Proceedings WIIS2020

Workshop on Intelligent Information Systems

Vladimir Andrunachievici Institute of Mathematics and Computer Science
December 04-05, 2020, Chisinau

CZU 004+519.7(082)

I-58

VLADIMIR ANDRUNACHIEVICI INSTITUTE OF MATHEMATICS AND COMPUTER SCIENCE
5, Academiei street, Chisinau, Republic of Moldova, MD 2028
Tel: (373 22) 72-59-82, Fax: (373 22) 73-80-27
E-mail: imam@math.md
WEB address: http://www.math.md

Editors: Dr. I.Titchiev, Prof. C.Gaindric

Authors are fully responsible for the content of their papers.

Descrierea CIP a Camerei Naţionale a Cărţii

Proceedings of Workshop on Intelligent Information Systems WIIS2020, December 04-05, 2020, Chisinau, Republic of Moldova / editors: I. Titchiev, C. Gaindric. – Chişinău : Vladimir Andrunachievici Institute of Mathematics and Computer Science, 2020 (Tipogr. "Valinex"). – 217 p. : fig., tab.

Antetit.: Vladimir Andrunachievici Inst. of Mathematics and Computer Science. – Referinţe bibliogr. la sfârşitul art. – 100 ex.

ISBN 978-9975-68-415-6.

004+519.7(082)

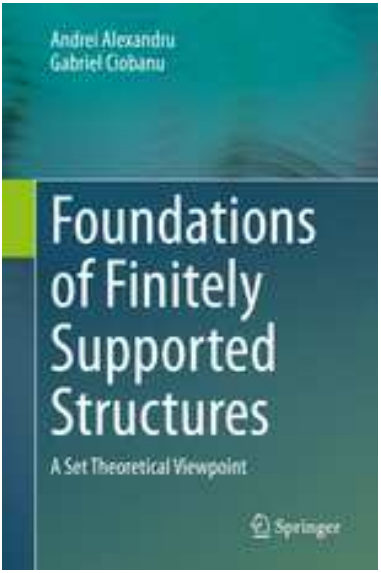
ISBN 978-9975-68-415-6

This issue is supported by research project 166/1- PS, 20.80009.5007.22, “Intelligent information systems for solving ill-structured problems, processing knowledge and big data”.

Finitely Supported Structures: A Framework for Computing within Infinite Structures

Andrei Alexandru

The goal of this presentation is to describe finitely supported sets and structures that were studied in the book "Andrei Alexandru and Gabriel Ciobanu. *Foundations of Finitely Supported Structures: a set theoretical viewpoint*, **Springer**, 2020", available at the address `https://www.springer.com/gp/book/9783030529611` [4] and in the papers [1, 2, 3, 5, 6, 7] published by the author in the last two years. This work is a joint work with Dr. Gabriel Ciobanu.



The theory of finitely supported sets (generally called Finitely Supported Mathematics, or, shortly, FSM) is related to Fraenkel Mostowski permutation models of Zermelo Fraenkel set theory with atoms (which is a refinement of Zermelo Fraenkel set theory, obtained by weakening the axiom of extensionality), to the theory of admissible sets of Barwise - particularly to the theory of hereditary finite sets, to the theory of nominal sets of Pitts used to study the binding, scope, freshness and renaming in programming languages and related formal systems [9], and to the concept of 'logical notions' of Tarski. Invariant sets are defined as usual sets equipped with actions of the group of permutations of infinitely many basic elements (called atoms), that additionally satisfy a finite support requirement which states that every element in an invariant set must be finitely supported (i.e. for each element there exists a finite subset of atoms such that the element is left unchanged under the effect of each permutation of atoms that fixes the related subset of atoms pointwise). The least finite set supporting a finitely supported element is called 'support'. Finitely supported (sub)sets are finitely supported elements in the powerset of an invariant set. Finitely supported relations (and functions) are finitely supported subsets of the Cartesian product of two invariant sets. A set having the property that all its elements are supported by the same finite set of atoms is called uniformly supported.

Inductively defined finitely supported sets involving the name-abstraction together with Cartesian product and disjoint union can encode a formal syntax modulo renaming of bound variables. In this way, the standard theory of algebraic data types can be extended to include signatures involving binding operators. In particular, there is an associated notion of structural recursion for defining syntax-manipulating functions and a notion of proof by structural induction. Certain generalizations of finitely supported sets are involved in the study of automata, programming languages or Turing machines over infinite alphabets; for this, a relaxed notion of finiteness called 'orbit finiteness' was defined; it means 'having a finite number of orbits (equivalence classes) under a certain group action' [8]. Informally, the theory of

finitely supported sets allows a discrete study of structures which are possibly infinite, but contain enough symmetries such that they can be clearly/concisely represented and manipulated.

The motivation for studying finitely supported structures also comes from the idea of modeling infinite algebraic structures, that are hierarchically constructed from atoms, in a finitary/discrete manner, by analyzing the finite supports of these structures. Thus, in the new framework we admit the existence of infinite atomic structures, but for such an infinite structure we remark that only a finite family of its elements (i.e. its 'finite support') is "really important" in order to characterize the related structure, while the other elements are somehow "similar". As an intuitive/straightforward motivation, in a lambda-calculus interpretation, the finite support of a lambda-term is represented by the set of all "free variables" of the term; these variables are those who are really important in order to characterize the term, while the other variables can be renamed (by choosing new names from an infinite family of names) without affecting the essential properties of the lambda-term. This means we can obtain an infinite family of terms starting from an original one (by renaming its bound variables), but in order to characterize this infinite family of terms it is sufficient to analyze the finite set of free variables of the original term. We proved in our book that the theory of finitely supported structures provides important tools for studying infinite algebraic structures in a discrete/computable manner.

Our goal is to study the foundations of finitely supported structures that are defined as finitely supported sets equipped with finitely supported internal operations or with finitely supported relations. In this sense we analyze whether a classical result (obtained in the framework of non-atomic sets) can be adequately reformulated by replacing 'non-atomic element/set/structure' with 'atomic, finitely supported element/set/structure'. Another goal is to present various properties of atomic structures that do not have a non-atomic correspondent.

Translating classical Zermelo Fraenkel results into the world of finitely supported structures is not an easy task since the family of

finitely supported sets is not closed under Zermelo Fraenkel subsets constructions, meaning that there exist subsets of finitely supported sets that fail to be finitely supported. For proving results in the framework of finitely supported sets, one cannot use related results from the Zermelo Fraenkel framework without priory reformulating them with respect to the finite support requirement. A proof of a result in the new framework should involve only finitely supported constructions (even in the intermediate steps). The meta-theoretical techniques for the translation of a result from the world of non-atomic structures to atomic structures are based on the 'S-finite supports principle' introduced by the author which claims that for any finite set S of atoms, anything that is definable in higher-order logic from S-supported structures by using S-supported constructions is also S-supported'. The formal involvement of this principle implies a hierarchical construction of the support of a structure by employing, step-by-step, the supports of the substructures of a related structure.

Furthermore, not even the translation of the results from a non-atomic framework into an atomic framework such as Zermelo Fraenkel set theory with atoms, in order to obtain results in FSM by using model theory, is an easy task. This is because results from Zermelo Fraenkel set theory may lose their validity when reformulating them in Zermelo Fraenkel set theory with atoms. For example, it is well known that multiple choice principle and Kurepa's maximal antichain principle are both equivalent to the axiom of choice in Zermelo Fraenkel set theory. However, Jech proved that multiple choice principle is valid in the Second Fraenkel Model, while the axiom of choice fails in this model. Furthermore, Kurepa's maximal antichain principle is valid in the Basic Fraenkel Model, while the axiom of choice fails in this model. This means that the following two statements (that are valid in Zermelo Fraenkel set theory) 'Kurepa's principle implies axiom of choice' and 'Multiple choice principle implies axiom of choice' fail in Zermelo Fraenkel set theory with atoms. Similarly, there are also results that are consistent with the axioms of Zermelo Fraenkel set theory, but fail to be consistent with the finite support requirement of FSM (we par-

ticularly mention choice principles and Stone duality).

We study the consistency of various forms of choice (and equivalent results), as well as the consistency of results regarding maximality, infinity and stability, in the framework of finitely supported sets. Formally, we prove that the choice principles **AC** (axiom of choice), **HP** (Hausdorff maximal principle) **ZL** (Zorn lemma), **DC** (principle of dependent choice), **CC** (principle of countable choice), **PCC** (principle of partial countable choice), **AC(fin)** (axiom of choice for finite sets), **Fin** (principle of Dedekind finiteness), **PIT** (prime ideal theorem), **UFT** (ultrafilter theorem), **OP** (total ordering principle), **KW** (Kinna Wagner selection principle), **OEP** (order extension principle), **SIP** (principle of existence of right inverses for surjective mappings), **FPE** (finite power set equippolence principle) and **GCH** (generalized continuum hypothesis) are not valid in the framework of finitely supported structures (i.e. they are in contradiction with the finite support requirement). Although choice principles are not valid in the newly defined atomic framework, they still remain consistent in the non-atomic framework, meaning that no non-atomic result is weakened under the new approach. Furthermore, in the new framework Banach-Tarski paradox is eliminated.

We introduce and study the notion of 'cardinality' in the framework of finitely supported sets (without involving ordinals). Two finitely supported sets have the same cardinality if there exists a finitely supported bijection between them. We define specific order relations on cardinalities ($\leq$ where the cardinalities are ordered via finitely supported injections, and $\leq^\star$ where the cardinalities are ordered via finitely supported surjections), and we study the properties of these relations. Particularly, we prove that the relation $\leq$ is equivariant, reflexive, anti-symmetric and transitive, but it is not total, while the relation $\leq^*$ is equivariant, reflexive and transitive, but it is not anti-symmetric, nor total.

We prove that the arithmetic properties and some order properties of cardinalities are naturally translated from the usual Zermelo Fraenkel set theory, while we are also able to prove some specific or-

der properties for atomic cardinalities. We also introduce finitely supported partially ordered sets, lattices and Galois connections, and we present fixed point, calculability and approximation properties. Several classical fixed points theorems for partially ordered sets, namely Tarski Theorem, Tarski-Kantorovitch Theorem, Bourbaki-Witt Theorem and Kleene-Scott Theorem, can be adequately reformulated in the framework of finitely supported structures. A generalization of Tarski's theorem where the condition of having a least upper bound is imposed only for those uniformly supported subsets of an invariant lattice (and not for all the finitely supported subsets of the related invariant lattice) is also proved. Other fixed point properties for finitely supported functions defined on invariant sets containing no infinite uniformly supported subsets, including calculability properties for fixed points are proved. The authors prove that finitely supported order preserving self-mappings on the finite powerset and, respectively, on the uniform powerset of a set containing no infinite uniformly supported subsets have least fixed points. Particularly, finitely supported self-mappings defined on the finite powerset of atoms have infinitely many fixed points if they satisfy some particular properties (such as strict monotony, injectivity or surjectivity).

Some of these important fixed point formal results are listed below:

- Let $(X, \sqsubseteq, \cdot)$ be a non-empty finitely supported partially ordered set with the property that every finitely supported totally ordered subset of X has a least upper bound. Let $f : X \to X$ be a finitely supported function with the property that $x \sqsubseteq f(x)$ for all $x \in X$. Then there exists $x \in X$ such that $f(x) = x$.

- Let $(X, \sqsubseteq, \cdot)$ be a non-empty finitely supported partially ordered set with the property that every uniformly supported subset of X has a least upper bound. Let $f : X \to X$ be a finitely supported function with the property that $x \sqsubseteq f(x)$ for all $x \in X$. Then there exists $x \in X$ such that $f(x) = x$.

- Let $(X, \sqsubseteq, \cdot)$ be a non-empty finitely supported partially ordered set with the property that every uniformly supported subset of X

has a least upper bound. Let $f : X \to X$ be a finitely supported order preserving function with the property that there exists $x_0 \in X$ such that $x_0 \sqsubseteq f(x_0)$. Then there exists $x \in X$ with $x_0 \sqsubseteq x$ such that $f(x) = x$.

- Let $(X, \sqsubseteq, \cdot)$ be a finitely supported partially ordered set with the property that every uniformly supported subset has a least upper bound. If $f : X \to X$ is a finitely supported function having the properties that $f(\sqcup Y) = \sqcup f(Y)$ for every uniformly supported subset Y of X and there exist $x_0 \in X$ and $k \in \mathbb{N}^*$ such that $x_0 \sqsubseteq f^k(x_0)$, then f has a fixed point which is supported by $supp(f) \cup supp(x_0) \cup supp(X) \cup supp(\sqsubseteq)$, where $supp(x)$ generally denotes the least set supporting x.

- Let $f : \wp_{fin}(A) \to \wp_{fin}(A)$ be finitely supported and strictly order preserving (i.e. f has the property that $X \subsetneq Y$ implies $f(X) \subsetneq f(Y)$). Then we have $X \setminus supp(f) = f(X \setminus supp(f))$ for all $X \in \wp_{fin}(A)$, where A is the set of all atoms and $\wp_{fin}(A)$ is the finite powerset of the set of atoms.

We present and study various examples of invariant lattices, and we analyze their properties. We particularly mention the family of those finitely supported subsets of an invariant set, the family of those finitely supported functions from an invariant set to an invariant complete lattice (i.e. the family of those finitely supported L-fuzzy sets with L being an invariant complete lattice), the family of those finitely supported subgroups of an invariant group, and the family of those finitely supported fuzzy subgroups of an invariant group. Formally we prove that all these families are invariant complete lattices, meaning that they have the property that every finitely supported subset has a least upper bound. In this way, the fixed point results presented above are applied to the related families.

The framework of finitely supported sets allows many pairwise non-equivalent definitions (of Dedekind type, of Mostowski type, of Kuratowski type, of Tarski type, of Levy type, and so on) for the concept of 'infinity'. We are able to establish comparison results between them

and to provide relevant examples of finitely supported sets which satisfy some forms of infinity, while do not satisfy others. We are particularly focused on uniformly supported sets. Uniformly supported sets are of interest because they involve boundedness properties of supports, meaning that the support of each element in a uniformly supported set is contained in the same finite set of atoms; in this way, all the individuals in an infinite uniformly supported family can be characterized by involving only the same finitely many characteristics. In this sense, we construct some atomic sets which seem to be 'large', but they do not contain infinite uniformly supported subsets, and for them we provide some important finiteness properties. For example, we prove an important result stating that if X is an invariant set that does not contain an infinite uniformly supported subset, then the set of all finitely supported functions from A to X does not contain an infinite uniformly supported subset. For some finitely supported mappings defined on sets containing no uniformly supported subset, the injectivity is equivalent with the surjectivity. For example, a finitely supported mapping $f : \wp_{fin}(A) \to \wp_{fin}(A)$ is injective if and only if it is surjective.

The most important results regarding the relationship between several forms of infinity are summarized below:

Let X be a finitely supported subset of an invariant set Y.

1. X is called *FSM classical infinite* if X cannot be represented as $\{x_1, \ldots, x_n\}$ for some $n \in \mathbb{N}$.

2. X is *FSM covering infinite* if there is a finitely supported directed family $\mathcal{F}$ of finitely supported subsets of Y with the property that X is contained in the union of the members of $\mathcal{F}$, but there does not exist $Z \in \mathcal{F}$ such that $X \subseteq Z$.

3. X is called *FSM Tarski I infinite (TI i)* if there exists a finitely supported one-to-one mapping of X onto $X \times X$, where $X \times X$ is the Cartesian product of X with itself.

4. X is called *FSM Tarski II infinite (TII i)* if there exists a finitely

supported family of finitely supported subsets of X, totally ordered by inclusion, having no maximal element.

5. X is called *FSM Tarski III infinite (TIII i)* if there exists a finitely supported one-to-one mapping of X onto $X + X$, where $X + X$ is the disjoint union of X with itself.

6. X is called *FSM Mostowski infinite (M i)* if there exists an infinite finitely supported totally ordered subset of X.

7. X is called *FSM Dedekind infinite (D i)* if there exists a finitely supported one-to-one mapping of X onto a finitely supported proper subset of X.

8. X is called *FSM ascending infinite (A i)* if there is a finitely supported increasing countable chain of finitely supported sets $X_0 \subseteq X_1 \subseteq \ldots \subseteq X_n \subseteq \ldots$ with $X \subseteq \cup X_n$, but there does not exist $n \in \mathbb{N}$ such that $X \subseteq X_n$.

9. X is called *FSM non-amorphous (N-am)* if X contains two disjoint, infinite, finitely supported subsets.

Some properties of FSM Dedekind infinite sets are listed below:

1. Let X be a finitely supported subset of an invariant set. Then X is FSM Dedekind infinite if and only if there exists a finitely supported one-to-one mapping $f : \mathbb{N} \to X$.

2. Let X be a classical infinite, finitely supported subset of an invariant set. Then $\wp_{fs}(\wp_{fin}(X))$ is FSM Dedekind infinite.

3. Let X be a finitely supported subset of an invariant set such that X does not contain an infinite uniformly supported subset. Then $\wp_{fin}(X)$ is not FSM Dedekind infinite.

In Figure 1 we present some of the relationships between the FSM definitions of infinity. The 'ultra thick arrows' symbolize *strict* implications (of the form p implies q, but q does not imply p), while 'thin

dashed arrows' symbolize implications for which we have not proved yet if they are strict or not (the validity of the reverse implications follows when assuming choice principles over non-atomic ZF sets). 'Thick arrows' match equivalences.

The forms of infinity satisfied by classical atomic sets are presented in the table below.

Set	TI i	TIII i	D i	M i	Asc i	TII i	N-am
A	No	No	No	No	No	No	No
$A + A$	No	No	No	No	No	No	Yes
$A \times A$	No	No	No	No	No	No	Yes
$\wp_{fin}(A)$	No	No	No	No	Yes	Yes	Yes
$T_{fin}(A)$	No	No	No	No	Yes	Yes	Yes
$\wp_{fs}(A)$	No	No	No	No	Yes	Yes	Yes
$\wp_{fin}(\wp_{fs}(A))$	No	No	No	No	Yes	Yes	Yes
A_{fs}^{A}	No	No	No	No	Yes	Yes	Yes
$T_{fin}(A)_{fs}^{A}$	No	No	No	No	Yes	Yes	Yes
$\wp_{fs}(A)_{fs}^{A}$	No	No	No	No	Yes	Yes	Yes
$A \cup \mathbb{N}$	No	No	Yes	Yes	Yes	Yes	Yes
$A \times \mathbb{N}$	No	Yes	Yes	Yes	Yes	Yes	Yes
$\wp_{fs}(A \cup \mathbb{N})$	No	Yes	Yes	Yes	Yes	Yes	Yes
$\wp_{fs}(\wp_{fs}(A))$	?	Yes	Yes	Yes	Yes	Yes	Yes
$A_{fs}^{\mathbb{N}}$ and $\mathbb{N}_{fs}^{A}$	Yes	Yes	Yes	Yes	Yes	Yes	Yes

In the above table, $+$ denotes disjoint unions, $\times$ denotes Cartesian products, $\wp_{fin}(X)$ denotes the finite powerset of an invariant set X, $\wp_{fs}(X)$ denotes the family of those finitely supported subsets of an invariant set X, $T_{fin}(X)$ denotes the family of those finite injective tuples of an invariant set X, X_{fs}^{Y} denotes the family of all finitely supported functions between two finitely supported sets X and Y.

Acknowledgements. The author is grateful to the co-author of his papers, Dr. Gabriel Ciobanu without whom this research would not have been possible, and to the leadership of Romanian Academy, Institute of Computer Science, especially to Dr. Vasile Apopei, for the support in carrying out this research.

References

[1] A. Alexandru and G. Ciobanu. *Fuzzy sets within Finitely Supported Mathematics*. Fuzzy Sets and Systems, vol. 339, 2018, pp. 119–133.

[2] A. Alexandru and G. Ciobanu. *On the foundations of finitely supported sets*. Journal of Multiple-Valued Logic and Soft Computing, vol. 32, no. 5-6, 2019, pp. 541–564, 2019.

[3] A. Alexandru and G. Ciobanu. *Fixed Point Results for Finitely Supported Algebraic Structures*. Fuzzy Sets and Systems, vol. 397, 2020, pp. 1–27.

[4] A. Alexandru and G. Ciobanu. *Foundations of Finitely Supported Structures: A set theoretical viewpoint*, Springer, 2020.

[5] A. Alexandru and G. Ciobanu. *Properties of the Atoms in Finitely Supported Structures*. Archive for Mathematical Logic, vol. 59, no. 1-2, 2020, pp. 229–256.

[6] A. Alexandru and G. Ciobanu. *Relaxing the Fraenkel-Mostowski Set Theory*. Journal of Multiple-Valued Logic and Soft Computing, vol. 34, no. 5-6, 2020, pp. 499–526.

[7] A. Alexandru and G. Ciobanu. *Uniformly Supported Sets and Fixed Points Properties*. Carpathian Journal of Mathematics, vol. 36, no. 3, 2020, pp. 351–364.

[8] M. Bojanczyk, B. Klin, S. Lasota. *Automata with group actions*. 26th Symposium on Logic in Computer Science, pp. 355–364, 2011.

[9] A. Pitts. *Nominal Sets Names and Symmetry in Computer Science*, Cambridge University Press, 2013.

Andrei Alexandru

Romanian Academy, Institute of Computer Science
E–mail: andrei.alexandru@iit.academiaromana-is.ro

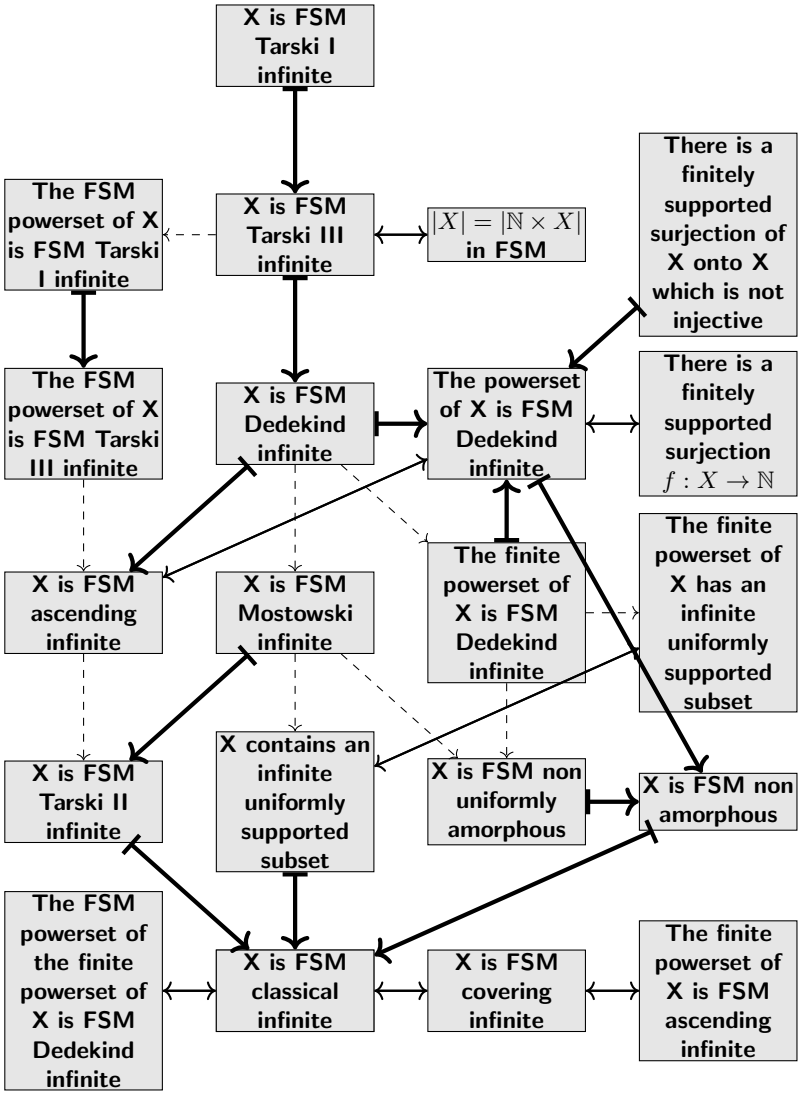


Figure 1. Relationships between various forms of infinity in FSM

A Road Map for an Informational Ontology of Information/Energy/Matter' Homeostatic Unity Inspired by Information Catastrophe

Veaceslav Albu

Abstract

The paper presents a possible framework of ontology of information/energy/matter unity inspired by recently published data on possibility of Informational Catastrophe on our planet based on information/energy/matter equivalence principle. Nevertheless, theoretical physicist proponents take the reductionist stand on such equivalence principle from which information does not represent the same fundamental entity as matter and energy do. For purpose of demonstrating the information/matter/energy fundamental unity, an account of Milky Way galaxy as galactic quantum computing system is presented. The proposed hypothesis of Informational gravity and Emergent Information based on quantum effects helps to build the possible quantum computational entanglement of the Sun and the Earth.

Keywords: informational ontology, emergent information, informational gravity, density of emergent information, galactic quantum computational system.

1. Introduction

History of human knowledge acquisition counts from the first dated written texts of any culture or from first cave paintings dated almost 50 000 years ago. All that texts deposited in libraries all over the world

represent information about humans glorifying struggles with own ignorance and lack of knowledge about the surrounding world. From Greek epoch, we gain first attempts to scientific formulation of knowledge and appearance of Philosophy. Up to middle of the 20th century all the amount of such information, we can regard as epistemological information, as the main purpose of it was to structure the knowledge.

After Norbert Wiener opened the new epoch of cybernetic systems, a new kind of information entered the world's stage – the digital information, with unknown exact properties and development potential. The digital status of information became of double nature. From one side it prolongs to possess the previous epistemological status of structuring knowledge. On the other side, from the appearance of mobile smart-phones and portable computers a new status of information has reviled itself. Human society was invaded by digital information and the uncontrollable Artificial Intelligence was prioritized as main danger of human beings in the closest future. Nevertheless, the publication in August 2020 of the article on topic of Information Catastrophe admits that in the next 150 years each second Earth's atom will be used for purpose of digital information in cybernetic systems. "Essentially, in the extreme case scenario when our digital information production growth is sustained at 50% per year, by the year 2070, we will have 1 kg of digital bits content on the planet stored on all the traditional and cloud data storage centers and endpoints such as PCs, smart-phones, and Internet of Things (**IoT**) devices. Similarly, at 50% growth per year, by the year 2245, half of the planet's mass will be made up of digital bits" [1]. Following, the Information Catastrophe become the main danger not only for humans, but also for phenomenon of Life on our planet. That fact unveils the new ontological status of information for phenomenon of life not only as an artefact of human activity but also as fundamental as matter and energy does. In this paper, we will formulate a road map for the ontological status of information in the Universe, with the same degree of fundamentality as matter and energy from the perspective of science of cybernetics.

Theoretical physicist J.A. Wheeler best expresses the modern history of convergence of scientific debates on topic of ontological status of information: *"In the first period, I was in the grip of the idea that*

Everything Is Particles… I call my second period Everything Is Fields… Now I am in the grip of a new vision, that Everything Is Information" [2].

2. The Methodology of the Proposed Road Map from Perspectives of Cybernetics

For the purpose of the present research, the following methodology is proposed. In philosophy of science, the ontology refers to the branch of philosophy that studies what are the things that make reality and how these entities are related with each other at the most fundamental level of reality. From the existing reductionist scientific paradigm, the concept of information represents an emergent property of the more fundamental concepts as matter and energy does. In other words, the information represents either an artifact created by humans or knowledge of how nature works. In computer science, ontology stands for the description of a domain of a discourse that contains its representation, categorization by definition and formal naming, relations and their properties between the concepts, data and entities that substantiate that specific domain.

The proposed methodology will encompass the ontological status of the concept of information first, for the realm of digital information in existing cybernetic systems, second, for the realm of living organisms and all life forms on our planet and Solar system, and finally, the realm of our Milky Way galaxy, where our planet Earth nests. Why those exactly three domains only taken for the research purpose? Because for the majority of scientists that sustain the existing reductionist scientific paradigm the fact of real existence and of the ontological status of these three realms does not represent a matter of debates. It is common sense such existing factual realities as virtual worlds of cybernetic systems, conscious and unconscious behavior of living systems on our planet as well as the Milky Way galaxy with its billions of stars and monster black hole. All these three realms we take as our undoubtable reality. For the purpose of our methodology, we do not yet need to take into account the Universe as a whole.

We will observe the ontology and ontological status of information for each from the above three realms and we will study how the information, matter and energy and the nature of their relations and properties are related. We will follow the sameness of such properties and

their nature for each from those specific realms and from that we will try to conclude the ontological framework for the information/energy/matter unity for our reality that perhaps can be hypothesized for the entire Universe.

3. The Main Players of the Ontology of the Information/Energy/Mater Unity

For the ontologies of the three realms of reality under the examination, the main fundamental entities are all known components of state of the arts of modern scientific paradigm with one exception. From the reductionist approach in modern scientific paradigm despite of declared energy/matter/information equivalence, the ontological status of matter/energy and information at quantum level is different. The concept of information possesses a status of an artefact or emergent property from the interaction of the energy and matter via laws of quantum realm. Except that three main players for the sake of this work we will underlie the *gravity* as a universal force that transcends all three mentioned realms of reality. Having this in mind, we can proceed to taxonomy of the proposed ontology.

4. Tripartite holomorphic unity of information/matter/energy

We use the term "holomorphic unity" in both senses, as philosophical concept introduced by Aristotle and as concept from computer science as well. Indeed, in any computer system there persists an information/matter/energy unity as interchange of bits "0" and "1" in the same processor cell and data channel. The state of the processor is related with information from the precedent state, so named retro causality of the given state of computer memory from the previous one. Same things happen in recent developed quantum computing systems, where *bits* are relocated by quantum *cubits*.

On the realm of life, the same unity on the fundamental level possesses the transport chains in living cells as well as in the neurons of the brain. That holomorphic unity manifests at all levels of manifestation in living organisms. In the present work here, we will not discuss the mind/body problem. The discussed tripartite unity for the galactic realm is backed by the link between phenomenon of life and photosynthesis effect

on plants that become the main cause of flourishing life on our planet. In favor of that unity, we take the homeostatic unity of star formation in galaxy and the role of central black hole in star and galaxy formation, well described in [3].

5. Homeostasis as manifestation of information/matter/energy unity

Another important property of ontology of information/matter/energy unity is the homeostasis based on different cases of organized feedback control as fundamental property for existence of such realms as dynamical systems. For cybernetic systems as well for living systems, Norbert Wiener employed feedback control as main property observed in living organisms [4]. For realm of Milky Way galaxy the scientific description in modern astronomy about homeostasis in star and galaxy formation, we refer also to the [5]. As a conclusion, we would like to draw attention to the fact that homeostasis plays the role of unifying principle of dynamical evolution of all three realms within Milky Way galaxy as described in [6].

6. The fractal nature of tripartite information/matter/energy unity

The fractals as a powerful scientific tool was introduced by B. Mandelbrot [7] and about the fractal nature of the Universe, we will refer to J. Gaite [8]. The dynamical fractals revolutionized the scientific paradigm in almost any field of modern science. That became possible only by means of computers in the second half of the last century. Such fractals as Mandelbrot set is not possible to imagine without the performing on the computer of millions of recursive iterations with complex numbers. One can say that Mandelbrot set exists in virtual worlds as pure mathematical and informational structure given by very simple equation: $z_{n+1} = z_n^2 + c$. The fractal nature of reality is backed also by recently published Wolfram's models [9], which represent an attempt to describe the physical world by means of fractal models.

7. The information/matter/energy singularities that seed on quantum realm

Singularities in modern science do not have an explicit definition but in each case refer to the state of extreme physical conditions that dominate the process of birth, vanishing or of existence of described entity.

Definition. As information/matter/energy singularity in given volume of space, we will take a state described by the extreme and dominating conditions of one of the given three entities – information, matter or energy.

The Big Bang theory rests on the primordial energetic singularity of the Bing Bang. At the Milky Way galaxy level, as such energetic singularities, one can take the example of quasars, supernova and the life span of all stars within galaxy. As matter/mass singularity, we have the well-known black holes entities, dominated by mass of collapsed matter in the given volume of space due to cosmological conditions. Finally, the Informational Catastrophe represents a description of singularity in our Milky Way galaxy with clearly described domination of the information on Earth by two reasons. Firstly, the Earth represents an informational singularity dominated by the phenomenon of life that evolutionary emerged on Earth and represents a dynamically evolving fractal of the Tree of Life of millions of species [10]. Secondly, the Earth represents a potentially informational singularity due to the described danger of the Informational Catastrophe. To explain this term we will claim the following hypothesis.

8. Hypothesis of Informational Gravity or *Infogravity*

Under the condition of dynamically evolving Universe as fractal entity, the dynamical fractal of evolving phenomenon of Life in given volume of space, represents a pattern of increasing of embedded information in complex living systems. *Infogravity* as dynamical process follows the principle of the information/energy/matter equivalence with a potential to converge towards Information Catastrophe.

As an explanation of informational gravity in living systems, we will introduce the definitions of the Density of Emergent Information (DEI) and respectively the Embedded Information (EI) in complex living systems. To present the above definition, we need to refer to quantum biology [11] that explains how the known events from quantum mechanics are embedded in living organisms and are responsible for many

processes that determined the evolution of the given organism. Let us say that we can present the following definitions. With the same degree of generality, we claim that all three kinds of above described singularities rest on quantum realm of the given volume of space.

Definition of Emergent Information (EI). By Emergent Information, we understand all quantum processes that occur in given quantum event in given point of matter, followed by changes in previous state of affairs and increasing embedded in matter information.

Definition of Density of Emergent Information (DEI). The Density of Emergent Information in the given volume of matter represents the total amount of Emergent Information in that volume per unit of time.

Having the above definition in place, we can present the following Hypothesis of Realm of Tripartite singularities.

Hypothesis 1. The Milky Way galaxy persists in homeostatic unity that is influenced and occurs between three kinds of information/matter/energy singularities via quantum realm.

To support the above hypothesis, we will employ the following facts. Firstly, recently there was made the scientific discovery that in our Milky Way galaxy there are up to one billion of black holes, with the monster central black hole. Secondly, the recent NASA research evaluates that one from ten stellar planetary system in Milky Way galaxy probably has a planet with similar Earth conditions where can emerge the phenomenon of life. Finally, the following calculations bring the evidence that total emergent information per second that emerges on Earth is almost the same with the one that emerges on the Sun.

Total mass of hydrogen atoms used per second to feed the Sun's core fusion reactions equals as 650 000 000 000 kg/s. The mass of one hydrogen atom equals 1.673×10^{-27} kg. Following that total, EI per second on the Sun's core equals 1.08×10^{39} of single tunneling quantum effects per second that penetrate the quantum realm of space occupied by Sun's core. In their turn, on Earth due to dynamically evolving phenomenon of life the numbers of eukaryotic [12] and prokaryotic [13] cells count together as $3{,}0 \times 10^{30}$ cells. Taking into account that average number of chemical reactions per cell per second equals 10^8, the results of chemical reactions happening in all living cells on Earth represents 3.0×10^{38}.

Because each chemical reaction presupposes a respective quantum effect responsible for each specific reaction, we conclude that number of quantum effects per second happening on the Sun and on the Earth possesses the same levels of magnitude of 10^{38}.

Hypothesis 2. Our Sun planetary system possesses a unique property that represents an entangled system of local quantum effects (QE) with the same amount of Emergent Information of magnitude 10^{38} QE/sec.

Taking into account that quantum computers relay on quantum effects, we claim the hypothesis that described similarity between Emergent Information on Sun and Earth may represent a part of evolutionary quantum computing system by natural kinds.

Hypothesis 3. According to the last scientific astronomical data research [14] in the Milky Way galaxy, there are $6x10^9$ or six billion Earth-like planets surrounding their Sun-like stars. Simultaneously, NASA' Science Mission Directorate reported [15] as much as up to 10^9 or one billion black holes lurking in our galaxy. Based on this data and from the Hypothesis of Emergent Information described in this work, we are claiming the possible existence of the galactic quantum computational system evolved by natural kinds with power of Emerging Information approximate equal to 10^{47} magnitude of quantum events per second. According to Hypothesis 3, such galactic quantum computing system is based on tripartite entanglement between three types of information/matter/energy singularities – habitable planets, black holes and stars of our galaxy.

That hypothesis fits with many hypothesized theories of Computational Universe. From the information/matter/energy principle, we can approximate the energy equivalence of galactic Emergent Information. That can imply reasoning on black matter and black energy findings, but that we will live for next stages of this research.

9. Conclusion

In the present paper, we presented an attempt to elucidate the possible grounds that can help to build a credible framework for scientific community that supports the famous view of Norbert Wiener on equality of the ontological status of information/energy/matter in the Universe.

Important output of this research represents the proposed notion of Emergent Information and Density of Emergent Information because of quantum effects in given volume of matter/space. From such approach, we constructed the Informational Gravity Hypothesis. From this framework and from last scientific data we claimed the Hypothesis of galactic quantum computing system for Milky Way galaxy that finally represents a cybernetic system based on principle of homeostatic feedback control. The ontological information/matter/energy unity results as their holomorphic unity in the informational feedback process by analogy with classical cybernetic system.

Acknowledgments. The research was supported by the project 20.80009.5007.22. Intelligent information systems for solving ill-structured problems, processing knowledge and big data.

References

1. Melvin M. Vopson. *The information catastrophe*. AIP Advances **10**, 085014 (2020); doi: 10.1063/5.0019941, Published Online: 11 August 2020.
2. Israel Belfer. *Informing Physics: Jacob Bekenstein and the Informational Turn in Theoretical Physics*. Phys. Perspect. 16 (2014) 69–97, © 2014.
3. Joseph Silk. *Feedback in Galaxy Formation*. https://arxiv.org/pdf/1102.0283v1 , 01 of February 2011, Accessed on 01 of November 2020.
4. Norbert Wiener. *Cybernetics: Or Control and Communication in the Animal and the Machine*, (Paris: (Hermann & Cie) & Camb. Mass. (MIT Press), 1961).
5. NASA report. *Stellar Interiors - Hydrostatic Equilibrium and Ignition on the Main Sequence*, accessed: http://apod.nasa.gov/apod/astropix.html on 01 of November 2020.
6. John S. Torday. *Homeostasis as the Mechanism of Evolution, Biology* 2015, *4*, 573-590; doi:10.3390/biology4030573.
7. Mandelbrot B.B. *The fractal geometry of nature* (rev. ed. of: *Fractals*, 1977); W.H. Freeman and Co.: NY, 1983.
8. Jose Gaite, *The Fractal Geometry of the Cosmic Web and its Formation*, Published in: Adv.Astron. 2019, (2019) 6587138.
9. Stephen Wolfram, *A Class of Models with the Potential to Represent Fundamental Physics*, Complex Systems, 29(2), 2020 pp. 107-536.
10. Evogeneao, *Tree of Life, Explorer* https://www.evogeneao.com/explore/tree-of-life-explorer, [accessed 01 of October, 2020].

11. Werner Gitt. *Information in Living Organisms*, https://answersingenesis.org/genetics/information-theory/information-inliving-organisms/ [accessed 01 of October 2020].
12. Yinon M. Bar-On, Rob Phillips, Ron Miloa. *The biomass distribution on Earth*, Proceedings of the National Academy of Sciences, 115(25):201711842, 2018, DOI: 10.1073/pnas.1711842115.
13. William B. Whitman, David C. Coleman, and William J. Wiebe, *Prokaryotes: The unseen majority*, Proc. Natl. Acad. Sci. USA 95, pp. 6578–6583, June 1998.
14. Michelle Kunimoto and Jaymie M. Matthews. *Searching the Entirety of Kepler Data. II. Occurrence Rate Estimates for FGK Stars*, The Astronomical Journal, 2020; 159 (6): 248 DOI: 10.3847/1538-3881/ab88b0.
15. *Black Holes in the Milky Way.* NASA Reports 2020, https://science.nasa.gov/astrophysics/focus-areas/black-holes#:~:text=Most%20stellar%20black%20holes%2C%20however,in%20the%20Milky%20Way%20alone, accessed on 15 of October 2020.

Veaceslav Albu

Vladimir Andrunachievici Institute of Mathematics and Computer Sciernce
E-mail: vaalbu@googlemail.com

Proceedings of Workshop on Intelligent Information Systems
WIIS2020, December 04-05, 2020, Chisinau, Republic of Moldova

Non Standard Treebank Parsing

Victoria Bobicev, Cătălina Mărănduc

Abstract

We report on our ongoing work on Dependency Treebank development. The corpus has been affiliated with Universal Dependencies (UD) project as UD Romanian-Nonstandard Dependency Treebank; it contains 26225 sentences, 572436 tokens, part of them being collected and annotated in the Republic of Moldova. The paper concentrates on the corpus parsing using MaltParser tool.

Keywords: text corpus, treebank, dependency grammar, non-standard language, syntactic parser, automate parsing evaluation.

1 Introduction

Text corpora are the first and most essential linguistic resources in the Natural Language Processing (NLP) field of research. Corpus value is determined by its annotation, the additional information that can be as diverse as the corpora themselves. The advantages of corpus annotation are multi-functionality, re-usability, and easy of exploitation both by humans and computers. Morphological information is considered as a basic one. As a rule, the next level of annotation is syntactic one which indicates syntactic relations among words in sentences. These relations form a graph in the shape of a tree and such annotated corpus is referred to as a treebank.

The corpus presented in the paper is the current version of Alexandru Ioan Cuza University (UAIC) Romanian Dependency Treebank (UAIC-RoDia-DepTb). It has obtained the International Standard

Language Resource Number (ISLRN 156-635-615-024-0)1 and it is not merely syntactically annotated but it is presented in several formats: (1) the syntactic classic, containing more than 39,000 sentences, (2) the UD syntactic, and (3) a new syntactic-semantic one. Moreover, it is the biggest syntactically annotated Romanian corpus.

2 Related Work

Several corpora have been created for the Romanian language. CoRoLa [1] is a representative corpus of Contemporary Standard Romanian. Now, it contains more than 1,200 million words in written form and about 300 hours of oral texts with associated transcripts. It contains only Contemporary Standard Romanian and has only morphological annotation. Our corpus is annotated syntactically; it increases the corpus value.

Currently, the international community increased its interest in preserving and digitizing the cultural heritage, i.e. in the processing of old texts. In Romania as well as in Moldova, multiple old Romanian texts have been written using a specific Old Romanian Cyrillic alphabet. Texts with Cyrillic letters have been scanned and recognized using an Optical Character Recognizer (OCR) adapted and trained for these specific texts at the Institute of Mathematics and Computer Science in Chisinau [2]; then they have been transliterated using a special tool for the various versions of Old Romanian Cyrillic in the modern Latin. Our corpus contains the texts obtained by transliteration.

UAIC-RoDia-DepTb contains several parts with folklore. The Folk corpus of verses in Romania and Republic of Moldova is unique by combining folk texts and dependency annotation [3].

3 Corpus Annotation Format

The Romanian Non-standard UD corpus has been developed starting from the Alexandru Ioan Cuza University Romanian Diachronic Treebank (UAIC-RoDia). At present, the UAIC Dependency Treebank is the largest one for Romanian.

RoDia (Romanian Diachronic) corpus contains non-standard types of texts [4]. The standard language is rarely used in human communication; in our corpus we concentrate on the annotation of non-standard text types such as oral regional fiction, social media communication, poetry, Old Romanian texts and others.

The whole corpus has morphological and syntactic annotation using the dependency grammar conventions. There are 544 labels used for the complete morphological analysis in our corpus, part of the labels used in the MULTEXT EAST project [5], and 44 syntactic labels are used for the dependency relations.

The next step of our corpus development was its affiliation with Universal Dependencies (UD) [6]. UD annotation is coded in CoNLLU format[1]. It presents text files in which each line contains one word of the text with all its description separated by tabulation. The UD annotation convention highlights words with full meaning and the relational words are subordinated to them. In the UAIC convention, the relational words are the heads for the dependent words. In the UD system, it is easier to compare texts in very different languages and to emphasize the relation between the morphology and the syntax. UD annotation uses 17 very general PoS tags and 37 dependency labels.

The current numbers of sentences in both formats described above are shown in Table 1.

Table 1. The Current Statistics of Two Formats of UAIC Treebank

Nr.	Format	Sentences	Tokens
1	UAIC Syntactic (basic)	39,575	836,196
2	UD Syntactic	21,403	449,959

4 Parsing Evaluation

There are two Romanian UD corpora: RRT which contains modern Romanian texts and Non-Standard Corpus with old texts and folklore.

[1] http://universaldependencies.org/format.html

By May 2020 the volumes of these corpora were: 218,511 tokens total in RRT and 449,959 tokens in Non-Standard. Statistics of morphology tags as well as some other statistics for both corpora is presented on the UD page[2].

