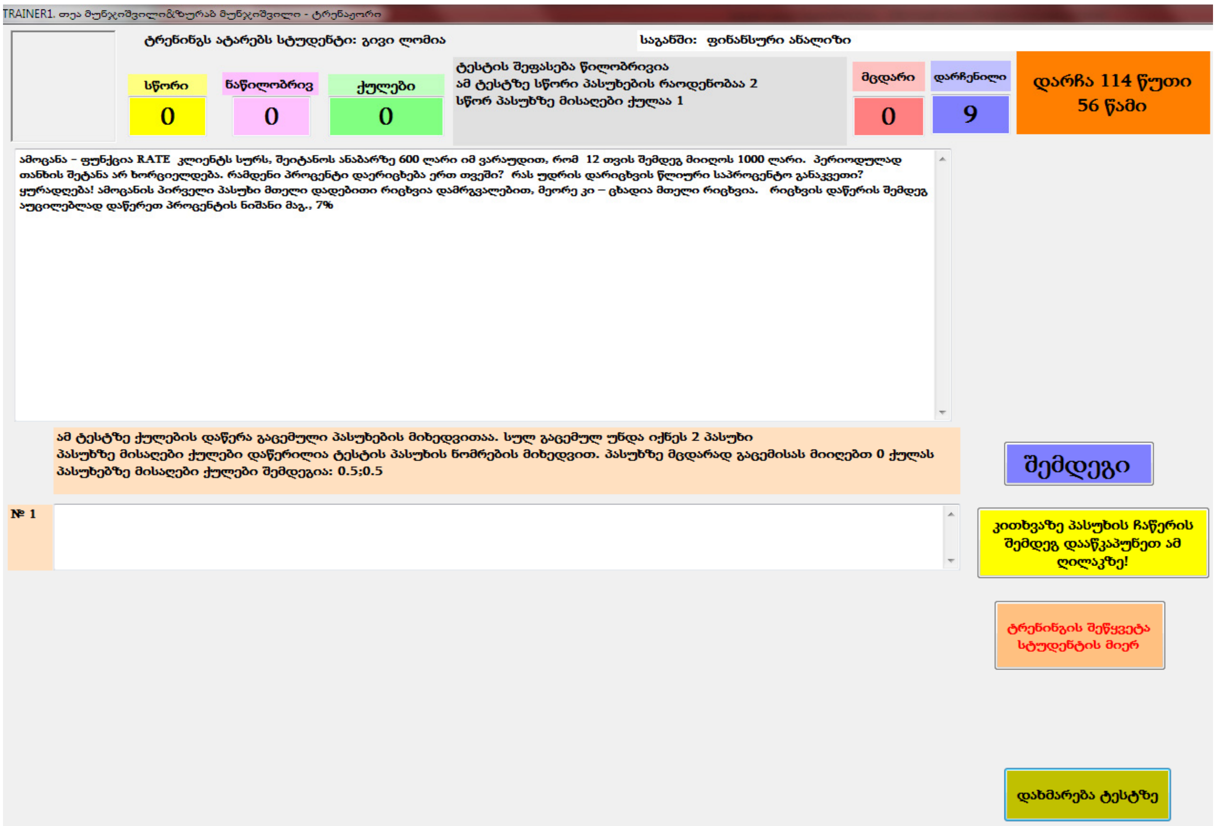


Pic. 2. Individual task for training formed by the student

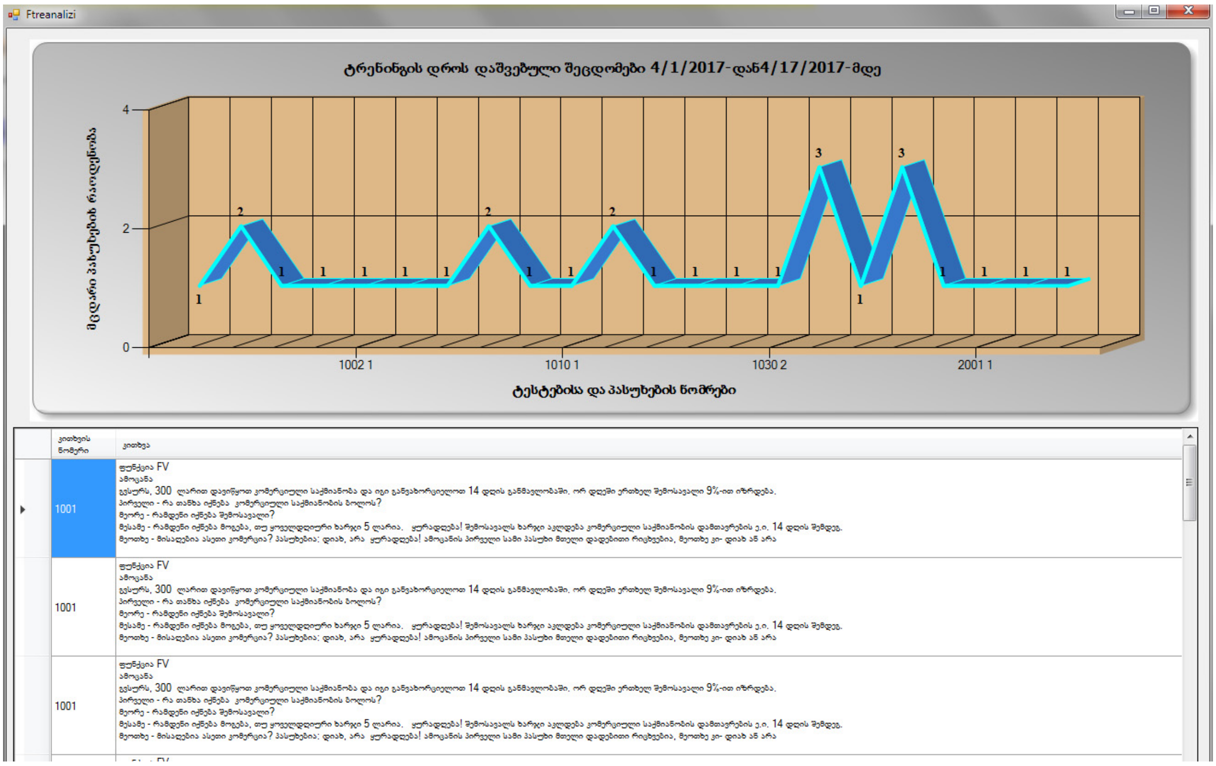
It is showed on the picture that time for the training is 7000 seconds, from the 6 tests of third sub-topic of second topic 5 tests are brought on the training. In total 9 tests are brought on the training, possible points to get is 15. 5 tests are open and 4 tests are closed (assessment of the open tests are partial, while that of closed tests isn't).

After pressing the button “start training” the training starts. On the /pic.3/ open type test is shown with two answers for it. Respective points for answers are also shown. “help on test” button is active in the same window. Pressing on it window will appear providing solution (solution method, algorithm, model and etc.) for the problem.