The main aim of our work is the creation of the gold standard corpus to be used for future training of part of speech taggers and syntactic parsers; its volume should be enough for reliable parsing with minimum errors. We tested MaltParser[3] which is considered the basic one for UD corpora on our corpus. We evaluated nine parsing algorithms of the parser[4] training on train part of our UD corpus and testing on the testing part. The results are presented in the Table 2.

Table 2. The results of the Malt-parser evaluation on our UD corpus

Nr.	Model	UAS	LAS
1	planar	82.28	74.36
2	stackeager	82.54	74.48
3	2planar	82.61	74.82
4	nivreeager	82.81	74.82
5	stacklazy	82.99	74.94
6	stackproj	82.99	74.94
7	nivrestandard	82.91	74.97
8	covnonproj	83.10	75.43
9	covproj	83.39	75.68

We measured the accuracy using two most commonly used metrics for dependency parsers: the labeled attachment score (LAS) and the unlabeled attachment score (UAS) on the word level [7]. LAS is calculated as the percentage of correct links with correct labels of all links in the test corpus; UAS is calculated as the percentage of correct links

[2]http://universaldependencies.org/treebanks/ro-comparison.html
[3]http://maltparser.org
[4]http://maltparser.org/guides/opt/quick-opt.pdf

ignoring their labels.

Our previous (autumn 2018) version of Non-Standard corpus resulted in LAS = 72.45 and UAS = 80.73. Since then, the train part of the corpus grew from 7965 to 10144 sentences. We experimented with different training part sizes; the learning curve is shown in Figure 1. The figure shows that the accuracy growing speed is slowing down while the corpus reaches 3000 sentences. It means that further gain in accuracy will cost us more and more additional text.

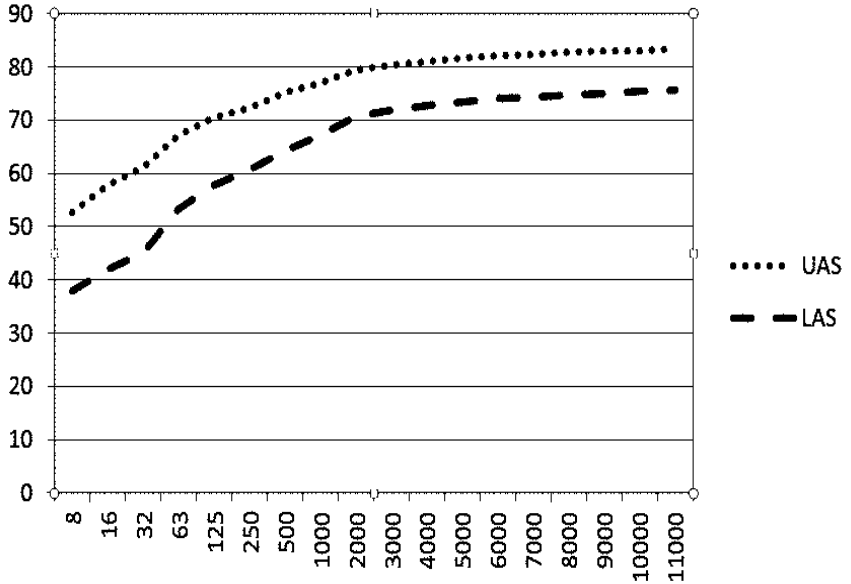


Figure 1. Dependency between the training corpus size and parsing accuracy

Tables 3 and 4 present more details of the the parsing errors. Table 3 contains two columns: Part of Speech (PoS) accuracy; syntactic link label accuracy. Table 4 presents triple's (PoS - link label - PoS)

accuracy. There are 41 syntactic link labels for Romanian UD corpora, including several specific labels such as labels for the Romanian reflexive pronouns and its types written as "expl": expl:pv, expl:poss, expl:impers. expl:pass. We also collected statistics for 253 types of so-called triples to see which pairs of PoS are connected by which links; Table 4 contains triples with the best and worst accuracy.

Table 3. The best and the worst parsing accuracy for various parts of speech and link types

PoS	Acc.	Link	Acc.
ADP	0.96	det	0.97
DET	0.94	aux	0.97
AUX	0.89	case	0.97
PART	0.88	expl:pv	0.94
...	...	...	...
ADV	0.73	expl:poss	0.19
INTJ	0.72	expl	0.18
NOUN	0.64	advcl:tcl	0.08
VERB	0.59	orphan	0.04
X	0.16	expl:pass	0.03

The first column of Table 3 presents accuracy for various parts of speech and it is not exactly correlated with their frequency. For example, NOUN and VERB, one of the most frequent PoS have almost the worst accuracy; it can be explained by the multitude of relations these PoS have. The second column presents link types and it is seen that it actually correlates with the first column with PoS. The triples presented in Table 4 explain some correlations from the previous table. For example, **ADP** and **case** appear in **ADP case PRON** triple; **DET** and **det** appear in **DET det PROPN** and **DET det PRON**, all these triples having accuracy 1.

Table 4. The best and the worst parsing accuracy for various triples: parts of speech - link type - part of speech

Triple	Accuracy
ADP case PRON	1
DET det PROPN	1
DET det ADJ	1
ADP case NUM	1
...	...
NOUN orphan NOUN	0.1
VERB advcl:tcl VERB	0.09
NOUN nsubj:pass VERB	0.09
PROPN nsubj:pass VERB	0.08
PRON expl:pass VERB	0.02

5 Conclusion

The paper presents an ongoing work on the development of the corpus of non-standard texts with syntactic annotation. Several efforts have been made to enrich the corpora with various more difficult for the annotation examples such as folklore, chat and Old Romanian. The annotated and manually corrected texts would serve as a training corpus for the statistical parser. Our current goal is to increase the volume of Old Romanian and folklore parts of our corpus until we get the acceptable accuracy of automate parsing for these types of texts. Our affiliation to UD increased the visibility of our common efforts and opened the perspectives of participating in the international projects.

References

[1] V. Barbu Mititelu, E. Irimia, and D. Tufiş. *CoRoLa – The Reference Corpus of Contemporary Romanian Language*. Proceedings

of the Ninth International Conference on Language Resources and Evaluation. (LREC-2014), pp. 1235-1239, 2014.

[2] S. Cojocaru, A. Colesnicov, and L. Malahov. *Digitization of Old Romanian Texts Printed in the Cyrillic Script*. Proceedings of DATeCH 2017, pp. 143-148, 2017.

[3] V. Bobicev, T. Bumbu, V. Lazu, V. Maxim, and D. Istrati. *Folk poetry for computers: Moldovan Codri's ballads parsing*. Proceedings of the 12th International Conference Linguistic Resources and Tools for Processing the Romanian Language, CONSILR-2016 pp. 39-50, 2016.

[4] C. Mărănduc, C.-A. Perez, L. Malahov, and A. Colesnicov. *A Diachronic Corpus for Romanian (RoDia)*. Annual Conference of the German Linguistic Society, p.151, 2017.

[5] Tomaž Erjavec. *MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages*. Language Resources and Evaluation, 46/1, pp. 131-142, 2012.

[6] J. Nivre. *Towards a Universal Grammar for Natural Language Processing*. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2015. Lecture Notes in Computer Science, vol 9041. Springer, Cham., 2015.

[7] J. Nivre and C.-T. Fang. *Universal Dependency Evaluation*. Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies. (UDW 2017), pp. 86-95, 2017.

Victoria Bobicev[1], Cătălina Mărănduc[2]

[1]Faculty of Computer Science,"Al. I. Cuza" University, Iasi, Romania
E–mail: catalinamaranduc@gmail.com

[2]Tehnical University of Moldova
E–mail: victoria.bobicev@ia.utm.md

A criterion for estimating the favoring of beneficiaries in apportionments

Ion Bolun

Abstract

There is proposed a criterion to quantitatively estimate the favoring of large or small beneficiaries in particular apportionments and on the whole (on an infinity of cases) by apportionment methods. By computer simulation, it is shown that the favoring of beneficiaries by d'Hondt method can be considerable, in total overpassing 10 of apportioned entities.

Keywords: quantitative criteria, favoring a beneficiary, favoring large or small beneficiaries, computer simulation.

1. Introduction

Integer character of the proportional apportionment (APP) problem usually causes a certain disproportion of the representation of deciders' will in the decision [1, 2], some beneficiaries (parties, states, schools, etc.) being favored at the expense of others. Therefore, reducing the favoring in question is one of the basic requirements when choosing the APP method to be applied under concrete situations (bias condition [1, 3]). But there are also other aspects (see, par example, [3, p. 9]), which eventually led to the application of various APP methods, such as Hamilton (Hare), Jefferson, Webster, d'Hondt, Sainte-Laguë, Huntington-Hill [1, 2] and others.

So, when selecting an APP method, the property of non-favoring of beneficiaries is useful. It is well-known that d'Hondt method favors large beneficiaries, and Huntington-Hill method favors the small ones [1, 2]. But which of the two favors beneficiaries to a larger extent? Namely, criteria to investigate such aspects are examined in this paper.

2. Preliminary considerations

The departure point of APP methods is to minimize the disproportion of apportionment of entities (seats, computers, tickets, etc. – a-entities) to

beneficiaries (parties, states, schools, etc.). In order to estimate this disproportion, various indices were proposed. Starting from the value *d* of a decider will (*d = M/V*, where *M* is the total number of a-entities and *V* is the total number of deciders; *d* reflects unequivocally the rights of each decider in the decision), and basing on a comparative multi-aspectual analysis of 12 indices, in [2] the opportunity of using the Average relative deviation index *I* for this purpose is argued:

$$I = 100\sum_{i=1}^{n}\left|\frac{V_i}{V} - \frac{x_i}{M}\right| \to \min. \qquad (1)$$

Here n is the number of beneficiaries, V_i is the number of deciders and x_i is the number of a-entities apportioned to beneficiary i.

Evidently, a beneficiary i is considered larger than a beneficiary k, if $V_i > V_k$.

To quantitatively estimate the favoring of beneficiaries, the formalization of the notion of beneficiaries favoring is needed. It will be distinguished three notions of favoring of beneficiaries by an APP method:

a) favoring of a beneficiary in an apportionment;
b) favoring of large or small beneficiaries in an apportionment;
c) favoring of large or small beneficiaries on the whole, on an infinity of apportionments.

Also, each of the specified above three aspects can be characterized by:

A) identifying the fact of favoring;
B) quantitatively estimating the favoring of beneficiaries.

Unfortunately, all quantitative criteria, along with the respective quantitative assessments (aspect B), can be used also to identify the fact of favoring of large or small beneficiaries by an APP method (aspect A).

3. Formalizing the notion of favoring of a beneficiary

In case of proportional apportionment, knowing the value of *d*, it is easy to determine the expected rights D_i of beneficiary i in the decision, namely, $D_i = dV_i$. In other words, D_i is the influence power of beneficiary i in the decision, delegated to it by their V_i deciders. Let's transform (1) as follows

$$I = 100\sum_{i=1}^{n}\left|\frac{V_i}{V} - \frac{x_i}{M}\right| = \frac{100}{M}\sum_{i=1}^{n}\left|dV_i - x_i\right| = \frac{100}{M}\sum_{i=1}^{n}\left|D_i - x_i\right| \to \min. \quad (2)$$

So, if $x_i = D_i, i = \overline{1,n}$, we have $I = 0$, that is if the number of a-entities, apportioned to each beneficiary i, is equal to the expected value D_i for $i = \overline{1,n}$, then $I = 0$ and there are no favored beneficiaries. The disproportion in an apportionment can occur because D_i is a real number, and x_i is an integer.

Definition 1 [2]. In an apportionment, a beneficiary i is favored, if it gets an excess of a-entities ($\Delta D_i = x_i - D_i > 0$); is disfavored, if it obtains a deficit number of a-entities ($\Delta D_i < 0$); and is neutral (neither favored nor disfavored), if it gets a number of a-entities equal to the expected one ($\Delta D_i = 0$).

So, the A aspect, for a beneficiary i:

1) favored, occurs $x_i > a_i$, where $a_i = \lfloor D_i \rfloor \le D_i$;
2) disfavored, occurs $x_i \le a_i$ at $D_i > a_i$;
3) neutral, occurs $x_i = a_i$ at $D_i = a_i$.

Aspect B for a beneficiary i is characterized by the number of a-entities in excess in the apportionment: $\Delta D_i = x_i - D_i$; unfortunately, if $\Delta D_i < 0$, then the beneficiary i is disfavored, because it has a deficit of a-entities.

It remains to define the other two notions of favoring of large or small parties in an apportionment and on the whole by an APP method – cases (b) and (c) from Section 2.

4. Formalizing the notion of favoring of large/small beneficiaries

First of all, it is useful to mention that, because of $D_1 + D_2 + \ldots + D_n = M$ and $x_1 + x_2 + \ldots + x_n = M$, if some beneficiaries are favored, the other ones are mandatory disfavored. Evidently, there are no alternatives and it is easy to formalize the notion of favoring of large or small beneficiaries if $n = 2$. But there is not the case for $n > 2$. One of the well-known alternatives is done in [1].

Definition 2 (according to [1, p. 125]). An apportionment method favors large parties if

$$\frac{\sum_{i\in L} x_i}{\sum_{i\in L} V_i} > \frac{\sum_{j\in S} x_j}{\sum_{j\in S} V_j} \quad (3)$$

and it favors small parties if

$$\frac{\sum_{i\in L} x_i}{\sum_{i\in L} V_i} < \frac{\sum_{j\in S} x_j}{\sum_{j\in S} V_j}, \qquad (4)$$

where L and S are subsets of $\{1, 2, \ldots, n\}$ such that $x_i > x_j$ whenever $i \in L$ and $j \in S$ [3].

Without diminishing the universality of the approach, further it is considered that n beneficiaries are ordered in non-ascending order of V_i, $i = \overline{1,n}$, i.e. $V_1 \geq V_2 \geq V_3 \geq \ldots \geq V_n$. Let's consider the apportionments for which $x_1 > x_2 > x_3 > \ldots > x_n$. For such an apportionment and $|L| + |S| = n$, there are $n - 1$ variants of different pairs of subsets L and S: $L_1 = \{1\}$, $S_1 = \{2, 3, \ldots, n\}$; $L_2 = \{1, 2\}$, $S_2 = \{3, 4, \ldots, n\}$; …; $L_{n-1} = \{1, 2, 3, \ldots, n-1\}$, $S_{n-1} = \{n\}$. In general, in each of subsets L and S there may be beneficiaries with the same value of x. In such a case, the number of variants of different pairs of subsets L and S is larger than $n - 1$, but it is easy to show that examination of only the described above $n - 1$ variants on the subject in question is also sufficient.

If for all mentioned above $n - 1$ variants of different pairs of subsets L and S the relation (4) takes place, then it is simple to decide that large beneficiaries are favored, and vice versa – if the relation (5) takes place. But there may be cases, when for some pairs of subsets L and S the relation (4) takes place, and for the other pairs of subsets L and S the relation (5) occurs. Let's consider the following example.

Example 1. Let $M = 9$, $n = 4$, $V_1 = 500$, $V_2 = 495$, $V_3 = 395$, $V_4 = 390$ and for apportionment the d'Hondt method is applied. Then $Q = 197.7(7)$ and $\Delta M = 3$. The other calculations for the apportionment are shown in Table 1.

Table 1. Calculations for the apportionment to Example 1

i	V_i	a_i	$V_i/(a_i + 1)$	$V_i/(a_i + 2)$	Δx_i	x_i	x_i/V_i
1	500	2	166.(6)	125.00	1	3	0.006000
2	495	2	165.00	123.75	0	2	0.004040
3	395	1	197.50	131,(6)	1	2	0.005063
4	390	1	195.00	130.00	1	2	0.005128

The three variants of subsets L and S are: $L_1 = \{1\}$, $S_1 = \{2, 3, 4\}$; $L_2 = \{1, 2\}$, $S_2 = \{3, 4\}$; $L_3 = \{1, 2, 3\}$, $S_3 = \{4\}$. The results of calculations by formulas (4) and (5) for subsets L_j and S_j ($j = 1, 2, 3$) are systemized in Table 2.

Table 2. Results of calculus for subsets L_j and S_j (j = 1, 2, 3) to Example 1

j	$\sum_{i\in L_j} x_i / \sum_{i\in L_j} V_i$	Relation	$\sum_{i\in S_j} x_i / \sum_{i\in S_j} V_i$	Favored beneficiaries
1	0.006000	>	0.004688	large
2	0.005025	<	0.005096	small
3	0.005036	<	0.005128	small

According to the $\{L_1, S_1\}$ variant, the requirement (4) takes place, but for the $\{L_2, S_2\}$ and $\{L_3, S_3\}$ variants the requirement (5) takes place (see Table 2). So, in this case, to decide whether the obtained apportionment favors large or small beneficiaries or it is neutral by using the Definition 2 is not possible.

Therefore, taking into account Definition 1, it is useful also another approach to the subject in question. Let's note $\Delta D_i = x_i - D_i$, $i = \overline{1,n}$. Also, basing on relation (6), it is opportune to redefine the L and S subsets of large and, respectively, small beneficiaries as follows:

$$L = \{1, 2, \ldots, \lfloor n/2 \rfloor\} \quad (5)$$

$$S = \{\lceil n/2 \rceil + 1, \lceil n/2 \rceil + 2, \ldots, n\}, \quad (6)$$

where $x_i \geq x_j$ whenever $i \in L$ and $j \in S$.

According to (5) and (6), one has $|L| = |S| = \lfloor n/2 \rfloor$ and, if n is even, the pair of subsets $\{L, S\}$ coincides with one of $n - 1$ variants of subsets $\{L, S\}$ used by Definition 2.

Definition 3. An apportionment favors large beneficiaries, if the summary a-entities in excess, obtained by large beneficiaries (L), is greater than that, obtained by small beneficiaries (S) and vice versa, that is it favors large beneficiaries if $F_{a1} > 0$, it favors the small ones if $F_{a1} < 0$ and it is neutral if $F_{a1} > 0$, where

$$F_{a1} = \sum_{i=1}^{\lfloor n/2 \rfloor} \Delta D_i - \sum_{i=\lceil n/2 \rceil + 1}^{n} \Delta D_i. \quad (7)$$

In addition to identifying the favoring of large or small beneficiaries in an apportionment (aspect A), criterion F_{a1} also allows quantitative estimation of absolute favoring in question, measured in a-entities (aspect B).

When applying (7) to Example 1, one has $F_{a1} = 3 - 500d + 2 - 495d - (2 - 395d + 2 - 390d) = 1 - 210d = 1 - 210 \times 9/1780 = -0.0618$ a-entities

< 0. Thus, according to Definition 3, the apportionment of Example 1 favors small beneficiaries.

To mention that, according to Table 2, if to take into account only the pairs (5) and (6) ($L = L_2 = \{1, 2\}$ and $S = S_2 = \{3, 4\}$), the apportionment of Example 1 also favors small parties. But it is easy to show that conditions (3) and (4), when using the interpretation of pair $\{L_2, S_2\}$ as defined by (5) and (6), are not equivalent to stipulations of Definition 3; they are not interchangeable even for the identification or the fact of favoring of large or small beneficiaries in a particular apportionment.

Moreover, with refer to Definition 2, there are particular apportionments, for which the inequality (4) takes place for all the $n - 1$ variants of subsets L_j and S_j (see Example 2), but on the whole (on an infinity of apportionments) the method favors large beneficiaries and vice versa, the inequality (4) takes place for all the $n - 1$ variants of subsets L_j and S_j (see Example 3), but on the whole (on an infinity of apportionments) the method favors small beneficiaries. So, Definition 2 refers to the favoring of beneficiaries (parties) in particular apportionments.

Example 2 [2]. D'Hondt method favors small beneficiaries. Let $M = 9$, $n = 2$, $V_1 = 500$ and $V_2 = 390$. Then $Q \approx 98.9$ and $\Delta M = 1$. The other results of calculations are shown in Table 3.

Table 3. Calculations to Example 2

i	V_i	a_i	$V_i/(a_i + 1)$	Δx_i	x_i	Δx_i	x_i/V_i
1	500	5	83,3	0	5	0	0.0100
2	390	3	97,5	1	4	1	0.0103

We have $x_1 > x_2$ and $x_1/V_1 = 5/500 = 0.0100 < x_2/V_2 = 4/390 \approx 0.0103$. So, in this apportionment, according to relations (3) and (4), d'Hondt method favors the small beneficiaries (beneficiary 2). The same result is obtained following the stipulations of Definitions 1 and 3. Thus, even the d'Hondt method, which is considered to be strongly favoring large beneficiaries, sometimes favors small beneficiaries.

Example 3 [2]. Huntington-Hill method favors large beneficiaries. Let $M = 26$, $n = 2$, $V_1 = 1000$ and $V_2 = 900$. Then $Q \approx 73.08$, $q^* = 73$ and $\Delta M = 1$. The other results of calculations are shown in Table 4, where $z_i = \lfloor V_i/q^* \rfloor$.

Table 4. Calculations to Example 3

i	V_i	a_i	V_i/q^*	Relation	$\sqrt{z_i\,(z_i+1)}$	x_i	x_i/V_i
1	1000	13	13.70	>	13.49	14	0.014
2	900	12	12.33	<	12.49	12	0.01(3)

We have $x_1 > x_2$ and $x_1/V_1 = 14/1000 = 0.014 > x_2/V_2 = 12/900 = 0.013(3)$. So, in this apportionment, according to relations (3) and (4), Huntington-Hill method favors the large beneficiaries (beneficiary 1). The same result is obtained following the stipulations of Definitions 1 and 3. Thus, even the Huntington-Hill method, which is considered to slightly favoring small beneficiaries, sometimes favors the large ones.

Thus, we have to distinguish between the favoring of beneficiaries in a particular apportionment and the favoring of beneficiaries by an apportionment method on the whole. It can happen that in particular apportionments the method favors large beneficiaries (par example, d'Hondt method), but on the whole, on an infinity of apportionments, it favors small beneficiaries and vice versa (par example, Huntington-Hill method). At the same time, when grouping n parties in subsets L (large beneficiaries) and S (small beneficiaries) according to (5) and (6), the comparative value, in pairs ("larger", "smaller"), of the beneficiaries' number of deciders is taken into account. Therefore, in particular cases it can happen that $V_{\lfloor n/2 \rfloor} > V_{\text{avi}}$ or $V_{\lceil n/2 \rceil+1} < V_{\text{avi}}$, where $V_{\text{avi}} = (V_1 + V_2 + V_3 + \ldots + V_n)/n$. But, in case of an infinity of apportionments and uniform distribution of values V_i, $i = \overline{1,n}$, relations $\text{avrg}\{V_{\lfloor n/2 \rfloor}\} < V_{\text{avi}}$ and $\text{avrg}\{V_{\lceil n/2 \rceil+1}\} > V_{\text{avi}}$ take place.

Definition 4. An apportionment method favors large beneficiaries, if the average summary number of a-entities in excess, obtained by large beneficiaries (L), is greater than that obtained by small beneficiaries (S), and vice versa, that is, it favors large beneficiaries if $\overline{\mathrm{F_{a1}}} > 0$, it favors the small ones if $\overline{\mathrm{F_{a1}}} < 0$ and it is neutral if $\overline{\mathrm{F_{a1}}} = 0$, where $\overline{\mathrm{F_{a1}}}$ is the average of $\mathrm{F_{a1}}$ on an infinity of apportionments.

So,

$$\overline{\mathrm{F}_{a1}} = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \left(\sum_{i=1}^{\lfloor n/2 \rfloor} \Delta D_{ik} - \sum_{i=\lceil n/2 \rceil+1}^{n} \Delta D_{ik} \right) = \sum_{i=1}^{\lfloor n/2 \rfloor} \overline{\Delta D_i} - \sum_{i=\lceil n/2 \rceil+1}^{n} \overline{\Delta D_i}\,, \tag{8}$$

where $\overline{\Delta D_i}$ is the average of ΔD_i on an infinity of apportionments ($K \to \infty$). Criterion $\overline{F_{a1}}$ is measured in a-entities.

In addition to identifying the APP method favoring of large or small beneficiaries (aspect A), criterion $\overline{F_{a1}}$ also allows the quantitative estimation of the absolute favoring in question, measured in a-entities (aspect B).

5. A case study: favoring of beneficiaries by d'Hondt method

To determine the value of criterion $\overline{F_{a1}}$, computer simulation with the SIMAP application was used. The initial data for calculations are: M = 6, 11, 21, 51, 101, 201, 501; n = 2, 3, 4, 5, 7, 10, 15, 20, 30, 50; $n \le M - 1$; $V = 10^8$; uniform distribution of values V_i, $i = \overline{1,n}$; sample size 10^6. So, we have 58 variants of values for the pair {M, n}: 4 + 6 + 8 + 10 × 4 = 58. The graphs of criterion $\overline{F_{a1}}$ dependence to M and n, when using the d'Hondt method, are presented in Figure 1.

From Figure 1 it can be seen that $\overline{F_{a1}}$(d'H) value is increasing to n and slightly increasing to M, especially at $M \ge 2n$. For $6 \le M \le 501$, $2 \le n \le 50$ and $n < M$, the $\overline{F_{a1}}$(d'H) value at n = 2 is in the range of 0.32÷0.41 a-entities, and at n = 50 it is in the range of 9.0÷12.1 a-entities, being considerable, especially at relatively high values of n.

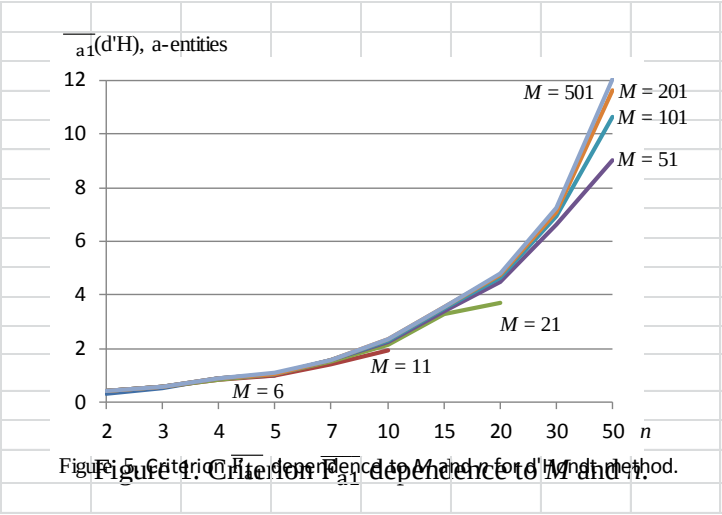


Figure 1. Criterion $\overline{F_{a1}}$ dependence to M and n.

Figure 5. Criterion $\overline{F_{a1}}$ dependence to M and n for d'Hondt method.

5. Conclusion

There is a clear distinction between favoring of beneficiaries in an apportionment and favoring of beneficiaries on the whole by an APP method. The proposed criteria F_{a1} and $\overline{F_{a1}}$ (see Definitions 3 and 4) can be used to quantitatively estimate the favoring of large or small beneficiaries in an apportionment or, respectively, on the whole by an APP method. The calculations carried out for the d'Hondt method show that the favoring of large beneficiaries can overpass 10 a-entities.

References

[1] M.L. Balinski, H.P. Young. *Fair Representation: Meeting the Ideal of One Man, One Vote*. 2nd ed. Washington, DC: Brookings Institution Press, 2001.

[2] I. Bolun. *Optimization of multi-optional decisions*. Chisinau: ASEM, 2016.

[3] Niemeyer, H.F., Niemeyer,A.C. *Apportionment Methods*. arXiv:1510.07528v1 [math.HO], Oct. 27, 2015, pp. 1-24. (https://arxiv.org/pdf/1510.07528.pdf, accessed 25.07.2020).

Ion Bolun

Technical University of Moldova
E-mail: ion.bolun@isa.utm.md

TinyAPL - Agent Oriented Programming Language Architecture

Vasili Braga, Dumitru Ciorbă, Irina Cojuhari

Abstract

An Agent Oriented Programming Language (AOPL) represents a quite complex structure containing at least 3 parts: an interface, meaning a Domain Specific Language (DSL), an object oriented structure with included knowledge representation and reasoning capabilities (KR&R), and the simulation medium for a Multi-Agent System (MAS). The purpose of this paper is to provide a possible implementation for an Agent Oriented Programming Language Framework.

Keywords: artificial intelligence, intelligent agents, domain specific languages, agent oriented programming, knowledge representation and reasoning.

1. Introduction

Designing an AOPL [1, 2, 5], one must consider the principles of agent-oriented programming, also the known Multi Agent Systems (MAS) architectures with more than 20 years of evolution [4] and the unsolved problems.

The implementation of MAS and intelligent agent's (IA) knowledge representation and reasoning [3] have to be in synergy with same language design and grammar. It can be inappropriate to have different grammar structure for different purposes, even for inter-agent communication [6] the approach must be the same.

This paper will present a common frame for this abstract agent oriented and domain specific language - *tinyAPL* grammar, architecture and implementations tools and mechanisms.

2. Related Work

First step in choosing a grammar begins with the supposition, that the language has to be practical to represent at least three things: object oriented [7, 8] structures, knowledge & reasoning and inter-agent communication, also it must definitely be able to allow event driven programming [9] for the simulation phase.

The biggest conflicts of the early approaches were the common practices of representation for AOPLs, KR&R and Agent Communication Languages (ACL). If for AOPL the main base of inspiration was object oriented languages like C++ or Java, for KR&R and ontologies the common practice was to use XML schema [10] and finally the main concern for inter-agent communication was a semantic agnostic protocol.

Also for the survival and the spreading of the language the *tinyAPL* grammar is imperative to have a low entry barrier for programmers, thus it must be based on very common principles for markup languages. The analysis of latest practices permitted to identify the JSON-LD [11] knowledge representation framework. JSON-LD is initially object oriented and allows a pretty decent ontology representation. For the reasoning capacities this language has to be augmented with framing capabilities [12]: IF_ADDED, IF_NEDDED and for the event oriented approach was added IF_CHANGED and IF_TIMER triggers.

The chosen tool for the parser implementation is the ANTLR [13] (ANother Tool for Language Recognition) - a parser generator for reading, processing, executing, or translating structured text, used to build languages, tools, and frameworks. The C++ language was chosen for the implementation of rendered MAS.

3. Architecture Definition

The *tinyAPL* architecture consists of three big modules: PREPROCESSOR, RENDER and SIMULATION units (Figure1). The PREPROCESSOR unit must prepare all the text for a future parsing: including all files with the libraries, replacing the text from the #DEFINE statements, also it has to make the first syntactic check.

The RENDER unit is definitely more complex, but roughly it consists of two modules: the rendering of the abstract object library and the instantiation of the MAS.

The Simulation module has the purpose of running the simulation for the designed and rendered Multi-Agent System, and it is implementing the reasoning, communication and interactions.

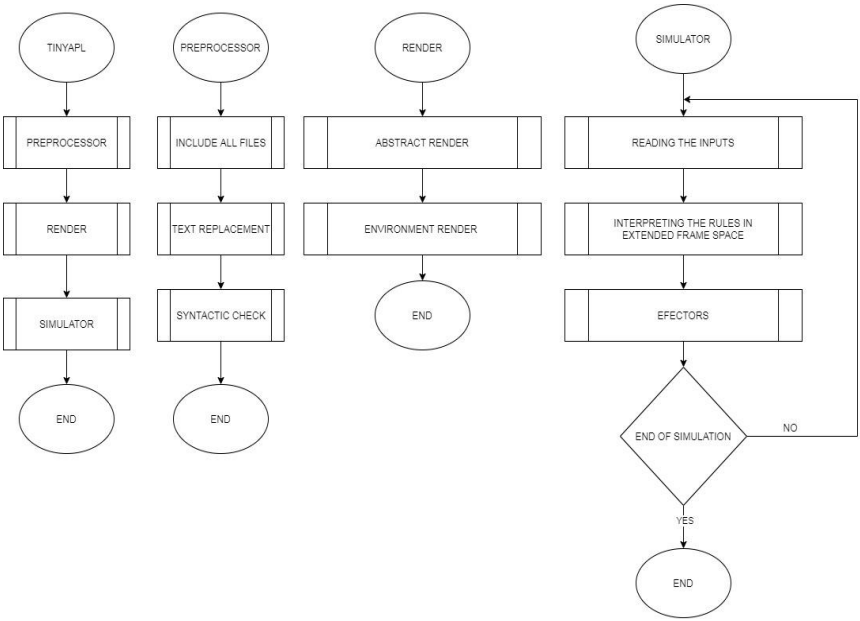


Figure 1. TinyAPL architecture diagram

3.1 The Rendering Unit Architecture

The Rendering unit must transform all parsed tree into linked memory objects, resolve the inheritance issues and build the environment and the communication channels.

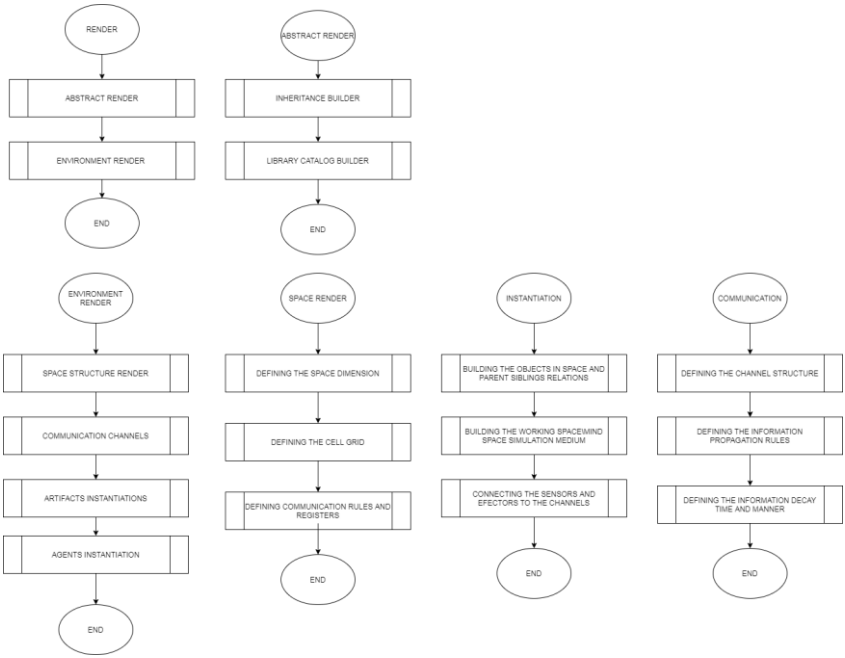


Figure 2. The Rendering unit architecture diagram

4. The Code Example

An example of code can be presented in the following way:

```
{     ABSTRACT :{
      ORGAN: {NAME : olfactive_sensor, PASSIVE & SENSOR,
      MEMORY : { NAME : short_term_memory, DECAY : 10, CLOCK
: 1, VOLUME : 100,
     REGISTER :{NAME : olfactive_history, INTERN, ARRAY [2]:
[10,9], READ & WRITE},
              RULES:{
                    IF_TIMER {
                       this->olfactive_history.push_back(VOLUME)   }
   } }  },
      AGENT: { NAME : being},
      AGENT: { NAME : human, IS_A : being },
      AGENT: { NAME : doctor, IS_A : human }
```

```
},
ENV:{
    AGENT: { NAME : Vasile, IS_A : human, POSITION : {[1,2,3], ABSOLUTE}, ROTATION:{[0,0,0],RELATIVE}},
    ARTIFACT: { NAME : Vasile, IS_A : human, POSITION : {[1,2,3], ABSOLUTE}, ROTATION:{[0,0,0],RELATIVE}},
    AGENT: { NAME : Vasile, IS_A : human, POSITION : {[1,2,3], ABSOLUTE}, ROTATION:{[0,0,0],RELATIVE}}
} } }
```

5. Future Work and Conclusion

The presented abstract framework for the *tinyAPL* language must be extended for the believable agents [14] representation through the Artificial Endocrine System [15] and an agent society structure [16]. Even so the architecture is quite solid for any kind of intelligent agent system representation and implementation.

Acknowledgments. This work is supported by the 154/23.10.19A project.

References

[1] MB. Van Riemsdijk. *20 Years of Agent-oriented Programming in Distributed AI: History and Outlook*. Proceedings of the 2Nd Edition on Programming Systems, Languages and Applications Based on Actors, Agents, and Decentralized Control Abstractions. New York,USA: ACM; 2012. pp. 7–10.

[2] T. Kosar, PE. Martinez López, PA. Barrientos, and M. Mernik. *A preliminary study on various implementation approaches of domain-specific language.* Information and Software Technology, 2008, pp. 390–405.

[3] M. Gelfond and Y. Kahl. *Knowledge Representation, Reasoning, and the Design of Intelligent Agents: The Answer-Set Programming Approach.* Cambridge University Press, 2014.

[4] V. Mascardi, D. Weyns, A. Ricci, CB. Earle, A. Casals, M. Challenger, et al. *Engineering Multi-Agent Systems: State of Affairs and the Road Ahead.* SIGSOFT Softw Eng Notes. 2019, pp. 18–28.

[5] LS. Sterling and K. Taveter. *Agent Programming Platforms and Languages.* The Art of Agent-Oriented Modeling. The MIT Press, 2009.

[6] T. Finin, R. Fritzson, D. McKay, and R. McEntire. *KQML As an Agent Communication Language*. Proceedings of the Third International Conference on Information and Knowledge Management. New York, NY, USA: ACM; 1994, pp. 456–463.

[7] O. De Troyer and R. Meersman. *A logic framework for a semantics of object oriented data modelling. OOER '95: Object-Oriented and Entity-Relationship Modeling.* Springer Berlin Heidelberg; 1995, pp. 238–249.

[8] A. Ricci and A. Santi. *Concurrent Object-oriented Programming with Agent-oriented Abstractions: The ALOO Approach*. Proceedings of the 2013 Workshop on Programming Based on Actors, Agents, and Decentralized Control. New York, NY, USA: ACM, 2013, pp. 127–138.

[9] C. Elliott. *Declarative event-oriented programming.* Proceedings of the 2nd ACM SIGPLAN international conference on Principles and practice of declarative programming. New York, NY, USA: Association for Computing Machinery; 2000, pp. 56–67.

[10] K. Aberer, T. Catarci, P. Cudré-Mauroux, T. Dillon, S. Grimm,M.-S. Hacid, et al. *Emergent Semantics Systems. Semantics of a Networked World Semantics for Grid Databases*. Springer Berlin Heidelberg, 2004, pp. 14–43.

[11] M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, and N. Lindström. *JSON-LD 1.0.* W3C Recommendation, 2014.

[12] R.B. Atman and N.F. Abernethy. *Frame-based knowledge representation system and methods.* US Patent, 6442566, 2002.

[13] T. Parr. *The definitive ANTLR 4 reference.* Pragmatic Bookshelf, 2013.

[14] A. Bryan Loyall. *Believable Agents: Building Interactive Personalities.* Ph.D., School of Computer Science Carnegie Mellon University, 1997.

[15] H. Samani. *Multimodal cognitive processing using artificial endocrine system for development of affective virtual agents*. Труды СПИИРАН,2018, pp. 56–75.

[16] V. Braga. *Un concept de agent credibil în cadrul unei societăți bazat pe sisteme de valori.* In: Ştiinţele socioumanistice şi progresul tehnico-ştiinţific: conf. șt. interuniv., 2019, pp. 166-172.

Vasili Braga, Dumitru Ciorbă, Irina Cojuhari

Technical University of Moldova
E-mails: vasili.braga@ati.utm.md, dumitru.ciorba@fcim.utm.md,
irina.cojuhari@ati.utm.md

Decision support systems for cyber risk management

Aurelian Buzdugan[1], Gheorghe Capatana[2]

Abstract

In this paper we will outline the results of the literature review on use of decision support systems for cyber risk management in critical infrastructures. We will briefly present the identified research goals as well as the opportunity of using the support systems by decision makers and operators. Based on the elements that we have identified so far, as well as the analysis of previous research, we will draft a conceptual design on the use of decision support systems for managing cyber risks. We will outline the previous results on this topic and propose future steps on developing a proof of concept.

Keywords: decision support systems, cyber security, risk management, critical infrastructures.

1. Introduction

Decision making process has become more complex with the growing amount and complexity of data, as well as short time required in making a decision. This is also applicable in the activities of managing cyber risks in critical infrastructures, where data sets consisting of risks, implications and mitigations have to be assessed in order to aid the decision makers and operators. Existing approaches in risk management are general and cover different types of risks, however are not fully focused on the cyber security risks. The emergence of IT components in almost all sectors of the industry, creates new types of risks that require specialized knowledge and understanding of the systems in order to assess, evaluate and apply adequate mitigating controls.

The developments in the past decade have led to a new state of the context driving IT systems, where the number of decision-making

activities has increased, the problems became more complex as well as the need for computer aided tools has become a necessity [3].

This sets the premises upon which we ascertain that decision support systems (DSS) are adequate solutions for managing cyber risks in critical infrastructures.

2. Current status of DSS use in cyber risk management

In previous papers we have reviewed existing scientific literature on the use of DSS in cyber risk management [1]. The main results show that this type of systems has not been fully explored in cyber risk management and we proposed future research goals in this direction. Among the highlights, we noticed a focus on elements such as resilience and contextual awareness, denoting the importance of these factors in the critical infrastructure domain.

This contributes to the idea that a holistic approach is required in order to identify and propose a concept of using DSS for our scope, where domain specific elements need to be taken into consideration. We have also observed that:

- All papers focus on a concept of DSS, however none was yet proven in practice.
- Various concepts and algorithms were embedded in the DSS, such as Bayesian networks or ontology description of actions.
- Actions such as modelling, categorization and correlation were common in the system descriptions.
- No DSS focused on cyber risks has been identified yet.

The critical infrastructure domain is one that does not have extensive coverage in the use of DSS. We believe that dedicated DSS for cyber risk management would be a novelty, as well as a modern solution for the emerging risks.

There are specific requirements that have to be taken into account when developing a concept, and mainly in terms of target users. In critical infrastructures we foresee both operators as well as decision makers making use of the risk mitigations and recommendations created by the DSS. Therefore, the data analyzed could considerably grow, as it needs to include both strategic as well as technical details. Respectively, the output data will be adjusted and formatted based on the target users.

We believe that a concept of DSS for managing cyber risks would support the existing research in this type of systems, and also integrate other emerging disciplines such as machine learning, artificial intelligence and data mining.

Incidents or events related to the physical aspects of a critical infrastructure, such as theft, intentional damages or natural hazards were the focus for years in the risk management process. Certain frameworks and systems for supporting risk management in the identification, assessment and evaluation of risks have been developed, however these in our opinion do not cover exhaustively the emerging risks such as cyber security risks. Our goal is to explore how DSS explicitly focused on cyber risks could be developed in this context. Taking into account that generic risk management frameworks exist, having DSS as interoperable modules would make this concept likely to be adopted and implemented. This is also supported by the fact that developing DSS for all types of risks could be time and resource consuming.

We believe that the conceptual DSS can be also assumed to be a type of version 8.0 [2], which will be built using computer cognitive systems, and making use of cloud and mobile computing [3]. We would like to add the potential of using artificial intelligence in certain processes, such as searching the best solution, or identifying interdependencies and modelling of the system state after a proposed action would be taken.

The response time, taking into account the peculiarity of risk management and then the response, should be as swift as possible. The design should accommodate multi-participant use, as risk management is an organizational issue. The same applies to cyber risks, where at least decision makers and operators are involved, and potentially third parties as well [4]. The multi-participant model, as per Nunamaker, ensures that requirements such as parallelism and anonymity would also be part of the design. These can be applied to multiple use cases, such as cooperating on the risk mitigation, either gathering input on potential mitigation measures from both internal, as well as external sources.

3. Conclusion

We believe DSS possess adequate abilities to help tackle emerging cyber risks in the critical infrastructure domain. The features, concepts and

technologies used by a modern DSS are an opportunity in support decision makers by analyzing large and complex sets of data and extracting relevant data.

The future goals are to develop the architecture of the proposed type of DSS for managing cyber risks in critical infrastructures. This would improve the overall process of managing cyber risks in this domain.

This concept highlights a potential application of DSS for cyber risk management in critical infrastructures. The previous results, combined with the conceptual features of the DSS, constitute a baseline for a future DSS for our scope. The insights gained serve as input for the future research agenda and will be integrated in the architecture of a decision support system.

References

[1] A. Buzdugan. *Review on use of decision support systems in cyber risk management for critical infrastructures*, Journal of Engineering Science, vol. XXVII, no. 3, 2020, pp. 134 - 145, DOI: 10.5281/zenodo.3949684.

[2] D.J. Power and G. Phillips-Wren. *Impact of social media and Web 2.0 on decision-making*. J Decis Syst, 20(3), 2011, pp. 249–261.

[3] F.G. Filip. *DSS—A Class of Evolving Information Systems*. In: Dzemyda G., Bernatavičienė J., Kacprzyk J. (eds) Data Science: New Issues, Challenges and Applications. Studies in Computational Intelligence, vol 869. Springer, Cham. 2020, https://doi.org/10.1007/978-3-030-39250-5_14.

[4] G.L. Kolfschoten and J.F. Jr. Nunamaker. *Organizing the theoretical foundation of collaboration engineering.* In: Collaboration systems: concept, value, and use. Routledge, Taylor and Francis Group, London, 2015, pp 27–41.

Aurelian Buzdugan[1], Gheorghe Capatana[2]

[1]Moldova State University
E-mail: aurelian.buzdugan@yahoo.com

[2]Moldova State University
E-mail: gh_capatana@yahoo.com

Proceedings of Workshop on Intelligent Information Systems
WIIS2020, December 04-05, 2020, Chisinau, Republic of Moldova

Generation and visualization of graphical representations of finite automata

Constantin Ciubotaru

Abstract

Algorithms are proposed for automatic generation of graphical representations of finite automata (FA) and their visualization. Known methods are modified to explore the specific properties of FA: existence of a single source node, well-defined information flow, absence of isolated nodes. LaTeX/TikZ files are generated for drawing and visualization.

Keywords: finite automata, graph drawing, LaTeX, TikZ.

1 Introduction

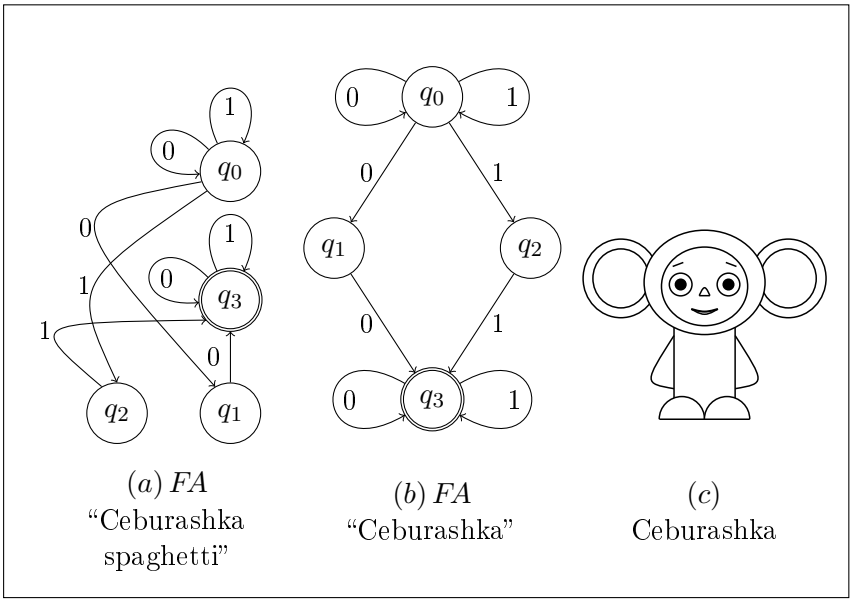
Theory and practical application of automata, in particular finite automata (FA), is one of the oldest and most actively studied fields of computer science.

Along with traditional applications related to compiler design, artificial intelligence, word processing, modern applications are currently being developed for natural language processing, speech recognition, software modeling, testing, probabilities (Markov chains), video games, image processing, cryptography etc. The development of the applications inevitably implies the need to carry out equivalent transformations on FA [1], such as:

- elimination of inaccessible and nonproductive states,
- removing ε-transitions,
- converting nondeterministic FA to an equivalent deterministic FA,
- minimizing FA and others.

For most of these transformations it is very useful to represent FA as a graph. If analytical and tabular representations do not involve difficulties for visualization, the graphical representation and visualization of AF is more difficult.

One graph may be drawn in several ways. Some images may be simple, comprehensible, having an attractive aesthetic appearance, others – more difficult to notice, with a failed structuring. For the finite automaton that recognizes the language $L = \{0,1\}^*\{00,11\}\{0,1\}^*$, there may be drawn several graphical representations. For example, the graph in Figure 1(a) contains multiple intersections of edges, that interfere with visualization and recognition tracking process. Such representations due to the intersections of the edges are called "spaghetti" representations.



(a) FA "Ceburashka spaghetti"

(b) FA "Ceburashka"

(c) Ceburashka

Figure 1. Graphical representation of the "Cheburashka" FA

In Figure 1(b) the graph represents the same *FA*, but, unlike the first graph, this one is easier visualizated and obviously illustrates the process of recognizing language strings. This *FA* is named as "Cheburashka" in association with the hero of the famous cartoon movie [1] (Figure 1(c)). Such comparation helps to memorize this automaton.

Having several representations of the same graph, it is quite easy to choose the most suitable variant, especially having at its disposal a series of appreciation criteria, usually of an aesthetic nature. For example, the graph must fit into a given and limited space, to contain as few intersections of the edges as possible, to avoid sharp curves, to respect the proportions regarding the length of the edges and the values of the angles of incidence, to favor the elements of symmetry and concentration of nodes, to use suitable shapes for nodes, to respect the orientation of the information flow (from top to bottom or from left to right). For finite automata the general flow of the information will always be oriented from initial state to final states. It is quite difficult to transmit these criteria to the computer. Here intervenes not only the problem of formalizing the criteria, but also the fact that some of them are contradictory. Inevitably there arise compromise situations.

2 Proposed solutions

Starting with the *FA* definition (or analytical representation), representation of the *FA* is generated automatically in the form of a graph $\Gamma = (Q, E, F)$, where Q is the set of nodes (states of the *FA*), E – the set of edges (q_i, a, q_j), and F – the set of final state nodes. In order to obtain the graphical representation, a compiler was developed that generates for this graph an LaTeX/Ti*k*Z [2, 3, 4, 5] file.

The compiler randomly generates and uniformly distributes the node coordinates. This means that between any two nodes there will be respected the minimum distance specified in advance according to the number of nodes.

[1] `https://en.wikipedia.org/wiki/Cheburashka`

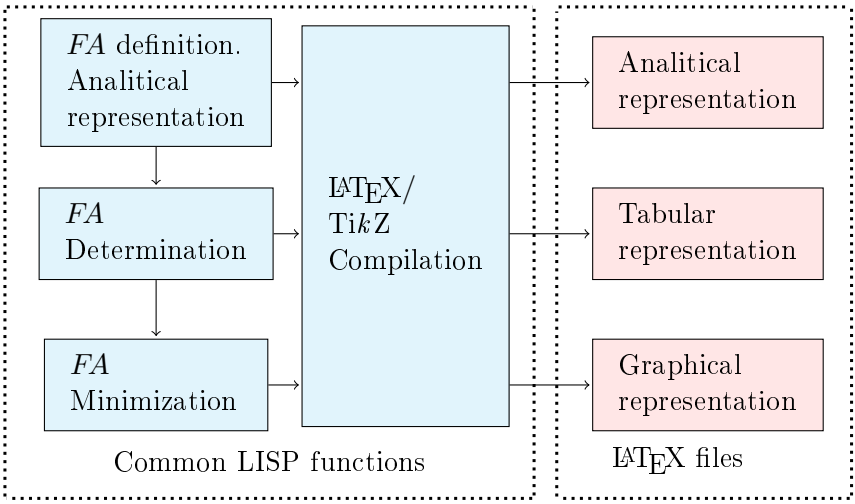


Figure 2. Scheme of the developed applications

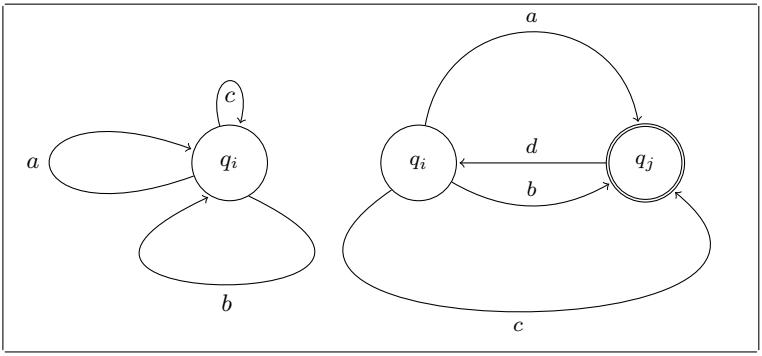


Figure 3. Visualization of the generated fragments

Figure 2 shows the schema of the developed applications and their interaction.

```
\node[state](q_i)at (3.5,3.5){\scriptsize{$q_{i}$}};
\node[state,accepting](q_j)at(6.5,3.5){\scriptsize{$q_{j}$}};
```

a) Templates for nodes.

```
(q_i)edge[to path={(\tikztostart.-60).. controls
    (4.5,1.5)and (-2.6,1.5) .. node[below]
    {\footnotesize $b$} (\tikztotarget.-120)}](q_i)
(q_i)edge[out=200,in=160,looseness=20]node[left]
    {\footnotesize $a$} (q_i)
(q_i)edge[loop above]node[below]{\footnotesize $c$} (q_i);
```

b) Templates for loops.

```
(q_i) edge [out=80,in=100,looseness=1.5] node [above]
 {\scriptsize $a$} (q_j)(q_i) edge [bend right=30] node
 [above]{\scriptsize $b$} (q_j)
(q_i) edge [to path={(\tikztostart.-135).. controls (0.0,1.0)
  and (9.5,1.0) .. node [below]{\scriptsize $c$}
 (\tikztotarget.-45)}](q_j)
(q_j) edge node [above]{\scriptsize $d$} (q_i);
```

c) Templates for edges.

Figure 4. Generated templates

Generating LaTeX files for analytical and tabular representations is done using `\tabto` and `\tabular` packages and is relatively simple. Several difficulties arise during generating graphical representations. It conserns the way the nodes are placed, the loops and edges are drawn, and the tendency to obtain an aesthetic and comprehensible drawing. Full automation of this process is virtually impossible, because the formulated criteria are difficult to formalize, and sometimes they are even contradictory [6]. The solutions, proposed in this paper, will help the

user to build an acceptable graphic structure.

The developed compiler exploits, along with the possibilities of the LaTeX system, the package `\tikz` and its libraries, especially the libraries `automata, positioning, arrows`. It should be mentioned that there exists the possibility to highlight the final and the initial states and rich arsenal for drawing loops/edges.

Several possible variants are inserted in Figure 3. For each edge (loop) functional templates are generated, one is active, the others are commented and offered to the user in case he wants to intervene manually.

These templates are inserted in Figure 4.

3 Example

An FA was selected as an example for which there were automatically generated the deterministic and minimized equivalent models. LaTeX files for tabular and graphical representations were generated for all these constructs. All these are inserted in Figures 5, 6.

4 Conclusion

The applications proposed in the paper are used in the training process for studying finite automata. Of course, manual intervention is still required, but the functional version LaTeX, obtained as a result of the compilation, is very useful. Upcoming the involvement of the Sugiyama framework is envisaged, which broadly minimizes the number of the edges intersections.

EXAMPLE

Analitical definition

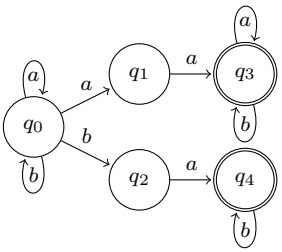
$AF=(Q,\Sigma,\delta,q_0,F)$.
$Q=\{q_0,q_1,q_2,q_3,q_4\}$,
$\Sigma=\{a,b\}$,
$F=\{q_3,q_4\}$
$\delta(q_0,a)=\{q_0,q_1\}$,
$\delta(q_0,b)=\{q_0,q_2\}$,
$\delta(q_1,a)=\{q_3\}$,
$\delta(q_2,a)=\{q_4\}$,
$\delta(q_3,a)=\{q_3\}$,
$\delta(q_3,b)=\{q_3\}$,
$\delta(q_4,b)=\{q_4\}$

Tabular reprezantation

$AF=(Q,\Sigma,\delta,q_0,F)$.
$Q=\{q_0,q_1,q_2,q_3,q_4\}$,
$\Sigma=\{a,b\}$,
$F=\{q_3,q_4\}$

q	$\delta(q,a)$	$\delta(q,b)$
q_0	$\{q_0,q_1\}$	$\{q_0,q_2\}$
q_1	$\{q_3\}$	$\{\}$
q_2	$\{q_4\}$	$\{\}$
q_3	$\{q_3\}$	$\{q_3\}$
q_4	$\{\}$	$\{q_4\}$

Graphical reprezentation



Deterministic FA, generalized states

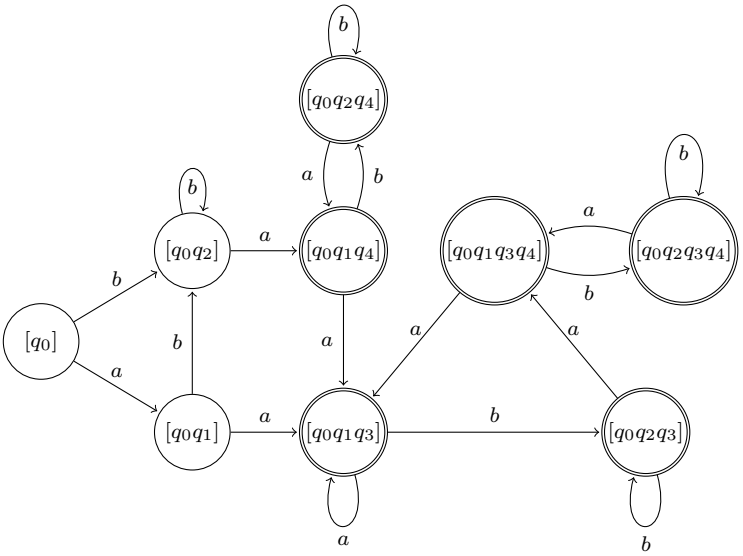


Figure 5. Example. Page 1

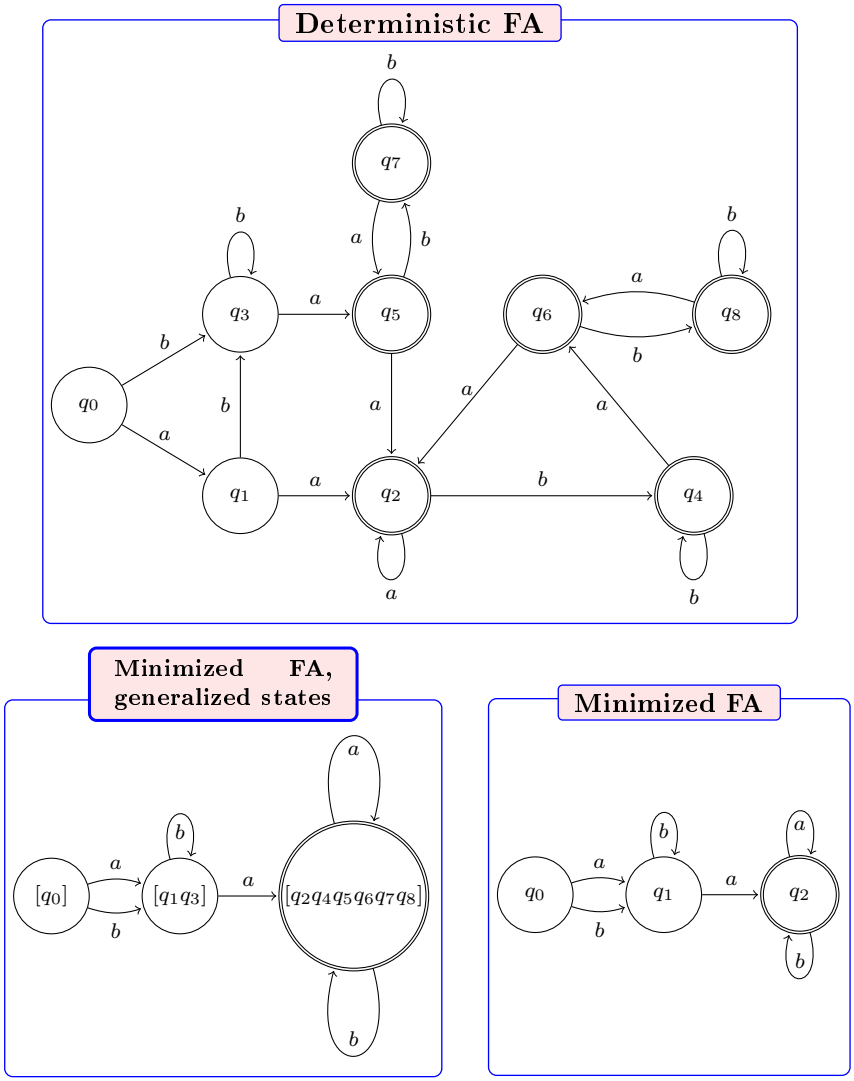


Figure 6. Example. Page 2

Acknowledgments. This article was written as part of the research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data".

References

[1] J.E. Hopcroft, R. Motwani, and J.D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison Wesley, 1979, 427 p.

[2] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion.* Addison-Wesley, Reading, Massachusetts, 1993.

[3] Stefan Kottwitz. *LaTeX Cookbook.* Packt Publishing, 2015, 387p.

[4] *The TikZ and PGF Packages. Manual for version 3.1.6.* Institut für Theoretische Informatik Universität zu Lübeck, 2020, 1320p.

[5] Till Tantau. *Graph Drawing in TikZ.* Journal of Graph Algorithms and Applications, 17 (4), 2013, pp.495–513.

[6] G. Di Battista, P. Eades, R. Tamassia, and I.G. Tollis. *Graph Drawing. Algoriithms for the visualization of graphs.* Prentice Hall, 1999, 397 p.

[7] K. Sugiyama, S. Tagawa, and M. Toda. *Methods for visual understanding of hierarchical system structures.* IEEE Trans. Syst. Man. Cybern., 11(2), 1981, pp.109–125.

[8] N.S. Nikolov. *Sugiyama Algorithm.* Encyclopedia of Algorithms - 2016 Edition, Springer 2016, pp. 2162–2166.

Constantin Ciubotaru

Vladimir Andrunachievici Institute of Mathematics and Computer Science
E–mail: chebotar@gmail.com

On convergent technology in development of information systems for processing of documents with heterogeneous content

Alexandru Colesnicov, Svetlana Cojocaru, Ludmila Malahov, Lyudmila Burtseva

Abstract

Recognition of heterogeneous documents is a challenging problem. To solve it, a toolkit is proposed in the form of a platform for processing documents containing heterogeneous content. For the development, we used convergent technology, according to which both existing software tools, as well as those developed by us, can provide all stages of the recognition process. The development of a platform for solving the problem is done in a virtual environment. The paper describes a functional subset of the platform for solving one subtask: to analyze a scan of a document and to cut it into segments.

Keywords: Recognition of heterogeneous documents, convergent technology, virtualization.

1. Introduction

Development of information technologies led to a variety of specialized systems and software packages for solving specific problems. Currently, there are problems that can be solved using convergent technologies, which represent convergence of existing material, information and cognitive technologies.

We define convergent technologies through the general principle of physical and/or logical pooling of resources based on any resemblance, belonging on similarity of tasks being solved in some aggregate or document processing platform, resource pools, communications, systems and/or data storage or backup networks.

During document digitization, we met the necessity to use for processing the tools that provide support for recognition not only of the text but of other elements of heterogeneous nature: for example, really huge archives [2], [3] of scanned newspapers with heterogeneous content. Encyclopedia is also a good example of a document with heterogeneous content because we meet on its pages not only text, but a variety of content types including images, mathematical and chemical formulas, musical scores, technical drawings, chess notation, electronic circuits, etc., with varied geometric shapes.

In our work, we propose a toolkit in the form of a platform for processing documents containing heterogeneous content. We suppose also that any heterogeneous content is associated with the possibility of its presentation in a scripting language. The main features of such content are the following: the document is not exclusively in natural language; there is one or more scripting languages for presenting its components; the graphic representation can be re-rendered from the scripts [1].

The problem of document image processing includes tasks that have already become almost classical, such as segmentation, noise filtering and image extraction from the background, determination of object boundaries, and pattern recognition.

The following describes one of the subtasks of the platform for document processing, namely segmentation, i.e., the process of selecting existing objects in the image by using existing software.

2. Development of a platform for solving the problem

Analysis of the work cycle for processing historical documents with a heterogeneous context showed that only a semi-automated workflow organization is possible. For implementation, a convergent technology for assembling complex software systems from ready-made modules a single platform is used; each of modules performs a small part of the task.

Let's consider the main ready-made systems used to develop the platform.

The main principle of our development is the platform virtualization. For this purpose, Docker[1] is used.

[1] https://www.docker.com/

The Docker system consists of a service that performs all the basic operations, two shells, the command line and the graphical ones, and the docker-compose extension.

Docker runs programs in an environment that is isolated from the operating system. It works on Linux, MacOS and Windows 10+. Docker *images* are quickly built and start. Each image, when launched, spawns its runtime instance called *container*.

An image is a collection of immutable *layers* that represent differences in the states of the filesystem. Images are immutable, and containers, in addition to the immutable layers of the image, contain a writable layer that reflects its changes. The *docker commit* command creates a new image from the container, adding its writable layer on top of the base immutable layers and making it a new immutable layer.

The developer is provided with a version tracking system similar to GitHub, with the ability to rollback.

A group of containers launched via docker-compose has access by default to the internal network that connects these containers. Each container can also connect to host networks either through port forwarding or directly. Finally, you can build an additional network from selected containers.

The containers' RAM is isolated from the host and from each other. Container disk storage can be modeled on host disks, on another machine, or in the cloud. You can also permit the container direct access to the host disk.

A library of ready-made images Docker Hub[2] is available.

The commercial system ABBYY FineReader Engine (FRE) is used to analyze the scanned document.

Image processing is performing with ImageMagick batch utility.

Another method used in platform development is Deep Learning ([4], [5]) whose implementation templates are widely available. This approach seems very suitable for our goals.

The language of implementation is Python. Its rich libraries provide a lot of ready-made solutions for subtask of the implemented platform.

[2] https://www.docker.com/products/docker-hub

3. Semi-automated workflow for recognition of heterogeneous documents

Despite a lot of achievements, automated recognition of the heterogeneous content remains a difficult problem. The problem is to maximize the support of semi-automated work.

To solve this problem, we propose a platform for recognition of heterogeneous documents, which uses existing and newly developed programs, and can perform all stages of the process.

The platform is created to integrate all existing software to maximize the degree of recognition automation. The recognition of heterogeneous documents involves many processes. Some may be performed automatically (green in Figure 1) using specialized software. Some processes need slight manual intervention or manual control. If the specialized software does not exist, the processing is executed manually (red in Figure 1) under the general purpose software. The integrating platform should facilitate the manual operations if they are necessary.

In Figure 1, after the preparatory stage the scan of the document is processed by a Python script developed by us, which analyzes the image. As a result of partitioning into regions, we obtain segment files with a uniform context (Scan segments) and page maps.

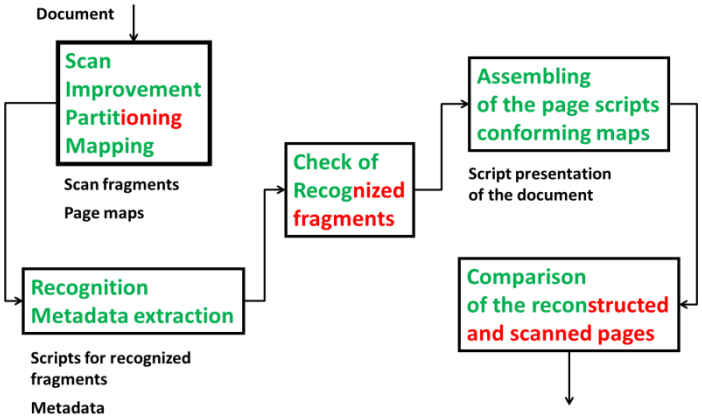


Figure 1. Structure of platform for recognition of heterogeneous documents

Using software for recognition and metadata extraction, we get scripts for the recognized segments and metadata.

The next stage is the check that can be automatic, or semi-automatic, or manual, depending on the content.

Finally, the assembly of the page scripts conforming to maps follows.

The result is a script presentation of pages of the document that can be used as a logical representation both for storage and for further processing.

Therefore, we can group all involved processes as follows:

- Automated: scan; segment recognition according to types of segments; assembly of script presentation of pages; metadata integration; reconstruction of page images from the script; automated verification.
- Semi-automated: image quality improvement; page layout analysis; task distribution for manual verification.
- Manual: human verification and manual correction.

Manual verification will be performed by experts in the corresponding areas. It implies that the platform will be Web-based like Wikipedia or version control systems.

The coordination of this activity is also necessary. Web platform for recognition of heterogeneous documents could be implemented to integrate all used tools.

4. Subtask: analysis of a scanned document and cutting it into segments with the same content type

ABBYY FineReader Engine (FRE) includes a ready-made command line interface (FRE CLI) that performs the full recognition cycle for one page at a time. The result is returned in XML format and contains the coordinates of the page segments, their type (text, image, table, separator, etc.), and the recognized text for text segments. Thus, it is possible to cut the page into parts with homogeneous content.

To process several pages, the utility can be called in a loop from a command script in any suitable language, including Python. It is possible to use container technology to build a platform from separate scripts.

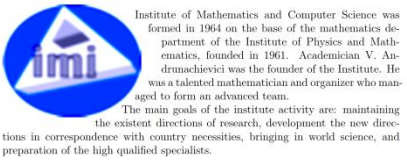
A Python program was developed and written to cut a scan of a document into segments with the same content type. The algorithm is as follows:

1. Using the FRE CLI utility, we analyze a page scan getting an XML file with the coordinates of the upper and lower corners of the segment rectangles and the segment type (text, picture, table, etc.).

2. A batch of scans of a multi-page document is processed in a cycle. Names of files with page images are set in the command line, with the ability to use regular expression elements (placeholders * and ?). For each image, a subdirectory is created with a name derived from the image name, into which the XML file and page segments are placed in a format that matches the page image format (Figure 2).

The Python script reads an XML file, selects the segment metadata (coordinates, segment type) and calls batch utility ImageMagick for image slicing. Separators are excluded from further processing.

3. After processing with FRE, each fragment on the page is described in an XML <block> element. The fragment's geometry is set by the <block> tag parameters: coordinates of the upper left and lower right corners of the minimal enclosing rectangle. A more complete description of the fragment's geometry is contained in a nested <region> element, which, in its turn, contains <rect> elements, each of which describes one rectangle. A fragment (“region”) consists of one rectangle or of several rectangles in more complex cases (Figure 3). For example, one line in a rectangle is such a region, if the analyzed fragment is text. If there are several such rectangles in the region, then the program provides a restructuring module to process such fragments. The module uses ImageMagick in a loop to compose a fragment from the constituent rectangles. The glued image fragment is written to a file for further processing.




```
<region>
<rect l="682" t="102" r="1884" b="159"/>
<rect l="683" t="159" r="1884" b="160"/>
<rect l="740" t="160" r="1884" b="463"/>
<rect l="682" t="463" r="1884" b="521"/>
<rect l="683" t="521" r="1884" b="523"/>
<rect l="623" t="523" r="1884" b="583"/>
<rect l="90" t="583" r="1884" b="762"/>
</region>
```

Figure 2. Image and set of coordinates for the text fragment produced by FRE CLI

The restructuring module allows one to solve the problem of reconstructing the geometric shape of a fragment. After executing the script, the following result is obtained.

For each page image, a subdirectory is created with the name generated from the image file name.

These subdirectories contain images of page segments recovered after restructuring. The subdirectory also contains two XML files. One of them is generated by the FRE CLI utility and describes the entire page. There are coordinates of all blocks, regions and rectangles, as well as the recognized text. The second XML file is generated in Python and contains a list of page segments with segment types and coordinates of the enclosing blocks.

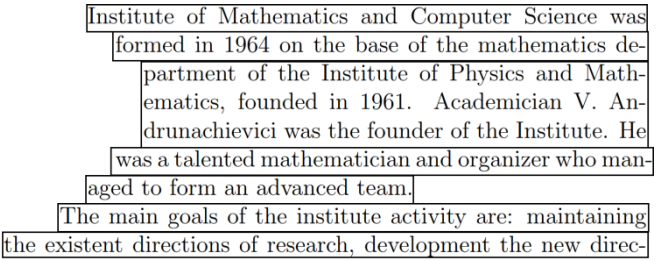


Figure 3. Fragment of irregular shape composed from several rectangles

Thus, a script developed in Python that uses FRE and the ImageMagick batch utility, generates files with scans of document segments and XML files with metadata for further processing by the platform, implementing a subset of the platform functionality.

4. Conclusions

Despite a lot of achievements, automated recognition of the heterogeneous content remains a difficult problem. We proposed a design of Web platform to maximize the support of semi-automated work of all used tools for recognition of heterogeneous documents.

In implementation, a convergent technology for assembling complex software systems from ready-made heterogeneous modules on a single platform is used. Each of modules performs a small part of the task using inside a container.

The program for partitioning and mapping of heterogeneous documents into homogeneous segments with different shapes was developed using ABBYY FineReader Engine, ImageMagick and Python. Scans with text, music, images, etc. were analyzed.

This is an example solution of segmentation and markup problem. The problem of complete classification for types of heterogeneous content isn't solved being the next stage of our work.

Acknowledgments. This article was written as part of the research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data".

References

[1] A. Colesnicov, S. Cojocaru, and L. Malahov. *Recognition of heterogeneous documents: problems and challenges.* Proceedings of the 5th Conference on Mathematical Foundations of Informatics. 3-6 July 2019, Iasi, Romania, pp. 231-245, Iasi: „Alexandru Ion Cuza" University Publishers.

[2] Hui-Yin Wu and Pierre Kornprobst. *Multilayered Analysis of Newspaper Structure and Design.* [Research Report] RR-9281, UCA, Inria. 2019. hal-02177784

[3] Lee, B.C., Mears, J., Jakeway, E., Ferriter, M.M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., and Weld, D.S. *The Newspaper Navigator Dataset: Extracting And Analyzing Visual Content from 16 Million Historic Newspaper Pages in Chronicling America,* 2020. ArXiv, abs/2005.01583.

[4] Simon Haykin. *"Neural Nemworкs: A Comprehensive Foundation"* Second Edition. [In Russian.] Саймон Хайкин. *«Нейронные сети: Полный курс»* Второе издание : Пер. с англ. М. – Издательский дом «Вильямс», 2006. 1104 с.

[5] Nikolenko S., Kadurin A., and Arkhanghelskaya E. *Deep Learning.* [In Russian.] Николенко С., Кадурин А., Архангельская Е. *Глубокое обучение.* – СПб.: Питер, 2018. – 480 с.

A. Colesnicov[1,2], S. Cojocaru[1,3], L. Malahov[1,4], L. Burtseva[1,5]

[1] V.A.Andrunachievici Institute of Mathematics and Computer Science
[2]E-mail: acolesnicov@gmx.com
[3]E-mail: svetlana.cojocaru@math.md
[4]E-mail: ludmila.malahov@math.md
[5]E-mail: luburtseva@gmail.com

Upgrading Cloud Infrastructure for Research Activities Support

Nichita Degteariov, Petru Bogatencov, Nicolai Iliuha, Grigore Secrieru

Abstract

In the paper, approaches for upgrading the heterogeneous distributed computing infrastructure that integrates various types of computing resources are described. It is shown the necessity of development of computer infrastructures and technologies, which is focused on creating conditions for solving complex problems with high demands of computing resources. Analysis and trends of development of tools for automation of complex cloud infrastructures configuration and administration are presented. Problems that restricted scalability of the existing Cloud infrastructure are identified and solutions to overcome existing limitations by application of new tools for cloud infrastructure configuration and administration are suggested.

Keywords: cloud computing, information technologies, e-infrastructure & services, deployment tools.

1. Introduction

In the past years, development of distributed and high-performance computing (HPC) technologies for solving complex tasks with specific demands of computing resources, creating abilities to store and access increasing amounts of research data are actively developing, including in Moldova [1]. New European Open Science Cloud Initiative (EOSC), aimed at the accumulation of various scientific information in cloud for organization of open access, has a further significant impact on the intensification of the use of distributed computing resources. The initiative is oriented at creation of open research data repositories to support open

science and development of technologies for the accumulation and use of FAIR (Findable, Accessible, Interoperable and Reusable) data based on wide utilization of cloud computing resources. New areas of works in this direction focused on deployment of new types of cloud infrastructure that will integrate Grid and HPC computing resources and gain benefit to end users from uniting computational resources of multiprocessor clusters with effective application platforms, users' interfaces and infrastructure management tools offered by Cloud infrastructure.

2. Overview of Cloud Infrastructure development for research activities support at the Vladimir Andrunachievici Institute of Mathematics and Computer Science (VA IMCS)

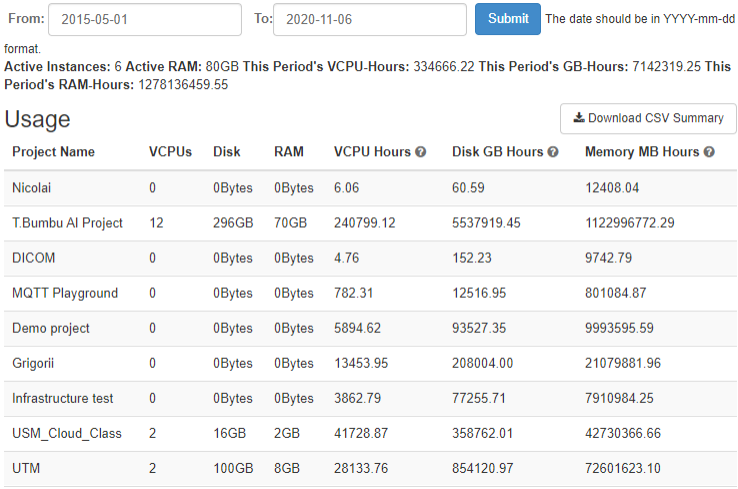
Work on the implementation of cloud infrastructure at the Institute of Mathematics began in 2014-2015, as a result of the participation of VA IMCS and partner organization RENAM engineers in the regional project Experimental Deployment of an Integrated Grid and Cloud Enabled Environment in BSEC Countries on the basis of gEclipse (BSEC gEclipseGrid) supported by Black Sea Economic Cooperation Programme (http://www.blacksea-Cloud.net) [2].

The experience and results accumulated in the course of this project, in 2015, resulted in the Cloud infrastructure available for evaluation and application testing, based on OpenStack version 13 Mitaka (release 2016), installed and accessible at https://cloud.renam.md/. It was deployed using two computing nodes of the multiprocessor cluster of the Vladimir Andrunachievici Institute of Mathematics and Computer Science, which at that time were taken out of mainstream operation in this computing facility. The total amount of resources on these two servers was quite modest even by those standards - only 16 CPU cores, 32 Gb RAM, 750 Gb HDD, 1 Gbit/s network. Despite this, this infrastructure was widely used, both for evaluation purposes, and in several short-term projects and tasks in which it was necessary to deploy quickly a small virtual infrastructure to test various scenarios and products.

In 2018, the infrastructure was supplemented with a modern high-performance server with 32 CPU threads, 128 Gb RAM, 3 Tb storage, 1Gb network. This upgrade has opened a new stage in the development of Cloud computing facility at VA IMCS – new resources for several

institutional projects in the field of Machine Learning and Neural Language Processing were provided there.

In 2019, the Cloud infrastructure of the Institute of Mathematics was used for new service – support of on-line lectures organization for the State University and the Technical University of Moldova. The experience gained over the years and user feedback made it clear about the need for further development of Cloud infrastructure in VA IMCS (see Figure 1), but a further increase in resources was no longer an option - the technologies used in the existing Mitaka release no longer met modern security standards, and the manual installation process used to implement OpenStack created insurmountable difficulties for further upgrade and system administration.


From: 2015-05-01 To: 2020-11-06 Submit The date should be in YYYY-mm-dd format.

Active Instances: 6 **Active RAM:** 80GB **This Period's VCPU-Hours:** 334666.22 **This Period's GB-Hours:** 7142319.25 **This Period's RAM-Hours:** 1278136459.55

Usage — Download CSV Summary

Project Name	VCPUs	Disk	RAM	VCPU Hours	Disk GB Hours	Memory MB Hours
Nicolai	0	0Bytes	0Bytes	6.06	60.59	12408.04
T.Bumbu AI Project	12	296GB	70GB	240799.12	5537919.45	1122996772.29
DICOM	0	0Bytes	0Bytes	4.76	152.23	9742.79
MQTT Playground	0	0Bytes	0Bytes	782.31	12516.95	801084.87
Demo project	0	0Bytes	0Bytes	5894.62	93527.35	9993595.59
Grigorii	0	0Bytes	0Bytes	13453.95	208004.00	21079881.96
Infrastructure test	0	0Bytes	0Bytes	3862.79	77255.71	7910984.25
USM_Cloud_Class	2	16GB	2GB	41728.87	358762.01	42730366.66
UTM	2	100GB	8GB	28133.76	854120.97	72601623.10

Figure 1. Cloud in VA IMCS 2015-2020 resource usage

3. Identified problems in the existing Cloud infrastructure and suggested solution

The fact is that the OpenStack Cloud infrastructure is a very flexible and, as a result, a very complex product, consisting of hundreds of open-source "bricks" components combined into a single system, which have many dependent on each other services and components, with their own settings,

which must be written into the configuration files before executing the commands to "dock" the components into a single working environment. Mistakes when installing such a complex system are inevitable and sometimes irreversible and necessitate reinstalling all components from scratch. This approach makes further administration and scaling of the system absolutely impractical and complicated process.

Installing OpenStack manually is great for getting familiar with the internals of a system and understanding how its components interact. However, when installing the system on large infrastructures, where clustering of more than 3-5 servers is used, manual installation is already absolutely inappropriate and unpromising in terms of further scaling.

To overcome these limitations, increase the reliability and the possibility of further upgrading and scaling the Cloud system, a completely different approach was required.

To solve this problem, it was decided to apply a modern approach to the administration of cloud systems - using Deployment Tools, which allow creating scripts to automate the installation of the system. It is difficult to imagine a modern IT project without such solutions. There are already known about a dozen of such kind of tools to automate configuration processes, the main ones and mostly used being Ansimble, Puppet, Chef, Juju. To deploy a new improved Cloud system on the basis of modern equipment, we have chosen a combination of opensource tools such as MAAS (Metal-As-a-Service) and Juju Charms.

MAAS is designed to deploy quickly and easily Ubuntu configurations across multiple servers using techniques used in cloud platforms. But unlike cloud platforms, resource allocation on such kind of cluster occurs at the level of physical servers, not virtual environments. At the heart of MAAS there is the simple idea of Preboot eXecution Environment (PXE) booting and a tool for deploying and maintaining Juju environments, which turns the installation and configuration process into an extremely simple task, performed using two or three commands [3]. It would take too long to manually configure the OS and services on each server node in the cluster, whereas tools like MAAS can deploy an entire cluster in just a few minutes.

Juju is an orchestrator that can be used to declaratively describe the infrastructure configuration of an application: which applications are

running, on which machines, in how many copies, and how they are linked to other services. The custom code for configuring individual virtual machines with Juju is called Charm [4].

Having the installation of automation system, we get not only a gain in man-hours for deploying a ready-made production-ready infrastructure, but also flexibility in its administration and ease of scaling. For example, having a Compute Node setup script written, you can start provisioning any number of new nodes with a single command, using an existing debugged configuration. This allows to minimize the occurrence of errors when commissioning new system components and to carry out maintenance with minimal delays or even without downtime.

Creation of new infrastructure will allow us to eliminate many of the limiting factors of the infrastructure currently operating in our research Cloud. The main advantages of the updated system are: more computing resources, block-storage for creating a backup, a new network model that allows users independently to create self-service local networks with local addresses and use the mapping of floating IP address. This, in turn, will increase security and significantly reduce the use of public IP addresses – for example, now only 16 real IPs are available and used in the current system.

4. Future plans of research Cloud infrastructure development

Introduction of new technologies, ideas and planned improvement of the computer and network infrastructure in the VA IMCS and RENAM data centers gave impetus to the creation and transition to a new, more modern, productive and large-scale Cloud platform.

The new Cloud platform is designed to eliminate bottlenecks in the current system, provide users with more resources using modern high-performance servers, more bandwidth by migrating infrastructure to 10 Gbps connectivity, and being more reliable, flexible and resilient by using automated deployment tools.

This work began at the end of 2020, with the transition of the VA IMCS and RENAM infrastructures to new high-performance servers and new 10Gbps network equipment. Completion of the work is planned in 2021, and then the system will become available for testing and after a

successful evaluation of results the obtained VA IMCS resources will also be integrated in the upgraded Cloud infrastructure.

5. Conclusion

Today it is already impossible to imagine life without Cloud systems. They penetrate all areas of our lives and continue to gain popularity, and the current state of affairs with the coronavirus pandemic will continue to increase the demand for the provision of more and more virtual resources and services in the Cloud. They will be an excellent aid for supporting scientific and educational activities.

Acknowledgments. This work was supported by research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data" and EU H2020 project "National Initiatives for Open Science in Europe – NI4OS Europe" (Grant Contract 857645).

References

[1] P. Bogatencov, G. Secrieru, N. Degteariov, and N. Iliuha. *Scientific computing infrastructure and services in Moldova*. Springer Link, Journal Physics of Particles and Nuclei Letters LNCS, vol. 13, Issue 5 (2016), pp. 685-688, DOI: 10.1134/S1547477116050125.

[2] H. Astsatryan, A. Hayrapetyan, W. Narsesian, P. Bogatencov, N. Iliuha, R. Kvatadze, N. Gaamtsemlidze, F. Florian, G. Neagu, and A. Stanciu. *Deployment of a Federated Cloud Infrastructure in the Black Sea Region, Computer Science and Information Technologies*. Proceedings of the CSIT Conference, Sep. 23-27 — Erevan, Armenia (2013), pp. 283–285.A.A. Waksman. *Permutation Network.* Journal of the ACM, vol. 15, no 1 (1968), pp. 159-163.

[3] *MAAS documentation* — https://maas.io/docs.

[4] *Juju documentation* — https://juju.is/docs.

Nichita Degteariov[1], Petru Bogatencov[1], Nicolai Iliuha[1], Grigore Secrieru[1]

[1] Vladimir Andrunachievici Institute of Mathematics and Computer Science
E-mail: nichita.degteariov@math.md

An approach to structure information regarding patient diagnostics in the form of taxonomy in management of mass casualty disasters

Constantin Gaindric, Olga Popcova, Sergiu Puiu, Iulian Secrieru,
Elena Gutuleac, Svetlana Cojocaru

Abstract

Processing of poorly structured data and knowledge remains very important, as processing methods greatly depend on the application domain. It is particularly difficult to provide activities with data, information and knowledge of good quality. This article presents main features of ill-structured problems, and describes the experience of structuring information regarding patient diagnostics in the form of taxonomy (based on vital and sonographic signs) in the process of management of mass casualty disasters. Using the described approach, during the working sessions of expert and "knowledge engineer" the task of acquisition and formalization of facts and decision rules was performed.

Keywords: ill-structured problems, knowledge acquisition, data structuration, taxonomy, mass casualty situation, sonographic diagnostics.

1. Introduction

The existence of society today, but even more so in the future, depends on information and communication technologies, which are integrated with the traditional ones, gradually replacing them.

Integrated technologies are increasingly applied in the real production environment, as well as in the artificial one, but also in services. Information technologies ensure fast and easy transfer of data in the form of text, images and voice, making services more efficient. Virtual reality

will make possible interactive virtual modelling and design of information systems at a more advanced level.

Predominantly unstructured nature of the data, on the basis of which the information systems operate, with those databases are populated, and which are traditionally manipulated, remains a permanent actual problem. Therefore, the processing of poorly structured data and knowledge remains important, especially since these processing methods at present, but also in the near future, greatly depend on the application domain. It is particularly difficult and important to provide activities with data, information and knowledge of good quality.

In order to improve the quality of data, information and knowledge in the daily activity and for the information systems, there is a need of new solutions and specific tools to be used.

Recently more and more attention is paid to the quality of data and information. Many researchers, including Pierce, Kahn, and Melkas [1], examine the relationship between data quality, information quality, and knowledge quality. It is stated that improving the data quality should lead to an improvement in the information quality, generated from this data. Therefore, it seems reasonable that improving information should, in its turn, improve the knowledge quality.

We are convinced that information systems will become the predominant factor in the progress of any field of activity, including in the help of doctors, especially to provide a correct, fast and efficient diagnostics.

2. Features of ill-structured problems

Ill-structured problems are the ones that everyone commonly faces in his/her everyday life. These include important social, political, economic and scientific issues.

The solution for ill-structured problems, usually, requires the following activities: a) definition, description, problem formalization; b) generating possible solutions; c) evaluation of alternative solutions, taking into account the end-users' preferences; d) implementation of the most viable solution; e) monitoring of the implementation.