Student can conduct training everyday N times, get advices and recommendations for error correction made in the test. After finishing any training student can see training report, advices, recommendations and statistics of the errors made by him/her. /pic. 4/



Pic. 3. Window for listing problem solution results for open tests



Pic. 4. Statistics of errors made at training

Errors made in the certain time of trainings conducted by the students with the test number and correct answer are given on the picture. In the table in same window contents of the tests given on the diagram are shown.

Below we will discuss modeling of financial stability of enterprise by FINSIM1.1 simulator Altman model. In FINSIM1.1 three models of Altman are given. One of them named Altman1 is following /7/:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$$

where:

X_1- working capital/ total assets

X_2- undistributed profits/total assets

X_3- profit before interest and taxes/total assets

X_4- market value of capital/liabilities

X_5- sales/total assets

As it can be seen from the model, coefficient X1 represents the measure of liquidity. According to the author Altman losses in the company reduces this coefficient first of all.

Coefficient X2 is a kind of financial leverage, because in the firms with high level of this coefficient assets are financed more by own sources than by borrowed financial resources.

Coefficient X3 shows actual profitability without taxes and interest rates paid.

X4 market value of capital includes common and preferred stocks, short-run and long-run liabilities. Actually, using this coefficient lower threshold of the value of company's assets is defined, when it is below the liabilities.

X5 shows number of sales company's assets can ensure.

Actually 7 variables are included in the model. After systematization of names of variables formula (1) gets following form:

$$Z = (1.2 * X2 + 1.4 * X3 + 3.3 * X4 + X7) / X1 + 0.6 * X5 / X6 \quad (2)$$

Where:

X1 – total assets;
X2 – working capital;
X3 – undistributed profit;
X4 – profit before interest and taxes;
X5 – market value of capital;
X6 – liabilities;
X7 – sales.

Based on the values of Z we can talk over financial conditions of the firm, particularly:

Firm is in a safe zone		$Z > 2.99$
Firm is in a grey zone	=	$1.81 < Z < 2.99$
Firm is under the risk of buncrupcy in two years		$Z < 1.81$

We make simulation using actual data of actual enterprise. In the table 1. Values of $X_i \ \mathrm{i} = \overline{1,7}$ variables for “X” JSC 2010-2016 are given:

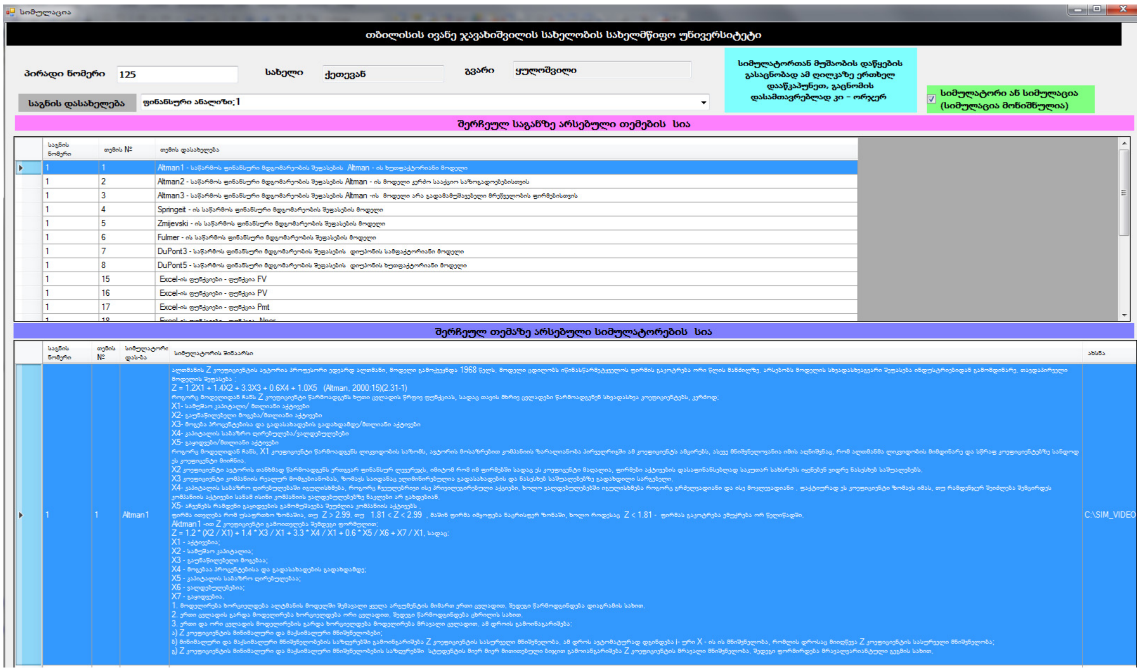
Name	Years						
	2010	2011	2012	2013	2014	2015	2016
Total assets (GEL)	29334	39107	45953	40776	21 762	20 089	14 002
Working capital(GEL)	12198	15879	17776	14482	10 182	9601	6454
Undistributed profits(GEL)	10136	13228	17177	16294	11 580	10 488	7 548
Profit before interest and taxes(GEL)	2128	2690	4339	4325	6 403	6 195	3 368
Market value of capital(GEL)	736	1886	3424	2511	4 788	2961	2530
Liabilities(GEL)	439	1523	2521	1767	519	699	741
Sales(GEL)	5 597	6 324	4 233	5 983	6760	3040	4 1644

After calculating minimum, maximum and mean values of variables in program based on the table 1 we get new table with these results.

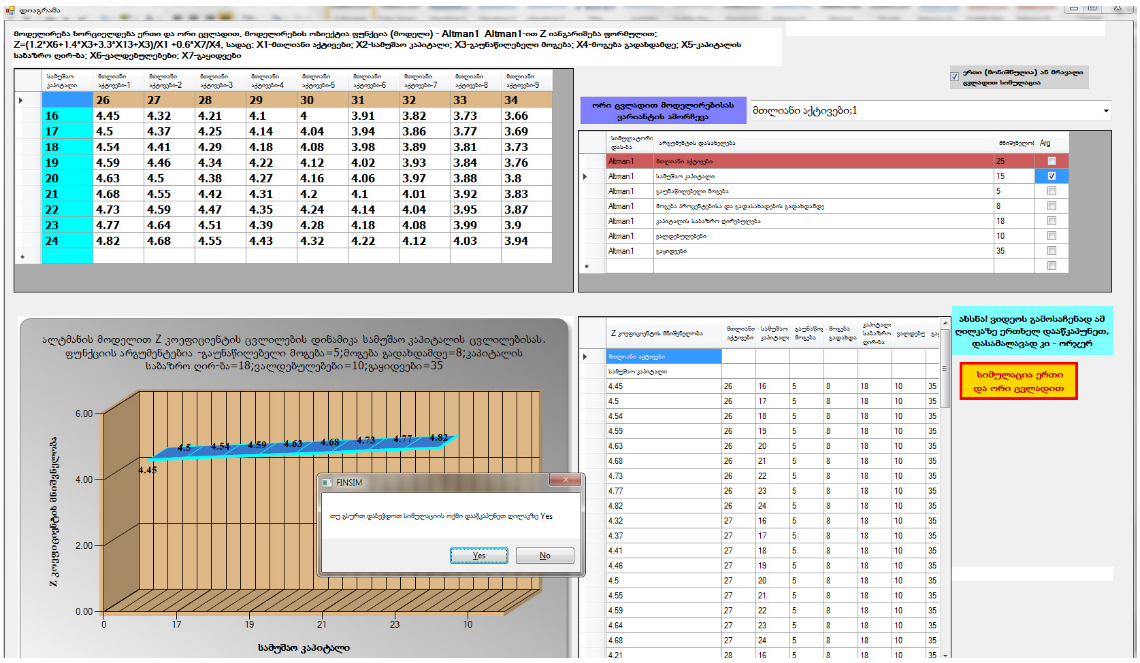
Using Altman model FINSIM1.1 simulation of financial stability of the enterprise is done in a following way:

1. After calling out the program window will be showed up in which working language with the system can be selected – Georgian. After selecting the language window will appear in which after student identification and selection of subject and topic the list of simulators regarding selected subject and topic will be scrolled down /pic. 5/.
2. After clicking on the name of the simulator respective window will appear.

Basis of the simulation is Altman model. Values of arguments (variables) can be changed by the student. Modeling the process is made by one or two variables, because of this, indication of $\vee x_i$ variable and modeling with this variable are taken into account. Result of the modeling will be received in a form of diagram and table. /pic. 6./



Pic.5. Window after student identification and selection of subject and topic.

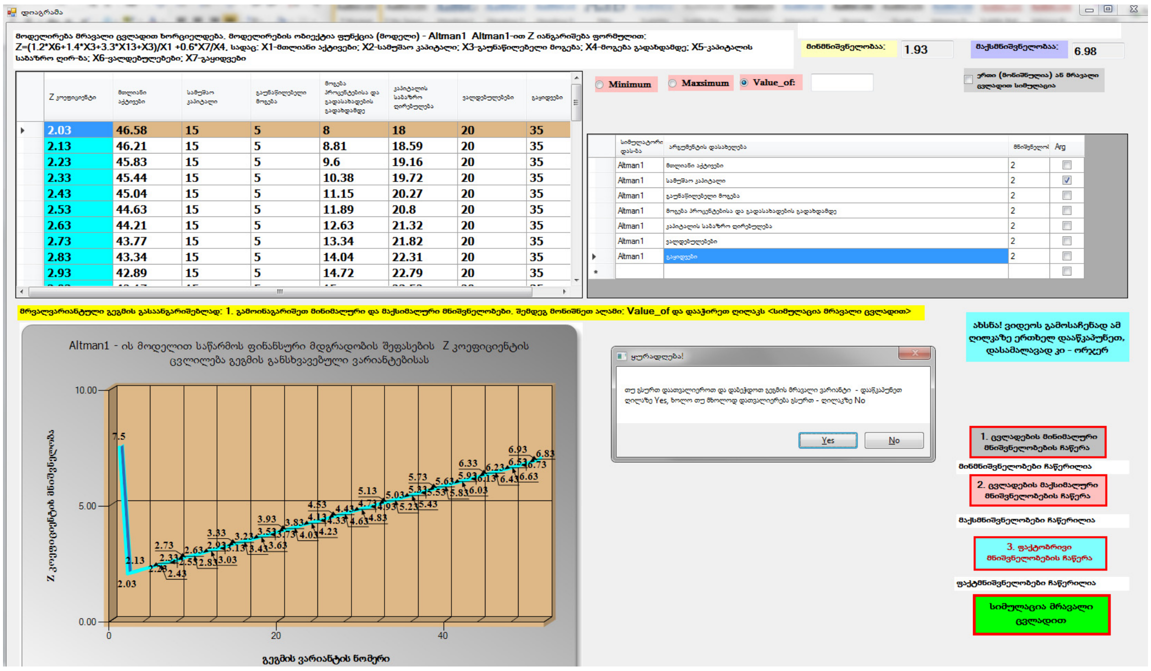


Pic. 6. Result of the simulation by one and two variables

On the diagram dependence of Z coefficient (function) on the changing of value of variable X_i is shown, while in the table dependence of change in values of Z coefficient on the changes in X_i and X_{i+1} variables is shown. After clicking on an “explain” button window will appear showing explanation of simulation result.

Modeling by multiple variables

Minimum, maximum and actual data of $X_i\ i=\overline{1,7}$ variables calculated based on table 1 are given as a basis plan for modeling. In case of absence of the data minimum, maximum and initial data of variables are brought in by the student. Value of Z coefficient is calculated for the minimum and maximum values of variables and multi-variant modeling is being made by certain step in $Z_{min} < Z < Z_{max}$ range. As a result of modeling $\forall Z_J \rightarrow X_i\ , i = 1 \dots 7$ can be seen, i.e. value of any Z coefficient from selected values is defined by certain values of the variables. This allows us to define optimal values of the parameters of financial stability of enterprise using Altman model and after that control and manage their values in the process of functioning of enterprise. On a picture 10. Values of the variables respective to the different values of Z coefficient taken by modeling are shown. Multiple variants of plan is given on the picture 8.



Pic. 7. Results got by multiple variable modeling using Altman's 5 factors model

	Current Values:	2.03	2.13	2.23	2.33	2.43	2.53	2.63	2.73	2.83	2.93
Changing Cells:											
მთლიანი აქტივები	25.00	46.58	46.21	45.83	45.44	45.04	44.63	44.21	43.77	43.34	42.89
სამუშაო კაპიტალი	28.96	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
გაუნაწილებელი მოგება	10.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
მოგება პროცენტებისა და გადასახადების გადახდამდე	15.00	8.00	8.81	9.60	10.38	11.15	11.89	12.63	13.34	14.04	14.72
კაპიტალის საბაზრო ღირებულება	30.00	18.00	18.59	19.16	19.72	20.27	20.80	21.32	21.82	22.31	22.79
ვალდებულებები	10.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00
გაყიდვები	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00
Result Cells:											
O1											

Pic. 8. Multiple variances of plan got by multiple variable modeling using Altman's 5 factors model