Knowing the domain and its good description are the main factors in solving ill-structured problems. In addition, professional skills and

knowledge, involved in generating solutions in the decision-making process, should be identified. All solutions should also have the justification component in order to be evaluated later on. Other two important components of solving ill-structured problems are: taking into account the decision maker's view and selecting the solution based on a personalized end-user approach.

Usually, solutions for ill-structured problems are rarely correct or incorrect, but they should fall within a range of acceptances. As a result, in order to be judged there are needed the stages of testing, implementation and evaluation based on the arguments.

The stages of solving ill-structured problems are very similar with the five stages of the development of knowledge-based systems: 1. Problem identification; 2. Knowledge acquisition; 3. Knowledge structurization; 4. Knowledge formalization; 5. Prototype development, testing and implementation. This is due to the fact that both processes are based on human reasoning logic.

As in the case of the development of knowledge-based systems, the main stage in solving ill-structured problems is considered the phase of knowledge acquisition, structurization and formalization. The aim is to obtain an informal description of the knowledge regarding the studied domain in the form of a graph, table, diagram or formatted text.

Taxonomy is the most commonly used form of representation of the structured description of the problem area.

The following section describes the experience of structuring information regarding patient diagnostics in the form of taxonomy (based on vital and sonographic signs) in the process of management of mass casualty disasters, an eloquent example of a domain with poorly structured and heterogeneous data and knowledge.

3. Knowledge structurization and formalization in the form of taxonomy during expert-"knowledge engineer" interaction

As quality of the acquired knowledge is the determining factor for the successful solution of any ill-structured problem, the decision was made to use the traditional method of knowledge acquisition – with participation of the "knowledge engineer" [2]. The main objective of this method is that during the sessions of professional knowledge acquisition, the competence

and expert knowledge should be transferred to the "knowledge engineer", in order to obtain the most complete possible representation of the problem area.

Later, the information obtained from medical experts was structured, formalized and introduced by the "knowledge engineer" into the knowledge base (a pyramid of meta-concepts and a set of rules created on their basis), using the ExpShell tool [3].

The main characteristics/notions (vital and sonographic signs) used in the description of the domain of patient diagnostics in the process of management of mass casualty disasters have formed the main nodes. The other notions have formed the nodes of higher levels, being connected to the main ones through hierarchical links – in this way forming a tree structure of "attribute" and "value" nodes.

The obtained hierarchical structure is nothing but a taxonomy of the problem area.

Common work of the "knowledge engineer" and experts has shown that in medical examination domain the reasoning with meta-concepts (facts) and knowledge representation as a pyramid/taxonomy completely corresponds to the experts' mentality and thinking.

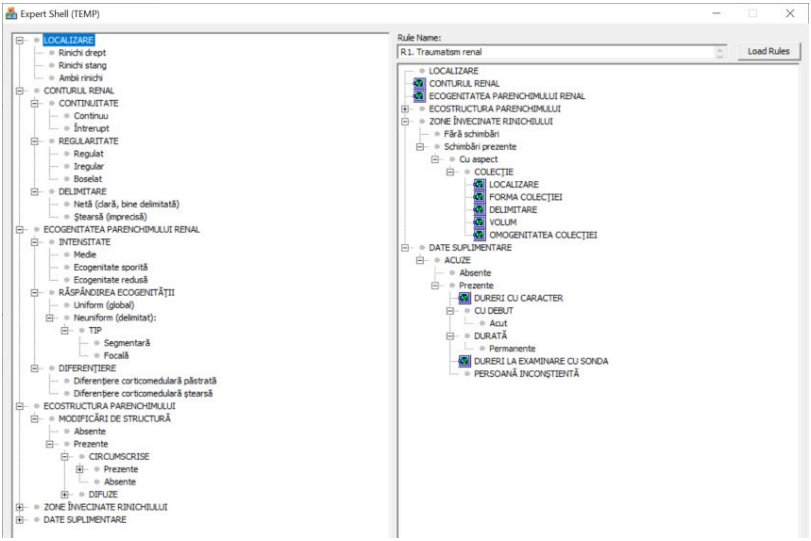


Figure 1. Domain formalization – kidneys (prototype).

The domain of "mass casualty disasters" was studied as a domain with poorly structured and heterogeneous data and knowledge. In particular, information about the injuries of abdominal organs (liver, pancreas, kidneys, and spleen) was acquired. The acquired information was structured in the form of facts (see Figure 1) – preparatory action for formalization and creation of decision rules. Based on these facts the decision rules were created. Later, the created decision rules were validated.

Information about fluid volume and thoracic air volume, based on features captured by sonographic scanners, were acquired. The acquired information was structured in the form of facts. The structure of facts was revised, in according to the Extended Focused Assessment with Sonography in Trauma (EFAST) (see Figure 2), an extended version of FAST protocol, used for sonographic examination in case of mass casualty situations [4-5].

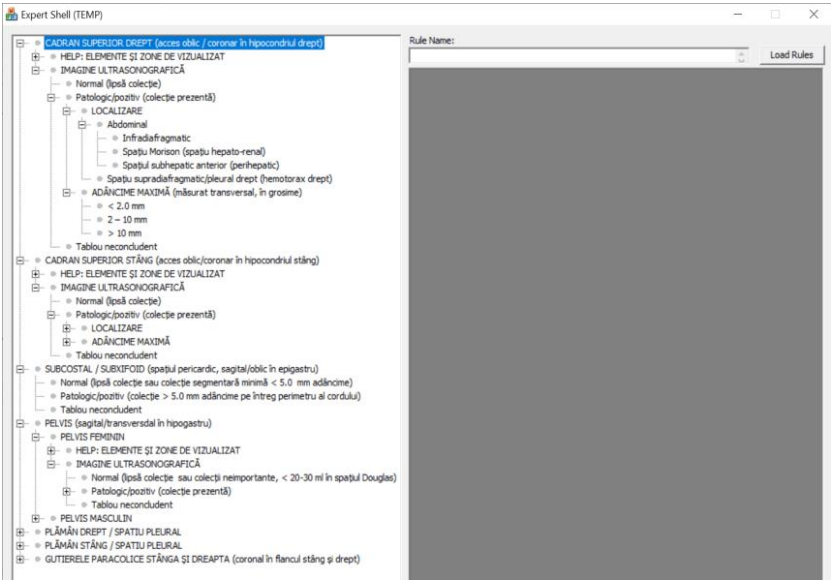


Figure 2. EFAST protocol (prototype).

Information about casualty state based on the vital signs (state of consciousness, pulse, respiratory rate, blood pressure) was acquired. The

acquired information was structured in the form of facts (see Figure 3). Based on these facts the decision rules for 4 classes of hemorrhagic shock (Class 1-4) were created. Later, the created decision rules were validated.

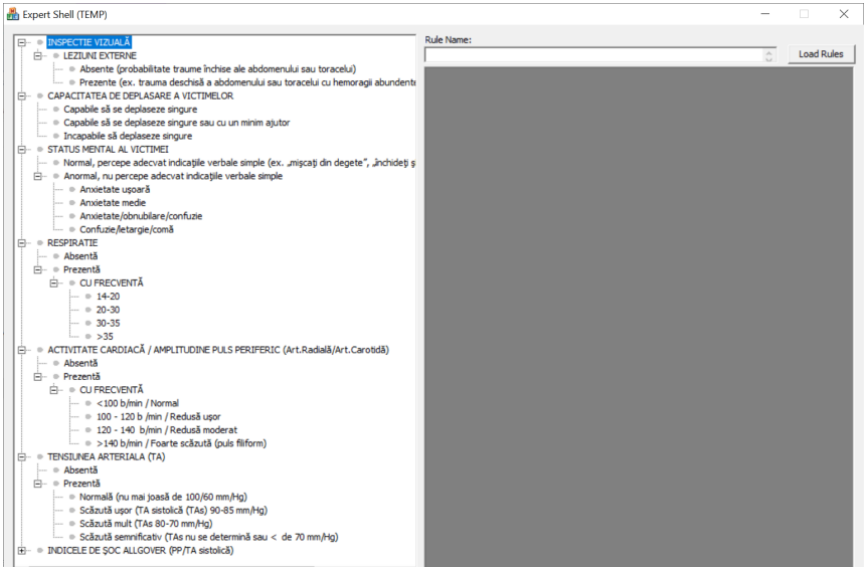


Figure 3. Casualty state based on the vital signs (prototype).

As a result, a prototype of the formalized domain was created.

Table 1. Knowledge base results

	Root nodes	Main nodes	Depth levels
Liver	5	129	10
Pancreas	6	149	10
Kidneys	6	115	11
Spleen	5	114	11
Shock	7	30	4
EFAST	7	63	6

5. Conclusion

Within the described approach, information solutions were proposed, with an increased level of intelligence, to assist the decision makers of ill-structured problems at the stage of problem definition, description and

formalization. These solutions were based on methods and algorithms in the fields of decision support systems, knowledge-based systems, advanced methods of new professional knowledge acquisition and identification, logical inferences oriented on ill-structured problems, and elements of artificial intelligence.

The proposed solution takes into account the fragmentary and heterogeneous structure of information, data and knowledge in the problem area – patients diagnostics in the process of management of mass casualty disasters. This involves studying how to integrate different data sources by using the taxonomies/ontologies associated with these data sources in order to define standardized structures, providing interoperability and consistency of stored data.

The implementation of the proposed approach allows one to follow the reasoning of the decision maker(s) from the initial stage until the solution of the concrete case/precedent. In the case of the solution confirmation and its argumentation, it can be disseminated as "good practice" or "malpractice" – a very important issue in poorly structured domains.

Acknowledgments. The research for this paper have been supported by the project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data" in the framework of the State Program and G5700 NATO Science for Peace and Security Programme.

References

[1] E. Pierce, B. Kahn, H. Melkas. *A comparison of quality issues for data, information, and knowledge.* In: M. Khosrow-Pour (Ed.), Emerging Trends and Challenges in Information Technology Management: Proceedings of the 2006 Information Resources Management Association Conference, 17th IRMA International Conference, Washington, DC, USA, May 2006, pp. 21-24.

[2] L. Burtseva, S. Cojocaru, C. Gaindric, E. Jantuan, O. Popcova, I. Secrieru, D. Sologub. *SONARES – A decision support system in ultrasound investigations.* Computer Science Journal of Moldova, vol. 15, nr. 2 (44) (2007), pp. 153-177.

[3] Iu. Secrieru, D. Sologub. *Expert shell aimed at creation of the knowledge base for ultrasonic research intelligent system.* Revista de inventica, nr. 48, vol. IX (XV-2005), pp. 7-12.

[4] American Institute of Ultrasound in Medicine, American College of Emergency Physicians. *AIUM practice guideline for the performance of the focused assessment with sonography for trauma (FAST) examination*. J Ultrasound Med, Nov. 33 (11) (2014), pp. 2047-2056.

[5] A.W. Kirkpatrick, M. Sirois, K.B. Laupland, D. Liu, K. Rowan, C.G. Ball, et al. *Hand-held thoracic sonography for detecting post-traumatic pneumothoraces: the Extended Focused Assessment with Sonography for Trauma (EFAST)*. J Trauma, 57(2) (2004), pp. 288-295.

Constantin Gaindric[1], Olga Popcova[2], Sergiu Puiu[3], Iulian Secrieru[4], Elena Gutuleac[5], Svetlana Cojocaru[6]

[1]Affiliation/Institution: Vladimir Andruhachievici Institute of Mathematics and Computer Science
E-mail: constantin.gaindric@math.md

[2]Affiliation/Institution: Vladimir Andruhachievici Institute of Mathematics and Computer Science
E-mail: oleapopcova@yahoo.com

[3]Affiliation/Institution: Medical Center "Ana Maria", State University of Medicine and Pharmacy "Nicolae Testemitanu"
E-mail: puiusv@yahoo.com

[4]Affiliation/Institution: Vladimir Andruhachievici Institute of Mathematics and Computer Science
E-mail: iulian.secrieru@math.md

[5]Affiliation/Institution: Vladimir Andruhachievici Institute of Mathematics and Computer Science
E-mail: elena.gutuleac@math.md

[6]Affiliation/Institution: Vladimir Andruhachievici Institute of Mathematics and Computer Science
E-mail: svetlana.cojocaru@math.md

DICOM image optimization for CLOUD solutions, Open science platforms and mobile systems

Alexandr Golubev

Abstract

Latest trends in software development for all sectors of enterprise solutions are porting and migration of the applications to the cloud based environments. Besides the basic advantage of cloud application, that is scalability, there are many other advantages like specific tools for monitoring and data protection. This article will describe how should be changed "DICOM Network" architecture in order to fit cloud infrastructure and there will be listed advantages and disadvantages of using cloud solutions for medical images. The last but not the least, there will be described new features that could be accessed thanks to cloud migration like secure access for mobile applications via REST API, federated identity and data share for open science.

Keywords: DICOM, Cloud, Mobile, Medical Images, Networking, Optimization, Open Science.

1. Introduction

The main scope of this article is to show the possible scenarios and present the optimal solution for migrating "DICOM Network" platform to the cloud-based infrastructure. Digital Imaging and Communications in Medicine (DICOM) is the standard for the communication and management of medical imaging information and related data [1]. DICOM is the most used for storing and transmitting medical images enabling the integration of medical imaging devices such as scanners, servers, workstations, printers, network hardware, and picture archiving and communication systems (PACS) from multiple manufacturers.

The "DICOM Network" project was launched in Moldova in 2012, whose goal is to provide access to the collected data for medical staff with the appropriate access rights and as well as patients to the personal radiography investigations [2]. As the result of the project, eleven medical equipment were connected to the system [3]. Today the system collects and processes more than 1TB of data per month.

In the second chapter of this article there will be presented an overview of existing cloud solutions and medical image processing based on those cloud solutions. The third chapter will show how the architecture should be optimized to make possible installation on a cloud infrastructure and usage of cloud tools. In the last chapter of the article we describe the methods and solutions available for data optimization based on cloud technologies.

2. Overview for the existing solutions in national and international market

First of all, it is important to highlight that criteria for selecting a cloud infrastructure is based on the following features: scalability; high performance; data security; shared data storage (some institution do not want to store personal data outside their network due to the security reasons, so in this case cloud solution should offer possibility only for data processing); data access and low storage costs, because, as it was mentioned above, one hospital can generate over 1TB of data every month.

In order to select optimal solution for "DICOM Network" cloud infrastructure, we made a study that allowed identifying a list of commercial, national and scientific cloud vendors. There were analyzed about 20 different solutions but at last there were selected 4 of them for further consideration:

- Commercial:
    - Microsoft Azure – a cloud computing service created by Microsoft for building, testing, deploying, and managing applications and services through Microsoft-managed data centers;
    - Amazon Web Services (AWS) – a subsidiary of Amazon providing on-demand cloud computing platforms and

APIs to individuals, companies, and governments, on a metered pay-as-you-go basis.

Both solutions have good tools and scalability options to cover the requirements. In addition, the storage price is low that makes it possible using this infrastructure for medical images repository. The only small advantage of Azure is additional tools for software testing and development.

- Governmental
    - "M-Cloud" platform capitalizes on government spending and consolidates data centers in a joint management form. Thus, the costs are significantly reduced, the officials' work becomes more efficient and ultimately generates quality public services.

Main advantage is trusted data storage for the personal data that is provided by Government. On the other hand, there are less functionalities and services comparing to commercial analogs.

- Scientific
    - RENAM Scientific Cloud Infrastructure – a platform built for the National Research and Educational community. Platform provides resources that could be used for research and educational scope.

One of the cloud solutions that was not described above, but is important to overview, is "DICOM stores" from Google. This is built in store with API for a DICOM image storage. Unfortunately, this feature does not have specific features for Open Science and data impersonalization, so it could be used without enhancements for "DICOM Network" system [4].

There are existing DICOM solutions on the market that have partially implemented features, but all of them have no all the required functionality, the most important one is shared data, and usually have high price for data storage that will not allow one to support these solutions in our region. For example, "Cloud PACS" by postDICOM have 800$ fee for month for 10TB, that is not enough even for keeping one-year archive [5].

Based on the research above the optimal solution for DICOM Network could be Microsoft Azure and Custom national clouds that have

all the required functionalities, tools to cover the community requirements and have the necessary data securing level.

3. Architecture solutions for "DICOM Network" Cloud

"DICOM Network" informational system was developed for medical treatment and diagnostics institutions for collection, processing and visualization of medical images.

This system covers all necessary workflows for processing and documentation of medical investigations – from collecting images directly from equipment to archiving investigation in the patient medical record. "DICOM Network" offers extended functionality for enhancing quality of medical management and secured access to investigations. This helps doctors, specialists and patients to gain access to structured database of medical images, allows documenting images that are collecting from various medical apparatus. At institutional level, the system helps to reduce costs of investigation, rise the quality of provided services.

This system is installed on 6 servers and connects both public and private hospitals (see Figure 1). International partners express the interest to join the network and use the services and data.

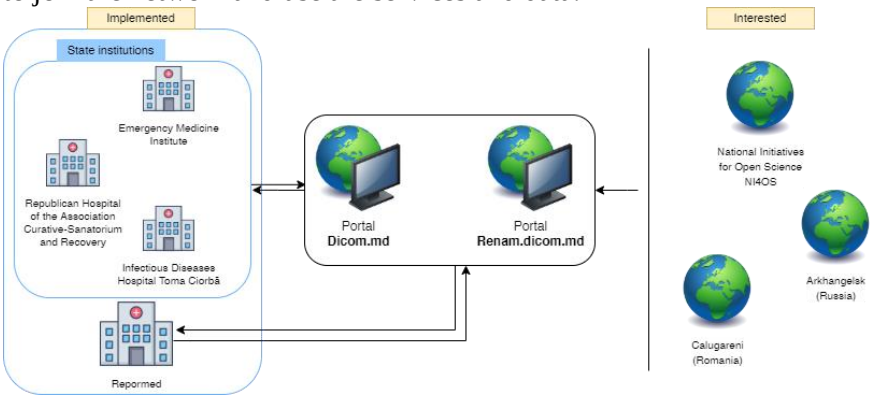


Figure 1. Connections

The existing "DICOM Network" is represented by four layers of applications that are shown in the diagram below (see Figure 2).

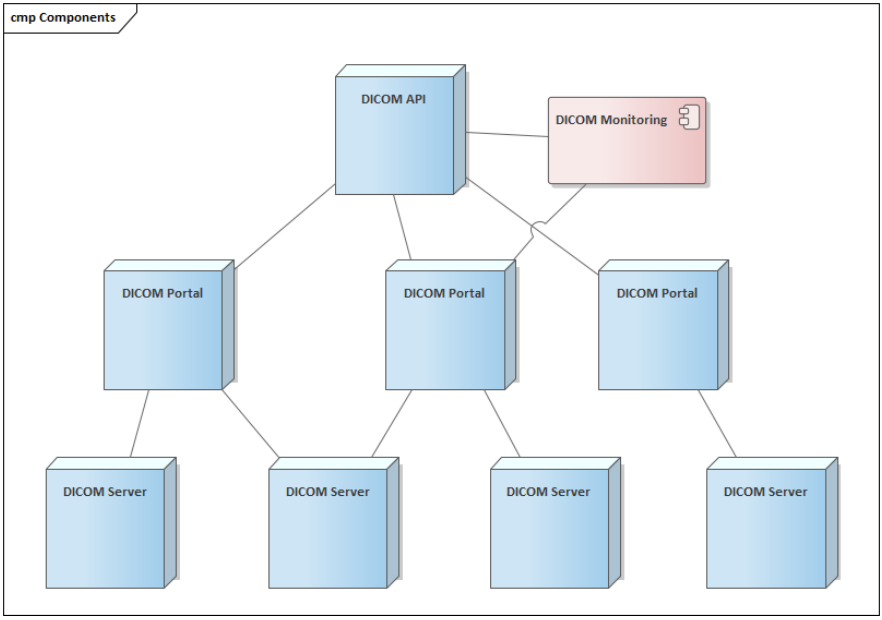


Figure 2. Layers of applications

DICOM Viewer – a set of client applications that are responsible for medical images visualization. Now the full desktop version and light web version are implemented. In future releases it is planned to add mobile version for IPhone and Android.

DICOM Server – This layer could be considered as "Advanced PACS" and is responsible for data import from various types of equipment. This layer is also responsible for raw data storage and access. Important is that this layer stores personal patient images, that in most cases should have restricted access.

DICOM Portal – Portal layer represents a set of modules responsible for medical data collection and integration with different HIS and LIS. Each customer (hospital) can install custom portal where there will be accessible only specific DICOM investigations or use Generic portal available at http://www.dicom.md/.

DICOM API – is a specialized API and list of tools that is responsible for data interoperability between different DICOM portals.

All three “server based” layers could be migrated on cloud. The main problem is that storing personal data to the external systems is usually prohibited that creates some limitations for using cloud infrastructure. Another issue that we should consider is backward compatibility with the “non-cloud” realizations. That makes possible to move to cloud “DICOM API” and “DICOM PORTAL” layers without any personalization of the data issues; for “DICOM Server” there could be used the following approaches:

- Some data could be anonymized and encrypted before sending to the cloud storage. That will make impossible to extract personal data from DICOM investigations;
- Another option is using the Cloud infrastructure exclusively for scientific scope. This will require functionalities for requesting access rights direct from patient and medical institution.

4. Data Optimization for “Open Science” and mobile devices based on cloud technologies

“DICOM Network” realization based on cloud platform opens many possibilities for using this system for various types of activities. DICOM investigations could be added to some other datasets that are collecting and available from cloud infrastructure. However, additional functionalities require realization of supplementary solutions for Data anonymizing [6].

The diagram below shows different data processing options that make it possible to store collected data at the cloud storage (see Figure 3).

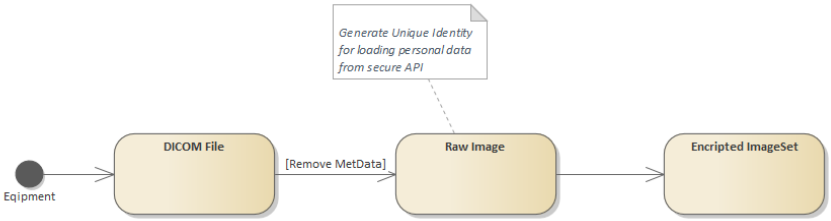


Figure 3. Data processing options

Initially all the data is collected in DICOM (.dcm) format that contains raw image and xml data with personal information. The proposed optimisations could be devided in 2 steps:

First is removing the XML with peronal data from the image file. Extracted data will be sent to DICOM Portal as metadata and stored under the Unique identifier for this imageset investigation. The impersonalized this way data could be stored within the local network.

Second is encripting the full imageset that will compeletely eliminate the possibilitry of restoring the personal data based on the image. This option will make possible to store the encrypted images using storage in external cloud.

This will require additional image processing for data storage, data accessing and data visualization, but it makes all personal data protected. At the same time this approach makes possible to use various cloud based API and services for data exchange.

5. Conclusion

Based on research of different cloud solutions, the conclusion is that migration of the DICOM Network to cloud is important and necessary for future development and data access optimization. The main problem is personal data storage in cloud that could be solved by images additional processing and archiving.

Each year more and more medical institutions require additional resources for data archiving and new effective mechanisms for collected medical images processing. That makes reasonable for more hospitals to become connected to DICOM Network. The main problem is that Moldavian hospitals in most cases do not have reliable server infrastructure and IT specialists to maintain this system inside the institution. That creates necessity to allow those institutions to join "DICOM Network" that is hosted outside internal network and ensure that personal data couldn't be shared outside. All this makes important to install Cloud based "DICOM Network" that can be used for integrating of new customers. On the other hand, installing "DICOM Network" in cloud can make it more accessible for scientific communities for research and educational purposes.

In addition, it is reasonable to mention that COVID 19 pandemic has revealed several problems in the global medical system. "DICOM Network" solutions can help patients, medical specialists and institutions in the context of offering special facilities in the period of pandemic.

During this year 3 new hospitals were connected to “DICOM Network” that allowed them to have significant savings for XRAY investigations for COVID-19 patients.

All the described above shows that setting up “DICOM Network” cloud will provide real support development of the modern medicine, will offer new useful tools for medical intuitions and research community.

References

[1] *DICOM format description* - http://dicom.nema.org/standard.html

[2] Alexandr Golubev, Peter Bogatencov, Grigore Secrieru, and Nicolai Iliuha. *DICOM Network - Solution for Medical Imagistic Investigations Exchange*. International Workshop on Intelligent Information Systems. Proceedings IIS. 13-14 September, Chisinau, IMI ASM, 2011, pp. 179-182. ISBN 978-9975-4237-0-0.

[3] Peter Bogatencov, Nicolai Iliuha, Grigore Secrieru, and Alexandr Golubev. *DICOM Network for Medical Imagistic Investigations Storage, Access and Processing*. “Networking in Education and Research”, Proceedings of the 11th RoEduNet IEEE International Conference, Sinaia, Romania, January 17-19, 2013, pp. 38-42. ISSN-L 2068-1038.

[4] *VI-SEEM - Virtual Research Environment (VRE) in Southeast Europe and the Eastern Mediterranean* - https://vi-seem.eu.

[5] S. N. Becker and R.L. Arenson. *Cost and benefits of Picture Achieving and Communication System*. Journal of the American Medical Informatics Association, Volume 1, pp. 361-371, Number 5, Sep / Oct 1994

[6] A. Anagnostaki, S. Pavlopoulos, E. Kyriakou, and D. Koutsouris, *A Novel Codification Scheme Based on the VITAL and DICOM Standards for Telemedicine Applications*, IEEE Transactions on Biomedical Engineering, vol. 49, no. 12, pp. 1399–1411, 2002.

Alexandr Golubev

Affiliation/Institution: RENAM Association, Chisinau, Republic of Moldova
E-mail: galex@galex.md

Incident Handling and Personal Data Protection in Medical Images systems

Alexandr Golubev, Petru Bogatencov, Grigore Secrieru, Ecaterina Matenco

Abstract

Modern e-Health systems require innovative solutions for data protection and overall information system security on the one hand and mechanisms for data share, taking into account personal data protection laws and regulations on the other hand. This article will show specific problems for data protection for Medical images collected in DICOM format and propose solutions for security incident reporting and tracking. Main issues that should be taken into account and addressed is the fact that not only metadata could be treated as personal data, but so is for any data that can be used to identify a person, because the latest software solutions could restore human face based on set of X-ray films that can be used to identify a person. Those tasks for data protection can be solved by using innovative ticketing and monitoring systems that are described in this article.

Keywords: Information Security, Ticketing system, Medical Images, DICOM, Personal Data, CERT.

1. Introduction

E-health systems working with medical images play important role for modern hospital information systems (HIS) all over the world. In Moldova, there exist three popular HIS realizations and many custom small medical informational systems installed both in private and public hospitals and diagnostics centers. Most of those systems have integration

with various types of medical images collections that are connected to the patient medical record.

Personal patient data is the most sensitive and important information that should be secured when it is stored in these systems. All patient personal data are protected by NCPDP (The National Center for Personal Data Protection of the Republic of Moldova) regulation that is based on national legislation. That means that patient data could not be shared or transferred without special agreement with the person or other owners of his data. That's why patient should sign a special agreement when he is visiting hospital, where he/she agrees that the hospital will use the data inside the institution to treat the patient, but will not share any patient information outside the institution.

That creates patient's data security problems in most cases when any medical information is used for external consultations, when analysis or medical images should be transferred to another medical institution/doctor or even when the patient wants to take medical images away. That obviously creates preconditions for necessity to open public hospital portal where patient should confirm that he allows sharing his information online without issuing any additional paper-based agreements.

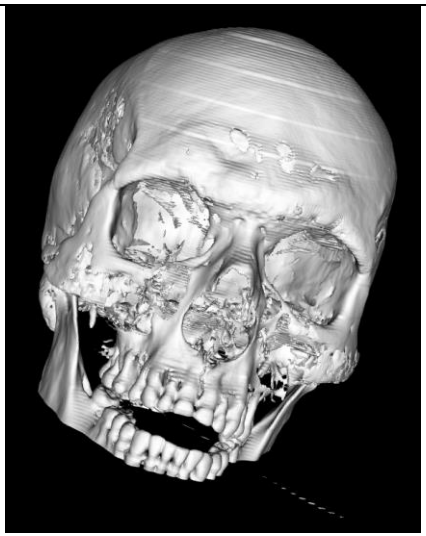
This article is targeted mainly on medical images security, so the attention will be focused on technical aspects of medical image sets processing for secure transfer and procedures required in case if any data security incident occurs.

2. Specific issues for DICOM images security and data protection

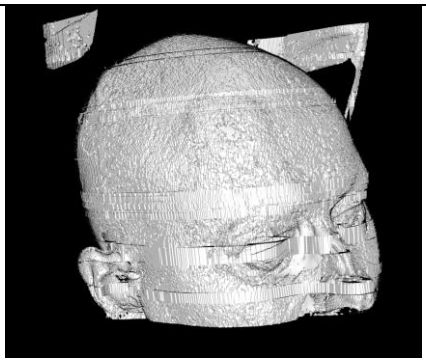
The common standard for medical images storing and transferring is DICOM format [1]. This format includes metatags that are built in the file using XML format and image itself in raw high-resolution format.

Based on conclusions made by authors of this research, it is highly recommended to remove metadata information from medical images after image is processed and stored in HIS or in Laboratory information system (LIS), as far as this data is not important for medical purpose data processing, but contains personal patient information that is not possible to protect in case if file will be stolen or lost.

By removing the metadata, it is possible to reduce available volume of personal data, but unfortunately, it will not exclude possibility of patient identification. In the Figure 1 below it is shown the "Bones" layer that is built based on simple data transform and more sophisticated "Soft Tissues" that could be used for face reconstruction. That is possible because each raw data pixel contains information about the "material" that it represents in the human's body:



Bones



Soft Tissues

Figure 1. "Bones" layer simple data transformation for "Soft Tissues" restoring

The other specific issue is that data share and transfer is possible, allowed and in many cases even required for modern e-health system, so it is the real scenario when a patient will require his data to continue the treatment in another medical institution [2]. That requires realization of a mechanism and specified procedures for hospital and patient that will allow request data transfer and trusted approval mechanism that will ensure correct identification of the requester and verified approval procedure. One of the best mechanisms could be digital signature based

on personal digital certificate like “MSign”, that is linked to the person and have confirmed juridical value.

The last but not the least is data anonymization that will make possible share data for research and educational purposes. As it was described in the beginning of this section, any image can be considered as personal data, so patient should approve using his data before it could be widely used by researchers and for educational purposes [3]. This will also require creation of specific approval mechanism that will use trusted and protected tools and procedures [4].

3. Incident Handling in modern Ticketing systems

The issues described in the second section could and should be solved by specific algorithms and tools built in Medical Information System (MIS), but also should be monitored and analyzed both automatically by monitoring tools and by security officers. This will make necessary to collect and handle all security incidents. By term “incident” we mean any issue or request that could not be solved automatically. This could be a request to close access, or alert from monitoring tool for unauthorized attempt to access the data. As far as HIS working with sensitive personal data, then all data manipulations should be logged to allow handle security incidents. This specific feature of the system will require installing and configuring one of the modern ticketing systems that should be customized to fit the HIS security requirements.

There exist many ticketing systems on the market both commercial or public, that could be customized for this issue [5]. It is reasonable to highlight three suitable types of ticketing systems:

1. Commercial: like “Jira Service Desk” or “Service Now” – the powerful solutions that have multiple integration options. The main disadvantage is that those are not cost effective.
2. Public/Free: the most popular example is Request Tracker (RT), that is free distributed, also has good documentation and customization features.
3. Custom: that could be developed especially for a medical information system. The main advantage of this option is that it could be easily integrated directly with patient medical record.

For all three options listed above incident handling procedure is implemented by the following steps:

Step 1: Receiving incident reports. Incident reports reach via several channels, mostly by e-mail, but also by telephone or on-line messages. Notes are made for all available details in a fixed format while receiving the incident report.

Step 2: Incident evaluation. The authenticity and relevance of the reported incident is verified and the incident is classified (by category, criticality and sensitivity). Triage is on the critical path for understanding what is being reported throughout the organization.

Step 3: Actions. Usually triaged incidents go into a request queue in an incident handling tool that is used by one or more incident handlers.

3.1 Start the incident ticket handle. Create incident ticket number, if it hasn't been created automatically.

3.2 Incident lifecycle. This circle contains the following processes: analysis, obtain contact information, provide technical assistance, coordination.

3.3 Incident handling report.

3.4 Archiving.

One incident could be linked to one or many medical images that offer the traceability for whole process.

4. Conclusion

Information security and personal data protection is a critical functionality for any medical information system. This becomes more and more demanded in the modern world because hospitals and patients require tools for being able to share the data and have access to historic medical records from any place in the world.

Medical images are specific data that need solving many specialized tasks to make sure that the data is protected. In practice, this means that each MIS, HIS and LIS (Laboratory informational system) requires development of additional modules that could on one the hand protect the data and on the other hand make it possible the data sharing based on request/approval mechanism.

All medical data manipulations should be tracked, and in case of any security issues a ticket should be generated and sent to specialists for

analysis. Based on the number of patients and estimated number of incidents it is not possible to handle all the security tasks without optimization and automatization of incidents handling processes. That will require installing a specialized helpdesk that should be integrated with the patient medical records.

The activities described above should make the entire patient treatment process more secure and eliminate possible data leaks.

Acknowledgments. This article was written as part of the research project 20.80009.5007.22 “Intelligent information systems for solving ill-structured problems, processing knowledge and big data”.

References

[1] *DICOM format description* - http://dicom.nema.org/standard.html.

[2] Alexandr Golubev, Peter Bogatencov, Grigore Secrieru, and Nicolai Iliuha. *DICOM Network - Solution for Medical Imagistic Investigations Exchange*. International Workshop on Intelligent Information Systems. Proceedings IIS. 13-14 September, Chisinau, IMI ASM, 2011, pp. 179-182. ISBN 978-9975-4237-0-0.

[3] Peter Bogatencov, Nicolai Iliuha, Grigore Secrieru, and Alexandr Golubev. *DICOM Network for Medical Imagistic Investigations Storage, Access and Processing*. “Networking in Education and Research”, Proceedings of the 11th RoEduNet IEEE International Conference, Sinaia, Romania, January 17-19, 2013, pp. 38-42. ISSN-L 2068-1038.

[4] A. Anagnostaki, S. Pavlopoulos, E. Kyriakou, and D. Koutsouris, *A Novel Codification Scheme Based on the VITAL and DICOM Standards for Telemedicine Applications*, IEEE Transactions on Biomedical Engineering, vol. 49, no. 12, pp. 1399–1411, 2002.

[5] *The Best Help Desk Software for 2020* - https://www.pcmag.com/picks/the-best-help-desk-software.

Alexandr Golubev[1], Petru Bogatencov[2], Ecaterina Matenco[1], Grigore Secrieru[2]

[1]Affiliation/Institution: RENAM Association, Chisinau, Republic of Moldova
E-mail: galex@galex.md
[2]Affiliation/Institution: Vladimir Andrunachievici Institute of Mathematics and Computer Science, Chisinau, Republic of Moldova
E-mail: bogatenc@asm.md

Proceedings of Workshop on Intelligent Information Systems
WIIS2020, December 04-05, 2020, Chisinau, Republic of Moldova

Analysis of similarity between artificially simulated time series with Dynamic Time Warping

Dmytro Krukovets

Abstract

Paper presents a suite of the model that finds similarity in dynamics between time series and groups them by this property; and an artificial data generator that builds those time series that have issues, close to the real ones. These two parts open a rich field for the further analysis of both real-life data and new algorithms that are able to find and distinguish these real-life issues for the more comprehensive analysis.

Keywords: Dynamic Time Warping, clustering, distance matrix, artificial dataset

MSC 2010: 37M10, 91B84, 62H30, 51K05.

1 Introduction

In this paper, I would like to introduce a model that helps to split time series into several groups and helps with an analysis of the underlying relationship that puts these series into similar groups. This is about a Dynamic Time Warping algorithm that is widely used to find distances between series. Distance is a measure of similarity to some extent, thus series that are akin by dynamics would have a low distance between them. This algorithm captures situations where there are some distortions between series: lags, stretching, contractions and other.

Among major use-cases for the algorithm, I'd like to mention sound recognition and sound motion analysis, where you'd compare audio by

several dynamic properties [5]. Another example is an analysis of cardiograms, where there are some "healthy" patterns and "problematic" ones. They can be compared and splited into healthy and unhealthy for further ability to recognize the latter ones fast [1]. One more example is financial pattern recognition that is highly used to analyze stock prices [4].

A common characteristic of the abovementioned use-cases is high-frequency data and a decent amount of observations. But what if these time series are rather short? The major example of such a situation is macroeconomics and corresponding forecasting. Recently, the Data Science algorithms start to infiltrate into the field and become a popular tool despite the fact that series are short [7]. This invasion was accompanied by modern ways to enlarge current macroeconomic datasets with support of web-scraping techniques, Google Trends and other. But also, another way was to adapt traditional Data Science algorithms to fit the case of shorter series without much loss of efficiency.

The paper will be focused on the development of a model that creates some artificial data with several parameters, that represent real data and then the other part of the model that divides this dataset into several groups. There will be a comparison of DTW with simpler benchmarks (Euclidean distance, correlation-based distance) to prove a better performance of the former one in terms of capturing the real dynamics.

2 Data

The research is highly dependent on the data quality because we would like to find the similarity of the series and subjectively evaluate whether results correspond to the initial hypotheses. Also, the research is not bound by some particular topic and has no requirements for the data except to be realistic. There are several options of what kind of dataset to choose in order to get the most appropriate result: 1) open dataset with short time series; 2) artificially created set of series. I have decided to follow the second approach for several reasons: 1) It allows a full control over issues that should be analyzed; 2) We are not forced to

control for endogeneity; 3) We can formulate the hypothesis with a relative ease. In the case of real data, these problems become a subject of higher amount of biases and errors, based on subjective judging of the data dynamics.

The dataset is created according to the TimeSynth project by [8]. This is a tool for simpler creating the data series that might be easily replicated in the Excel or whatever other application that allows working with the data series. In the case of the paper, I've created a set of forty series with ninety observations each.

These series have got several properties. First of all – it is a basic function. It could be a sine, a cosine, AR(1), AR(12), ARMA(1,1) or ARMA(12,1) process. Then the white noise, generated as a normally distributed variable with mean 0, is added. Noise can have standard deviation 0.1 (small), 1 (medium) and 5 (large). Also, there is an option to do not add the noise. The last property is an outlier that can take multiple forms: small outlier (1), large outlier (5) and a set of two small and one large outlier. All of them can be both positive and negative.

This routine helps to create twenty-five series. Fifteen more are created in a similar manner, but they have got no outliers and have a structural break, which means a new basic function starting from the thirty-first observation. A full set of variables is given in Appendix A.

3 Model

As for a model, the pipeline is as follows: to find distances between time series, to build a distance matrix and represent series as a point on the two-dimensional plane with corresponding distances between them, to cluster these series into groups with the corresponding algorithm.

3.1 Distance algorithm

3.1.1 Euclidean distance

As a first and the easiest algorithm, we'll use simple Euclidean distance between series. The algorithm finds distances as follows. If we have

time-series S_1 and S_2, $S_1(t) = p_t$ and $S_2(t) = q_t$, then:

$$dist(S_1, S_2) = \frac{\sqrt{\sum_{t=1}^{n}(p_t - q_t)^2}}{n}. \quad (1)$$

In other words, it's an average of squared deviations of one series from another. Obviously, this approach does not count for any distortions such as lagged series (those, who has similar functional form, but one of them with a lag). The approach is not of great use for real-life time series analysis as long as it cannot count popular properties of series, but it is a good benchmark to compare with.

3.1.2 Correlation

The next step is a correlation-based measure that allows representing high correlation as a short distance and vice versa. A distance between two series is given as a

$$dist(S_1, S_2) = 1 - |corr(S_1, S_2)|. \quad (2)$$

The design is given in such a form because a high absolute correlation (whenever it is positive or negative) means a high level of similarity [3]. There are a plethora of other forms for correlation-based distance, but we will stop on the simplest one as a great benchmark and also because an investigation of these sub-methods is out of the scope for the current paper.

3.1.3 Dynamic Time Warping

Then we'll go with a simple Dynamic Time Warping (DTW) algorithm [2]. In the original form, the DTW builds a matrix with distances from each point of one series to each of another. The left bottom cell is a distance between the first points, while the right upper is between last points. Then, the path between these cells is built in a way to minimize the sum of distances. It helps to produce no intersections between correspondences and connect all points between each other. Such a design gives an opportunity to deal with lagged reaction (correspondence

might be with a time shift), stretching (one point might correspond to several) and contraction (inverse to the previous one).

The most famous expansion of the algorithm, FastDTW [10], is called to speed it up. The original algorithm complexity is $o(n^2)$, thus the time to perform it grows quadratically with increased number of observations, because there is a necessity to build an n by n matrix of distances. FastDTW shows that there is no necessity to calculate the whole matrix, but only a part of it. This comes from the fact that path lies in the central area mostly. Thus, few "masks" are used to "shadow" the area that has relatively low chances to contain a part of the path, thus should not be calculated at all. FastDTW utilizes few more additions, however abovementioned one is the most important for the further work because it gives the idea with limitations of the distance matrix. Moreover, despite the overwhelming usage of the FastDTW approach, it is not necessary in our case because this work is concentrated on a relatively small series that does not hold even a thousand observations.

3.2 Distance Matrix

After finding distances with all abovementioned methods, we can put them into corresponding matrices where a cell in a row p and column q means a distance between series p and q. This matrix is symmetric because the distance between p and q is equal to the distance between q and p. A theorem [6] suggests that we can build a unique (up to the rotation) two-dimensional set of points, distances between which are equal to those in the distance matrix. It is important in order not to obtain several different groupings for a set.

3.3 Clustering

During the previous stage, we obtained a two-dimensional plane with a set of points and now we can use a clustering technique to group similar series into a single cluster. An algorithm of choice was a simple K-Means clustering [9]. The reason is the simplicity and interpretability

that helps with hypothesis checking, corresponding to the basic visual analysis results etc.

4 Results

As for the result, the main one is a plane of points, that correspond to series, grouped via K-Means algorithm. We'll do this for all cases and analyze corresponding results.

Let us start with the simple Euclidean distance case. In Figure 1, we can see two main clusters (black and blue), so as several pairs of relatively close series like in the ltblue and green clusters. The black cluster consists mainly of AR-type series with different noises, while blue is a cluster of Sine/Cosine. Green cluster is AR(12) with large noise and outlier, while ltblue is a combination of cosine and AR(12) with large noise. The major plus is that algorithm is able to find the basic difference between series (divides on AR-type and sine-type functions), despite its straightforward nature. On the other hand, it gives no better insight into the data, which is unfavourable for the future applying.

The next will be a sub-exercise with standardized data rather than the raw one. As we can see in Figure 2, this case has less strongly marked clusters. The strongest one is blue, which contains ARMA(1,1) processes with different noises. The ltblue cluster seems to be not bad too, but it contains sine-based functions (even without cosine). Other clusters are not that tight and do not give much of the additional information. This exercise repeats the previous conclusion that there are relatively no conclusions, it seems like it works worse than the previous one, but on the other hand, standardized data removes "levels" of series that plays a role as long different series have a different constant term.

Figure 3 shows a correlation-based distance case, which is already more interesting as long as it has several pretty nice groups. The first one is black, which consists of ARMA(1,1) models and several combinations of ARMA with AR (structural break). The red one seems

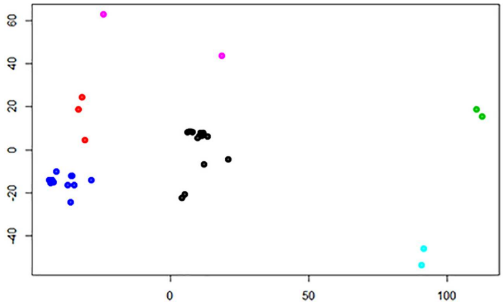


Figure 1. Euclidean distance

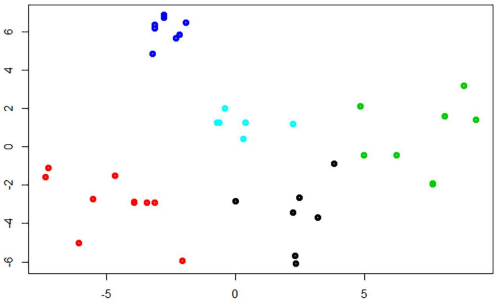


Figure 2. Scaled Euclidean distance

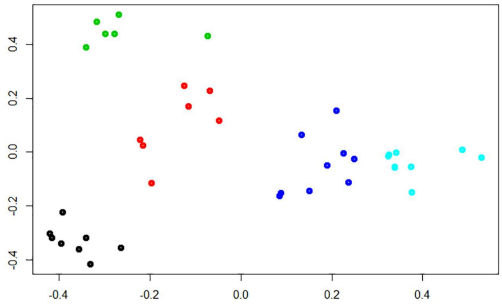


Figure 3. Correlation-based distance

to be slightly divided into two sub-parts and it consists of Sine/Cosine plus one ARMA(12,1) with large noise. The green cluster is a bunch of sine series with small-to-medium noises. The blue one (that seems to be united with the ltblue) is full of series with structural breaks, where one part is AR-based and the other is Sine-based. It is quite interesting as long as previous methods were unable to find this correspondence. The ltblue cluster continues the tradition of blue and consists of many broken series as long as several AR(1) with relatively little noise. The case with scaled data leads to the same result as long as standardization does not affect correlation.