Conclusions and recommendations

1. It is not a panacea to use simulator in the teaching process including our simulator. It is one of the means for deep understanding of the issue in a short time. Effect of the simulator is achieved with the use of using simulator with respectively designed electronic textbook. Under electronic textbook we consider a textbook designed using hypertext technology, subtle search system, represented by the graph and video;
2. Integral part of training, simulation is discussion and analysis of the results – debriefing. For conducting group debriefing by the teacher using FINSIM1.1 the report describing training, simulation is made;
3. Knowledge acquisition system TRAINER1 is attributed to the class of intellectual systems for knowledge acquisition. It is an integral part of distanc learning;
4. Defining individual exercises by the students within the training allows taking into account abilities of student as much as possible and paying attention by the student to the key issues for him/her;
5. Giving advices and recommendations timely in case of mistakes allows repeating solution of the problem several times that makes the process of knowledge acquisition easy and fast;
6. Semantic analyzer of indicative sentence in TRAINER1 allows for taking into account peculiarities of problem solution;
7. Adaptation of TRAINER1 to specific subject is made by inclusion of information describing this subject in the database of TRAINER1;
8. Necessary condition for using TRAINER1 is preparation of information describing the subject, particularly: single out topics, sub-topics; form typical problems in any sub-topic; explain problem solution method, algorithm; experience for solving such problems; characterize typical errors made in problem solving and show the ways of their improvement;
9. TRAINER1 allows the information describing the same subject to be represented by different teachers (indicating their authorship): problems, solution methods and etc.

Bibliography

1. https://www.bpsimulator.com/ru/business/
2. А.В. Трухин, АНАЛИЗ СУЩЕСТВУЮЩИХ В РФ ТРЕНАЖЁРНО-ОБУЧАЮЩИХ СИСТЕМ, http://ido.tsu.ru/files/pub2008/8.pdf
3. Методы и принципы симуляционного обучения, Свистунов Андрей Алексеевич, http://rosomed.ru/kniga/metodi_i_principi_simulationnogo_obucheniya.pdf
4. qumcpo.minsk.edu.by/ru/main.aspx?quid=1761

5. Гипертекстовые тренажеры для обучения по экономическим дисциплинамтема диссертации и автореферата по ВАК 08.00.13, кандидат экономических наук Ельцин, Андрей Владимирович, http://www.dissercat.com/content/gipertekstovye-trenazhery-dlya-obucheniya-po-ekonomicheskim-distsiplinam#ixzz439UL0S43
6. Рышкевич В. М. Симуляционные формы обучения в практике преподавания экономических дисциплин и повышения квалификации [Текст] // Теория и практика образования в современном мире: материалы III междунар. науч. конф. (г. Санкт-Петербург, май 2013 г.). — СПб.: Реноме, 2013. — С. 162-163. http://www.moluch.ru/conf/ped/archive/70/3864/
7. Altman Edward I., Predicting Financial Distress Of Companies: Revisiting The Z-Score And Zeta® Models, July 2000
8. Tea Munjishvili, Zurab Munjishvili, Knowledge demonstration and assessment system “Cyber1”, international Journal "Information Technologies & Knowledge" Volume 8, Number 3, 2014, pp. 271-279.
9. Tea Munjishvili, Zurab Munjishvili, Knowledge demonstration and assessment system “Cyber1”, international Journal"Information Technologies & Knowledge" Volume 8, Number 3, 2014, pp. 271-279.
10. Z.Munjishvili, problem-oriented method of semantic analysis for sentence of natural language. Collection of Knowledge, dialogue, decision, Kiev, Ukraine, “Naukova Dumka”, 1990.
11. Munjishvil T., Munjishvil Z., Nakashidze V. System of knowledge revealing and rating – “Cyber 2”. 9th MIBES ANNUAL INTERNATIONAL CONFERENCE 2014 THESSALONIKI, GREECE, 30/5–1/6 .CD ISBN# 978-960-93-6161-3. pp. 111-121.
12. Thea Munjishvili, Zurab Munjishvili. "The system of Discovery and Estimation of Knowledge “Cyber2”", Scholars' Press, Saarbrucken HRB 18918.Published on: 2015-01-15 Number of pages: 108. Book language: English. ISBN-13: 978-3-639-76094-1.
13. Tea Munjishvili, Zurab Munjishvili. The semantic analysis method and algorithms of open tests answers on "Cyber-2" patternin the Knowledge revival and evaluation systems. 2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems(ICICIS 2015), Volume 3, 12 - 14 December, 2015, Cairo, Egypt pp. 50-55

Authors' Information

Tea Munjishvil- *Iv.Javakhishvili Tbilisi State University; Chavchavadze Av.1, 0129, Tbilisi, Georgia.*

e-mail : tmunjishvili@gmail.com

DEVELOPMENT OF INTELLECTUAL MODELS AND METHODS OF EXPERT SYSTEMS OF CLINICAL MEDICINE

Oleksandr Kuzomin, Oleksii Vasilenko, Tetiana Tolmachova

***Abstract**: The study examines the processes of modeling clinical medicine, the development of analytical methods, the synthesis of biomedical data and knowledge, as well as the creation of expert systems of clinical medicine. We propose ways of developing models and methods for analyzing biomedical Big Data and developing expert-analytical systems of clinical medicine.*

***Keywords**: Open EHR, Web Mining, OLAP, Data Mining, diagnostic criteria, quality data model, natural language processing, cTAKES, conditional random fields.*

***ITHEA Keywords**: J.3 Life and Medical Sciences, D.2.11 Software Architectures, H.3.3 Information Search and Retrieval*

The problem of clinical diagnosis and statement of the research task

The main issue that arises before a doctor when he meets a patient is the diagnosis - a brief medical report on the nature of the disease and the patient's condition.

In the literature [Yager and Mcintyre, 2013, Organization WH, 2011, CMS Quality Data Model, 2015, EHR Incentive Programs, 2017] there is no unified conception of the diagnostic process, its logical structure often describes only the external formal aspects of clinical thinking and logic (Table 1). Textbooks and manuals do not always have a strict system of exposition of the semiotics of diseases and factual material, often there are no clear indications of which symptoms are significant and which are secondary, to what extent these symptoms are permanent and specific for a particular disease.

There are many varieties of clinical diagnoses, they can be formed by the method of construction, by the time of detection, by the degree of validity (Figure 1). In practical medicine, only three types of diagnosis are most often used: preliminary, basic (clinical) and final, which reflect certain stages of diagnosis.

Table 1. Problems of Clinical Medicine

#	The problem of clinical medicine	Direction of research in work
1	Extraction of data and knowledge from the Internet, social networks and medical practices necessary for analytical and algorithmic solutions to the problems of clinical diagnosis	1. Use the electronic health record (EHR) for data normalization. 2. Combining the best results from the clinical archetypes of the Open EHR, guides and ontologies. 3. Using Web Mining: intellectual analysis of data on the Internet. 4. Logical programming to identify knowledge.
2	Choice of models and methods for qualitative, reliable and timely analysis and algorithmization of clinical diagnostics	1. Big Data (using accurate and reliable large samples, IBM Watson Analytic). 2.OLAP 3.Data Mining
3	Multi-criteria analysis and optimization under uncertainty	The use of the Bayesian network of trust
4	Selection of models and methods for processing biomedical Big Data	Hadoop Technologies: 1. Map Only 2. Classic MapReduce 3. Iterative Map Reducse or Map-Collective
5	Ensuring reliable storage of biomedical Big Data	Backup Modification

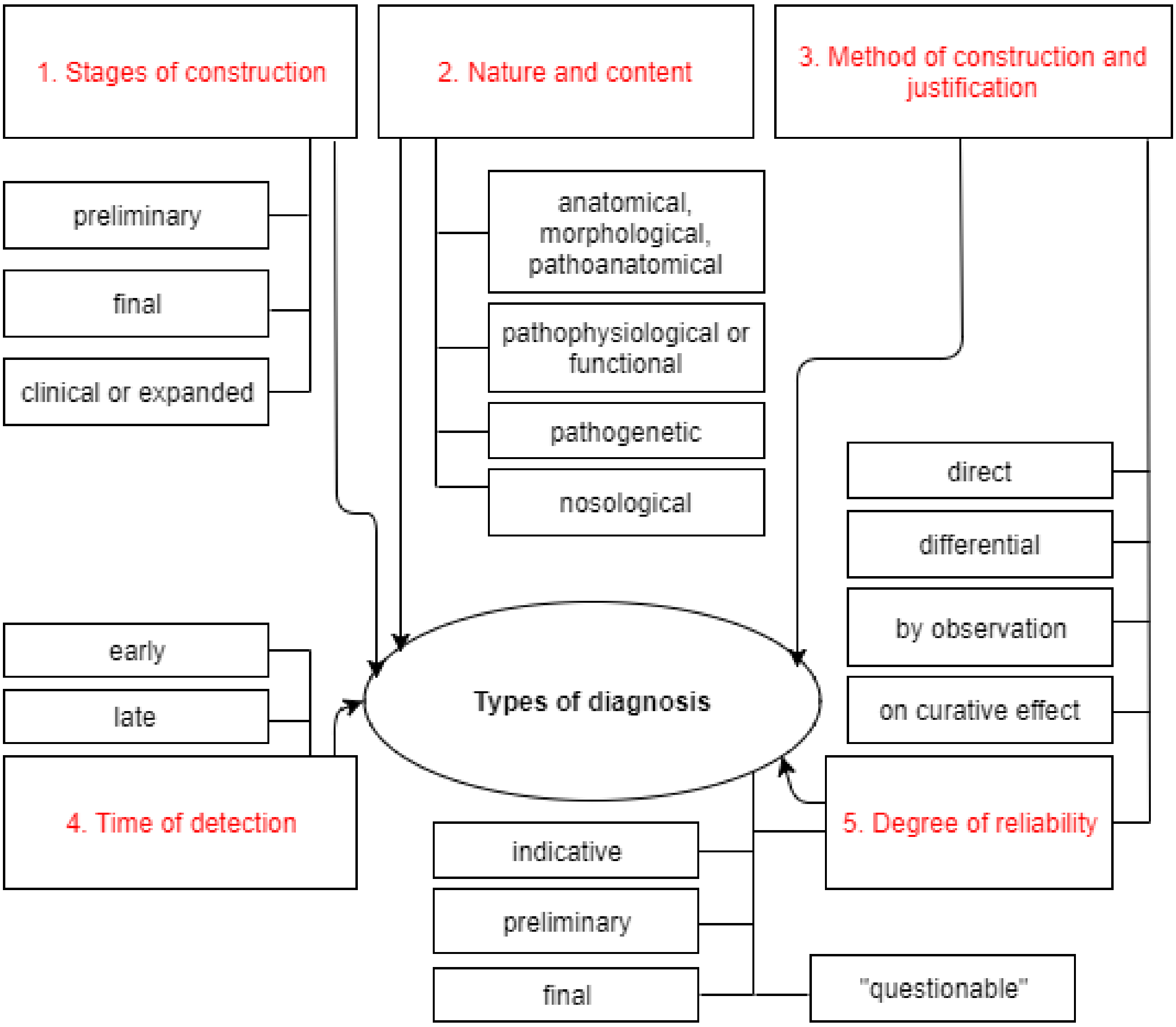


Figure 1. Varieties of medical diagnosis

Logical activity of the doctor-diagnostician is carried out in such forms as concept, judgment, inference, induction and deduction, analysis and synthesis, the creation of ideas and hypotheses. Most often in the diagnosis used inductive method, the method of analogy. If the doctor moves from simple to complex with the use of an inductive method of diagnosis, in the case of analogy, he seeks to learn and recall what he already knows that he has met before, comparing the similarities and differences of symptoms in the observed patients with the symptoms of known diseases.

From the above, we can draw the following conclusion about the problem of diagnosis by traditional methods: the uncertainty of the initial data and the methods of formulating the diagnosis depend to a large extent on the ontological experience of the doctor, his ability to diagnose using an inductive method, the analogy method. If the doctor moves from simple to complex with the use of an inductive method of diagnosis, in the case of analogy, he seeks to learn and recall what he already knows that he has met before, comparing the similarities and differences of symptoms in the observed patients with the symptoms of known diseases.

The next problem of diagnosis is the ambiguous conformity of symptoms and disease to humans. In addition, diagnosis is often difficult, because many signs and symptoms are nonspecific. Diagnosis is often determined by related symptoms. Often, to model the patient’s condition is used Bayesian models. For example, redness of the skin (erythema) in itself is a sign of many disorders and thus does not tell the health care professional what is wrong.

Currently, the most important conclusion about the possibilities of choosing the direction for solving problems in practical medical diagnostics is associated with the use of ontological, precedent Big Data, which are used for their processing with the help of Data Manning analysis, prediction and diagnosis of diseases. As a result, in addition to qualitative diagnostics, it is possible to develop machine learning and diagnostic algorithms.

Ontological, case-law and knowledge databases Big Data analyze textual positive and negative descriptions from medical clinical experience, the Internet and social networks.

Currently, the natural language processing tool (NLP), known as cTAKES [Yager and Mcintyre, 2013, Organization WH, 2011, CMS Quality Data Model, 2015], is used to detect sentences and annotate events in diagnostic criteria. cTAKES consists of several components. Each of them has unique qualities and capabilities. Each component includes at least one analysis engine (annotator), some of which include more. You will want to evaluate the utility of each component for you. UIMA provides a toolkit for selecting which annotators are used together, and how annotators are executed. Each section here covers one component.

cTAKES provides two versions of the source cTAKES pipeline that detects named objects and assigns them attributes:

- For processing text notes: cTAKESdesc / cdpdesc / analysis_engine / AggregatePlaintextProcessor.xml;
- For processing formatted notes in the field of clinical documentation (CDA): cTAKESdesc / cdpdesc / analysis_engine / AggregateCdaProcessor.xml.

Both versions use the same set of components, except that the document preprocessor is not used for plain text.