Finally, we're coming to the Dynamic Time Warping algorithm, which case is depicted in Figure 4. Here we can observe several extremely tight clusters (ltblue, part of the red and part of the violet). Green and black clusters seem to be outliers. All these series are fully or partially (due to the structural break) are AR(12) with large noise. As for the red cluster, the tightest part contains sines with small noise, another are cosines or combinations of sine and cosine. It is interesting because in this case sines and cosines became closer to each other, which was not observed earlier. And it is a correct move because basically sine and cosine dynamics are the same. The ltblue cluster is basically very tight and consists of most of the AR-type series with

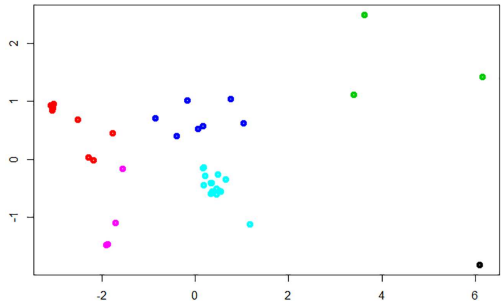


Figure 4. Dynamic Time Warping distance

no-to-medium noise. It also seems better than in the previous cases because DTW captures the AR-type dynamics despite the noise (if it's not too large and leading in the total dynamics). Moreover, even AR-type with structural break takes its place here. Blue cluster is mostly about very noisy series, so as a violet one, that contains mostly sine-based functions with or without a structural break.

5 Conclusion

The model, developed and described in the paper, supports a deep investigation of the relationship between time series, especially those that imitate structural breaks, outliers, a small number of observations etc. These particular issues are rather popular in the macroeconomics, where the exercise finds its good use as one of the possibilities to investigate economic relationship based on the data.

Results show us that the more sophisticated algorithm, the more information we can obtain out of it. For example, simple Euclidean case so as the correlation-based method could not understand that sine and cosine are basically the same dynamics, while the DTW was much more able to do this despite the noise. However, the large noise plays its role and distract the algorithm results. As for the outlier,

there was no decent sign that algorithm was able to find it and it could not make any difference, but the result regarding the structural break is more interesting. In the case of DTW, there is more tendency of those series to be with other simple series, especially when the break is homogeneous (sine to cosine or AR(1) to ARMA(12,1)). That seems very promising for a further and deeper investigation.

As for further development, there are two major ways. The first one is to focus on the development of DTW extensions and a richer set of clustering models or even more advanced standardization in order to obtain a better result and squeeze more information from the data. The second one is to move forward in the development of the actual tool for this comparison and artificial data generation process, make it even more automatic, consider other possible basic function and issues that arise in the real data and should be modelled in the artificial case properly.

References

[1] Annam J. R., Mittapalli S. S., and Bapi R. S. *Time series Clustering and Analysis of ECG heart-beats using Dynamic Time Warping*, Annual IEEE India Conference, Hyderabad, 2011, pp. 1–3, DOI: 10.1109/INDCON.2011.6139394. Available at: https://ieeexplore.ieee.org/abstract/document/6139394.

[2] Berndt D. and Clifford J. *Using Dynamic Time Warping to Find Patterns in Time Series*, 1994, AAAI Technical Report WS-94-03. Available at: https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf.

[3] Chatterjee D. *Log Book – Guide to Distance Measuring Approaches for K-Means Clustering*, 2019. Available at: https://towardsdatascience.com/log-book-guide-to-distance-measuring-approaches-for-k-means-clustering-f137807e8e21.

[4] Coelho M. *Patterns in Financial Markets: Dynamic Time Warping*, 2011. Available at: https://run.unl.pt/bitstream/10362/9539/1/Coelho_2012.pdf.

[5] Cowling M. and Sitte R. *Comparison of techniques for environmental sound recognition*, Pattern Recognition Letters, Volume 24, Issue 15, November 2003, pp. 2895–2907. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0167865503001478?via%3Dihub.

[6] Dokmanic I., Parhizkar R., Ranieri J., and Vetterli M. *Euclidean Distance Matrices. Essential Theory, Algorithms and Applications*, 2015. Available at: https://arxiv.org/pdf/1502.07541.pdf.

[7] Krukovets D. D*ata Science Opportunities at Central Banks: Overview*, Visnyk of the National Bank of Ukraine, 249, 13-24, 2020. https://doi.org/10.26531/vnbu2020.249.02.

[8] Maat J. R., Malali A., and Protopapas P. *TimeSynth: A Multipurpose Library for Synthetic Time Series in Python*, 2017. Available at: http://github.com/TimeSynth/TimeSynth.

[9] MacQueen J. *Some methods for classification and analysis of multivariate observations*, 1967. Available at: https://projecteuclid.org/euclid.bsmsp/1200512992.

[10] Salvador S. and Chan P. *FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space*, 2004. Available at: https://www.semanticscholar.org/paper/FastDTW%3A-Toward-Accurate-Dynamic-Time-Warping-in-Salvador-Chan/05a20cde15e172fc82f32774dd0cf4fe5827cad2.

Appendix A

Without structural break	**With structural break**
AR(1)_Small_noise_Small_up	ARMA(12,1)_COSINE_Medium_noise
SINE_Large_noise_SLS_up	AR(12)_AR(1)_No_noise
COSINE_Small_noise_Small_up	COSINE_AR(12)_Large_noise
AR(1)_Medium_noise_Small_down	ARMA(1,1)_SINE_Large_noise
ARMA(1,1)_Medium_noise_SLS_up	AR(1)_ARMA(1,1)_Medium_noise
SINE_Large_noise_Large_up	AR(1)_SINE_No_noise
ARMA(1,1)_Small_noise_SLS_up	COSINE_AR(12)_Large_noise
SINE_No_noise_SLS_down	AR(12)_AR(1)_No_noise
COSINE_Large_noise_Large_up	AR(1)_ARMA(1,1)_No_noise
SINE_Small_noise_SLS_down	ARMA(1,1)_SINE_No_noise
COSINE_Large_noise_Small_up	AR(12)_SINE_Large_noise
AR(12)_Large_noise_Small_down	COSINE_ARMA(1,1)_No_noise
ARMA(12,1)_Large_noise_Small_up	SINE_COSINE_Medium_noise
ARMA(1,1)_Medium_noise_No_outlier	COSINE_ARMA(1,1)_Large_noise
AR(12)_Large_noise_SLS_up	COSINE_ARMA(1,1)_Large_noise
AR(1)_Small_noise_SLS_up	
ARMA(1,1)_No_noise_SLS_down	
AR(1)_No_noise_No_outlier	
ARMA(1,1)_No_noise_Small_down	
SINE_No_noise_Small_up	
COSINE_Medium_noise_Large_up	
ARMA(12,1)_No_noise_No_outlier	
ARMA(1,1)_Medium_noise_Small_up	
ARMA(12,1)_Small_noise_Small_down	
SINE_Small_noise_Large_down	

Dmytro Krukovets
Received October 24, 2020

Institution: Taras Shevchenko National University of Kyiv
Address: Akademika Hlushkova Ave, 4D, Kyiv, Ukraine, 03680
E–mail: dkrukovets@kse.org.ua

Experimental Identification of the Mathematical Model of the DC Motor based on the Genetic Algorithm

Vladimir Melnic, Irina Cojuhari

Abstract

In this paper it is presented the experimental identification of the mathematical model of the DC motor. The experimental curve of the DC motor was raised and it was proposed to be approximated with transfer function with inertia of the second and the third order. For identification procedure it was used the genetic algorithm. As objective functions there are proposed to be used the Mean of the Squared Error, Integral of Time multiplied by Absolute Error, Integral of Absolute Magnitude of the Error, and Integral of the Squared Error.

Keywords: experimental identification, transfer function with inertia, DC motor, genetic algorithm, objective functions.

1. Introduction

The practice of the automation of the various technological processes demonstrates that a lot of processes are complex and the experimental identification is the only one solution to obtain the mathematical model that describes the control process. Experimental identification supposes that mathematical model is obtained based on the experimental data achieved from the control process as input and output signals.

There are developed many methods that can be used in the experimental identification of the mathematical model, which approximate the industrial process by the transfer functions with the second and the third order inertia.

Nowadays the artificial intelligence approaches such as evolutionary algorithms are started to be used widely in different areas for solving diverse engineering problems [1].

In this paper, it was proposed to use the genetic algorithm for estimation of the mathematical model of the DC motor. It was proposed to be done the comparation of the different fitness functions for estimation of the mathematical model by the Mean of the Squared Error, Integral of Time multiplied by Absolute Error, Integral of Absolute Magnitude of the Error, and Integral of the Squared Error.

2. Genetic Algorithm Description

Evolutionary algorithms represent the search and optimization algorithms that are inspired from the biological principles as: natural selection and genetics [1]. Genetic algorithms are a particular type of evolutionary algorithms that have strong search capabilities, being stochastic global search methods that mimic the process of natural evolution, where a population of candidate solutions to a target problem is evolving.

Genetic algorithm manipulates not just with one potential solution to a problem but a collection of potential solutions that is known as population. Every solution corresponds to a chromosome and each parameter represents a gene, where genetic algorithm evaluates the fitness of each individual in the population using a fitness (objective) function.

The genetic algorithm uses the genetic operators such as *selection*, *crossover* and *mutation* for the generation of the new chromosomes from the existing population.

The *selection mechanism* for parent chromosomes takes the fitness of the parent into account and this will ensure that the better solution will have a higher chance to procreate and donate their beneficial characteristic to their offspring.

After selection mechanism the *crossover operation* goes, that swaps certain parts of the two selected chromosomes in a bid to capture the good parts of old chromosomes and create better new ones.

The last evolutionary operator is *mutation*, in which one or multiple genes from chromosome are altered. The mutation operator maintains the diversity of population by introducing another level of randomness. In

fact, this operator prevents solutions to become similar and increase the probability of avoiding sub-optimum solutions in the algorithm.

3. The Objective Function in Identification Process

The implementation of the procedure of the identification of the mathematical model, based on the genetic algorithm that approximates the control process, starts with definition of the chromosome representation. It was proposed to use in identification procedure the two types of transfer functions:

1. Model of object with inertia of the second order:

$$H(s) = \frac{k}{(T_1 s+1)(T_2 s+1)}, \tag{1}$$

where k is the transfer coefficient of the model, T_1, T_2 – time constants.

In this case the chromosome is formed by three parameters that correspond to the transfer coefficient - k and time constants – T_1, T_2.

2. Model of object with inertia of the third order.

$$H(s) = \frac{k}{(T_1 s+1)(T_2 s+1)(T_2 s+1)}, \tag{2}$$

where k is the transfer coefficient, T_1, T_2, T_3 – time constants.

In this case the chromosome is formed by three parameters that correspond to the transfer coefficient – k and time constants – T_1, T_2, T_3.

The objective is to minimize the error between real output from the process and the output from the model.

In the process of estimation the mathematical model of the control process uses different objective functions as Mean of the Squared Error (MSE), Integral of Time multiplied by Absolute Error (ITAE), Integral of Absolute Magnitude of the Error (IAE), and Integral of the Squared Error (ISE) [2]:

$$MSE = \frac{1}{t}\int_0^t (e(t))^2\, dt, \tag{4}$$

$$ITAE = \int_0^t t|e(t)|\, dt, \tag{5}$$

$$IAE = \int_0^t |e(t)|\, dt, \tag{6}$$

$$ISE = \int_0^t e(t)^2\, dt, \tag{7}$$

$$ITSE = \int_0^t te(t)^2\, dt, \tag{8}$$

where $e(t)$ – is the error between real output from process and output from model.

4. Experimental Data Acquisition

As DC motor it was chosen the FK130SH motor for testing several reaction wheels. The system was implemented based on the NUCLEO-F303K8 platform from ST Microelectronics and the reaction wheel is coupled directly to the motor. The speed of the motor coupled with the reaction wheel is controlled by the STM32F303K8 microcontroller.

The experimental identification involves the acquisition of data, so that the experimental variation of the DC motor speeds at the reference speed of 7330 rpm was obtained as presented in the Figure 1.

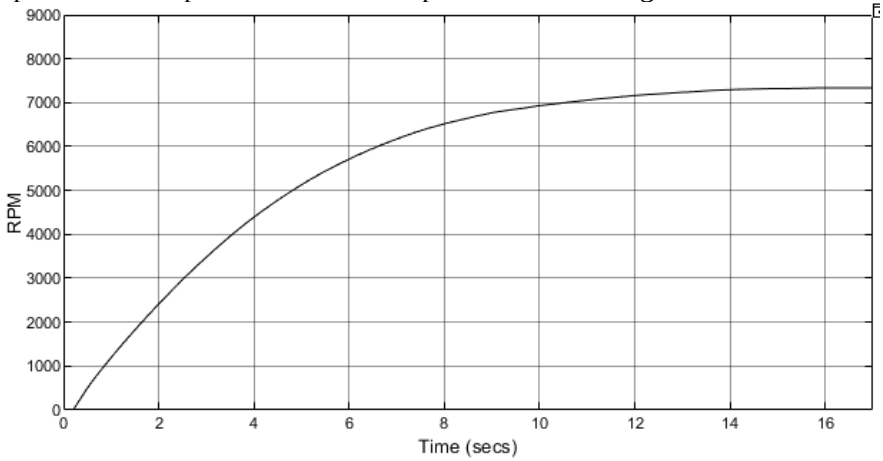


Figure 1. Experimental curve

5. Experimental Identification based on Genetic Algorithm

Based on the experimental curve presented in Figure 1, it was performed the experimental identification using the genetic algorithm. It was proposed to approximate the experimental curve with model of object with inertia of the second order and the third order (transfer functions (1) – (2)). For identification and evaluation of the obtained mathematical models there were used different objective functions as MSE, ITAE, IAE and ISE. The obtained results of experimental identification according to the genetic algorithm and different objective functions are presented in Table 1 and Table 2.

Table 1. Model of Object with Inertia of the Second Order

No.	Objective functions	k	T_1	T_2	Error
1	MSE	1.042	3.749	0.682	1.7%
2	ITAE	1.078	0.209	6.056	8.3%
3	IAE	1.814	73.084	0.669	divergent
4	ISE	0.973	1.458	2.35	2.7%
5	ITSE	1.027	0.522	3.871	0.95%

Table 2. Model of Object with Inertia of the Third Order

No.	Objective functions	k	T_1	T_2	T_3	*Error*
1	MSE	1.026	0.595	3.656	0.283	2.14%
2	ITAE	1.772	0.272	86.3	0.306	divergent
3	IAE	1.879	0.035	78.444	0.043	divergent
4	ISE	1.435	0.196	0.016	32.782	divergent
5	ITSE	1.436	48.791	0.59	0.937	divergent

In Figure 2 and Figure 3 there is presented the obtained transient process in comparison with the original experimental curve. The numbering of the curves corresponds to the numbering from Table 1 and Table 2.

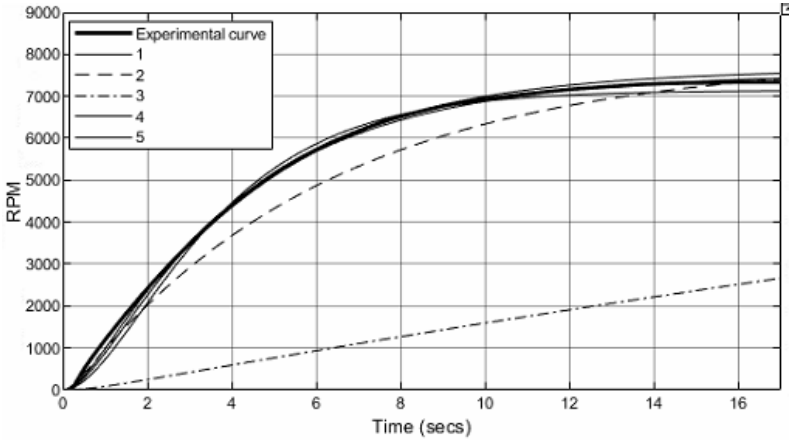


Figure 2. Transient processes in case of approximation with model of object with inertia of the second order

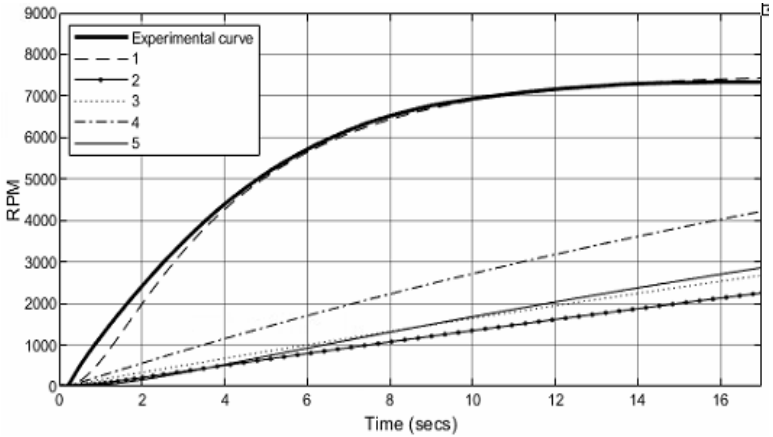


Figure 3. Transient processes in case of approximation with model of object with inertia of the third order

5. Conclusion

In this work it was proposed to be performed the experimental identification of the mathematical model of the DC motor by the genetic algorithm. It was proposed to approximate the experimental curve with model of object with inertia of the second and the third order and it was done the comparison between different objective functions. It was observed that the best results were obtained for the case of approximation of the experimental curve with model of object with inertia of the second order and the best results are obtained for the case of using the MSE and ISE objective functions.

References

[1] L.V. Stepanov *et al*. *Mathematical modeling method based on genetic algorithm and its applications.* Journal of Physiscs: Conference Series1203 (2019) 012082, pp. 1-10.

[2] A. Mirzal, S. Yoshii, and M. Furukawa. *PID Parameters Optimization by Using Genetic Algorithm.* STECS Journal, Vol. 8, 2006, pp. 34-43.

Vladimir Melnic, Irina Cojuhari
Technical University of Moldova
E-mails: vladimir.melnic@mate.utm.md, irina.cojuhari@ati.utm.md

Scheduler for the Management of Medical Treatments

Monica-Lavinia Nedelcea, Adriana Albu

Abstract

This paper proposes a solution for the management of the pills, which is composed by a desktop application and an electronic kit integrated into a pill box. The software solution is able to receive and to send information from/to the peripheral equipment, to store the information into a database, and to display all the actions performed by the user. Through the desktop interface, the user, which can be a patient or a nurse, can set his alarms at fixed hours by filling in the corresponding fields. Also, the application has the functionality to display all the alarms registered by the user, providing a history of the alarms.

Keywords: medication administration, alarms database storage, pills management, warning system.

1. Introduction

This software solution was created for the medical field, especially for the patients and for the nurses. The proposed application is part of a complex warning system [1], which has an electronic kit incorporated into a pill box. The project presents interest because it helps nurses or the persons who are concerned with the administration of the medical pills to a patient, by the fact that once the administration program is registered, it will warn the patient that the time to take his treatment has come. It has four main functions:

— registers the alarms by a user interface, which is performed by filling in two fields corresponding to the day and the hour of the alarm;
— gives the opportunity to cancel the “Save” command by clearing the completed fields;

— displays the history of the registered alarms, providing to the user the possibility to check them;
— provides a section for information about the status of the system.

The medication alert system is composed by a software part and a hardware one. The aim of this paper is to describe the software application, which consists of: i) a part dedicated to storing alarm data represented by the Oracle database and ii) a part dedicated to user interaction through an interface made in the Visual Studio 2015 development environment using the C# programming language.

The project was divided in two essential subsystems that perform the proposed functions, namely to alert visual and acoustic the patient to take the medication depending on the time of day (morning, lunch or evening), along with the need for the user to record each alarm through the user interface.

Nowadays, the medical system is more and more automated, helping the patients and the medical stuff to perform their treatment, respectively their work. Considering the healthcare field, there are mobile applications, which notify the patients or the nurses that it is the time for the medication. The patents described in [2], [3], and [4] have also an incorporated pill box with sections for storing the medication. Comparing the patents with the warning solution proposed by the current paper, it can be outlined some advantages:

— the patients or the nurses can register the medical program using an intuitive user interface;
— the communication between the application and the system is easy to be performed by the USB connection;
— the system warns the patient by indicating the pills which he needs to take;
— a sound alarm provides a reminder that it is the time for medication;
— the patient can stop the visual and acoustic alarm by pressing a button;
— the system informs the patient about the hour and the calendar date.

The system for recording and storing the data related to an alarm is using the desktop user interface and the Oracle package. When the desktop application starts, the connection to the database is automatically made. The user will choose the day and time of an alarm, through the interface made in C# and he will save them into the database stored

locally on the Oracle server from the package installed on the personal computer. If he chooses the wrong day or time, he has the opportunity to cancel the operation and to choose the desired data again. If the user does not choose one of the two data or even neither of them, but triggers the "Save" operation, a message will be displayed to warn him that he has omitted either the choice of both data or one of them.

After recording all the desired alarms, the user will choose the connection port to the alert subsystem, through which the connection will be made. He will be able to view the recorded alarms, having displayed in the interface a history of them.

2. The Architecture of the System

The described application is part of a warning system solution [1], which has a pill box with an integrated electronic kit based on an Arduino Mega 2560 board and other peripherals: Real Time Clock (RTC), Liquid Cristal Display (LCD), Light Emitting Diodes (LED), button and mini-speaker. The operating flow of the patient alert system for medication administration is shown in Figure 1.

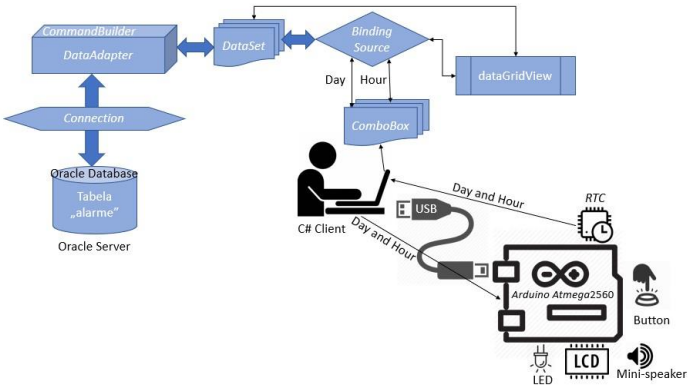


Figure 1. System operation diagram

Once all the connections are established, the process consists in filling in the fields corresponding to an alarm (by choosing the desired day and time), then saving it into the Oracle database. After recording the

desired alarms, the connection to the Arduino development board will be made, considering that the sketch has been loaded.

The RTC module will broadcast the current day and time on the Serial at the 59-th minute of each hour of the day. Based on this information, a query of the database will be performed to check if there is a record for the day and time received. If so, the information extracted from the database will be transmitted to the Arduino board via serial communication, it will be processed and compared with the information received from the clock module, checking the time interval corresponding to the time of day (morning, lunch, evening) of which the received hour is part. This will trigger the acoustic and visual functionality of the alarm, thus alerting the patient that the time has come to take his pills.

The acoustic warning will be made through the mini-speaker, and the visual one through two components: the bright, red LED for the day compartment and the screen that will display intuitive messages for the time of day corresponding to the compartment where the user has to take out the pills. If no alarm is recorded for the time received from the RTC module, the system will remain in the initial display state of the current time and calendar date.

3. The Implementation

The design of the interface has been made in an intuitive way, suggesting to the user the actions he has to perform in order to register an alarm. This interface contains: a *BindingNavigator* (for easy movement between recorded alarms), a first *groupbox* entitled "Set new alarm" (it contains two *comboboxes* for choosing the day and time of an alarm, and two *buttons*: one for saving and one for cancel), a second *groupbox* entitled "Serial connection" (it contains a *combobox*, automatically populated with the available serial ports and a connection button), a *DataGridView* (positioned on the right side to view the recorded alarms).

To detail the functionality of the desktop application, the next flow is described: the first step is the user's choice of the desired day and time for recording an alarm. The next step is to press the "Save" button in order to record the information. In case of a wrong choice of data, the user has the option to press the "Cancel" button, through which the boxes corresponding to the selected data are emptied, and thus the alarm is not

saved in the database, providing the opportunity to choose new information (time and day). After recording the desired number of alarms, the user connects to the subsystem mounted in the pill box by choosing the port and pressing the "Connect" button. Previously, a serial connection between the computer and the Arduino board was established via the permanent USB connection. This cable is also used to power the development board.

Recording data in a way that allows it to be used later in the electronic environment is done using an Oracle database server, which supports SQL instructions to access the data. Therefore, a table called "alarms", with two fields called "day" and "time" was created.

The connection to the Oracle database takes place when the desktop application is launched, because the instantiation of the connection object and the calling of the `Open()` method, specifying the creation of the connection, is done in the executed section before the form is displayed for the first time. Instantiation of the object is done using a string, received as a parameter, which contains information that identifies the database to which the connection takes place and the password which is extracted from the connection string, as it can be seen in the below code. The password for the user is specified by its Oracle id and username [5].

```
private void Form1_Load(object sender, EventArgs e){
      try{
            conexiune = new OracleConnection("DATA SOURCE =
            localhost:1521/XE; PASSWORD = student; PERSIST
            SECURITY INFO = True; USER ID = STUDENT");
            conexiune.Open();
            //...
      }
      catch (Exception ex){
            conexiune.Close();
            label3.ForeColor = Color.Red;
            label3.Text = "Eroare la conectarea cu baza de
            date." + ex.ToString();
      }
      finally{
            conexiune.Close();
      }
}
```

The try-catch section has been completed with a finally branch to close the connection regardless of whether or not an exception occurs

while connecting to the database. If so, an intuitive message will be displayed at the bottom of the user interface.

For database operation, an *OracleCommand* object and a string variable which will contain the DML command (INSERT, SELECT) [6] are declared globally. Two *OracleParameter* objects will retain the inserted data (day and time). An *OracleDataReader* object will provide a way to read a stream of rows of data from the table only for transmission.

The table will be populated with an alarm, when, through the interface, the user will choose the desired data (day and time) and will press the "Save" button. Objects `p1` and `p2` are the parameters for `OracleCommand`, used to place data in the `DataSet` [7]. These retain the data chosen via the text box that allows the user to select a value from the drop-down list. The next set of instructions consists in the actual execution of the insert command in the "alarms" table.

```
p1 = new OracleParameter();
p2 = new OracleParameter();
p1.Value = comboBox2.SelectedItem;
p2.Value = comboBox3.SelectedItem;
str_cmd = "INSERT INTO alarme(zi,ora) VALUES(:1,:2)";
comanda = new OracleCommand(str_cmd,conexiune);
comanda.Parameters.Add(p1); comanda.Parameters.Add(p2);
comanda.ExecuteNonQuery();
```

If the insert command is not executed successfully, an intuitive message will be displayed informing the user about this problem. In this case the user has to restart the system.

Once the information is registered into the database and the user chooses the port to which the Arduino board is connected (from the autocomplete list when the desktop application is running), the "Connect" button can be pressed. After performing this operation, the information related to the current day and time is received from the real time clock module in order to extract the alarm registered into the database.

The instructions for extracting the information related to an alarm are the following:

```
str_cmd = "SELECT * FROM alarme WHERE zi='" + ziua + "' AND
ora=" + oraa;
comanda = new OracleCommand(str_cmd, conexiune);
conexiune.Open();
extras = comanda.ExecuteReader(); extras.Read();
string zi = extras.GetString(0);
int ora = extras.GetInt32(1);
```

Once the database server successfully completed the previous command, the system can proceed to the instructions for transmitting the data to the Arduino Mega2560 development board.

To connect the two subsystems (the one that interacts with the physical environment and the one that records the information related to alarms), it was necessary to attach the Arduino Mega2560 development board to the computer via the USB type A to type B cable, which has a data transfer rate of up to 480 Mbps and is resistant over time to manipulation and repeated connections. The transfer of data from one subsystem to another is done by serial communication.

After pressing the "Connect" button, the connection between the Arduino and the application is made. In order to make this connection, an object of the *SerialPort* type was declared at the beginning of the program. It was used to read the received data and to transmit the data extracted from the database back to Arduino into an array of strings, in which the valid serial ports are retained. Ports are obtained from the PC system registry using the `GetPortNames()` method.

Receiving serial data is done using the `ReadLine()` method. The transmission of the extracted data from the database to the development board is done by calling the `Write()` method. It receives as a parameter the string composed of the character "#", followed by the command to be executed "START", "TEXT" or "STOP" and ending with the string "\n", which indicates to the development board the fact that the command has been fully received. The "TEXT" command is also followed by the data to be transmitted (day and time) extracted from the database.

The current data (day and time) is transmitted from the RTC module to the desktop application in order to process the received data so that it can be possible to query the database. The received data should be processed, because the received day is in English, and in the database the day related to an alarm is registered in Romanian. This change is made by the private string function `day_in_romanian()`. Also, the received time is in the format "hh:mm:ss", representing the 59th minute of the time before the one set for an alarm. For example, if the alarm is set to "16:00:00", the time received will be "15:59:00". This time will be processed by the private function `int correct_time()` to extract the desired alarm from the database.

4. Results

The solution for the management of medical treatments is a complex one, having the functionalities such as recording a new alarm and viewing the recorded ones as shown in Figure 2, besides the fact that it also has an electronic warning system, which communicates with the stored alarms. The user interface has a section, which displays the state of the system in accordance with the commands given by the user, informing him about each action that is performed.

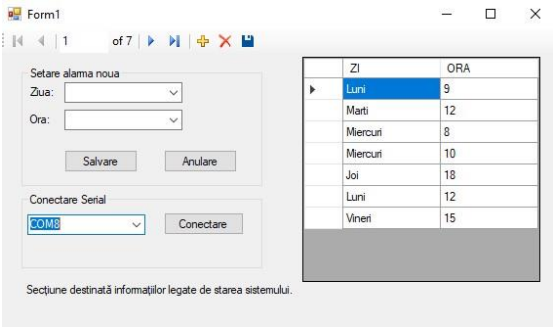


Figure 2. User interface

Figure 3. The medication warning system

The day and the hour can be chosen from *ComboBox* fields, to prevent the introduction of the erroneous data. The user interface provides the possibility to view the registered alarms and to navigate through them.

The communication between the desktop interface and the hardware system is made through a USB cable. To connect both systems, the user has a section, whence the serial port can be chosen.

The electronic kit is incorporated into a pill box, which has LEDs that indicate the day of the week, an LCD which displays an intuitive message about the period of the day (morning, lunch or afternoon) represented by Figure 3. Also, the system has the functionality to warn the patient by a sound provided by a mini-speaker.

5. Directions of Further Development

There are some functionalities of the desktop application that can be further improved, such as:

- the possibility to modify an alarm, even if it has already been saved;
- the option of deleting an alarm;
- the opportunity to add photos to each pill, in order to be more easily identified;
- the possibility to add more information related to an alarm, such as: the name of the pills and their quantity.

On the hardware part, there are also some features that can be improved, such as:

- adding a LED to each chamber, which corresponds to the period of the day, providing more accurate visual indication;
- adding a box open sensor and a weight sensor to ensure that the patient takes the pills;
- adding a new pill section for each day, if the user needs extra medication.

6. Conclusions

The solution for the management of medical treatments presented in this paper brings remarkable benefits for the healing process of the patients who take a long-time treatment, or for the persons who take care of them, such as nurses. Using it ensures that the treatment is taken at the right time, and the medicine takes effect, especially for the chronical sick persons, thus increasing the adherence of the treatment.

The system described by this paper is functional: the data acquisition from the user interface and from the peripheral equipment is accomplished, the communication between the storing software solution and the electronic board is working, and all the information from the sensors is processed. As the present paper shows, this solution for the management of medical treatments has many advantages that bring benefits for the patient, once all the scheduled pills are registered into the database.

References

[1] M. L. Nedelcea, A. Albu, *Warning Solution for Medication Administration*, The IEEE 14-th International Symposium on Applied Computational Intelligence and Informatics, May 2020.

[2] C. Lehmann, J. Brittelli, C. Hong Chen, T. Kehoe, R. Barnikel, *Home medication manager*, United States Patent US 9554969 B2, Jan. 31, 2017, last accessed on: Jan. 2020, available from: https://patents.google.com/patent/US9554969B2/en?q=~patent%2fUS20020026330A1.

[3] P. Panzini, H. Tetreault, R. Kulis, *Medication notification system*, United States Patent Application Publication US 2016/0217273 A1, Jul. 28, 2016, last accessed on: Feb. 2020, available from: https://patents.google.com/patent/US20160217273A1/en?oq=US20160217273A1.

[4] P. Panzini, H. Tetreault, *Method and apparatus for medication storage and delivery*, United States Patent US 9808403 B2, Nov. 7, 2017, last acces on: Feb. 2020, available from: https://patents.google.com/patent/US9808403B2/en?q=~patent%2fUS20020026330A1.

[5] *Introduction to the Oracle Server*, last accessed: 2019, available from: https://docs.oracle.com/cd/A57673_01/DOC/server/doc/SCN73/ch1.htm

[6] *ADO.NET*, last accessed 2019, available from: https://en.wikipedia.org/wiki/ADO.NET.

[7] *STATION - Lesson 01: Introduction to ADO.NET*, last accessed: 2019, available from: https://csharp-station.com/Tutorial/AdoDotNet/Lesson01.

Monica-Lavinia NEDELCEA[1], Adriana ALBU[2]

[1]Affiliation/Institution: Faculty of Automation and Computers, Politehnica University Timisoara, Romania
E-mail: monica.nedelcea@gmail.com

[2]Affiliation/Institution: Department of Automation and Applied Informatics, Politehnica University Timisoara, Romania
E-mail: adriana.albu@aut.upt.ro

Introducing an AI-based Response Framework for Mass Casualty Management

Marian Sorin Nistor, Van Loi Cao, Truong Son Pham,
Stefan Pickl, Constantin Gaindric, Svetlana Cojocaru

Abstract

Advances made in Artificial Intelligence over the last couple of years have revealed certain limitations to traditional Mass Casualty Management (MCM).

This paper introduces a MCM response framework using state-of-the-art OR-based models of existing AI solutions aimed to optimize each stage (extrication, triage, and transportation) of the response phase during a Mass Casualty Incident (MCI).

Keywords: Mass Causality Management, AI-based solutions, extrication, triage, and transportation.

1. Introduction

An MCI refers to a multiple-casualty situation in which Emergency Medical Services (EMS) resources, such as personnel and equipment, are overloaded. This occurs when the number of casualties is often much higher than the available resources in a particular area [1]. MCI events often result from transportation accidents, terrorism, fires, or natural disasters.

MCM is a widely used model for managing victims in MCIs. It often employs a multi-sectional approach for managing the strong connection between triage, field stabilization, and evacuation [2]. There are two key components to this approach: (1) a command post that coordinates the incorporated links between the field and health care facilities, (2) levels of special knowledge attributed to responders (e.g., policeman, firefighter, search and rescue, pre-hospital team).

2. Key Components in MCM

An optimal MCM should consider the four phases of the disaster cycle: mitigation, planning, response, and recovery [1].

In the first phase, some of the devastating effects of disasters can be reduced by acting before the actual event happens. In the second phase, realistic disaster planning and practicing are finally revised. The response consists of a series of necessary procedures from notification of an MCI, searching, rescuing, and sending casualties to hospitals. Finally, the recovery phase rebuilds and reconstructs the infrastructure while taking some actions to reduce future disasters.

This paper focuses on the response phase, more specifically on extrication, triage, and transportation. It is a sequential process from collecting and classifying victims to sending them to appropriate hospitals. Extrication is the process that prioritizes patients based on the severity of their conditions for further actions, such as immediate movement or treatment. Triage aims to provide the most efficient aid to as many casualties as possible and prioritizes treatment and transportation of victims. This requires a dynamic balance between needs (types and number of injuries) and resources (infrastructure, equipment, and competent personnel ability).

During this stage, patients are triaged and transported based on different variables, such as the number of victims, the type of incident, the available resources, and the existing infrastructure capability. Thus, triaging errors may lead to worse outcomes in later processes and an increasing number of fatalities.

The patients are typically tagged as red (critical cases - major life-threatening injuries), yellow (urgent cases, non-threatening injuries), green (non-urgent cases), and black (unlikely survival).

3. AI-based Decision Support Systems for MCM

Recently, AI is applied to build decision-support tools used in the various phases of MCM models [4-7]. The AI-based decision support tools aim to assist in resource management for disaster response with a broad range of objectives and decision variables.

One usage is, e.g., in the search and rescue phase, where Mishra et al. [8] proposed a state-of-the-art detection method based on computer vision and developed a large dataset for searching and rescuing in natural disasters using drone surveillance; Perry et al. [9] introduced a triage method based on computer vision to provide real-time casualties information at the disaster scene for the MCI commander and the EMS dispatch.

Moreover, effective decisions regarding the evacuation of mass casualty patients to hospitals should involve the assessment of the facility capacity, such as the availability of beds and the ability to deal with the patient overload. Information regarding real-time hospital bed capacity is key to controlling the flow of patients from an MCI. In this case, it allows for evidence-based decision making regarding the evacuation of patients.

Various decision support models are proposed for resource management in disaster response [3]. Several factors should be considered for integration into the triage process, such as available resources and the scale of the disaster.

There are three categories of AI-based solutions for MCM models [3]: traditional optimization-based decision support, AI-based optimization, AI-based transportation congestion detection. However, these solutions are often proposed to address issues of different components in the MCM response framework. There is no MCM response framework supported by AI-based solutions for almost all its components, including the MCI area, triage stage, and transportation stage, to the best of the authors' knowledge. Therefore, this paper introduces an MCM response framework with extended support from recent AI-based solutions.

4. Proposed Response Framework for MCM

The recent developments of AI-based support tools for MCM can be employed in one MCM model. The proposed framework aims to inherit the advantages of different AI-based solutions that have been proposed for MCM separately. The overview of the proposed framework is described in Figure 1. In the following, the AI-based solutions for the components of the response framework are discussed.

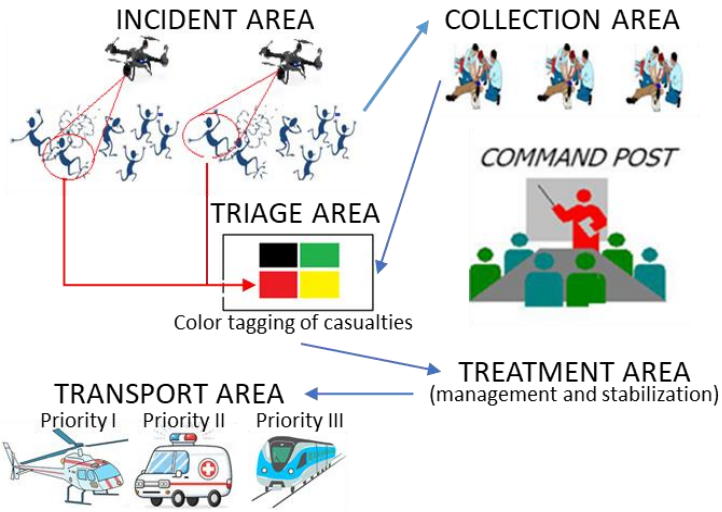


Figure 1. Response framework for an optimal MCM

(1) Search and rescue at the incident site: the task can be performed by applying the detection method introduced by Mishra et al. [8] or the triage method of Perry et al. [9]. In these methods, the AI-based detection models can analyze the video captured by drone surveillances to search for casualties. However, the AI-based triage method et al. [9] can be further used in the triage area.

(2) Search and rescue in natural disasters (i.e., floods): in this scenario, the victims are distributed in a large region. The search and rescue teams collect the casualties based on their requests. An optimization-based model proposed by Yan et al. [6] can be applied. In this model, the Support Vector Machine (SVM) and Reinforcement Learning (RL) work together to estimate the density of the regions of victim requests. Every time a victim is aided, the RL algorithm updates the density of patient requests.

(3) Triage process: the category information of each victim produced by the AI-based triage method of Perry et al. [9] for search and rescue automatically tags them into appropriate triage areas. Casualties with life-threatening injuries could ignore the collection area and be sent directly to triage for the quick delivery to hospitals. Here, expert knowledge is needed for final confirmation.

(4) Transportation: many AI-based models are proposed for real-time traffic congestion detection, as reviewed in [3]. These models typically use deep learning methods, such as Convolutional Neural Networks (CNNs), to identify vehicles and estimate their density on a given road segment. Among the reviewed models, the one proposed by Du et al. [10] is very promising. Besides traffic information, the model incorporates several factors, such as weather conditions, flying attitude of drones or unmanned aerial vehicles (UAVs), and vehicle category, to improve the detection accuracy.

5. Conclusion

AI is employed to address many phases in MCM, predominantly components in the MCM response phase. AI-based solutions, however, are often applied to address problems separately in MCM models. This paper introduces an MCM response framework in which most of its components are supported by robust AI-based models proposed in recent years.

Acknowledgments: This research was sponsored by the NATO Science for Peace and Security Programme under grant SPS MYP G5700. Proofreading of Jacob Ehrlich is gratefully acknowledged.

References

[1] Y. Kluger, F. Coccolini, F. Catena, and L. Ansaloni. *WSES Handbook of Mass Casualties Incidents Management*. Hot Topics in Acute Care Surgery and Trauma Ser. Springer, Cham, 2019.

[2] Pan American Health Organization: *Establishing a Mass Casualty Management System*. PAHO, 2001.

[3] Marian Sorin Nistor, Truong Son Pham, Stefan Pickl, Constantin Gaindric, and Svetlana Cojocaru: *A concise review of AI-based solutions for mass casualty management.* In Proceedings of the 1st International Workshop on Computational & Information Technologies for Risk-Informed Systems (CITRisk-2020), October 16-18, 2020, Kherson, Ukraine. Accepted.

[4] B. Kamali, D. Bish, and R. Glick. *Optimal service order for mass-casualty incident response*. EJOR 261, pp. 355–367, 2017.

[5] D.T. Wilson, G.I. Hawe, G. Coates, R.S. Crouch. *A multi-objective combinatorial model of casualty processing in major incident response*. EJOR, 2013.

[6] L. Yan, S. Mahmud, H. Shen, Y. Foutz, and J. Anton. *MobiRescue: Reinforcement Learning based Rescue Team Dispatching in a Flooding Disaster*. In Proc. ICDCS. IEEE, Singapore, July 8-10, 2020.

[7] Q. Meng, H. Song, Y. Zhang, X. Zhang, G. Li, and Y. Yang. *Video-Based Vehicle Counting for Expressway: A Novel Approach Based on Vehicle Detection and Correlation-Matched Tracking Using Image Data from PTZ Cameras.* Mathematical Problems in Engineering, 2020.

[8] B. Mishra, D. Garg, P. Narang, and V. Mishra. *Drone-surveillance for search and rescue in natural disaster*. Computer Communications, 2020.

[9] O. Perry, A. Bar-Hillel, and Y. Bitan. *Using Deep Learning to Provide Real Time Information During Mass-Casualty Incident*. In Proc. IPRED, 2020.

[10] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian. *The unmanned aerial vehicle benchmark: Object detection and tracking.* In Proc. ECCV, pp. 370–386, 2018.

Marian Sorin Nistor[1,a], Van Loi Cao[2], Truong Son Pham[1,b], Stefan Pickl[1,c], Constantin Gaindric[3,a], Svetlana Cojocaru[3,b]

[1]Department of Computer Science, Universität der Bundeswehr München, Germany

[a]E-mail: sorin.nistor@unibw.de

[b]E-mail: son.pham@unibw.de

[c]E-mail: stefan.pickl@unibw.de

[2]Department of Computer Science, Le Quy Don Technical University, Hanoi, Vietnam

E-mail: loi.cao@lqdtu.edu.vn

[3]Vladimir Andrunachievici Institute of Mathematics and Computer Science, Chisinau, Republic of Moldova

[a]E-mail: constantin.gaindric@math.md

[b]E-mail: svetlana.cojocaru@math.md

Proceedings of Workshop on Intelligent Information Systems
WIIS2020, December 04-05, 2020, Chisinau, Republic of Moldova

Wavelet based video scene cut detection

Yevhenii Orlov

Abstract

Video segmentation is a first step towards video analysis. Video segmentation is responsible for identifying scene boundaries, logically connected frames. Scene is a fundamental part of the video because it captures actions in a form of coherent sequence of frames. Although, many algorithms for scene cut detection were proposed, in most approaches only one visual feature was used. Here we propose a novel algorithm for scene cut detection that utilizes multiple visual features including color histogram, edge histogram and frame wavelet statistics. Experiments run on complex video data prove algorithm effectiveness compared to other scene cut detection algorithms.

Keywords: video processing, scene cut detection, color histograms, edge histograms, wavelet statistics.

1 Introduction

With rapid growth of electronic storage capacity and internet the growth of video libraries came. Modern video libraries can contain millions of hours of various video information. This created a demand for software that will help manage and will give access to video information like scenes, objects, heroes and eventually plot. Video analysis involves video scene cut detection, object detection and classification, optical character recognition and sound detection. Here in this paper we will focus on the scene cut detection topic.

The first step for video-content analysis, content-based video browsing and retrieval is the partitioning of a video sequence into scenes. A

scene is the fundamental unit of a video, it captures a continuous, logically connected, action from a camera and represents a spatially-temporal coherent sequence of frames. According to the literature, scene transitions can be classified in two types: abrupt cuts or gradual transition. Abrupt cut or hard cut, is a sudden change from one video scene to another one. The second type is gradual transition when the transition between two scenes accomplished gradually over several frames. Scene can have abrupt cuts, still form a single logically connected scene. Later, each scene is analyzed via object detection and classification algorithms; brief scene description is saved for further usage.

The basic idea behind scene cut detection consists of finding visual and audio discontinuities between neighboring scenes, measuring distance between these features and comparing them to threshold. Feature vectors are built between successive scenes, then we compare these feature vectors via similarity metrics. If this similarity is higher than a threshold, scene cut is detected. Threshold can be both global, with the same value across the frames, and adaptive, which depends on lightning conditions in different frames, the value of which depends on consecutive frames.

In the next section, we give an overview of current video scene detection methods. In Section 3, a method of segmenting video scenes based on wavelet statistics will be proposed. Section 4 will show the test results obtained by the proposed method compared to other methods. Here we have an overview of modern methods of video cut detection and present our approach to segment videos using semantic features. Finally, conclusion of the obtained results and future research issues will be discussed in the last section.

2 Previous work

Following the taxonomy presented in [1], we can group all scene cut detection methods into multiple groups on the basis of features that we extract from the frames and metrics that we use to measure the dif-

ference between extracted features. The first and the simplest methods to be introduced for scene cut detection were pixel based methods [2]. The distance between respective pixels color and hue from successive frames was measured and compared to the global threshold. Originally, measure was a simple difference between the sum of pixel differences. Drawback of this method is that difference measure accuracy depends on lightning conditions and motions. Later, other statistical measures were introduced in the literature utilizing entropy and normalized differences.

Histogram based methods were introduced by [3] and they utilized frame grey level histogram. Histogram is a picture feature composed of a number of pixels with specific grey level. Unlike pixel based methods, histogram methods utilize global features. Intra-frame difference is calculated as distance between each consecutive histogram as a sum of differences between each histogram column. [4] computed three histograms based on RGB space. Difference was calculated just like in [5], however maximum difference is selected from either R or G and B color channels. [6] utilizes cosine similarity measure to compute histogram difference. In this method 64 color bins are selected for each R, G and B color channel, three color histograms are calculated, then the final histogram is formed by concatenating R, G and B histograms. Then two successive frames are compared based on the calculated histograms using cosine similarity metric. Main drawbacks of these methods are the following: two different consecutive frames may have very similar color histogram.

Feature based methods consider wide range of features calculated from frames. Features included image contours, moments, Hough Transform points, edges, statistical features etc. Features can be selected according to a scene context. Methods presented in [7] utilize different visual features.

Hidden Markov Models were first presented for scene cut detection in [8]. Several features are calculated for each frame including audio features. Difference between successive frame's features is used to train HMM by Baum-Welch algorithm. Scene cuts are calculated by Viterbi

algorithm.

From this analysis emerges robust and accurate algorithm for scene cut detection. Most of the algorithms employ only one visual feature to describe the visual content of the frames and need a measure that will be responsible for detecting visual discontinuities. However, single features lack information about the frames, which makes scene cut detection unreliable. Frames with different images, that form different scenes, can have very similar color histograms. Similarly, frames that show objects from different points of view may have very different color histograms. Logically, we have to select different visual features to form a feature vector, and use composite different measure.

In this paper we propose an algorithm for scene cut detection that uses a combination of different visual features. Algorithms are compared to different algorithms in order to show that the proposed approach works better compared to one feature based algorithms as well as to state of the art algorithms.

3 Proposed approach

Here we will use three simple visual features: color histogram, edge detection histogram and wavelet statistics. Other features can be chosen, however we will focus on this as the most popular. Number of features is not restricted to three. However, adding new features to the feature vector will result in higher algorithm complexity.

Color histogram is composed of 64 color bins manually selected in HSV color space. Edge direction histogram is composed of 72 bins with 2.5 degree interval. Two Sobel filters are applied to obtain the gradient of the horizontal and vertical edges of the luminance frame image. These values are used to compute the gradient of each pixel and those pixels that exhibit a gradient above a predefined threshold are considered in computing the gradient angle and then the histogram.

Multiresolution wavelet analysis provides representations of image data in which both spatial and frequency information is present. In multiresolution wavelet analysis we have four bands for each level of

resolution resulting from the application of two filters, a low-pass filter and a high-pass filter. For an efficient representation a three-step Daubechies (16 coefficients) multiresolution wavelet decomposition is used. A frame is first decomposed into four sub-bands LL1, LH1, HL1, and HH1 (total of 4 sub-bands). The LL1 sub-band is substituted with its decomposition into the four sub-bands LL2, LH2, HL2, and HH2 (total of $4-1+4=7$ sub-bands). The same procedure is then applied to LL2 obtaining the four sub-bands LL3, HL3, LH3 and HH3 (total of $7-1+4=10$ sub-bands). To represent the energy distribution of the transformation coefficients, the mean and standard deviation are computed for each of the 10 sub-band obtained, resulting in a 20-valued descriptor. High order moments can also be used but we choose to limit the feature dimensionality. To compare a frame at time t with one at time $t-1$, a new difference measure is used to evaluate their color histograms, wavelet statistics and edge histograms visual descriptors. The difference between edge direction histograms is computed using the Euclidean distance as in the case of the wavelet statistics. We denote the three distances as $d_H(t,t-1), d_D(t,t-1), d_W(t,t-1)$:

$$d_H(t,t-1) = \sum_{j=1}^{64} min(H_t(j), H_{t-1}(j))$$

$$d_D(t,t-1) = \sqrt{\sum_{j=1}^{20} (W_t(j) - W_{t-1}(j))^2}$$

$$d_W(t,t-1) = \sqrt{\sum_{j=1}^{72} (D_t(j) - D_{t-1}(j))^2}$$

Where $d_H(t,t-1), d_D(t,t-1), d_W(t,t-1)$ represents color histogram, wavelet statistics and edge direction histogram feature vector computed on frame t.

Here we use an average as an aggregation function, however in the future we wouldn't limit ourselves to this.

$$d_{HDW} = (d_H + d_W + d_D)/3$$

4 Experimental results

Table 1. Test videos

Video Name	Format-W*H@FPS	Frames	mm:ss	Cuts
Dana Carvey-Full Unedited Interview With Stephen Colbert	MPEG4-720*480@25	29328	20:22	58
Full Extended Interview-Joe Biden Talks To Stephen Colbert	MPEG4-720*480@25	39048	27:07	78
Full Interview-Lady Gaga Talks To Stephen Colbert	MPEG4-720*480@25	41976	29:06	62

To evaluate the proposed cut detection algorithm, we have performed several experiments including the proposed pixel based, and Markov based algorithms. Details of the two experiments are reported in the next two sections. All tests are run on the videos listed in Table 1. Videos represent regular TV shows with different scene transitions.

Test results show that the proposed multi feature scene cut detection algorithm performs better that single feature algorithms. Compared to Markov chain algorithm, the proposed algorithm performs slightly better. However, in Full Interview- Lady Gaga Talks To Stephen Colbert test video, Markov chain algorithm performed better, due to ability to detect scenes that are logically connected but contain abrupt cuts.

Table 2. Test results

Video Name	Proposed algorithm	Color histogram	Markov chain
Dana Carvey-Full Unedited Interview With Stephen Colbert	Recall:0.93 Precision:1.00	Recall:0.69 Precision:0.79	Recall:0.91 Precision:0.93
Full Extended Interview-Joe Biden Talks To Stephen Colbert	Recall:0.95 Precision:1.00	Recall:0.71 Precision:0.68	Recall:0.92 Precision:0.99
Full Interview-Lady Gaga Talks To Stephen Colbert	Recall:0.87 Precision:0.97	Recall:0.63 Precision:0.72	Recall:0.94 Precision:1.00

5 Conclusion

In this paper we presented a novel multi feature algorithm for video scene cut detection. The algorithm is based on three visual features: color histogram, edge histogram, and wavelet statistics combined together into a feature vector. Number of features can very and include other visual descriptors that can be more suitable for specific video data. First of all, we showed that a measure which is built of multiple visual features is much more accurate than a scene cut detection algorithm that is built using only one visual descriptor.

Our algorithm was tested on real world complex video data and results were compared to several other wide spreaded algorithms. Results in Table 2 show that the proposed algorithm is far superior to Pixel-based algorithm and shows better results than the proposed Markov chain based algorithm. In the future, we plan to focus on adding audio features to the proposed algorithm.

References

[1] S. Lefèvre, J. Holler, and N. Vincent. *A Review of Real-Time Segmentation of Uncompressed Video Sequences for Content-Based Search and Retrieval*. Real-Time Imaging, vol. 9, no 1, 2003, pp. 73–98.

[2] A. Nagasaka and Y. Tanaka. *Automatic video indexing and full-video search for object appearances*. IFIP Working Conference on Visual Database Systems, 1991, pp. 113-–127.

[3] W. Ren, M. Sharma, and S. Singh. *Automated video segmentation*. Automated video segmentation. In International Conference on Information, Communication, and Signal Processing, 2001.

[4] Y. Tonomura and S. Abe. *Content oriented visual interface using video icons for visual database systems*. Journal of Visual Languages and Computing, vol. 1, no. 2, 1990, pp. 183–198.

[5] D. Pye, N.J. Hollinghurst, T.J. Mills, and K.R. Wood. *Audio-visual segmentation for content-based retrieval*. In International Conference on Spoken Language Processing, 1998.

[6] C. O'Toole, A. Smeaton, N. Murphy, and S. Marlow. *Evaluation of automatic shot boundary detection on a large video test suite.*. In Conference on Challenge of Information Retrieval, 1999.

[7] H. Yu, G. Bozdagi and S. Harrington. *Feature-based hierarchical video segmentation*. In IEEE International Conference on Image Processing, vol. 2, no. 3, 1997, pp. 498-–501.

[8] M. Ardebilian, X. Tu, and L. Chen. *Robust 3d clue-based video segmentation for video indexing*. Journal of Visual Communication and Image Representation, vol. 11, no. 1, 2000, pp. 58–79.

Yevhenii Orlov

Taras Shevchenko National University of Kyiv,
Faculty of Computer Science and Cybernetics.
E–mail: evgenyorlov1@gmail.com

The model of Web crawler for expansion the scope of initial search

Alexandr Parahonco, Mircea Petic, Corina Negara

Abstract

The paper discusses present obstacles of content formation and its maintenance while teaching process and offers an alternative modern way for this purpose. The article proposes the program model of the new application for creating dynamic content of training courses and briefly reviews the process of dynamic generation of teaching content and its integration into the Moodle learning platform. It discusses the construction of custom web-crawler and its functions within the program model.

Keywords: content generation, program model, Moodle, web-crawler, web-scraping, e-learning.

1. Introduction

One of the new educational technologies that has shown its undoubted effectiveness is electronic education, or in the original transcription - e-Learning. In developed countries, e-learning covers all levels of education and is widely used not only in universities, but also in high school and in the organization of corporate (postgraduate) education [1, pp. 1367-1368; 2, pp. 3–4]. Almost all universities and most American schools have implemented E-Learning. Increasing penetration of internet in many regions across the globe is a major factor driving the market growth [3, p. 8].

Such platforms require the elaboration of high-quality and relevant teaching resources, the constant updating of existing ones. This, in turn, is a complex process consisting of processing a variety of materials, their analysis, synthesis, creative development and processing of all elements to

build a single harmonious structure. Up to now, far too little attention has been paid to dynamic content generation for e-learning courses.

The aim of this work is to describe the developing application model for creating dynamic content of online training courses in Romanian, English and Russian languages, and its first part - the development of a search robot ("crawler") for expansion the scope of initial search.

An introduction part covers short description of the research, its structure and aims; literature review provides the problem actuality and solutions found by researches. Then we speak about our version of the problem sort out – new system of content generation for e-courses. Further we bring empirical model of our crawler along with experiment results and their analysis.

2. Literature review

The use of internet technology to deliver educational content is the latest trend in training and education development industry. A Learning Management System (or LMS) is often used to manage user learning processes. However, the majority of the current online based learning systems has two serious drawbacks: 1) non availability of ready content what leads to a dead end an instructor who begins to make up a course without the material to start up and 2) the rapid changes in the educational content, the vast amount of published papers, and the ever increasing training tutorials that necessitate the dynamic update of the existed courses in e-learning system.

The literature review led us to the solution proposed in [4, 5, 6]. These researches suggest one to use focused web crawler as a way to gather and process information from the internet. However, questions have been raised about the crawler architecture.

In [4] the researchers take as a basis for launching web crawler key words divided in two categories: included and excluded from the search query. The authors solve the first and the second problems by crawling the websites of open educational resources (OER) and mainly Massachusetts Institute of Technology (MIT) which first announced to make all of its course materials freely available. Nowadays this list is extended.

The web crawler is in charge of traversing the academic websites of open educational resources; retrieving the content and indexing it

according to the keywords. The administrator specifies both open academic web sites and keywords based on the e-course ID. The next step consists in parsing stored websites and extracting the content from the keywords related pages [4].

What is less clear is the nature of content selection approach. There is nothing said whether the downloaded e-courses are taken as a whole part or undergo some selection procedures of content. Thus, we can only assume that content is received entirely.

In [5] the authors delve into pedagogy and work out the upper mentioned problems by learning object metadata (LOM). Learning objects (LO) are learning resources named in conformance with the object-oriented paradigm that may be used, re-used or referenced during technology-supported learning. Thus, LOM is a data model, usually encoded in XML, used to describe a learning object and similar digital resources.

The scientists rely on digital libraries and learning object repositories, such as NEEDS and SMETE to search and download, at first, LOM storing in the local repository; then accessing its file with actual learning object content by reference (URL) within LOM. Such framework decreases the time for search and delivery of learning objects to learners.

The Focused Web Crawler system consists of two main applications – ID Web Crawler and LOM Downloader. ID Web Crawler takes responsibility for making a list of URLs containing LOM instances. Thereafter, the second application – LOM Downloader parses the pages, gets LOM instances and retains URL of the file. Whereby it obtains learning materials from different resources and composes one course.

Like in the previous article, this crawler cannot filter LO and gets only the necessary part of it and is compelled to use whole resource that is not the optimal solution [5].

Another article [6] throws light on educational content mined from University websites in the form of course pages. Researches claim that content can be mined from the following sources:

(a) Websites and open source course material like MIT Open Courseware, NPTEL India.
(b) Course websites of large international universities from the USA.
(c) Discussion Forums - Google Groups, Yahoo Answers

(d) Websites for animations/videos - Youtube, Google Video and metacafe
(e) Websites for general content - Wikipedia, Mathworld.

Their system tries to learn the navigation path by observing the user's clicks on as few examples searches as possible and then uses the learnt model to automatically find the desired pages using as few redundant page fetches as possible [6].

Considering the proposed crawler architecture, we decided to use similar web crawler model as it was suggested in [4, 5, 6]. The difference will consist of both expansion - keywords area and search zone. As our application should serve for multiple educational domains, neither open source courses nor digital libraries are available adequately in the Internet, on the one hand, and do not amount to considerable volume in Romanian and Russian languages, on the other. Therefore, we decided to use essential search zone – Google search engine. In order to simplify the search process for the user and expand keywords area we lean towards the idea of using dictionaries of synonyms, which may create various search requests and bring qualitative content. Consequently, we need two crawlers: one for parsing dictionaries of synonyms and another for working with Google engine. This article describes focused crawler of the first type.

The development of web crawler led us to web scraping technique. As a matter of fact, web scraping – or web crawling, were historically associated with well-known search engines like Google or Bing. These search engines crawl sites and index the web. Because these search engines built trust and brought back traffic and visibility to the sites they crawled, their bots created a favorable view towards web scraping. It is all about how you web scrape and what you do with the data you acquire. Nowadays there is no need to be an expert programmer to scrape web data. There are software solutions that render alike services: Impoprt.io, Octoparse, ScrapeSimple, ParseHub etc. Nonetheless, it should be mentioning that in case if it is compulsory to further process data (cleaning, deduplication, etc.) a web scraping software cannot really help [6, p. 2; 7, p. 2].

There are several techniques for using web scraping. Two categories can be distinguished among them: manual equipment (copy-paste), and automatic.

Manual scraping involves copying and pasting web content, which takes a lot of effort and is highly repetitive in the way it is carried out.

Automated scraping techniques shift from HTML Parsing [8, pp. 1 – 2, 9, p. 894], DOM Parsing, XPath to Google Sheets and Text Pattern Matching. Moreover, some semi-structured data query languages, such as XQuery and the HTQL, can be used to parse HTML pages and to retrieve and transform page content [10, p. 10].

All these techniques are good but we want to perform an automated scraping in most advanced and customized way using PHP server language. Hence, Google Sheets and Text Pattern Matching should be foreclosed from the list of candidates.

From the remained techniques most of all for our application fits "HTML Parsing" as we are planning to parse web pages by elements of the markup and CSS languages according to the template. Moreover, it is worth noting that even the automated web scraping process requires manual configuration to create a data extraction template [11, p. 365; 12, p. 136].

This article reports about elaboration of developing application model for creating dynamic content of online training courses in Romanian, English and Russian languages, starting from development of a crawler for expansion the scope of initial content search query.

3. Analysis of the proposed structure

Our literature review has led us to the idea of expansion search area of dynamic content generation. We started from the model of manual scraping and came finally to the program model represented in Figure 1.

As it can be seen at Phase 1 the operating principle of the developed model is to use synonymous connections to search for dictionaries that are similar in meaning with the help of a crawler, and use them at Phase 2 and 3 for advanced search using the Google search service. Thus, we obtain the behavior model of a user performing manual scraping.

According to phase 4 and 5, well merged content should be generated and further exported in Moodle. The content may be imported in Moodle

via Page and File standard plugins at Phase 6. Our application will allow to edit content (font, placement, color) and to download the files in html or pdf formats.

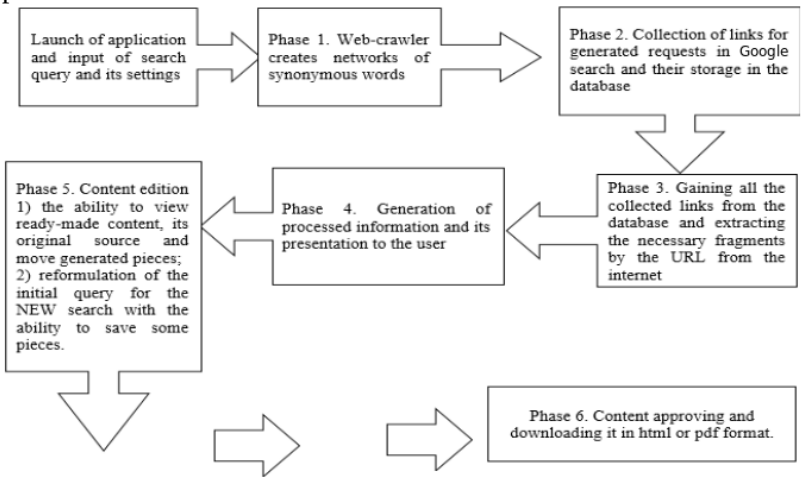


Figure 1. Scheme of the program model for the dynamic creation of training courses

4. Model of web crawler.

According to Phase 1 the first significant element in this model is the creation of a web crawler based on the server-side programming language PHP and Goutte screen scraping and web crawling library [13, p. 4]. This mechanism should create a semantically-related network of terms for a larger set of search query options.

Firstly, let's look at the main page that launches the application. As it can be seen from the picture (Figure 2) we have a "text input" for the search query, one "select input" for language type of dictionary and a start button. It is our first sample of the page and further it surely will be enriched with different options.

Secondly, we designed a table "dictionaries" that stores information about dictionaries with synonyms. When user launched the application, the dictionaries are selected by the language parameter type ("ru", "ro", "en").

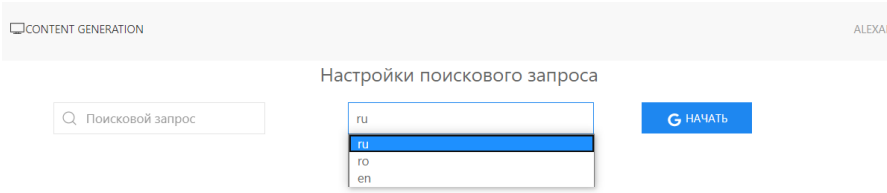


Figure 2. Main page of the application

The next step addresses crawler logic and elaboration of synonymous network. It can be divided into four phases:

1. extraction from query string prepositions and articles that do not make sense in finding synonyms (filtering);
2. reception of records from data table “dictionaries”;
3. passing through the loop of words from the filtered search string and loop of dictionaries (Figure 3);
    a. connection of crawling template;
    b. crawling the dictionary page by words iteration;
    c. extraction of synonyms;
    d. removal of non-unique words;
4. design of the data set structure.

```
foreach ($search_query_prepared as $index => $word) {

    foreach ($dictionaries as $key => $dictionary) {
        $param = ($dictionary->param_type == 'question')? '?' : '/';
        $url = $dictionary->url . $param .$dictionary->link_param .'=' . $
            word;
        $crawler = $client->request('GET', $url);
        $words = '';
        require(__DIR__. '/../Templates/' . $dictionary->template);
        if(!empty($words)){
            $extended_search_query[] = [
                'index' => $index,
                'synonyms' => implode(',',array_unique(explode(',', $words)
                    )),
                'original' => $word
            ];
        }
    }
    //check if there are synonyms
    if(strlen($words) > 0){
        //there are some synonyms
        Log::info('Found synonyms', ['synonyms' => $extended_search_query])
            ;
        Cache::put('synonyms', $extended_search_query);
    }
}
```

Figure 3. Crawler logic and design of data set

The data set structure represents an array of synonyms. Each element represents an array with index (position within search query), comma

separated synonyms and original word. This data set is further cached for future processing. We find it rational to store synonymous network in cache rather than in Data Base based on the resource economy, the speed of access and aim.

Generally, Data Base requires storing data for long time. Cache services were designed to store data for short time, although permanent storage is also available. In our case synonyms do not represent any value for retaining as their application will take long time and special algorithms. Online dictionaries already have words network for prompt search of synonyms. This is the main point to use them instead of Data Base. Furthermore, our aim is to build an application that does not consume huge resources for maintenance [14, p. 3 - 4].

As part of initial development, we have elaborated user interface for displaying synonyms and selected one dictionary – "Wordsmyth The Premier Educational Dictionary-Thesaurus" [15]. Our examination of the site and its structure persuaded us to take similar words category as a source for data set structure if synonyms category is absent. The application page provides accessible for understanding way to display original word from search query and its synonyms. It also shows the query itself. From it we can see that words as "of" and "in", being pronounced, were omitted from crawling approach, entering the list of exceptions.

Our experiments demonstrate that variations of our query expand the scope of the initial search. For example, the search for "abolition of serfdom in Russia" provides the following results (Figure 4) and Figure 5 provides the results for manual search for the word "abolition".

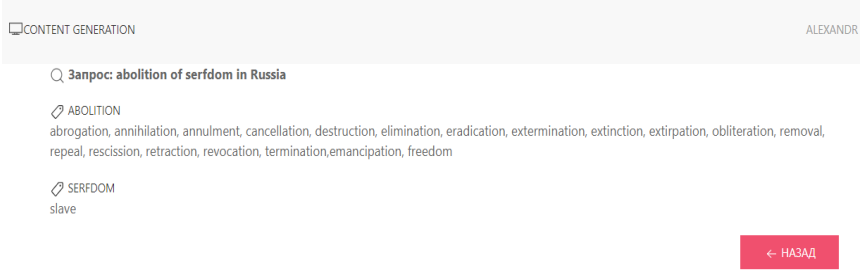


Figure 4. Web crawler and web scrapping results

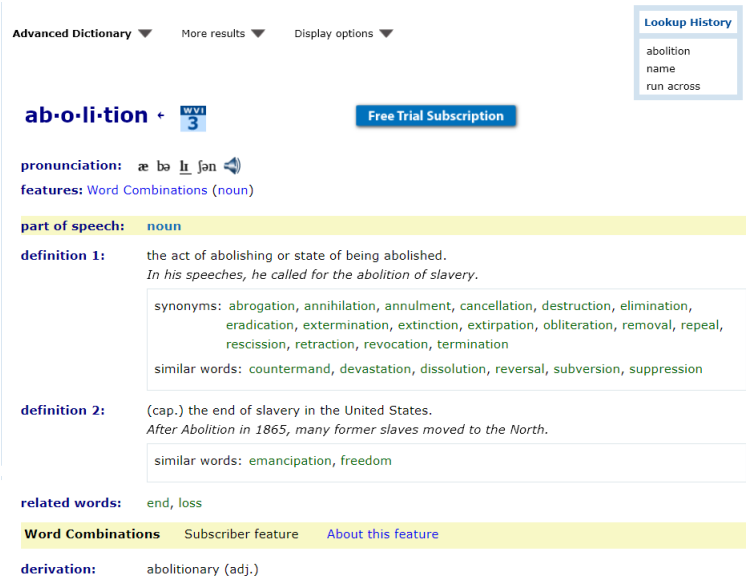


Figure 5. Wordsmyth page result for manual search for the word "abolition"

As it can be seen from the current example and others (Table 1), variations of our query are obtained that expand the scope of the initial search. However, it should be emphasized that there are also "irrelevant" combinations of inquiries (annulment, obliteration, etc.) which are non-commonly used in digital sources in the context of the search query and probably will lead to false searches. However, this effect is partly leveled at the stage of searching by Google, having at hand full query and can be reduced by preliminary selection of fragments with the desired text.

Table 1. Application examples of synonym network.

Initial query	Synonyms
design patterns in software engineering	DESIGN engineer, map out, plan, conceive, dream up, formulate, invent, earmark, intend, target, contrive, devise, blueprint, draft, layout, program, scheme, strategy, drawing, picture, pattern, art, drawing, fashion, graphic arts, graphics, painting, aim, goal, intention, objective, purport, purpose, intentions, plot

	PATTERN design, figure, motif, configuration, shape, structure, archetype, ideal, model, paradigm, prototype, standard, model, style, form, kind, rhythm, tempo, type, variety, fashion, figure
Introduction to linguistics	INTRODUCTION establishment, inauguration, initiation, insertion, institution, interposition, presentation, presentation, foreword, overture, preamble, preface, prelude, proem, prologue, innovation, overview, survey

5. Conclusions

In the course of our research we found out that today's e-learning environment suffers from the cold start problem – lack of readymade content for courses. Experts suggest building the focused web crawler for intelligent data extraction from web sources. We have adopted this idea in our application model and enhanced it by widening keywords area by composing semantically-related network of word and search zone by Google engine.

Further literature review brought us to the web-scraping technique that speaks about ways of content extraction and we selected "HTML Parsing" for our focused web crawler.

The empirical path starts from the elaborated program model of our application. It includes six phases. At Phase 1 we use synonymous connections to search for dictionaries that are similar in meaning with the help of a crawler, and use them at Phase 2 and 3 for advanced search using the Google engine. Thus, we obtain the behavior model of a user performing manual scraping. According to phases 4–6 the well merged content should be generated and further exported in Moodle via Page and File standard plugins.

As expected, we succeed at expansion the scope of the initial search query and produced semantically-related network of terms that is stored in the cache rather than in database. Insofar as cache is fast and resource-saving solution.

During the tests we observed that by crawling the dictionaries we gained many repeated words that we had to remove keeping only unique synonyms. The omitted words in the examples (Russia, software,

engineering, linguistics) do not have synonyms in the dictionary used. However, this does not matter as during the phase 2 the initial words will be included in search query at the beginning and then step by step will be substituted by their synonyms if they have.

Finally, our research is passing to the next stage – implementation of the Phase 2 and 3. We should now use our synonyms and collect links in database for their further crawling and extraction of "valuable" content.

Acknowledgments. This article was written as part of the research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data".

References

[1] D. Benta, G. Bologa, and S. Dzitac. *University Level Learning and Teaching via E-Learning Platforms*. Procedia Computer Science, no 55 (2015), pp. 1366–1373.

[2] J. Valverde-Berrocoso, M. Garrido-Arroyo, and C. Burgos-Videla. *Trends in Educational Research about e-Learning: A Systematic Literature Review (2009–2018).* Sustainability, volume 12, no 5153 (2020), pp. 1–23.

[3] L. Wood. *World Online Education Market, Forecast to 2025 by Type, Technology, Vendor, End-user and Region*. Global Online Education Market - Forecasts From 2020 To 2025, no 4 (2020), 124 p.

[4] H.W. Hijazi and J.A. Itmazi. *Smart Crawler Based e-Learning*. Hammamet-Tunisia, volume 11 (2013), pp. 209–2016.

[5] Y. Biletskiy, M. Wojcenovic, and H. Baghi. Focused Crawling for Downloading Learning Objects – An Architectural Perspective. *Interdisciplinary Journal of E-Learning and Learning Objects*, volume 5, no 1 (2009), pp. 169–180.

[6] U. Sajjanhar. *Focused Web Crawling for E-Learning Content: master's thesis. Indian Institute of Technology Kharagpur*, Master of Technology In Computer Science and Engineering, 2008. https://pdfs.semanticscholar.org/2da0/a0658ce31e2143cec9a050aebc3337cb188.pdf.

[7] A. Toth. *Is Web Scraping Legal? 6 Misunderstandings About Web Scraping*, *Scraping Authority, no 11*, 2017. https://www.import.io/post/6-misunderstandings-about-web-scraping/

[8] Bo. Zhao. *Web Scraping*. Springer International Publishing, Volume 5, (2017), 632 p.

[9] S. Ruihua. *Joint Optimization of Wrapper Generation and Template Detection*. KDD07: The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Minin, August, (2007), pp. 894–902.

[10] *Introduction to Web Scraping*, 2019 https://www.geeksforgeeks.org/introduction-to-web-scraping/

[11] Anand V. Saurkar, Kedar G. Pathare, Shweta A. Gode. *An Overview of Web Scraping Techniques And Tools*. IJFRCSCE, volume 4, no 4 (2018), pp. 363-367.

[12] S. Sirisuriya. *A Comparative Study on Web Scraping*. Proceedings of 8th International Research Conference, KDU, no 11 (2015), pp. 135 – 140.

[13] V. Bambuch. *Platform for Cryptocurrency Address Collection*. Excel 2020 Sbornik, no 6 (2020), pp. 3–10.

[14] I. Haber. *15 Reasons to use Redis as an Application Cache. White Papers, 2016.* https://redislabs.com/wp-content/uploads/2016/03/15-Reasons-Caching-is-best-with-Redis-RedisLabs-1.pdf

[15] *Wordsmyth The Premier Educational Dictionary-Thesaurus*, 2020 https://www.wordsmyth.net/

Alexandr Parahonco[1], PeticMircea[2,1], Corina Negara[1]

[1] Alecu Russo Balti State University, Republic of Moldova
E-mail: alexandr.parahonco@usarb.md, corina.negara@usarb.md

[2] Vladimir Andrunachievici Institute of Mathematics and Computer Science, Republic of Moldova
E-mail: mircea.petic@math.md

Important aspects in assessing the credibility of unstructured information

Mircea Petic, Adela Gorea, Ina Ciobanu

Abstract

In this paper we presented possibilities of credibility assessment tools. A large number of such tools have been developed for journalists. Another important subject of the paper is the actual approaches that are used in the process of assessment of credibility of unstructured information. In this sense, two points of view were presented.

Keywords: credibility, Web, social networks, fake news.

1. Introduction

With the development of online media, since the advent of the Internet, considered a democratic space, it has become much easier for anyone to express themselves freely, anytime and anyway. Beyond the media institutions, each with its own editorial policies and various frameworks that judge whether an event turns into news or not, there have been many sites that present themselves online as media products, posting various content that puts as doubt its credibility. Whenever we are about to accept or reject new information, we should ask ourselves what is the origin and reputation of the source. In the so-called "reputation era", critical appraisals should be directed not to the content of information, but rather to the network of social relations that has formed that content and offered to it a certain position, deserved or undeserved, in our knowledge system. But not every user is able to make an analysis and a distribution of content into a credible or untrusted one.

Social media networks also play an important role, because through them the content is distributed and redistributed daily by users, without being checked, often creating feelings of panic and revolt.

The issue of assessment of the credibility of unstructured information is one of the most important and is not just specific to digital media. Finding credible data is a task of great interest to those seeking information. The information is quite difficult to process due to its unstructured form. Even credibility coalition[1] was created which aims to understand the veracity, quality and credibility of online information.

The aim of this paper is to underline the possibilities of the credibility assessment tools and to detect the important approaches that are used in the process of assessment of the credibility of unstructured information.

That is why the paper is structured mainly in two parts. The first part is concerned to the tools for assessing information credibility. A large number of such tools have been developed for journalists. The second part is dedicated to the research methods for assessing data credibility. In this sense, two points of view are presented.

2. Tools for assessing the credibility of information

Several existing tools that assess the credibility of online texts and / or articles are analyzed, focusing mainly on user-generated evaluations as experts in journalism.

FactCheck.org is an information verification platform that was launched in December 2003. On this site, users can ask questions that are usually based on a rumor in politicians' statements. The site team conducts an investigation and provides a detailed explanation. The explanation includes information about who is the author of the statement, when it was released and how the team verified it. The site also has a special function for verifying scientific information.

Politifact.com is an information verification platform and one of the first fact-checking newsrooms in the US, founded in 2007. The group of reporters within this platform monitors the statements and speeches of politicians and denies false information [1].

[1] https://credibilitycoalition.org/

Snopes.com is an information verification platform developed in 1994 and aims to validate statements, articles, posts, photos on social media. This platform is not limited to simple statements (eg "true" or "false"), but uses more detailed categories ("true", "false", "partly true, partly false", "largely true"). , "Mostly false", "outdated information", "misunderstood information", etc.).

Fake Bananas is a tool developed by a group of Sworthmore College students. The tool is based on machine learning algorithms and defines credibility with 82% accuracy. The program searches for authorized online publications for articles with the context of the message, which must be verified and analyzes whether the authors of the articles agree with the formulated idea made in the statement. If trusted sources agree with it, the program evaluates the statement as true. Although the service is not publicly available, the program can be used in other projects [1].

Hoaxy is a tool developed in 2016 by a group of researchers at the Center for Complex Networks and Systems Research and the Indiana University Network Science Institute. The tool was developed to study how information is disseminated on social media. Focused on checking for fake news, the site generates interactive, color graphics so that users can see how messages are spread on Twitter.

NewsGuard is a tool dedicated to source-level evaluation, manually and methodically reviewing thousands of English-language news sources, mainly in the US. NewsGuard is available as a Chrome extension that can display this information when such news sources are open in the browser or appear in some web searches. These criteria are divided into two groups: credibility (does not repeatedly publish false content, regularly corrects or clarifies errors, collects and presents information responsibly, avoids misleading headlines, responsibly manages the difference between news and opinion) and transparency (site discloses ownership and funding, discloses who is responsible, clearly labels advertising, site provides name of content creators, along with contact information or biographical information) [2]. Websites also receive an overall score (the sum of points for each criterion) and a tag, which can be reliable, negative, satirical or platform (blogs, user-generated content or social networks). The total score of credibility and transparency is a maximum

of 100, and a news site is considered "safe" if it has accumulated at least 60 points.

My Web of Trust (WOT) is a reputable crowdsourced service that provides evaluations of websites through a browser extension. It offers two components - trust ("How much do you trust this site?") and security ("How appropriate is this children's site?") - in terms of a score and a measure of trust. Users can view general ratings and comments from the community and provide their own ratings. The user can also rate it and leave a review based on his personal impressions. WOT has two modes: real-time protection and manual mode. Real-time protection informs you about online threats [3].

If you find a site that does not yet have a reputation rating, you can ask the WOT community to rate that site. For ratings and reviews, WOT uses smart algorithms and manual verification to detect and remove fake reviews and can also use it to check blogs and social networks (Twitter, Facebook, Google+).

Although a significant number of scientific papers address many aspects of this topic, few have researched methods of measuring data credibility. There is almost no research that would propose algorithms for assessing the credibility of content that would allow the automation of solving this problem to an extent acceptable (accurate and useful) by users.

3. Research methods for assessing data credibility

According to [4], but also according to the analyzed tools from above, credibility assessment methods can be divided into three broad categories: methods based on automated approaches, methods based on human assessment and mixed methods, which in turn are divided into other subcategories. The combined method unifies main categories or different subcategories.

Speaking about automated-based approaches concerning credibility assessing we should speak about machine learning approaches, graph-based semi-supervised approaches and weighted algorithm and information retrieval algorithms.

Human based approaches consist of cognitive and perception approaches, voting approaches and manual verification approach.

Mixed methods use advantages of the post, topic or user level credibility assessments.

Moreover there is also another classification of the methods in credibility assessment of online information that is based on approaches related to social networks and linguistics approaches [5].

If we talk about the approaches related to social networks, then we take into account the notion of linked data[2] and the behavior of users on social networks. Just as users need to log in before using a social network, it gives them more confidence in the data that appears here.

In the case of computational linguistics approaches, we speak about the use of statistics on n-grams, the sentences transformation into more advanced forms of information representation (such as decision trees), to which the attached probabilities are then analyzed to identify anomalies, semantic analysis of information, relationships between linguistic elements, which help to determine the proximity to the centers of truth or deception, SVM classifiers or Bayesian Naïve classifiers are used to predict the outcome, use neural networks that identify false news.

However, the hybrid approach (combining machine learning in computational linguistics with social networking approaches) seems reasonable and promising [5].

5. Conclusion

Analyzing the web tools mentioned in the paper, we see a continuous effort to assess the credibility of information on the Internet. The use of the tools described in the paper confirms that the correct classification of the multitude of information is not a simple activity. Even if there are several applications, they cannot guarantee accurate results. Moreover, the results are modest for inhomogeneous documents that have a complex structure that contain not only texts but also images. Most of the researched applications work only for English, which proves that it is practically an unexplored field.

[2] https://www.w3.org/standards/semanticweb/data

Acknowledgments. This article was written as part of the research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data"[3].

References

[1] S. Molly *8 resources for detecting mis- and disinformation*, 2019, International jurnalists network, https://ijnet.org/en/story/8-resources-detecting-mis-and-disinformation.

[2] S. Wineburg, S. Mcgrew, J. Breakstone, and T. Ortega. *Evaluating information: The cornerstone of civic online reasoning*. In: Stanford Digital Repository. Retrieved January, 8:2018. DOI: 10.1145/3170427.3174355.

[3] F. Prochazka and W. Schweiger. *How to measure generalized trust in news media? An adaptation and test of scales.* In: Communication Methods and Measures, 13(1):26–42. DOI: 10.1080/19312458.2018.1506021.

[4] M. Alrubaian and M. Al-Qureshi. *Credibility in Online Social Networks: A Survey*. In: Digital Object Identifier 10.1109/ACCESS.2018.2886314. 2028-2055 pp. 2018. DOI:10.1109/ACCESS.2018.2886314.

[5] A. Iftene. *Exploiting Social Networks. Technological Trends*. In: Habilitation Thesis submitted at "Alexandru Ioan Cuza" University, December 2019. 163 p. ISSN 2668-1765, ISSN-L 1224-9327.

Mircea Petic[2,1], Adela Gorea[1], Ina Ciobanu[1]

[1] Alecu Russo Balti State University, Republic of Moldova
E-mail: adela.gorea@usarb.md, ina.ciobanu@usarb.md

[2] Vladimir Andrunachievici Institute of Mathematics and Computer Science, Republic of Moldova
E-mail: mircea.petic@math.md

[3] http://www.math.md/en/projects/20.80009.5007.22/

The Intelligent Robotic System for Text Typing based on the Image Processing

Victor Popa, Irina Cojuhari

Abstract

In this paper it is described the designed intelligent multifunctional robotic system for text typing on a computer. The text recognition is performed by the neural networks and then through digital image processing and photometry technologies the robotic arm performs the typing of each letter.

Keywords: machine vision, neural networks, photometry technologies, image processing, pinhole camera model.

1. Introduction

Nowadays, rapid progress of industrial technologies imposes new requirements and challenges to the different domains of human activity. In so way, the robotic systems, artificial intelligence, cloud computing and machine vision are becoming more integrated into different kinds of technological systems [5].

Due to the artificial intelligence domain, the control and implementation of different operations has become more effective, with possibility of further reusing.

As a research study, it was proposed to develop an intelligent system, based on visual detection of the keyboard and it is calculating the distance between the end point of the arm and the key corresponding to the required letter. All this is implemented by digital processing of images, that are captured by the video camera, the letters on the keyboard are being recognized and interpreted by the neural network. The distance and calculation of the movement trajectory of the robotic arm is done based on algorithms which implement photometry techniques.

2. Description of the Technologies Used in the Designed System

2.1 Machine Vision

Machine Vision is a scientific direction in the field of digital image processing technologies, this processing is performed by intelligent algorithms (neural networks), with the ultimate goal of using the processed data to solve various automation problems [1-2].

The Machine Vision areas of use:

- Large industrial production (maritime, automotive, chemical)
- Industrial security systems
- Visual control and management systems (bar reading metering)
- Automatic control of vehicles
- Product quality control
- Automotive Industry
- ADAS (Advanced driver-assistance systems).

2.2 Pinhole camera model

The **pinhole camera model** defines mathematical relationship between a 3D point and its projection onto the image plane of an ideal pinhole camera, where the camera aperture is described as a point and no lenses are used to focus light (Figure 1) [3-4].

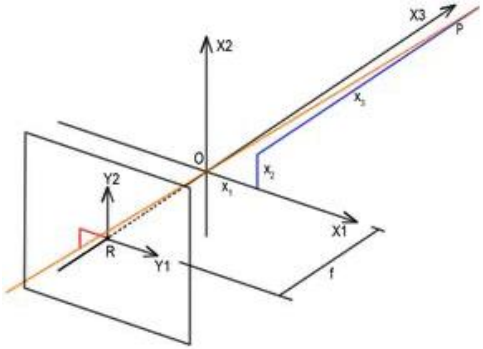


Figure 1.The geometry of pinhole camera model

2.3 Neural networks

A neural network is a network of circuit of neurons that acts based on the series of algorithms that aim to recognize underlying relationships in a set of data through a process that imitates the way the human brain operates (Figure 2) [6].

The use of artificial intelligence for the purpose of interpreting images has conditioned the popularization and reduction of the prices of computer vision systems, which has led to their widespread use in almost all branches of industry. The main cause is the high degree of reconfiguration of hardware and software itself. In turn, the possibility of self-learning of neural networks offers a wide range of possibilities depending on the appearance of each technological process of production.

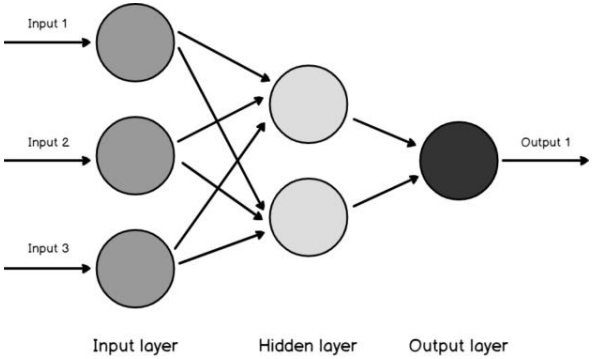


Figure 2. Example of neural network

3. Description of the Designed System

The designed system offers the possibility to automate the process of typing on the computer the text extracted from the image, where the acquisition of the text is done by a video camera, and the extraction and interpretation of the text is done with neural networks, the keyboard and letters are detected in real time by various techniques of digital image processing (Figure 3).

In order to increase the working speed of the program and to minimize the overload of the computer operating system, the interpretation of the image is performed on the web server of the "Microsoft" company. After the text is extracted from the image, the program starts the typing cycle of the text on the keyboard and the letters on the keyboard are also interpreted by the local neural network in real time. The program checks after each keystroke performed the current distance between the point arm terminal and the required letter on the keyboard, by the digitally processing of the image and converting the number of pixels into the

distance that is measured in millimetres, the algorithm generates the trajectory of the arm's movement from the current position point to the required position of the letter on the keyboard, the cycle is repeated until all letters are typed.