Suggested solutions

The object of the study are the processes of modeling clinical medicine and the development of methods for analysis and synthesis of biomedical data and knowledge and the creation of expert systems of clinical medicine.

The subject of the study are models and methods of analysis of biomedical Big Data and the development of expert systems of clinical medicine.

The 11th revision of the International Classification of Diseases (ICD-11) was officially published by the World Health Organization (WHO) in March 2007 [EHR Incentive Programs, 2017]. Within the WHO Thematic Advisory Group on Health Informatics and Modeling, a three-level content model was proposed and discussed [Organization WH, 2011]. The purpose of the ICD-11 content model is to present the knowledge underlying the ICD object definitions. Starting in May 2012, the beta phase of the revision of ICD-11 intends to accept public input through a distributed authoring model. One of the main uses is the creation of textual definitions for each category of ICDs. Textual definitions are written down by WHO as follows: "Each ICD concept will be accompanied by a written definition of its descriptive characteristics. This full text definition allows human users to understand the meaning of the concept for classification, translation and other reasons".

The term "diagnostic criteria" refers to a specific sequence of signs, symptoms and test results that clinicians use to determine the correct diagnosis [Kuzomin and Vasylenko, 2014]. This is one of the most valuable sources of knowledge that can be used to support the adoption of clinical decisions and improve patient care [Kuzomin and Vasylenko, 2017]. However, existing diagnostic criteria are scattered across media, such as medical textbooks, literature and clinical practice guidelines, and they are usually described in unstructured free text without a single standard. This situation prevents the effective use of diagnostic criteria to support modern clinical decision-making, which requires an integrated system with interoperable and computable processes.

One of the solutions to better support the adoption of clinical decisions is the computerization of these diagnostic criteria; however, specialists and clinicians are expensive and time consuming to perform all tasks manually. For this purpose, on the one hand, natural language processing technology (NLP) can be used to automatically or semi-automatically convert diagnostic criteria into computable format. On the other hand, the data model for representing diagnostic criteria is equally important for its computerized implementation. Such data model will allow to display diagnostic criteria in a structured, standard and coded structure to support many clinical applications in a scalable manner. To explore the adaptability of the model to diagnostic criteria, [Kuzomin and Vasylenko, 2014, Kuzomin and Vasylenko, 2010] through a data-based approach in which manually analyzed the distribution and coverage of data elements extracted from a set of diagnostic criteria in QDM. The results showed that the use of QDM is possible when building an information model based on standards for the presentation of computable diagnostic criteria [Kuzomin and Vasylenko, 2010].

The purpose of the study is to develop and evaluate automated methods for converting text clinical diagnostic criteria into a structured format using QDM.

It is proposed to conduct research on the analytical and expert system using biomedical Big Data, obtained as a result of searching the Internet, social networks, literature and the results of practical clinical medicine. Given that the main problem in the formation of biomedical Big Data is the processing of large text documents to facilitate the computerization and standardization of diagnostic criteria, very known clinical tools of NLP and original methods and models are used in the initial stages of the statistical analysis of signs and symptoms of diseases, based on the so-called the concept of microsituations [Kuzomin and Vasylenko, 2017] and self-organizing analytical Data Maning algorithms for textual descriptions of diagnostic data. In particular, we use a combination of known methods of clinical text and knowledge analysis (cTAKES) - supported and rule-based methods for extracting an individual diagnostic criterion from full-text clinical diagnostic criteria and proposed in the study of original methods and models of analysis, forecasting and machine learning. We also develop an algorithm for machine learning based on conditional random fields (CRF) for automatic annotation and classification of the attributes of diagnostic events. Finally, we are developing an integrated web-based system that automatically converts text diagnostic criteria into a standard QDM template by implementing algorithms.

The WHO Content Model ICD-11

WHO developed a content model for presenting the knowledge that underlies the ICD object definition [Kuzomin and Vasylenko, 2010, Yager and Mcintyre, 2013]. The content model consists of three layers:

- The base layer;
- Linearization layer;
- Ontological layer.

The main layer is the main product of the revision of ICD-11, which stores the full range of knowledge about all the classification units in the ICD. Each ICD object can be seen from different dimensions. The content model represents each of these dimensions as a parameter. Currently, 13 basic parameters for the category description in ICD-11 are defined in the content model, as shown in Table. 2.

Table 2. The ICD11 Content Model Main Parameters

#	Parameters
1	ICD Entity Title
2	Classification Properties
3	Textual Definitions
4	Terms
5	Body System/Structure Description
6	Temporal Properties
7	Severity of Subtypes Properties
8	Manifestation Properties
9	Causal Properties
10	Functioning Properties
11	Specific Condition Properties

"Diagnostic criteria" is one of the main parameters for describing the ICD category.

NQF QDM

QDM consists of two modules: a data model module and a logic module. The data model module includes category concepts (eg, medicines), a data type (for example, a "drug controlled"), an attribute (for example, dose, route, strength and duration information) and a set of values containing conceptual codes from one or more terms. The logical module includes logical operators, functions, comparison operators, time operators, subset operators. As mentioned above, HQMF provides a standard format for displaying criteria based on QDM (i.e., instance data) in XML format using a set of templates. [Kuzomin and Vasylenko, 2010, Yager and Mcintyre, 2013, Organization WH, 2011] proposes an assessment of the feasibility of using QDM to represent diagnostic criteria using a data-based approach and the assumption that generic patterns informed by QDM are useful and feasible in constructing a standard-based information model for computable diagnostic criteria. In this study, we used generic templates and selected a set of QDM data types and attributes to develop an initial ontology.

The architecture of the expert-analytical system of clinical diagnosis

Based on the accepted general structure of the system for creating criteria for a clinical diagnosis based on rules is shown. The system architecture contains two main modules: one is a development module that uses a standard information model, and the other is a translation module that uses SWRL. The first architecture module contains an initial ontology that supports the organization of the elements of the diagnostic criteria. The collection of the collection of the manually selected elements of the ICD-11 content model and QDM elements is used, which were informed by an analysis of real diagnostic criteria. The first module also contains a unified web user interface that supports the collection and development of diagnostic criteria by clinicians or subject matter experts on the Internet. The standard QDM model serves as the base layer for translation and reasoning. All collected data items, value sets, and logical expressions of diagnostic criteria are formalized using HQMF templates based on QDM. The second architecture module contains a rule transformation mechanism that converts the diagnostic criteria presented in the QDM / HQMF format into an ontology of the diagnostic criteria for a particular domain and a set of rules using SWRL.