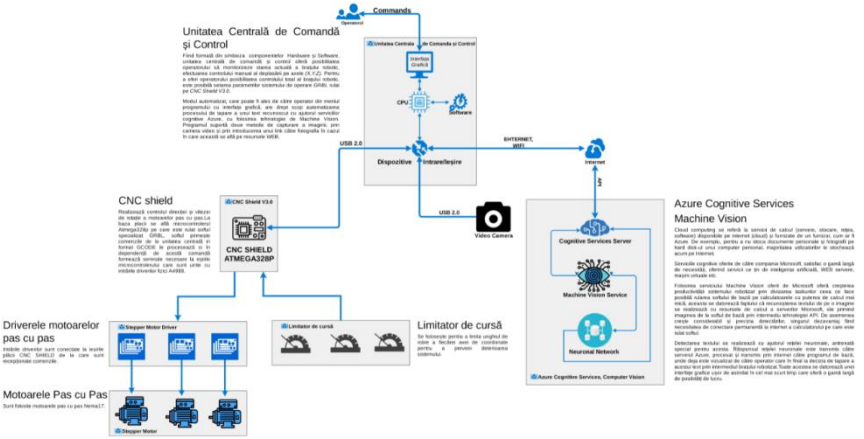


Figure 3. System overview

4. The Process of Acting the Intelligent Robotic System

4.1 Graphical User Interface (GUI) of the program

To make possible the interaction between the user and the intelligent robotic system, the graphical interface (Figure 4) has been developed, which facilitates monitoring and management of the process and it provides the main functionalities for a comfortable interaction.

The graphical interface has the following advantages and possibilities:

- Facilitates monitoring and management of the process of acquisition, detection and typing of the text
- Arrangement of the palettes in the intuitive and clear way
- Possibility to view the detected text
- Categorization of processes
- Possibility to control the whole process of detection the objects and web images.

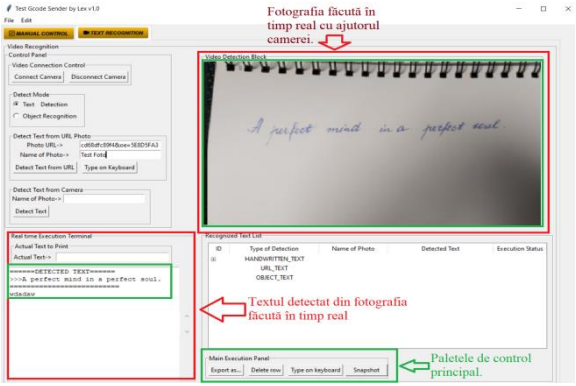


Figure 4. Graphical User Interface of the system

4.2 Text detection and interpretation by the Azure Cognitive Services

For the text to be typed by the robotic arm, it must first be acquired, processed, and then transmitted to the neural network for interpretation. In order to benefit from cognitive services (Figure 5) Microsoft needs to create a personal account on the Azure.com portal, after successful authentication the resource group is created that offers the possibility to detect and interpret the text on the image, the authentication between client and server is done by the API technology with the indication (ssh key, ssh endpoint) (Figure 6).

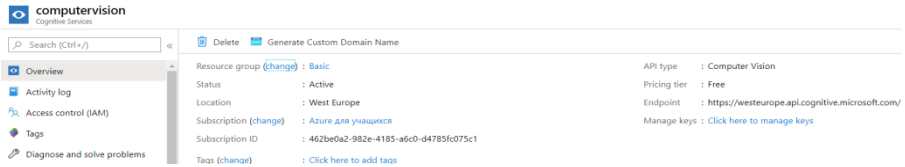


Figure 5. Azure Cognitive Services main control panel

```
In [1]: #----------IMPORTAREA BIBLIOTECILOR NECESARE----------
        from azure.cognitiveservices.vision.computervision import ComputerVisionClient          #Bibliotecile Azure
        from azure.cognitiveservices.vision.computervision.models import TextOperationStatusCodes#Bibliotecile Azure
        from azure.cognitiveservices.vision.computervision.models import TextRecognitionMode
        from azure.cognitiveservices.vision.computervision.models import VisualFeatureTypes      #Bibliotecile Azure
        from msrest.authentication import CognitiveServicesCredentials

        from array import array                                                                  #Biblioteca pentru lucrul cu vectorii
        import os                                                                                #Biblioteca pentru lucrul cu SO
        from PIL import Image                                                                    #Biblioteca pentru prelucrarea imaginii
        import sys
        import time                                                                              #Biblioteca de lucrul cu timpul
```

Figure 6. Import libraries in order to use Azure Services

After the text has been extracted (Figure 7) from the image, it is transmitted to the basic program, so that it can be typed by the robotic arm.

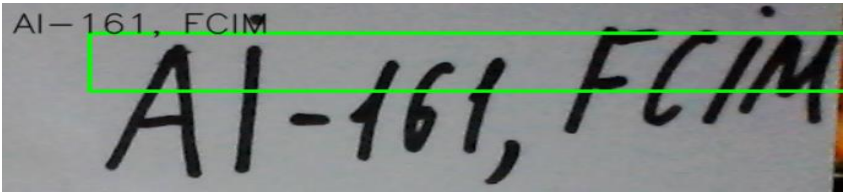


Figure 7. Extracted text by the neural network

The next step is detection of letters on the keyboard, using artificial intelligence.

4.3 Detection of letters on the keyboard

Visual detection of letters on the keyboard is performed using specially trained neural networks to detect typefaces. The neural network used in this project has the role of interpreting the image (Figure 8) already processed by the digital image processing technologies. The image segmentation is performed by a separate script that is not part of the neural network, it is assumed that the image transmitted for interpretation to the neural network can only be a letter that corresponds to a single class of letters.

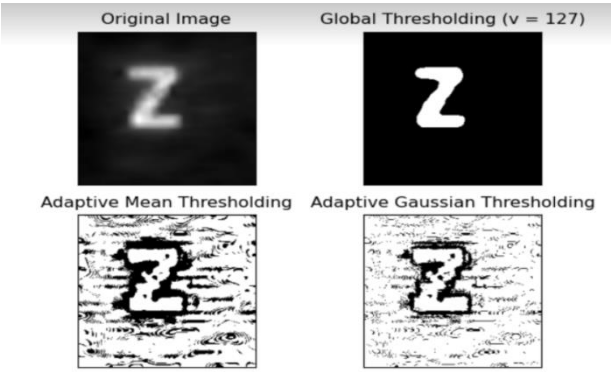


Figure 8. Digital processing of the image, which is sent to the neural network for interpretation

The digital image processing algorithm keeps the position of each letter transmitted to the neural network for interpretation, after the interpretation has been performed successfully, the algorithm assigns to each interpreted letter a position in the 2D axis system and draws this information on the real time screen view (Figure 9).

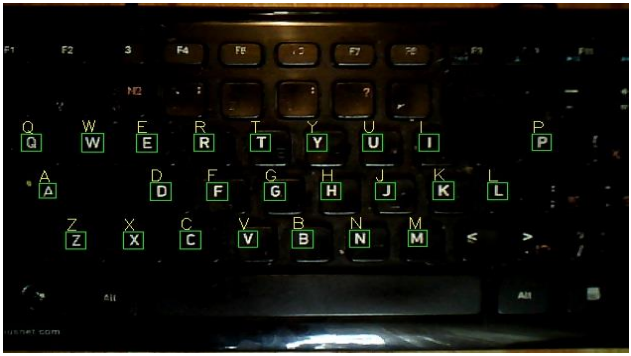


Figure 9. Result of letter identification by the keyboard detection algorithm

4.4 Calculating the distance between the letter and the robotic arm

In order to calculate the distance between the letters on the keyboard and the robotic arm, the pinhole model (Figure 1) was implemented, which offers the possibility to translate the position of the objects from the 3D space in the image plane (Figure10).

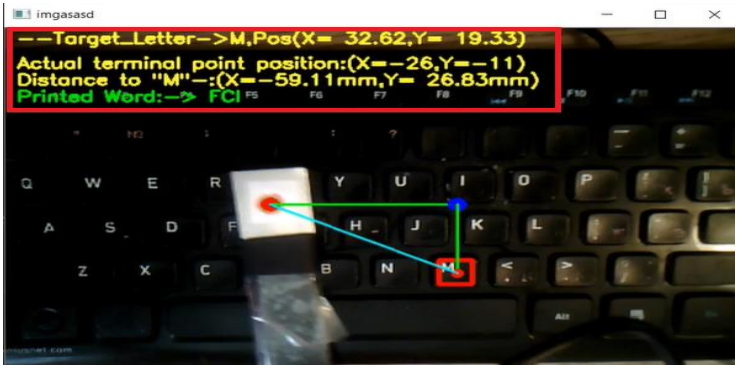


Figure 10. Calculating the distance between the letter and the robotic arm

5. Conclusion

This study proves the enormous possibilities of artificial intelligence in digital image processing, offering a wide range of possible variants of use depending on the wishes and needs of the engineer. The only one disadvantage is that it requires a modern technical material base and large computing resources. Machine vision is a branch that is increasingly used in all areas of life, which involves the development of this branch and as a result is an area of growing interest for subsequent research studies, which in relation to this study will focus on solving real problem.

References

[1] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer, September 3, 2010.

[2] D. Dechow. *The Fundamentals of Machine Vision*. FANUC America Corporation 3900.

[3] P. Sturm. *Pinhole Camera Model*. In: Ikeuchi K. (eds) Computer Vision. Springer, Boston, MA, 2014.

[4] R. Hartley, A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge, UK: Cambridge UP, 2003.

[5] K. M. Lynch, F. C. Park. *Modern Robotics Mechanics, Planning and Control.* Park May 3, 2017.

[6] D. Janglová. *Neural Networks in Mobile Robot Motion*. Inernational Journal of Advanced Robotic Systems Volume 1 Number 1 (2004), pp. 15-22.

Victor Popa [1], Irina Cojuhari[2]

[1]Politehnica University Timișoara
E-mail: victor.popa1@student.upt.ro

[2]Technical University of Moldova
E-mail: irina.cojuhari@ati.utm.md

Proceedings of Workshop on Intelligent Information Systems
WIIS2020, December 04-05, 2020, Chisinau, Republic of Moldova

About Gonsales-Couselo-Markov-Nechaev construction of orthogonal squares

Petr Radilov, Victor Shcherbacov, Irina Schvediuk

Abstract

We start to research Gonsales-Couselo-Markov-Nechaev construction of orthogonal squares.

Keywords: quasigroup, prolongation, Gonsales-Couselo-Markov-Nechaev construction.

1 Introduction

Definitions and elementary properties of quasigroups can be found in [1, 2, 4].

2 Orthogonal quasigroups of order ten

Using generalized Belousov construction of prolongation, it is possible to construct a pair of orthogonal quasigroups of order ten [3].

We start from quasigroup $(Z_7, \circ)$, with the form $x \circ y = 2 \cdot x + y \pmod 7$ for all $x, y \in Z_7$, where $(Z_7, +)$ is the cyclic group of order seven. Transversal $T^{(1)}$ is written using "normal" font, transversal $T^{(2)}$ is written using "italic" font, and transversal $T^{(3)}$ is written using "bold" font (see (1)).

0	1	2	**3**	4	*5*	6	□	□	□
2	3	4	5	6	**0**	1	□	□	□
4	5	*6*	0	1	2	3	□	□	□
6	0	**1**	2	*3*	4	5	□	□	□
1	2	3	4	**5**	6	*0*	□	□	□
3	*4*	5	6	0	1	**2**	□	□	□
5	**6**	0	*1*	2	3	4	□	□	□
□	□	□	□	□	□	□	7	8	9
□	□	□	□	□	□	□	8	9	7
□	□	□	□	□	□	□	9	7	8

(1)

In empty boxes of transversal $T^{(1)}$ (blue color) we write number 7, in empty boxes of transversal $T^{(2)}$ (yellow color) we write number 8, and in empty boxes of transversal $T^{(3)}$ (green color) we write number 9 (see (2) and (3)).

·	0	1	2	3	4	5	6	7	8	9
0	0	1	2	9	4	8	7	6	5	3
1	8	7	4	5	6	9	1	3	2	0
2	9	5	8	7	1	2	3	0	6	4
3	6	0	9	2	8	7	5	4	3	1
4	7	2	3	4	9	6	8	1	0	5
5	3	8	7	6	0	1	9	5	4	2
6	5	9	0	8	7	3	4	2	1	6
7	1	3	5	0	2	4	6	7	8	9
8	2	4	6	1	3	5	0	8	9	7
9	4	6	1	3	5	0	2	9	7	8

(2)

$\circ$	0	1	2	3	4	5	6	7	8	9
0	0	7	8	1	9	4	2	6	5	3
1	5	3	1	7	8	2	9	0	6	4
2	3	9	6	4	2	7	8	1	0	5
3	7	8	4	9	0	5	3	2	1	6
4	6	4	7	8	5	9	1	3	2	0
5	9	2	0	5	7	8	6	4	3	1
6	8	0	9	3	1	6	7	5	4	2
7	1	5	2	6	3	0	4	7	9	8
8	2	6	3	0	4	1	5	8	7	9
9	4	1	5	2	6	3	0	9	8	7

(3)

Quasigroup $(Q, \circ)$ is obtained from quasigroup $(Q, \cdot)$ using the following formula $x \circ y = y \cdot (x \cdot y)$.

Therefore we have the following pair of orthogonal quasigroups of order ten (see (4)).

00	17	28	91	49	84	72	66	55	33
85	73	41	57	68	92	19	30	26	04
93	59	86	74	12	27	38	01	60	45
67	08	94	29	80	75	53	42	31	16
76	24	37	48	95	69	81	13	02	50
39	82	70	65	07	18	96	54	43	21
58	90	09	83	71	36	47	25	14	62
11	35	52	06	23	40	64	77	89	98
22	46	63	10	34	51	05	88	97	79
44	61	15	32	56	03	20	99	78	87

(4)

3 Results

We start to research Gonsales-Couselo-Markov-Nechaev in more details. If we start from the cyclic group of order 10 $(Z_{10}, \cdot)$, then using formula $x \circ y = y \cdot (x \cdot y)$, we obtain the following groupoid and the following pair of orthogonal groupoids (see (5)–(7)).

$\cdot$	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	6	7	8	9
1	1	2	3	4	5	6	6	8	9	0
2	2	3	4	5	6	7	8	9	0	1
3	3	4	5	6	7	8	9	0	1	2
4	4	5	6	7	8	9	0	1	2	3
5	5	6	7	8	9	0	1	2	3	4
6	6	7	8	9	0	1	2	3	4	5
7	7	8	9	0	1	2	3	4	5	6
8	8	9	0	1	2	3	4	5	6	7
9	9	0	1	2	3	4	5	6	7	8

(5)

$\circ$	0	1	2	3	4	5	6	7	8	9
0	0	2	4	6	8	0	2	4	6	8
1	1	3	5	7	9	1	3	5	7	8
2	2	4	6	8	0	2	4	4	8	0
3	3	5	7	9	1	3	5	7	9	1
4	4	6	8	0	2	4	6	8	0	2
5	5	7	9	6	3	5	7	9	1	3
6	6	8	0	2	4	6	8	0	2	4
7	7	9	1	3	5	7	9	1	3	5
8	8	0	2	4	6	8	0	2	4	6
9	9	1	3	5	7	9	1	3	5	7

(6)

00	21	42	63	84	05	26	47	68	89
11	32	53	74	95	16	37	58	79	90
22	43	64	85	06	27	48	49	80	01
33	54	75	96	17	38	59	70	91	12
44	65	86	07	28	49	60	81	02	23
55	76	97	68	39	50	71	92	13	34
66	87	08	29	40	61	82	03	24	45
77	98	19	30	51	72	93	14	35	56
88	09	20	41	62	83	04	25	46	67
99	10	31	52	73	94	15	36	57	78

(7)

For construction of such groupoids, computer program was written and realised on language Pascal. We plan to improve this program.

References

[1] V.D. Belousov. *Foundations of the Theory of Quasigroups and Loops*. Nauka, Moscow, 1967. (in Russian).

[2] V.D. Belousov. *Elements of Quasigroup Theory: a Special Course*. Kishinev State University Printing House, Kishinev, 1981. (in Russian).

[3] S. Gonsales, E. Couselo, V. T. Markov, and A. A. Nechaev. *Recursive MDS-codes and recursively differentiable quasigroups*. Diskret. Mat., 10(2), pp. 3–29, 1998. (in Russian).

[4] H.O. Pflugfelder. *Quasigroups and Loops: Introduction*. Heldermann Verlag, Berlin, 1990.

Petr Radilov[1], Victor A. Shcherbacov[2], Irina Schvediuk[1]

[1]Ph.D. Student, Moldova State University
MD-2009, Chisinau, str. Mateevich, 60
Moldova
E–mails: illusionist.nemo@gmail.com, ira230396@mail.ru

[2] Dr., Vadimir Andrunakievich Institute of Mathematics
and Computer Science
MD-2028, str. Academiei, 5, Chisinau
Moldova
E–mail: vscerb@gmail.com

Proceedings of Workshop on Intelligent Information Systems
WIIS2020, December 04-05, 2020, Chisinau, Republic of Moldova

On weak-functionally complete systems of formulas containing paraconsistent negation in a paraconsistent logic

Andrei Rusu, Elena Rusu

Abstract

We consider a paraconsistent logic and establish the conditions for a system of its formulas containing paraconsistent negation to be functionally complete in it.

Keywords: paraconsistent negation, paraconsistent logic, functionally complete systems of formulas.

1 Introduction

A theory is called inconsistent if it has as its theorems formulas A and $\neg A$ [1]. If a theory derives all the formulas as its theorems, it is called trivial. These two notions are different, but they coincide in familiar to us systems. A theory is paraconsistent if it is inconsistent, but it is not trivial [2].

Paraconsistent logic is motivated not only by philosophical considerations, but also by its applications and implications. One of the applications is automated reasoning (information processing). Consider a computer which stores a large amount of information. While the computer stores the information, it is also used to operate on it, and, crucially, to infer from it. Now it is quite common for the computer to contain inconsistent information, because of mistakes by the data entry operators or because of multiple sourcing. This is certainly a problem for database operations with theorem-provers, and so has

drawn much attention from computer scientists. Techniques for removing inconsistent information have been investigated. Yet all have limited applicability, and, in any case, are not guaranteed to produce consistency. (There is no algorithm for logical falsehood.) Hence, even if steps are taken to get rid of contradictions when they are found, an underlying paraconsistent logic is desirable if hidden contradictions are not to generate spurious answers to queries [3].

Other motivation for investigations of paraconsistent logics is the part of artificial intelligence research called belief revision, which is one of the areas that have been studied widely. Belief revision is the study of rationally revising bodies of belief in the light of new evidence. Notoriously, people have inconsistent beliefs. They may even be rational in doing so. For example, there may be apparently overwhelming evidence for both something and its negation. There may even be cases where it is in principle impossible to eliminate such inconsistency. For example, consider the 'paradox of the preface'. A rational person, after thorough research, writes a book in which they claim $(A_1 \& \ldots \& A_n)$. But they are also aware that no book of any complexity contains only truths. So they rationally believe $\sim (A_1 \& \ldots \& A_n)$ too. Hence, principles of rational belief revision must work on inconsistent sets of beliefs. So, a more adequate account can be based on paraconsistent logic [4]. Other applications of paraconsistent logics are known in robot control [5], in air traffic control [6], in defeasible deontic reasoning [7], in information systems [8] and medicine. Connections between paraconsistent logics, adaptive logics and diagnosis are investigated in [9] and [10].

2 The problem

It is a well known class of problems in logic, algebra, discrete mathematics and cybernetics dealing with the possibility of obtaining some functions (operations, formulas) from other ones by means of a fixed set of tools. The notion of expressibility of Boolean functions through other ones by means of superpositions goes back to the works of E. Post [11], [12]. He described all closed (with respect to superpositions) classes

of 2 valued Boolean functions. The problem of completeness (with respect to expressibility), which requires to determine the necessary and sufficient conditions for all formulas of the logic under investigation to be expressible via the given system of formulas, is also investigated. In 1956 ([13, p. 54], [14]) A. V. Kuznetsov established the theorem of completeness according to which we can build a finite set of closed with respect to expressibility classes of functions in the k-valued logics such that any system of functions of this logic is complete if and only if it is not included in any of these classes. In 1965 [15] Rosenberg I. established the criterion of completeness in the k-valued logics formulated in terms of a finite set of pre-complete classes of functions, i.e. in terms of maximal, incomplete and closed classes of functions.

In the present paper we investigate the conditions of completeness with respect to expressibility of the systems of formulas of the modal logic $S5$ containing the paraconsistent negation of the logic $S5$.

The standard language of $S5$ is based on propositional variables and logical connectives: $\&$, $\vee$, $\rightarrow$, $\neg$, $\Box$, and $\Diamond$. We consider the paraconsistent negation $\sim$ of $S5$ [16] as follows:

$$\sim a =_{Def} \Diamond\neg a.$$

The logic $S5$ can be considered, according to [16], as a paraconsistent logic since it contains a paraconsistent negation. The logic $S5$ is characterized by the axioms and rules of inference of the classical propositional logic, the following axioms:

$$\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B),$$
$$\Box A \rightarrow A,$$
$$\Diamond A \rightarrow \Box\Diamond A,$$

and the necessity rule of inference: from A infer $\Box A$.

The formula F is said to be *weak-expressible in the logic L via a system of formulas* Σ, if F can be obtained from propositional variables, constants and formulas of Σ applying a finite number of times: a) the rule of substitution of equivalent formulas in the logic L, and b) the

rule of weak substitution, which permits, being given formulas A and B, to substitute one of them in another instead of a given corresponding propositional variable [17], [18], [19].

The system of formulas Σ is said to be *(weak-functionally) complete (with respect to the weak-expressibility) in the logic* L, if all formulas of the calculus of L are weak-expressible in the logic L via formulas of Σ [17],[18], [19].

3 Main result

The main result of the paper is the following.

Theorem. *There is an algorithm that decides weather a system of formulas containing the paraconsistent negation* $\sim$ *is weak-functionally complete in the modal logic* $S5$.

4 Conclusion

We can consider in the same manner the problems of (parametric, positive, implicit) expressibility of the systems of formulas containing the paraconsistent negation in the modal logic $S5$, too. Also we can investigate the problem of weak-functional systems containing paraconsistent negation in other logical systems.

Acknowledgments. National Agency for Research and Development has supported part of the research for this paper through the research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data".

References

[1] E. Mendelson. *Introduction to mathematical logic*, 4ed., CRC, 1997, 446 p.

[2] G. Priest. *The Logic of Paradox*. Journal of Philosophical Logic, vol. 8, pp. 47–50.

[3] G. Priest, R. Routley, and J. Norman. *Paraconsistent Logic: Essays on the inconsistent*. Philosophia, Munchen, 1989.

[4] I. Urbas. *Paraconsistency*. Studies in Soviet Thought, vol. 39, pp. 343–354.

[5] Nakamatsu, K., Abe, J. M. and Suzuki, A., *Defeasible deontic robot control based on extended vector annotated logic programming*, Computer Antecipatory Systems: SASYS 2001 - Fifth International Conference, ed. D. M. Dubois, American Institute of Physics, 2002, pp. 490–500.

[6] K. Nakamatsu, Y. Mita, T. Shibata, and J. Abe. *Defesiable deontic action control based on paraconsistent logic program and its implementation*. Proceedings of the 2003 International Conference on Computational Intelligence for Modelling, Control and Automation - CIMCA 2003, M. Mohammadia (Ed), pp. 233-246, 2003.

[7] K. Nakamatsu, J.M. Abe, and A. Suzuki. *Annotated semantics for defeasible deontic reasoning*. in W. Ziarko and Y. Yao (eds.), RCCTC 2000, Springer Verlag, 2001, pp. 470–478.

[8] S. Akama and J.M. Abe. *The role of inconsistency in information systems*. Proc. of the 5th World Multiconference on Systemics, Cybernetics and Informatics (SCI'2001), pp. 355–360, Orlando, USA, 2001.

[9] D. Batens, J. Meheus, D. Provijn, and L. Verhoeven. *Some adaptive logics for diagnosis*. Logic and Logical Philosophy, no. 11–12, pp. 39–65, 2003.

[10] A. Aliseda and L. Leonides. *Hypotheses testing in adaptive logics: An application to medical diagnosis*. Logic Journal of the IGPL, vol. 21, no.6, pp. 915–930, 2013.

[11] E. Post. *Introduction to a general theory of elementary propositions*. Amer. J. Math., vol. 43, pp. 163–185, 1921.

[12] E. Post. *The two-valued iterative systems of mathematical logic*, Princeton, 1941.

[13] S.V. Jablonskij. *Introduction to discrete mathematics*, Moscow, Nauka Ed., 1986, 384 p.

[14] A.V. Kuznetsov. *On problems of identity and criteria of functional completeness*. Proceedings of the 3rd Allunion Congress of Mathematics, Moscow, vol. 2, pp. 145–146, 1956.

[15] I. Rosenberg. *La structure des fonctions de plusieure variables sur un ensemble finit*. C.R.Acad.Sci., Paris, Groupe 1, vol. 260, pp. 3817–3819, 1965.

[16] Jean-Yves Beziau. *S*5 *is a paraconsistent logic and so is classical first-order logic*, Logical Studies, vol. 9, 2002.

[17] A.V. Kuznetsov. *On functional expressibility in the superintuitionistic logics*. Mathematical Researches, Chişinău, vol. 6, no. 4, pp. 75–122, 1971. (in Russian).

[18] A.V. Kuznetsov. *On tools for detection of non-deduction or non-expressibility*, in book: Logical deduction, Moscow: Nauka, 1979, pp. 5–33. (in Russian).

[19] M.F. Ratsa. *Expressibility in propositional calculi*, Chishinau, Shtiinca, 1991.

Andrei Rusu[1,2], Elena Rusu[3]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science,
E–mail: andrei.rusu@math.md

[2]Ovidius University of Constanta
E–mail: andrei.rusu@univ-ovidius.ro

[3]Technical University of Moldova
E–mail: elena.rusu@mate.utm.md

Proceedings of Workshop on Intelligent Information Systems
WIIS2020, December 04-05, 2020, Chisinau, Republic of Moldova

Modeling of the Process of Oil Reservoir Secondary Drilling-in with Explosive Charge

Boris Rybakin, Valery Goryachev,
Marina Kravchenko, Luben Stamov

Abstract

The paper presents the results of mathematical modeling of the process of the squib charge blast and cumulative jet formation. The influence of the squib and the liner shapes as well as the impact of liner material, its thickness and the type of explosive on speed and energy of cumulative jet was investigated. The generated cumulative jet forms a perforation in the metal pipe and cement sheath, which is formed during well drilling with further plugging-back. The use of squibs for oil reservoir secondary drilling-in allows creating a high–quality hydrodynamic connection well-reservoir and maintaining a stable inflow to the well.

Keywords: mathematical modeling, cumulative charges, resilience-plastic deformation.

1 Introduction

One of effective ways to create a high - quality hydrodynamic "well-reservoir" connection is cumulative perforation. The energy of an explosive in a cumulative charge is focused due to the special shape of the body frame, and a cumulative jet is formed [1]. The formation of cumulative jets with the necessary energy and speed depends on many different factors: the type and energy of the explosive, the shape of the external and internal shell, liner form and other factors. Steel is most often used for outer shell, while aluminum, copper, tungsten, and

alloys of tungsten and copper obtained by doping are used for the inner shell. The application of multilayer internal edgings with high-density powdery materials applied by detonation is suggested [1]. In [2], the influence of the liner material (steel, and aluminum) on the formation of the jet and the increase in its penetrating capability was studied. The processes of formation of a cumulative jet are studied by experimental and numerical methods. The results show that the aluminum liner gets the highest speed, but low penetration. The copper jet has significantly more energy, with the formation of a longer channel. The

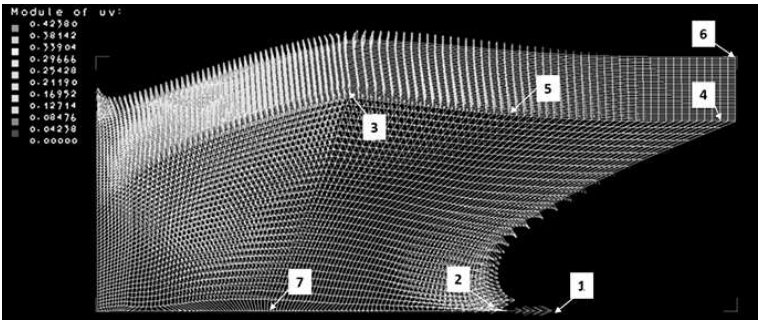


Figure 1. Computational grid, numbers of cells where values of density, velocity, stress and strain tensors are "recorded"

reservoir drilling-in with application of the energy of explosives is recognized as most effective. The negative point is the formation of a "sintering crust" around the perforation tunnel. This "crust" can be partially destroyed with further establishment of fractures network by interference of jets while several charges are applied. Mathematical modeling methods of the explosive charge initialization process, the occurrence of a high-temperature gas jet with subsequent penetration of the cumulative jet into the combined barrier allow optimizing the shape of the squib and choosing the material and thickness of the liner [3]. So far, relatively few studies have been conducted to examine samples of liner materials such as Fe, Ta, Al, and Cu and layered liners. The proposed mathematical model is a system of nonlinear partial dif-

ferential equations [4] describing the non-stationary process of solid explosive detonation, gas jet formation, and structures dynamic stress, considering the influence of various factors (geological and hydromechanical) [5].

Figure 1 shows the calculated grid and velocity field. The number of the "sensor" corresponds to the cell number where the values of density, velocity, and stress and strain tensors are "recorded". A graph is shown for the time instant of the detonation wave passing through cell No. 7. By this time, the detonation wave front had already assumed a spherical shape.

2 Basic equations

The oil-producing reservoir is considered with layering directions perpendicular to the well axis. The layer is assumed as isotropic. In this case, the stress state of the formation in the near-well zone will be determined by the rock pressure and hydrodynamic reservoir pressure. The standard cumulative perforator with squib placed in is located in the wellbore. The axis of the squib is assumed perpendicular to the wellbore wall. The aim of the study is to simulate the process of converting solid explosive into detonation products, as well as detonation development and high-temperature gas jet formation, the penetration of a metal layer and the formation of a perforation tunnel in a cement sheath. The following assumptions are made: mass forces are neglected; the task parameters do not depend on the spatial coordinate orthogonal to a reservoir surface; assume plane strain of the layer; the problem is solved in a two-dimensional formulation; the formation is modeled by a damaged thermo-resilience-plastic medium; the deformation process is considered adiabatic.

The system of defining equations, taking into account the accepted assumptions, can be written in a two-dimensional cylindrical system of r and z coordinates in the form of the laws of conservation of mass (1),

momentum (2)–(3) and energy (4):

$$\frac{1}{\rho}\frac{\partial \rho}{\partial t} = \Big[\frac{\partial u}{\partial r} + \frac{\partial w}{\partial z} + \alpha\frac{u}{r}\Big], \tag{1}$$

$$\rho\frac{\partial u}{\partial t} = \frac{\partial S_{rr}}{\partial r} + \frac{\partial S_{rz}}{\partial z} + \alpha\frac{S_{rr} - S_{\theta\theta}}{r} - \frac{\partial P}{\partial r}, \tag{2}$$

$$\rho\frac{\partial w}{\partial t} = \frac{\partial S_{rz}}{\partial r} + \frac{\partial S_{zz}}{\partial z} + \alpha\frac{S_{rz}}{r} - \frac{\partial P}{\partial z}, \tag{3}$$

$$\rho\frac{\partial E}{\partial t} = \frac{P}{\rho^2}\frac{\partial \rho}{\partial t} + \frac{1}{\rho}\Big[S_{rr}\frac{\partial u}{\partial r} + S_{zz}\frac{\partial w}{\partial z} + \alpha S_{\theta\theta}\frac{u}{r} + S_{rz}\Big(\frac{\partial u}{\partial z} - \frac{\partial w}{\partial r}\Big)\Big], \tag{4}$$

$$\frac{\partial r}{\partial t} = u, \quad \frac{\partial S_{rr}}{\partial t} = 2\mu\Big(\frac{\partial u}{\partial r} + \frac{1}{3\rho}\frac{\partial \rho}{\partial t}\Big),$$

$$\frac{\partial z}{\partial t} = w, \quad \frac{\partial S_{zz}}{\partial t} = 2\mu\Big(\frac{\partial w}{\partial z} + \frac{1}{3\rho}\frac{\partial \rho}{\partial t}\Big),$$

$$\frac{\partial S_{\theta\theta}}{\partial t} = 2\mu\Big(\alpha\frac{u}{r} + \frac{1}{3\rho}\frac{\partial \rho}{\partial t}\Big), \quad \frac{\partial S_{rz}}{\partial t} = \mu\Big(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial r}\Big).$$

where ρ — density, P – pressure, u, w — velocity vector components, t – time, E – specific internal energy, S_{lm} – stress tensor deviator components. Stress tensor is represented as follows:

$$\sigma_{ij} = -P\delta_{ij} + S_{ij}, \quad (i = 1, 2 \quad j = 1, 2),$$

$$P = \frac{1}{3}\Sigma_1^3\sigma_{ii}.$$

The parameter $\alpha = 0$ corresponds to a flat deformed state, and $\alpha = 1$ corresponds to an axisymmetric state.

The calculations use the Mises plastic flow model, i.e. the second invariant of the stress tensor deviator relating to plasticity, meets the Mises flow condition [4]

$$S_{rr}^2 + S_{zz}^2 + S_{\theta\theta}^2 + 2S_{rr} \leqslant \frac{2}{3}Y_0^2. \tag{5}$$

Several different pressure state equations (shock adiabat, The Mie-Grüneisen equation of state, etc.) will be used for different materials.

$$P = a_1(\eta - 1) + a_2(\eta - 1)^2 + a_3(\eta - 1)^3 + a_4\eta E. \tag{6}$$

Here $\eta = \rho/\rho_0$, ρ_0– initial density.

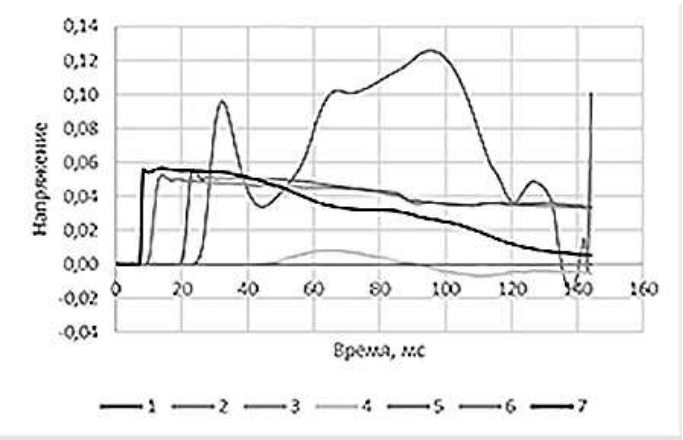


Figure 2. Stress tensor on time dependence, the numbers at the bottom of the chart correspond to the cell number

3 Numerical solution

The considered task is solved numerically by the Wilkins method [4]. Macrofracture modeling is performed by splitting the Lagrangian estimated grid. The entropy parameter of destruction is used as a macrofracture beginning parameter: the medium loses its entirety while the specific dissipation D has reached the limit value D*. At the point in the medium, where the destruction parameter is met, the rupture banks are explicitly established. Optimal disposal of estimated grid nodes plays an important role in solving mathematical modeling tasks [6]. Thus, the construction of good estimated grids is an actual task. In this paper, a program for constructing a quadrangular Lagrangian grid was applied [5].

Figure 2 shows charts that allow one to study the process of propagation of the detonation wave through the explosive. Stress tensor on time dependence chart is given. The numbers at the bottom of the chart correspond to the cell number shown in Figure 1. The blast wave initiated by the detonation wave propagates through the steel shell ma-

terial. To that point in time, the shape of the squib starts to deform (Figure 1) due to a reactive force which causes movement of the squib to the left and interacting with a solid boundary. In a steel shell, relief waves are observed, they are formed on its free surface.

4 Results of numerical studies

Figure 3 shows the generated cumulative jet velocities dependence on time for two points – 1 and 2 (Figure 1). The liner material is aluminum. The chart shows the simulation results, starting with liner acceleration by detonation wave. The shape of the squib and the stress values are shown in the rectangle at the top left. Figure 4 shows the

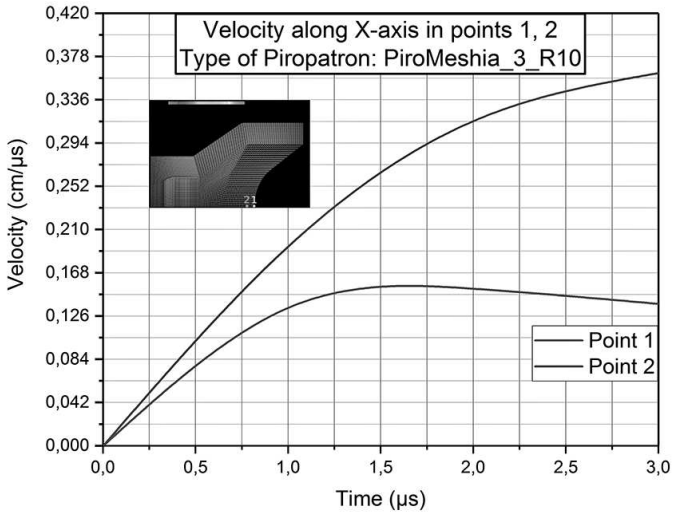


Figure 3. Cumulative jet velocities for two points – 1 and 2

process of the shell deformation and the liner cumulative jet formation. Figure 4 shows the vertical values of the stress tensor. The legend on the right shows the stress values and the corresponding level of the gray scale.

Figure 4 shows how the detonation wave begins to form a cumulative jet. The front of the detonation wave moving along the explosive (TNT, $\rho = 1.72\ g/cm^3$) reached the right border of the calculated area at the top. In the outer steel shell (steel, $\rho = 7.85\ g/cm^3$), strong stress waves spread out, which leads to strong resilience-plastic deformations of squib external body of up to destruction. A gas-metal jet is formed on the axis of symmetry, which moves with increasing speed (Figure 4). The entire process of detonation wave spread, shell deformation, jet formation and its speed increase takes about 40 ms considering this design of a squib.

The initiation of the explosive is taken as a time reference point. After the formation of the jet, the velocity of the substance in the center is faster than at the periphery. Since the elements of the jet formed at the top are faster, the speed of its top is constantly growing and reaches its maximum value in a time of about 10 ms. After that, the speed gradually decreases. The conducted mathematical modeling is in

Figure 4. The liner cumulative jet formation and stress tensor values disposal

accordance with the experimental data. Numerous calculations have been conducted, where there are various forms of squibs as well as

materials and angles of the liners. Studies have shown the cumulative jet velocity dependence on the squib shape [3].

5 Conclusions

The created original program code allows conducting numerical experiments in a wide range of parameters, and tracking all stages of the detonation process inside the shell device, followed by deformation of the squib shell and destruction of the complex target in real time. Numerical simulation calculations provide adequate results that are qualitatively consistent with laboratory experiments. This modeling is carried out in order to improve the quality of the perforation process and improve the hydrodynamic connection between the productive reservoir and the well.

Acknowledgments. This paper was prepared with support of Russian Foundation of Fundamental Research Grant №18-07-01303 A.

References

[1] P.V. Rogozhin. *Sposob izgotovleniya bimetallicheskih oblicovok kumulyativnyh zaryadov, primenyaemyh v perforatorah dlya vskrytiya produktivnyh neftyanyh i gazovyh plastov.* Metallurgiya i novye materialy, Samara, 2010. pp. 159–160.

[2] Cheng Wang, Jianxu Ding, and Haitao Zhao. *Numerical Simulation on Jet Formation of Shaped Charge with Different Liner Materials.* Defence Science Journal, vol. 65, no. 4, July 2015, pp. 279–286, DOI : 10.14429/dsj.65.8648.

[3] B.P. Rybakin, N.N. Smirnov, V.D. Goryachev at al. *Computer simulation of the formation drilling-in with cumulative charges application.* Vestnik kibernetiki [Bulletin of Cybernetics], 2018, vol. 31, no. 3, pp. 1–10.

[4] M.L. Wilkins. *Computer Simulation of Dynamic Phenomena*. Springer, 1999, 264 p.

[5] P.Z. Lugovoi, V.F. Meish, B.P. Rybakin, and G.V. Sekrieru. *Numerical simulation of the dynamics of a reinforced shell subject to nonstationary load*. Springer, International Applied Mechanics, 2008. vol. 44, no. 7, pp. 788–793.

[6] A.A. Korotkov, V.A. Pervichko, I.G. Plotnikova, and V.V. Chudanov. *Metod postroeniya ortogonal'nyh setok na ploskosti, na gladkoj trekhmernoj poverhnosti*. (Preprint). M.: IBRAE RAN, 2006, 47 p.

Boris Rybakin[1,2], Valery Goryachev[3], Marina Kravchenko[4], Luben Stamov[2]

[1]Lomonosov Moscow State University, Moscow, Russia
E–mail: rybakin@vip.niisi.ru

[2]SRISA RAS
E–mail: lyubenstamov@mail.ru

[3]Department of Mathematics, Tver State Technical University, Russia
E–mail: gdv.vdg@yandex.ru

[4]Gubkin Russian State University of Oil and Gas (NRU), Moscow, Russia
E–mail: dep.ngipg@yandex.ru

Proceedings of Workshop on Intelligent Information Systems
WIIS2020, December 04-05, 2020, Chisinau, Republic of Moldova

Regional intelligent data warehouse for DLD cases

Iulian Secrieru, Elena Gutuleac, Olga Popcova

Abstract

Diffuse liver diseases (DLD) is a continuing concern of internal medicine and a subject of numerous research and publications. Correct and early assessment of liver diseases combined with appropriate management of pathologies can certainly increase the patients' quality of life and its duration. The main aim of the research is to formalize and integrate data and scientific knowledge from the fields of diagnostics and treatment of DLD, which at the moment are unstructured, fragmented and heterogeneous into a unique informational space. Data and knowledge digital warehouse is used in order to allow the interoperability of the stored data contents and knowledge, for healthcare shareholders from EaP region (Moldova, Armenia, Azerbaijan, Belarus, Georgia and Ukraine).

Keywords: medical informatics, diffuse liver diseases, data digital warehouse.

1 Introduction

Chronic diffuse liver diseases play an important role in morbidity and mortality of the population of many economically developed countries, but also in developing and transition countries. Liver disease accounts for approximately 2 million deaths per year worldwide, one million due to complications of cirrhosis and one million due to viral hepatitis and hepatocellular carcinoma. Hepatic cirrhosis is currently the 11th most common cause of death globally and liver cancer is the 16th leading

cause of death; combined, they account for 3.5% of all deaths worldwide. Cirrhosis is within the top 20 causes of disability-adjusted life years and years of life lost, accounting for 1.6% and 2.1% of worldwide burden [1]. The burden of liver disease in Europe continues to grow [2]. We have witnessed the wide spread of DLD in the whole EaP region, which predominantly affect people of working age, having a significant negative impact on social and economic development of the countries [3].

Diagnostics of DLD requires special knowledge and use of laboratory and instrumental diagnostic methods, sometimes unique or rarely used in daily diagnostics. This fact causes impediments not only to novices, general practitioners, but also to experienced doctors. At the same time, international groups of experts have developed criteria, scales and diagnostic scores, based on certain parameters. Although these tools (criteria, scales and scores) are quite accurate, they are inconvenient in their use in everyday clinical practice.

There is a clear need to create and provide clinicians with information tools for collecting, storing and processing medical test data and instrumental investigations of patients predisposed or suspected of having a DLD.

2 Methods and Results

At the stage of elaboration of a single protocol for description of DLD the methods and practices from the domain of artificial intelligence (advanced technologies for acquisition of professional medical knowledge, structurization algorithms and creation of medical taxonomies, representation methods in the form of medical ontologies) have been used [4].

At the stage of information, data and expert knowledge collection and storage practices from the domain of relational databases have been used. For structurization and primary data processing classical methods of medical statistics have been used.

At the stage of development of tools for quantifying and assessing

DLD advanced algorithms from the domain of artificial intelligence, such as segmentation and clusterization, have been used [5].

Development of the user interface is based on existing practices, principles and approaches used in modern medical information systems. As a result, we created tools for quantifying and assessing DLD, which could be integrated in regional intelligent data warehouse.