The model of the diagnostic environment at time t:

$$P' \rightarrow Mod_{\Pi C} \leftrightarrow Opt_{\mathfrak{R}}(P'_1, P'_2, \dots, P'_j, \dots, P'_{13}) \quad (1)$$

where $P' = \{P'_1, P'_2, \dots, P'_{13}\}$ - a set of statistical and analytical procedures for the transformation of information; P'_1 - procedure of multifactor analysis and ranking of parameters of the diagnostic environment; P'_2 - procedure for regression analysis of medical signs; P'_3 - procedure for the variance analysis of medical features; P'_4 - procedure for classifying situations by referring to a confirmed diagnosis and unconfirmed diagnosis; P'_5 - procedure of cluster analysis of symptoms; P'_6 - Data Mining procedure; P'_7 - procedure for compiling a tree for selecting diagnostic solutions; P'_8 - the GMDH procedure; P'_9 - disease pattern recognition procedures; P'_{10} - procedure using fuzzy logic; P'_{11} - procedure for the determination and use of many micro-situations; P'_{12} - procedure for assessing the risk of disease diagnosis; P'_{13} - procedure for synthesizing methods of analysis, forecasting, machine learning.

$$P'_{11} \rightarrow \{Diag'\} \qquad (2)$$

where $\{Diag'\}$ - set of medical diagnostic micro-situations.

As a result of the conducted studies based on the results of preliminary analysis of a priori data built a diagnostic model environment for the symptom j and a parameter i:

$$Diag^{j}_{\Pi C_i} = f(X_1^{1,f}, X_2^{2,f}, \dots, X_i^{j,f}, \dots, X_N^{M,f}) \qquad (3)$$

where $j = \overline{1, M}$ - number of diagnostic micro-situations, M' - number of symptoms when $M \in M'$, $\{M'_1, M'_2, \dots, M'_f, \dots, M'_F\}$ - set of symptoms, $i = \overline{1, N}$ - number of medical signs of the disease, $\{X_i^{j,f}\}$ - a variety of medical signs of the disease diagnostic environment.

A generalized model of a problematic medical diagnosis $Diag_t$ at a controlled point in time t in:

$$Diag_t = \begin{cases} Diag'_{1,1} = f(X_{1,1}^{M'_1}, X_{2,1}^{M'_1}, \dots, X_{i,1}^{M'_1}, \dots, X_{N,1}^{M'_1}) \\ Diag'_{2,2} = f(X_{1,2}^{M'_2}, X_{2,2}^{M'_2}, \dots, X_{i,2}^{M'_2}, \dots, X_{N,2}^{M'_2}) \\ \dots \\ Diag'_{j,d} = f(X_{1,d}^{M'_f}, X_{2,d}^{M'_f}, \dots, X_{i,d}^{M'_f}, \dots, X_{N,d}^{M'_f}) \\ \dots \\ Diag'_{M,t} = f(X_{1,t}^{M'_F}, X_{2,t}^{M'_F}, \dots, X_{i,t}^{M'_F}, \dots, X_{N,t}^{M'_F}) \end{cases} \quad (4)$$

Development of a standard diagnostic criterion for the initial ontology

The goal here is to integrate existing standard information models relevant to the modeling of diagnostic criteria, through examination and expert editing. As mentioned earlier, we choose the content model ICD-11 and NQF QDM as reference models. The proposed study at this stage is to create an initial ontology of diagnostic criteria (DCUO) by integrating the ICD-11 content model with those QDM elements that are commonly used in diagnostic criteria. The distribution of QDM elements is evaluated using a set of textual diagnostic criteria. The selection of these QDM elements was confirmed by the results of the study, where 10 types of QDM data and 4 QDM attributes were selected and integrated with the ontology scheme based on the ICD-11 content model. In Table 3 provided a list of QDM data types and attributes used for integration.

Table 3. A list of selected QDM datatypes and attributes for developing the upper ontology

QDM Datatypes	QDM Attributes
Laboratory Test, Result	Result
Diagnostic Study, Performed	Method
Diagnostic, Active	Reason
Physical Exam, Performed	Severity

Symptom, Active	
Medication, Active	
Patient Characteristic Birth Date	
Patient Characteristic Race	
Patient Characteristic Sex	
Procedure, Recommended	

We used the Protégé ontology editing environment to manually integrate these two standard information models into the upper ontology of the diagnostic criteria.

The result of the work of the expert system in the form of a list of diagnoses using the ontology of the human phenotype (HPO) is presenting.

Modeling is based on four main stages: 1) a case reduced from HPO, by extracting from the ontology modules related to SARA, 2) an annotation from the free text description scoring scale with ontological modules, 3) development of two clinical archetypes, observation (for normalization content of the scale) and evaluation (for registration of clinical interpretations), 4) identification of information processing units for expressing the system for supporting clinical interpretation

Model approach:

1. Extraction of a shortened version of HPO;
2. Annotations to free text descriptions of elements and SARA estimates;
3. Development of two archetypes (observation and evaluation);
4. Definition of information processing.

Representation of knowledge using the Bayesian network of trust and conditional independence of events is shown on Figure 2.

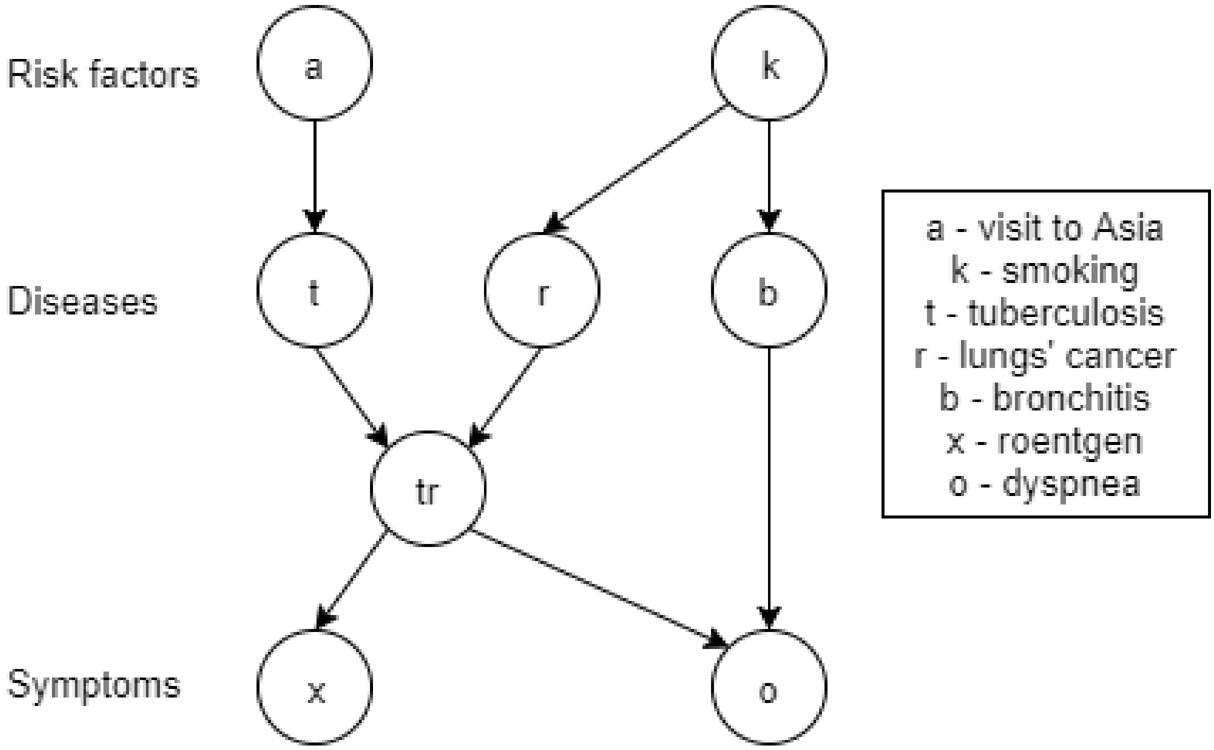


Figure 2. Knowledge representation

Representation of the fragment of the model of the medical DB in the form of BSD. This model corresponds to the following set of medical knowledge:

Dyspnea [o] may be due to tuberculosis [t], lung cancer [r] or bronchitis [b], as well as due to none of the listed diseases or more than one.

A visit to Asia [a] increases the chances of tuberculosis [t].

Smoking [k] is a risk factor for both cancer [r] and bronchitis [b].

X-ray results, determining the darkness in the lungs, do not distinguish between cancer [r] and tuberculosis [t], nor does it determine the presence or absence of dyspnoea [o].

The last fact is represented in the graph as an intermediate variable (event) [tr]. This variable corresponds to the logical function "or" for two parents ([t] and [r]) and it means the presence of either one or two diseases or their absence.

An important concept of the Bayesian network of trust is the conditional independence of the random variables corresponding to the vertices of the graph.

Two variables A and B are conditionally independent at a given third vertex C, if for a known value of C, the value of B does not increase the informativity about the values of A, that is,

$$p(A|B,C) = p(A|C) \quad (5)$$

If there is a fact that the patient is smoking, then we establish our trust regarding cancer and bronchitis. However, our trust regarding tuberculosis does not change. That is, [t] does not conditionally depend on [k] for a given empty set of variables

$$p(t|k) = 0 \quad (6)$$

Receiving a positive result of a patient's X-ray increases our confidence in tuberculosis and cancer, but not about bronchitis. That is, [b] - conditionally does not depend on [x] for a given k:

$$p(b|x,k) = p(b|k) \quad (7)$$

However, if we also knew that the patient had frequent breathing [o], then the x-ray results would also have an effect on our confidence in bronchitis. That is, [b] conditionally depends on [x] for given o and k. Thus, the inference in BDB means the calculation of conditional probabilities for some variables in the presence of information (evidence) about other symptoms.

The construction of standard and computable clinical diagnostic criteria is an important but challenging area of research in the community of clinical informatics. The Quality Data Model (QDM) is becoming a promising information model for standardizing clinical diagnostic criteria.

Conclusion

The solution of the set tasks allowed to obtain such results:

- Develop a method for distributing clinical medical situations to dangerous and safe classes that determines the most important factors of influence on the clinical state of the patient, which provides an opportunity to propose a technique for predicting the patient's condition using the proximity measure of micro-situations (with the greatest influence of parameters);

- A reliable backup method for storage of biomedical Big Data has been developed, which in practice has ensured high reliability in comparison with existing methods;
- A further development of the method for constructing the structure of expert systems for clinical medicine, which is distinguished by the repeated use of the ontology of successful outcomes, makes it possible to improve the reliability and speed of processing input data, as well as the effectiveness of decision-making;
- The situational model of clinical medicine has been improved for the analysis of the patient's condition, which, unlike existing approaches, uses situational presentation of the crisis situation on the basis of the triple "doctor-control action or solutions for re-use of the ontology-patient", which allows forecasting dangerous and safe situations for patient faster and with greater accuracy than existing models.

Bibliography

[Kuzomin and Vasylenko, 2014] Kuzomin, O.Ya., Vasylenko, O., Data loss minimization in situation's centrums databases. Chairman IDRC Davos 2014, Global Risk Forum GRF Davos, Davos, Switzerland. pp. 153-154

[Kuzomin and Vasylenko, 2017] Kuzomin, O.Ya., Vasylenko, O., Methods and models for building a distributed mobile emergency monitoring system. 17th International Multidisciplinary Scientific Geoconference SGEM 2017, Conference Proceedings, Informatics Geoinformatics, Vol.17. 2017. pp. 433 – 440.

[Kuzomin and Vasylenko, 2014] Kuzomin, O.Ya., Vasylenko, O., Obespechenie bezopasnosti ispolzovaniia baz dannyh v usloviiah chrezvychainyh situazii. International Journal "Information Technologies Knowledge", Vol. 8, Num. 2. 2014. pp. 173-187.

[Kuzomin and Vasylenko, 2010] Kuzomin, O.Ya., Vasylenko, O., Analiz estestvenno iazykovyh obiektov I predstavlenie znanii. Vostochno-Evropeiskii zhurnal peredovyh technologii, Vol. 6/2(48). 2010.

[Yager and Mcintyre, 2013] Yager, J., Mcintyre, J.S.. DSM-5 clinical and public health committee: challenges and considerations. Am J Psychiatr, 2014, 171(2). pp. 142–144.

[Organization WH, 2011] ICD-11 Alpha Content Model Reference Guide, 11th Revision. Geneva, Switzerland: World Health Organization. 2011.

[CMS Quality Data Model, 2015] CMS. Quality Data Model,Version 4.2. 2015. https://ecqi.healthit.gov/system/files/qdm_4_2_aug_2015.pdf.

[EHR Incentive Programs, 2017] Electronic Health Records (EHR) Incentive Programs. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html.

[ICD-11 Information Models, 2013] ICD-11 Information Models. 2013. http://informatics.mayo.edu/icd11model.

Authors' Information

Prof. Dr.-hab. Oleksandr Kuzomin *– Informatics chair of Kharkiv National University of Radio Electronics; Kharkiv, Ukraine; e-mail: kuzy@daad-alumni.de; tel.: +38(057)7021515*

Major Fields of Scientific Research: General theoretical information research, Decision Making, Emergency Prevention, Data Mining, Business Informatics.

Oleksii Vasilenko – *Data Strategy Manager, Australia, Aspirant of Kharkov National University of Radio Electronics; Kharkiv, Ukraine; e-mail: ichbierste@gmail.com ; tel.: +380 63 841 66 23*

Major Fields of Scientific Research: General theoretical information research, Knowledge Discovery and Engineering, Business Informatics

Tetiana Tolmachova *–Master student in ITIS of Leibniz University; Hannover, Germany; e-mail: tetiana.tolmachova@gmail.com*

Major Fields of Scientific Research: Big Data, Data Mining, Data Analyses.

TABLE OF CONTENT