The tools allow:

- to aggregate a large number of DLD cases in a standardized manner;
- to define criteria based on non-invasive measurements and laboratory tests for quantifying and assessing DLD;
- to establish thresholds and endpoints for onset and all stages of DLD progress.

As a regional data warehouse, we intend:

- to harmonize efforts of data stakeholders for in-depth DLD phenotyping;
- to promote clinically impactful new knowledge discovery and its translation into clinical practice.

3 Conclusion

The main criteria based on non-invasive measurements and laboratory tests for quantifying and assessing DLD have been identified. Based on these criteria taxonomy for DLD cases formalization in the frame of data warehouse have been created. Algorithms and informational tools for establishing the thresholds and endpoints for onset and all main stages of DLD progress have been developed. The regional intelligent data warehouse, as a computer-based shared platform, will intensify, simplify and facilitate knowledge exchange between healthcare shareholders from EaP region, that otherwise would be difficult or even impossible.

Acknowledgments. State Program Project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data"; G5700 NATO Science for Peace and Security Programme; EYR@EaP 2019 programme have supported the research for this paper.

References

[1] Asrani S.K., Devarbhavi H., Eaton J., et al. *Burden of liver diseases in the world.* Journal of hepatology, 2019.

[2] Pimpin L., Cortez-Pinto H., Negro F., Corbould E., Lazarus J.V., Webber L., Sheron N., Committee EHS. *Burden of liver disease in Europe: Epidemiology and analysis of risk factors to identify prevention policies.* Journal of hepatology. 2018; 69(3) pp. 718—735.

[3] *World Health Organization Regional Office for Europe.* European Detailed Mortality Database. 23 August 2016.

[4] Secrieru Iu. *Structured knowledge management techniques for the development of interactive and adaptive decision support system.* Computer Science Journal of Moldova, nr. 1 (49), vol. 17, 2009, Kishinev, pp. 58–73.

[5] Secrieru Iu., Popcova O., and Gutuleac E. *Quantification and Assessment of Diffuse Liver Diseases Using Deep Data Analysis.* Proceedings of the Conference on Mathematical Foundations of Informatics, Chisinau, Moldova, November 9-11, 2017, pp. 166–167.

Iulian Secrieru[1], Elena Gutuleac[2], Olga Popcova[3]

[123]Vladimir Andrunachievici Institute of Mathematics and Computer Science
E–mails: [1]iulian.secrieru@math.md, [2]elena.gutuleac@math.md, [3]oleapopcova@yahoo.com

Proceedings of Workshop on Intelligent Information Systems
WIIS2020, December 04-05, 2020, Chisinau, Republic of Moldova

On Degrading Systems Modeling via Finite Markov Chains

Volodymyr G. Skobelev, Volodymyr V. Skobelev

Abstract

Analytical models of degrading systems based on Finite Markov Chains are defined. These models are analyzed within the finite time horizon for recoverable and non-recoverable degrading systems. The set of critical states is identified, which is the basis for the implementation of a bounded probabilistic analysis of the investigated degrading system.

Keywords: degrading systems, Finite Markov Chains.

1 Introduction

A degrading system (DS) is any real system, the functioning of which deteriorates due to aging and accumulated wear. Therefore, its parameters gradually deteriorate, and if left unattended, then a failure can occur.

The framework for analysis of maintenance policies for recoverable DS has been laid in [1]. A significant contribution into this area has been made by the development of some base for DS analysis within the finite time horizon [2].

Remark 1. It should be emphasized that "to deal with the finite time horizon" is the essential condition for the selection of effective preventive maintenance policies for any DS.

The importance of solving the selection of effective preventive maintenance policies for any DS has been increased significantly due to the

extensive use of Cyber-Physical Systems (CPS) [3], since for these systems, in addition to recoverable DS, there is the urgent need to analyze the non-recoverable ones. The last DS have numerous applications in different applied areas. For example, economic evaluations in the health sector [4, 5] related to the development and treatment of any chronic disease naturally lead to analysis of some non-recoverable DS.

To simulate temporal variability of deterioration in DS different stochastic processes are usually used. Among them Finite Markov Chains (FMC) should be noted especially since transitions of the states in FMC can be naturally interpreted as measurements of the system's parameters when the fixed period of time expires. Exactly this feature takes place for a wide class of CPS.

In the given paper analytical models that are based on FMC and intended for analysis of recoverable and non-recoverable DS, are defined and investigated. It is shown that the proposed models can be applied for selection of any effective maintenance policy for the investigated DS on the basis of the methods of bounded reachability properties analysis that can be applied.

2 Proposed analytical models

It is well known that any FMC $\mathcal{C}_n$ with the states

$$s_1, \ldots, s_n \ (n \geq 2)$$

can be defined by the stochastic matrix

$$P_{\mathcal{C}_n} = \begin{array}{c|cccc} & s_1 & s_2 & \ldots & s_n \\ \hline s_1 & p_{11} & p_{12} & \ldots & p_{1n} \\ s_2 & p_{21} & p_{22} & \ldots & p_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_n & p_{n1} & p_{n2} & \ldots & p_{nn} \end{array},$$

where p_{ij} $(i, j = 1, \ldots, n)$ is the probability of transition from the state s_i to the state s_j.

When we are modeling the given DS $\mathcal{S}_n$ via the FMC $\mathcal{C}_n$, then in order to represent degradation and recovery, it is assumed that the following two assumptions hold:

Assumption 1. For a given positive integer k $(2 \leq k \leq n)$ it is fixed some partition

$$\pi = \{B_1, \ldots, B_k\}$$

of the set $\{s_1, \ldots, s_n\}$, such that

$$B_1 = \{s_{i_1}\},$$

$$B_j = \{s_{i_{j-1}+1}, \ldots, s_{i_j}\} \ (j = 2, \ldots, k-1),$$

$$B_k = \{s_{i_{k-1}+1}\},$$

where $i_1 = 1$, $i_{k-1} = n-1$, and $i_j < i_{j+1}$ for all $j = 1, \ldots, k-2$.

Assumption 2. Elements of the stochastic matrix $P_{\mathcal{C}_n}$ satisfy the following five conditions:

Condition 1. For all $i = 1, \ldots, n-1$ the inequality $p_{ii} > 0$ holds, and $p_{nn} = 1$.

Condition 2. For all $j = 2, \ldots, k-1$ the equality $p_{rh} = 0$ holds for all states $s_r, s_h \in B_j$ $(r \neq h)$.

Condition 3. For all $j = 1, \ldots, k-1$ and for any state $s_r \in B_j$ there exists some non-empty subset S_r^{dsc} of the set $\bigcup\limits_{m=j+1}^{k} B_m$, such that for all states $s_h \in S_r^{dsc}$ the inequality $p_{rh} > 0$ holds, and for all states $s_h \in \left(\bigcup\limits_{m=j+1}^{k} B_m \right) \setminus S_r^{dsc}$ the equality $p_{rh} = 0$ holds.

Condition 4. For all $j = 1, \ldots, k-2$ the following equality holds:

$$\bigcup_{s_r \in B_j} (S_r^{dsc} \cap B_{j+1}) = B_{j+1}.$$

Condition 5. For all $j = 2, \ldots, k-1$ and for any state $s_r \in B_j$ there exists some subset S_r^{anc} of the set B_{j-1}, such that $p_{rh} > 0$ for all states $s_h \in S_r^{anc}$, and $p_{rh} = 0$ for all states $s_h \in B_{j-1} \setminus S_r^{anc} \cup \bigcup\limits_{m=1}^{j-2} B_m$.

Remark 2. When we are modeling the given DS $\mathcal{S}_n$ via the FMC $\mathcal{C}_n$, the state s_1 represents the fully functional stage for $\mathcal{S}_n$, the states $s_2, \ldots, s_{n-1}$ represent the stages for partial functionality for $\mathcal{S}_n$, the state s_n represents the faulty stage for $\mathcal{S}_n$, and the set of states B_{k-1} represents the set of critical stages for $\mathcal{S}_n$.

Definition. *An FMC $\mathcal{C}_n$ is:*

1. A model of a recoverable DS $\mathcal{S}_n$, if for all $j = 2, \ldots, k-1$ and for any state $s_r \in B_j$ the disequality $S_r^{anc} \neq \emptyset$ holds;

2. A model of an non-recoverable DS $\mathcal{S}_n$, if for all $j = 2, \ldots, k-1$ and for any state $s_r \in B_j$ the equality $S_r^{anc} = \emptyset$ holds.

Example. *1. The FMC $\mathcal{C}_4^{(1)}$ with four states s_1, s_2, s_3, s_4, defined by the stochastic matrix*

$P_{\mathcal{C}_4^{(1)}} =$

	s_1	s_2	s_3	s_4
s_1	0.70	0.20	0.12	0.08
s_2	0	0.60	0.25	0.15
s_3	0	0	0.70	0.30
s_4	0	0	0	1

is a model of some disease progression for the given Person.

The state s_1 represents the stage of the Person before the disease starts, the states s_2 and s_3 represent possible stages of the disease, and the state s_4 represents the death of the Person.

In this case

$$\pi = \{\{s_1\}, \{s_2\}, \{s_3\}, \{s_4\}\}.$$

The single-element set of the states $\{s_3\}$ represents the set of critical stages for the Person. Due to Definition, the FMC $\mathcal{C}_4^{(1)}$ is a model of an non-recoverable DS.

2. The FMC $\mathcal{C}_4^{(2)}$ with four states s_1, s_2, s_3, s_4, defined by the stochastic matrix

$P_{\mathcal{C}_4^{(2)}} =$

	s_1	s_2	s_3	s_4
s_1	0.80	0.10	0.10	0
s_2	0.40	0.50	0	0.10
s_3	0.30	0.50	0	0.20
s_4	0	0	0	1

is a model of some net consisting of four computers.

The state s_1 represents the fully functional stage for the net, the states s_2 and s_3 represent possible stages of partial functionality for the net, and the state s_4 represents the faulty stage for the net.

In this case

$$\pi = \{\{s_1\}, \{s_2, s_3\}, \{s_4\}\}.$$

The two-elements set of the states $\{s_2, s_3\}$ represents the set of critical stages for the net. Due to Definition, the FMC $\mathcal{C}_4^{(2)}$ is a model of a recoverable DS.

3 Analysis of proposed models

It is supposed that the finite time horizon during which the behavior of the given DS $\mathcal{S}_n$ is investigated corresponds to l transitions of the states in the FMC $\mathcal{C}_n$. In this case, the sequence of vectors

$$\mathbf{u}_r = \mathbf{v}_0 P_{\mathcal{C}_n}^r \ \ (r = 1, \ldots, l)$$

must be analyzed, where $\mathbf{v}_0$ is the initial distribution of the probabilities of the states of the FMC $\mathcal{C}_n$.

When we are modeling the DS $\mathcal{S}_n$ behavior it is supposed that initially $\mathcal{S}_n$ is in the fully functional stage, as a rule, i.e.

$$\mathbf{v}_0 = (1, \underbrace{0, \ldots, 0}_{n-1 \text{ times}}).$$

For each vector $\mathbf{u}_r = (u_1^{(r)}, \ldots, u_n^{(r)})$ $(r = 1, \ldots, l)$ the probabilities

$$\mathrm{P}_r(s_1, B_{k-1}) = \sum_{j=i_{k-2}+1}^{n-1} u_j^{(r)}$$

and

$$\mathrm{P}_r(s_1, s_n) = u_n^{(r)}$$

can be computed.

The sequence of pairs of probabilities

$$(\mathrm{P}_r(s_1, B_{k-1}), \mathrm{P}_r(s_1, s_n)) \ (r = 1, \ldots, l)$$

forms some base for analysis of bounded reachability properties [6, 7] for the investigated DS.

Indeed, let the thresholds λ_1 $(0 < \lambda_1 < 1)$ and λ_2 $(0 < \lambda_2 < 1)$ be given. Then the sequence of systems of inequalities

$$\begin{cases} \mathrm{P}_r(s_1, B_{k-1}) \le \lambda_1 \\ \mathrm{P}_r(s_1, s_n) \le \lambda_2 \end{cases} \qquad (r = 1, \ldots, l)$$

forms some base for development of any effective maintenance policy for the investigated DS.

4 Conclusion

In the given paper analytical models for recoverable and non-recoverable DS are proposed. It is shown how the methods of bounded reachability properties analysis can be applied for selection of any effective maintenance policy for the investigated DS.

References

[1] R. Barlow and L. Hunter. *Optimum preventive maintenance policies*. Operations Research, vol. 8, no. 1, 1960, pp. 90–100.

[2] T. Nakagawa. *Advanced reliability models and maintenance policies*. Springer-Verlag London Limited, 2008. – 246 p.

[3] B. Dowdeswell, R. Sinha, and S. G. MacDonell. *Finding faults: A scoping study of fault diagnostics for Industrial Cyber-Physical Systems*. Journal of Systems and Software, vol. 168, 2020, https://doi.org/10.1016/j.jss.2020.110638.

[4] F. A. Sonnenberg and J. R. Beck. *Markov models in medical decision making: a practical guide*. Med. Decis. Making, vol. 13, no. 4, 1993, pp. 322–338, https://doi: 10.1177/0272989X9301300409.

[5] R. C. Sato and M. Zouain. *Markov models in health care*. Einstein, vol. 8, no. 3, 2010, pp. 376–379, https://doi.org/10.1590/s1679-45082010rb1567.

[6] E. Abraham, B. Becker, C. Dehnert, N. Jansen, J.-P. Katoen, and R. Wimmer, *Counterexample generation for discrete-time Markov models: an introductory survey*. In: M. Bernardo, F. Damiani, R. Hahnle, E.B. Johnsen, and I. Schaefer (eds), Formal methods for executable software models. SFM 2014. Lecture Notes in Computer Science, vol. 8483, 2014, pp. 65–121.

[7] A. Biere, A. Cimatti, E. M. Clarke, O. Strichman, and Y. Zhu. *Bounded model checking*. Advances in Computers, vol. 58, 2003, pp. 118–149.

Volodymyr G. Skobelev[1], Volodymyr V. Skobelev[2]

[1]V.M. Glushkov Institute of Cybernetics of NAS of Ukraine,
40 Glushkova ave., Kyiv, Ukraine, 03187
E–mail: skobelevvg@gmail.com

[2]V.M. Glushkov Institute of Cybernetics of NAS of Ukraine,
40 Glushkova ave., Kyiv, Ukraine, 03187
E–mail: volodimirvskobelev@gmail.com

Proceedings of Workshop on Intelligent Information Systems
WIIS2020, December 04-05, 2020, Chisinau, Republic of Moldova

Parallel architecture and software by workflow Petri nets

Inga Titchiev

Abstract

In this article we illustrate the possibilities offered by Petri nets in modeling the issues related to distributed data processing system.

Keywords: Petri nets, workflow Petri nets, parallel computing, High performance computing.

1 Introduction

High performance computing (HPC) infrastructure development offers many advantages for solving problems in various fields (bioinformatics, physics, chemistry, mathematical modeling, web servers and databases, optimization of business decisions and medicine). Development of HPC is characterized not only by increasing the number of elements involved in data processing, but by presenting the relationship between them and the management of interactions by very complex structure. Such interactions contributed to the occurrence of new problems related to the analysis, modeling and representation of causal relations in complex systems with many objects acting in parallel. In order to integrate the physical (architecture) and logical (software) components of a distributed data processing system, a method like Petri nets will be used.

2 Software by workflow Petri nets formalism

Petri nets formalism, during over the five decades since they were introduced, showed great flexibility in dealing with many types of practical

problems and a great capacity for expansion by incorporating some more complex points of view. Modeling of software component, represented by block scheme, as usually, is performed at the component level. Transitions are associated with program actions: calculations and decisions. For the interpretation of a Petri net, an interpretation must be made for each transition. The translation of the elements of block scheme through workflow Petri nets is presented in Figure 1.

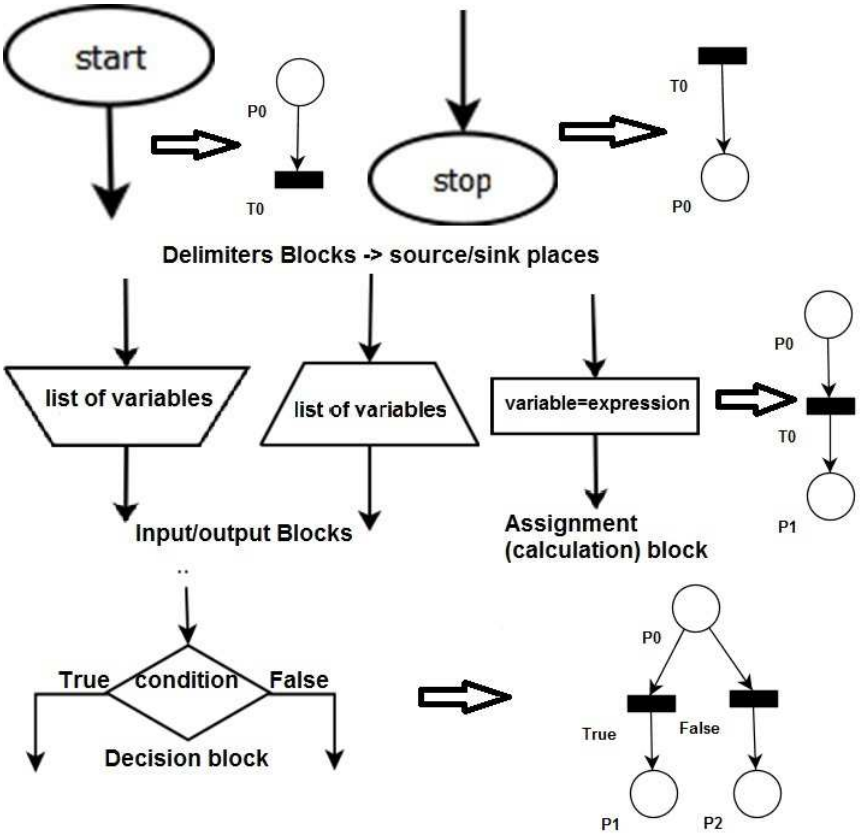


Figure 1. Block scheme elements by Petri nets

The transitions for the calculation actions have a single input and a single output, there can be no conflict for a transition representing a calculation, since its input location is no longer the input location for any other transition. Decision-making can bring conflicts into the net, but in a very narrow way: any choice can be made. The choice can be made either in non-deterministic way (i.e. randomly), or it can be controlled by the same external force (i.e. an agent) that calculates the truth or falsity of the decision and forces the triggering of the correct transition.

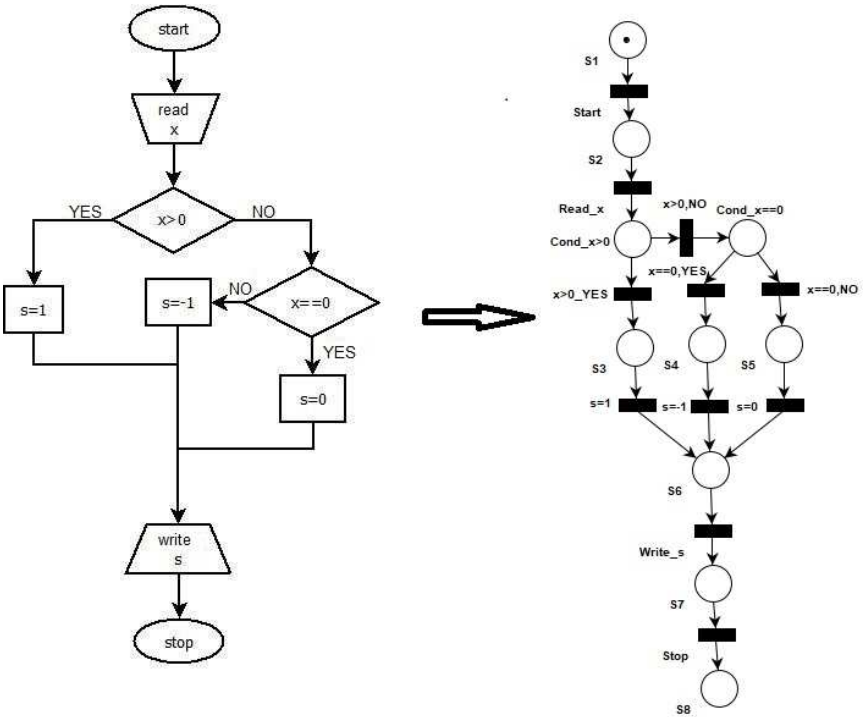


Figure 2. Bloch scheme and workflow Petri net

3 Parallel architecture

Ability to shape overlap and easily to combine subsystems using Petri nets, makes them useful for modeling of the more complex hardware components. Computational systems consist of many components, and many designers try to increase speed through parallel execution of certain functions. This is why Petri nets are a very suitable representation for such type of system. In order to convert application (or part of the application) from Sequential computing mode (when application runs on one computer and its run time depends on the computing capacity of the computer) to the parallel one, it is necessary to adopt one of the following parallelization technologies:

1. Parallel computing on one multicore computer – OpenMP technology.

2. Parallel computing on several computers, using one core on each (MPI Message Passing Interface technology).

3. Hybrid technology – using both OpenMP and MPI.

4. Distributed or Grid-computing – the ability to simultaneously run the same serial applications on multiple computers – distributed computing (computers of various capacities are united in a parallel computing system by the local and global networks).

Parallelism (competition) can be introduced in several ways. We consider the case of two concurrent processes. Each process can be represented by a workflow Petri net. Therefore, the composed Petri net, which is simply the union of Petri nets for each of the two processes, may represent concurrent execution of the two processes. Initial marking of compound workflow Petri net has two tokens, with one in each source place representing the beginning of each process. This fact introduces parallelism that can not be represented by a logical scheme and for which such representation as a Petri net is a very useful solution.

Another approach is to consider how parallelism can be introduced in a normal process in a computer system. Branching operations (FORK) and union (JOIN) ones, discussed in [1, 4], are considered.

Parallelism is useful in solving a problem if concurrent processes can cooperate in solving the problem. Such cooperation involves common processes information and resources. Shared access to information and resources must be controlled in order to ensure correct functioning of the system.

The Vladimir Andrunachievici Institute of Mathematics and Computer Science's cluster allows running Sequential, Parallel(Open MP, MPI and Hybrid) applications and Grid cluster.

A variety of coordination problems has been proposed in the literature to illustrate the types of problems [2, 3, 5] that may arise between cooperating processes.

4 Conclusion

In this article it was shown that Petri nets is a convenient formal method for modeling information flow represented by block scheme. Thus, it was shown how Petri nets have been used for representation of parallel processes in order to better understanding of these processes.

Acknowledgments. The research project 20.80009.5007.22, "Intelligent information systems for solving ill-structured problems, processing knowledge and big data" has supported part of the research for this paper.

References

[1] J.B. Dennis and E.C. Van Horn. *Programming Semantics for Multiprogrammed Computations*. Communications of the ACM, 9(3), pp. 143–155, March 1966.

[2] E. W. Dijkstra. *Solution of a problem in concurrent programming control.* Communications of the ACM, vol. 8(9), pp. 569, 1965.

[3] M. Raynal. *Algorithms for mutual exclusion.* The MIT Press, 1986. Translation of: Algorithmique du parallelisme, 1984.

[4] I. Titchiev, N.Iliuha, and M.Petic. *Workflow Petri nets used in modeling of Parallel architectures.* Proceedings of the International Conference on Intelligent Information Systems, August 20-23, Chisinau, Republic of Moldova, 2013, pp. 163–167.

[5] Syed Nasir Mehmood, Nazleeni Haron, Vaqar Akhtar, and Younus Javed. *Implementation and Experimentation of Producer-Consumer Synchronization Problem*, International Journal of Computer Applications (0975 – 8887), vol. 14, no.3, January 2011, pp. 32–37.

Inga Titchiev[1, 2]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science
[2] Tiraspol State University

E–mail: inga.titchiev@math.md

Methodology and models of assessment of costs necessary for development of electronic communications infrastructure

Grigore Varanița, Ilie Costaș, Anatol Godonoagă

Abstract

In this article, some cost assessment models are proposed and it is also emphasized that for the optimization of telecommunications networks, which is a problem with several parameters and variables, multiparametric and multi-criterium optimization methods and algorithms are usually applied. It is necessary that throughout the analysis of the results, some specific parameters, such as quality or reliability, are always controllable. It is worth mentioning that most variables in models are associated with quality parameters through a nonlinear relationship, often this relationship being of a default character.

Keywords: expenditure efficiency, electronic telecommunications networks, optimization methodology, quality criteria, electronic communications infrastructure.

Jel Classification: c3, c8

1. Introduction

The process of assessing the costs of a telecommunications network can be oriented in several directions. [1, p. 1361] One would be conditional on the design and creation of a new network. The second direction may be dictated by the need to reduce the consumption of resources, or the costs of an existing network, the third - to improve its emission quality and so on.

In this article, arising from the specific conditions in the Republic of Moldova [2, 3] generated by serious problems in the economic development of the country and, as a result, by the lack of necessary

financial resources, but also from the gap between different levels of development of telecommunications networks in different territorial areas of the country (especially in rural versus urban areas), a methodology is proposed to streamline the expenditure needed for the development of networks, but respecting modern requirements for the quality of the networks and services offered by them. At the same time, consideration is given to the technical aspects of the national information infrastructure, such as: the place and role of computer networks in modern society [4], macro synthesis of computer networks [5], the configuration of local server networks [6], models and algorithms for configuring the server set for local networks [7].

2. Methodology and Models for Evaluation of Costs which are for the Development of Telecommunication Networks

The issue of the development of national information infrastructure (IIN) cannot be solved uniformly throughout a country. It is all about the fact that the factors influencing the resolution of this problem are essentially different in different geographical areas. These factors relate to:

(a) intensity of local information flows;

(b) requirements for the quantity and quality of the services provided;

(c) density of public institutions and population in different areas;

(d) existing level of development of electronic communications networks in those areas;

(e) priorities conditional on the national economic development strategy of different areas;

f) budget available for such projects, etc.

In this context, in addressing the development of IIN, in terms of optimizing the expenditure needed for the development of the entire electronic communications infrastructure at national level, three steps can be highlighted:

1) Division of the entire territory of the country into geographical areas, each of which has specific and similar conditions and influence factors in terms of information needs, requirements for the quality criteria of electronic communications services, and other influence factors listed above.

This requires special investigations of these factors of influence and the development of concrete proposals for structuring of national territory with the approval of these areas. However, we believe that the starting point in this structuring must be recognition of the essential differences between the following levels of IIN:

a) Large administrative, scientific and cultural centers with high population density (Chisinau, Balti);

b) Middle cities (Orhei, Bender, Ungheni, etc.) and district centers;

c) Rural environment, which includes the rest of the country.

Admittedly, in determining geographical areas from the point of view of the development of INIs, in addition to these prior guidelines, the other factors of influence mentioned above must be considered. In particular, it is important that the current level of development of electronic communications networks is very different in different territories. As mentioned above, while users in urban areas benefit from a greater diversity of service provision, rural users face either limited choice or a total lack of access to the means of communication, even outside the coverage of mobile electronic communications networks.

So, even when the informational needs of users are similar in different areas of the country, the issue of the development of information infrastructure and the optimization of the necessary expenditure would be different because of the already different levels of development of electronic communications networks in these areas.

2) Assessment of the costs necessary for the development of electronic communications networks in each geographical area.

The aspects of mathematical cost assessment model in a separate geographical area will be listed. It is assumed that both wired and wireless networks could be developed in the same geographical area. Depending on the criteria and factors of influence, the cost assessment models described above are reflected in each related area. In the mathematical models already developed and described in [3], the geographic area index ZG_l is entered as an additional variable to those already discussed. The main component of the model for assessing the costs necessary for the development of the entire electronic communications infrastructure in the country is the model for calculating the costs required for the development of electronic communications networks in each geographical area**.**

Network cost relative to geographical area ZG_l:

$$CZG_l = K_l^1 \times CRCF_l + K_l^2 \times CRFF_l\,, \qquad (1)$$

where:

$CRCF_l$ expresses the cost of the wired network inZG_l;

$CRFF_l$ – the cost of the wireless network in ZG_l;

Coefficients $K_l^1, K_l^2 \in \{0,1\}$, $K_l^1 + K_l^2 \in \{1,2\}$.

Therefore, the possible variants for the (K_l^1, K_l^2) are: (1,0); (0,1); (1,1).

Formula (1) expresses the total cost of the network ($K_l^1 = 1, K_l^2 = 0$ *or* $K_l^1 = 0, K_l^2 = 1$), or networks ($K_l^1 = 1, K_l^2 = 1$) appropriate to ZG_l.

Remark: If a priori the network quality requirements are determined for the ZG_l, $QRZG_l$, then the problem of minimizing the total cost of the network for the given geographical area l has the aspect:

$$CZG_l \rightarrow min;$$
$$IQ^l \div QRZG_l\,,$$

where $CRCF_l$ has the aspect of the formula:

$$Cf_R = \sum_{i=1}^{n} Cf_i\,, \qquad (2)$$

adjusted to geographical area l;

$CRFF_l$ is presented by the formula:

$$C_R = \sum_{i=1}^{n(d)} C_i\,, \qquad (3)$$

the same adjusted in relation to the area$l$.

It's noted that the values $\boldsymbol{Cf_i}$ and $\boldsymbol{C_i}$ represent the total costs of implementing a base station in the given location *i* and geographical area for the corresponding wired and wireless networks, this area being made up of *n* locations.

Even if one of these types of networks were not included in the development strategy, the mathematical model would be the same as if both types of.

3) Assessment of the costs necessary for the development of the entire electronic communications infrastructure in the country.

After addressing the problem of optimizing expenditure for the development of electronic communications networks in each geographical

area taken separately, the total expenditure required for the development of electronic communications infrastructures throughout the country shall be calculated.

Given that not only expenditure for the development of electronic communications networks has been optimized at the previous stage, but also that it has been carried out in the assurance of all specific requirements in relation to the volume and quality of specific information services for those areas, we can conclude that the sum of the expenditure optimized in each geographical area of the country will present the total expenditure necessary for the development of the entire infrastructure of national electronic communications networks.

In this context, and as a result of the above-mentioned cost assessment and cost optimization models (wired and wireless), the total expenditure required for the development of the entire infrastructure of national electronic communications networks will be calculated according to an aggregated model. To do this, we will have the following notations:

ZG_l – geographical area l ($l = 1,2,3, \dots L$);
RZG_l – network of the geographical area l;
MZG_l – model for assessing network costs in the geographical area l;
CZG_l – cost of the remote network in the geographical area l;
$QRZG_l$ – network quality requirements for the geographical area l;
IQ^l – network quality indicators for the geographical area l.

The model for assessing the cost of the geographical area network can be presented succinctly in one of the following three forms:

$$MZG_l = \begin{cases} MCF_l \\ MFF_l \\ MCF_l \ \text{şi} \ MFF_l \end{cases}$$

where:

MCF_l expresses the model for assessing the cost of the wired network in the territorial area l;

MFF_l – model for assessing the cost of the wireless network in the territorial area l.

In some situations, it would be rational for each geographical area to be matched by a weight (priority of the given area) $ZG_l \rightarrow P_l > 0, l = \overline{1, L}$. By priority we can understand the order of execution of the works on

the respective areas depending on some objective and subjective factors, such as: development strategies, policy decisions, etc.

Let us have $1 \geq P2 \geq \cdots P_L$ *şi* $\sum_{l=1}^{L} P_l = 1$.

Let us admit that a clearly determined budget is allocated for the development of the national network B. A total cost will be allocated for each area C. If the area is *i*, for example, then it corresponds to the hexagonal coverage with the diameter of the hexagon determined by simulation and C_l is the cost of the network for ZG_l, based on Figure 1.

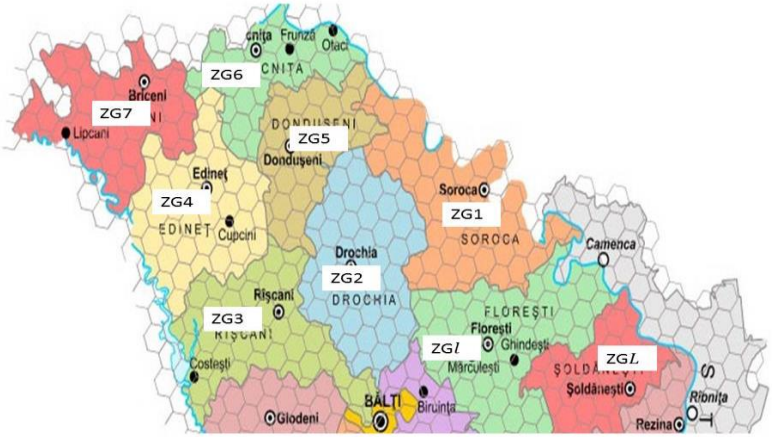


Figure 1. Representation of geographical areas
Source: author-produced

If the total cost of the national network simulated in the outcome of scenario S does not exceed budget B, then: $C_1^S + C_2^S + C_3^S + \cdots C_L^S \leq B$ and on any area the quality criteria submitted are met, then it can be said that budget B is sufficient to develop a functioning network throughout the national territory. In other words, the budget B allocated is sufficient for the development of the national network while being able to save part of this budget for other projects because it is obvious that in the best possible way, the total minimum cost of the national network will not exceed the budget B.

For the relation $IQ^l \div QRZG_l$ the following meaning is assigned: quality indicators IQ^l selected for the area l fully comply with the a priori

requirements for this area. When changing the radius of the hexagon in the direction of increasing its diameter, these indicators vary, and for some values of that diameter, some of these indicators may become inadmissible.

The general model, according to the notations can be rendered as follows: $CZG_1 + CZG_2 + \cdots CZG_l + \cdots CZG_L \rightarrow min$, provided that, $IQ^l \div QRZG_l$ $(l = 1,2,3, \dots L)$.

Remark: If a priori the network quality requirements are determined for the ZG_l, $QRZG_l$, then the problem of minimizing the total cost for all L geographical areas is: $\sum_{l=1}^{L} CZG_l \rightarrow min$; $IQ^l \div QRZG_l$ $(l = 1,2,3, \dots L)$, reduction to L optimization problems for the respective territorial areas:

$$CZG_l \rightarrow min; IQ^l \div QRZG_l \ (l = 1,2,3, \dots L).$$

In this obvious case, the relationship takes place:

$$min_{IQ^1,\dots,IQ^L} \sum_{l=1}^{L} CZG_l = \sum_{l=1}^{L} min_{IQ^l} CZG_l \ , \qquad (4)$$

which indicates that the total optimal cost of the telecommunications network per country is the sum of the minimum costs of the networks corresponding to the L geographical areas.

3. Conclusion

A mathematical cost assessment model for general case of development of telecommunications infrastructure for the entire territory of a country is proposed in this article. The following methodology has been proposed for the assessment of costs in accordance with this model and for the development of the entire national telecommunications infrastructure. Based on further research it is necessary to divide the entire territory of the country into areas with different requirements compared to the quality criteria of communications services. After that, depending on these criteria, the cost assessment models described above are fleshed out for each related area. Thus, the methodology examined will allow cost optimization problems to be addressed in accordance with the requirements in relation to the quality criteria.

References

[1] V.M. Vishnevskii and A.N. Dudin. *Queueing Systems with Correlated Arrival Flows and Their Applications to Modeling Telecommunication*

Networks. Automation and Remote Control, 2017. vol. 78, no. 8, pp. 1361–1403. ISSN 0005-1179.

[2] G. Varanița, A. Godonoaga, and I. Costas. *A model of wireline-based telecommunications network costs evaluation. (Model de evaluare a costurilor rețelelor de telecomunicații prin fir).* In: Competitivitate şi inovare în economia cunoaşterii: conferinţă ştiinţifică internaţională, *28-29 septembrie 2018.* Chișinău:ASEM, pp. 5.

[3] G. Varanița, A. Godonoaga, and I. Costas. *Models of wireline and wireless telecommunications network costs evaluation (modele de evaluare a costurilor rețelelor de telecomunicații prin fir și fără fir).* In: Economica, 2019, no. 2, pp. 10.

[4] I. Bolun. *Reţelele informatice în societatea modernă.* În: Tendinţe în dezvoltarea tehnologiilor informaţionale şi comunicaţionale în domeniul învăţământului şi managementului: conf.internaţ., 20-21 martie 2003. Chişinău: Editura ASEM, 2003. pp. 31-45.

[5] I. Bolun. *Macrosinteza reţelelor de calculatoare*. Chişinău: Editura Academia de Studii Economice din Moldova, 1999, 265 p.

[6] I. Bolun and A. Ciumac. *Configuration of local area network set of servers*. In: Computer Science Journal of Moldova, vol. 10, no. 2(29), 2002. Chisinau: IMCS. pp. 99-124 (1,6 c.a.).

[7] A. Ciumac. *Modele si algoritmi de configurare a setului de servere pentru reţelele locale.* Teză de doctor în tehnică. Teza a fost susţinută pe 30 noiembrie 2007 în CSS şi aprobată de CNAA pe 24 ianuarie 2008. ASEM. 2007. 142 p.

Grigore Varanița[1], Ilie Costaș[2], Anatol Godonoagă[3]

[1]Program Coordinator of the United Nations Conference on Trade and Development in Turkmenistan,
E-mail: grigore.varanita@gmail.com

[2]University professor, doctor habilitate ASEM, Chisinau, Republic of Moldova,
E-mail: costas.ilie@yahoo.com

[3]Associate Professor, Doctor ASEM, Chisinau, Republic of Moldova, 61, Banulescu Bodoni str., tel. (+373) 402 988, www.ase.md,
E-mail: anagodon22@yahoo.com

Researching and valorizing the lexicon of the Romanian Language in a general Romanian context

Olesea Caftanatov, Ludmila Malahov

Abstract

The aim of our research is to initiate the creation of lexicographic corpus consisting of dictionaries developed over the years by specialists from Institute of Philology. Additionally, we developed the project's website "LogosPlus", which includes online library, online dictionary, blog and other related resources.

Keywords: lexical borrowing, academic dictionaries, digitization, virtual library.

1. Introduction

Speaking about the active processes that occur in the lexicon of the current Romanian language, we mention the extension and restriction of the senses, metaphorizing, metonymizing, forming new words and meanings, and borrowings. *"The Romanian language is very open to novelty"*, thus, one of the priorities of the project is the creation of a bank of recent lexical acquisitions (based on the monitoring of the current print, electronic and audiovisual press in Moldova), in order to supplement the dictionaries with recent words and meanings.

At the same time, in today's computerized society, each language needs technological products to connect it to the international circuit of communicative techniques and tools. This is why, we developed a blog that will contain the post with results of our research (more about this is presented in Section 2).

Additionally, we intend to digitize dialectal texts and transliterate them in contemporaneous Romanian language alphabets. For this purpose, we digitize and transliterate the book "Тексте Диалектале" [1] (more about this see in Section 3).

2. Blog "Cultivating Romanian Language"

Due to the new technologies, the best solution for disseminating the results of our research is by creating digital text. Thus, we developed "Cultivating Romanian Language" as a blog, where each post contains the result of our investigation [2]. To date we created 36 posts that can be classified in seven main threads, such as lexical rules, grammar, morphology, orthography, semantics, syntax and stylistic, see Figure 1. Additionally, the content can be freely accessed and shared on other social media platforms such as Facebook, Twitter, LinkedIn etc.

Figure 1. The Blog's "Cultivating Romanian Language" interface

3. Digitization and transliteration

One of our project's goals is the creation of a lexicographic corpus consisting of dictionaries developed over the years by specialists from the Institute of Philology, especially from the sector of lexicology and

lexicography, which would complete the essential Romanian lexicographic corpus (100 dictionaries from the DLR bibliography).

Moreover, an important task regarding the creation of lexicographic corpus was to digitize and transliterate books from Soviet period [1, 3-5].

The most interesting in sense of technology development was the book "Тексте Диалектале" ("Dialectal text"), because in order to recognize phonetic features we created a special alphabet, that consists of Cyrillic alphabet, diacritics and conventional text.

Below we will describe few steps regarding whole digitization process. First, the digitization has several stages: scanning, post-processing of the scanned resources, preparation for recognition, optical character recognition (OCR), automatic and manual validation.

Regarding the first step, we carried out scanning with an accuracy of 600 dpi and the resulting files were placed in the Vladimir Andrunachievici Institute of Mathematics and Computer Science cloud. For post-processing step we used Scan Tailor application. It is an interactive post-processing tool for scanned pages. It performs operations, such as page splitting, deskewing, adding/ removing borders, binarising, cleaning, among others features [6].

Before OCR step, we prepared templates with special alphabet and extended existed dictionary. For recognition step, there was used ABBYY FineReader v.14, which was tuned to a specific phonetic alphabet of the Moldovan language (Romanian in Cyrillic). To improve recognition, we used a custom dictionary from the previously correctly recognized words from the book.

Automation validation was carried out in the editing mode with a hint from constructed dictionary. After that, only about 5-10% errors were corrected manually. To date, the recognition of volume 1 of the book has been completed. The texts are saved in the cloud: https://cloud.math.md/.

After recognition, the important task is transliteration into Latin alphabet. It should be noted, that there are dissimilarity of phonetic records of texts in Cyrillic and Latin, since in the phonetic alphabet of the Cyrillic alphabet there are Latin letters, Greek, thus, transmitting their non-standard pronunciation. Therefore, for transliteration we created 273 rules. Some of the rules for overlapping symbols can be seen in Table 1.

There was a problem of their classification, which was solved by successive partitions and recursion on test examples. However, for this process, manual validation is necessary in order to get good results.

Table 1. Examples of rules for overlapping symbols

а	→ a/o	э	→ ă/a	́э	→ ắ/P	т"	→ ä/W
о		а		Р		к"	
а	→ a/î	э	→ ă/e	¼	→ ¼/l	е	→ e/i
ы		е		л		и	
а	→ a/ă	́э	→ ắ/a	¡	→ ¡/n	é	→ é/i
э		а		н		и	
и	→ i/e	ы	→ î/i	ы	→ î/ă	h	→ M/h
е		и		э		х	

4. Virtual library

Another object of our research is creation of a virtual library (see Figure 2.A). This library will contain two data bases, one - for articles (see Figure 2.B) and another one - for monographs (see Figure 2.C).

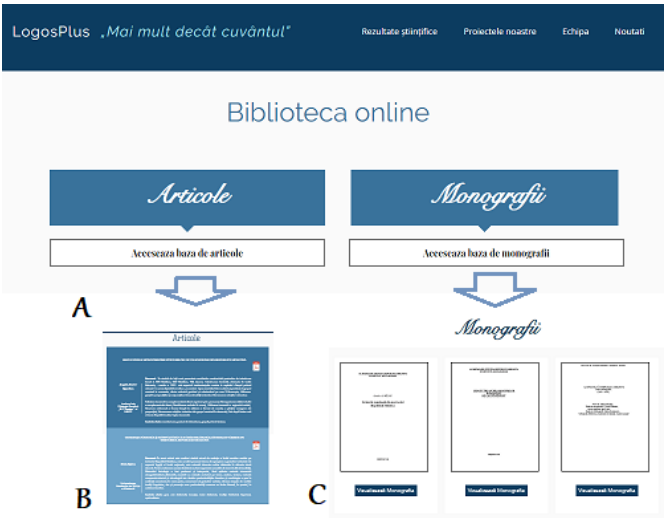


Figure 2. The Virtual Library Interfaces

5. Conclusion and future work

For our project, we recognize the book "Тексте Диалектале" vol. 1. Moreover, we elaborated 273 rules for transliteration. For future work we will continue to research and valorize the lexicon of the Romanian Language by digitizing and transliterating other books. The recognized books would be placed in our virtual library form our official project web site. In addition, we created a bank with lexical acquisition and will continue to extend it. The research results will be published on: "Cultivating Romanian Language" blog.

Acknowledgments. This article was written as part of the research project 20.80009.5007.22 "Intelligent information systems for solving ill-structured problems, processing knowledge and big data" and the project "Scientific valorization of the national linguistic heritage in the European integration context".

References

[1] V. Melnic, V. Stati, and P. Udler. *Texte dialectale*, vol 1. Chișinău, 1969, pp. 12-238.

[2] Project's web site Logos Plus. www.logosplus.org

[3] G. Buciușcanu. *Gramaticii limbii moldovenești*. Editura de Stat a Moldovei, Balta (1926).

[4] L. Madan. *Gramaticî moldovnescî* (1930). Partea I. Fonetica și Morfologhia. Tirișpolea, 1930.

[5] I. Cușmaunsa. *Gramatica limbii moldovenești*. Manual pentru școala necomplectă și mijlocie. Tiraspol, 1939.

[6] Scan Tailor official web site. https://scantailor.org/

Olesea Caftanatov[1], Ludmila Malahov[2]

[1]Vladimir Andrunachievici Institute of Mathematics and Computer Science
E-mail: olesea.caftanatov@math..md

[2]Vladimir Andrunachievici Institute of Mathematics and Computer Science
E-mail: ludmila.malahov@math.md

Contents