Proceedings MFOI-2017

Conference on Mathematical Foundations of Informatics

Institute of Mathematics and Computer Science
November 9-11, 2017, Chisinau, Moldova

CZU 51+004(082)

C 65

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: Prof. S.Cojocaru, Prof. C.Gaindric, Ph.D. I.Drugus.

Authors are fully responsible for the content of their papers.

Descrierea CIP a Camerei Naţionale a Cărţii

Conference on Mathematical Foundations of Informatics, November 9-11, 2017, Chisinau, Moldova : Proceedings MFOI-2017 / Inst. of Mathematics and Computer Science, Acad. of Sciences of Moldova ; ed.: S. Cojocaru [et al.]. – [Chişinău] : Institute of Mathematics and Computer Science, 2017 (Tipogr. "Valinex" SRL). – 182 p. : fig., tab.

Rez.: lb. engl. – Referinţe bibliogr. la sfârşitul art. şi în subsol. – 60 ex.

ISBN 978-9975-4237-6-2.

51+004(082)

ISBN 978-9975-4237-6-2

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

Results Regarding Cardinalities in FSM

Andrei Alexandru, Gabriel Ciobanu

Abstract

In this paper we present the consistency of various results regarding cardinalities and Dedekind-finiteness in a newly developed framework called Finitely Supported Mathematics.

Keywords: cardinality, Dedekind-Finiteness, FSM.

1 Finitely Supported Mathematics

Finitely Supported Mathematics (FSM) is a recently developed framework used for managing infinite structures by involving a finite number of characteristics. More exactly, in FSM we associate to each object a finite family of elements characterizing it, which is called its "finite support". The properties of each object are well described only by analyzing the properties of its "finite support".

Finitely Supported Mathematics has connections with the permutation models of Zermelo-Fraenkel set theory with atoms, with Fraenkel-Mostowski (FM) axiomatic set theory, with the theory of nominal sets (which is a Zermelo-Fraenkel (ZF) alternative to axiomatic FM set theory), and with the theory of admissible sets (because the sets constructed in FSM are hereditary finitely supported sets). More precisely, the theory of nominal sets originally developed over a fixed countable set of atoms is extended to a theory of invariant sets over a fixed infinite (possible non-countable) set of atoms. An invariant set $(X, \cdot)$ is actually a classical ZF set X equipped with an action $\cdot$ on X of the group of permutations of a (possible non-countable) fixed ZF set A (called the set of atoms), having the additional property that any element $x \in X$

is finitely supported. In a pair $(X, \cdot)$ formed by a ZF set X and a group action $\cdot$ on X of the group of all permutations of atoms, an arbitrary element $x \in X$ is finitely supported if there exists a finite family S of atoms such that any permutation of atoms that fixes S pointwise also leaves x invariant under the group action $\cdot$. A relation (or, particularly, a function) between two invariant sets is finitely supported if it is finitely supported as a subset of the Cartesian product of those two invariant sets. The theory of invariant sets allows us to define invariant algebraic structures (as invariant sets endowed with invariant algebraic laws) which are used in order to construct FSM [1]. Concretely, FSM represents a reformulation of the ZF algebra obtained by replacing '(infinite) structure' with 'invariant/finitely supported structure'. All the objects defined in FSM must be constructed according to the finite support principle. As proved in [1], not every ZF result can be directly rephrased in FSM. As a consequence, we cannot obtain a property in FSM only by involving a ZF result without an appropriate proof reformulated according to the finite support requirement (the proof should be itself internally consistent in FSM, and not retrieved from ZF). The algorithmic techniques for translating ZF results into FSM are described in [1]. By applying them we are able to present some results regarding cardinality and Dedekind-finiteness, in the world of finitely supported structures.

2 Cardinalities in FSM

In ZF, the trichotomy law for cardinalities states that any two cardinals are comparable. More precisely, the first ZF trichotomy principle states that for every two non-empty sets there is an one-to-one mapping from one into the other, while the second ZF trichotomy principle states that for every two non-empty sets there is a mapping of one onto the other. In ZF these principles are both equivalent with the axiom of choice. However, as we said before, the ZF relationship results (particularly those between choice and trichotomy) does not necessarily remain valid in FSM. Both trichotomy principles for cardinalities fail in FSM.

Theorem 2.1. *1. There may exist two non-empty invariant sets (particularly the set A of atoms and the set $\mathbb{N}$ of all positive integers) such that there is no finitely supported one-to-one mapping from one into the other.*

2. There may exist two non-empty invariant sets (particularly the set A of atoms and the set $\mathbb{N}$ of all positive integers) such that there is no finitely supported mapping of one onto the other.

In ZF set theory with the axiom of choice it is known that Tarski's theorem about choice holds, meaning that for any infinite set X there exists a bijection between X and $X \times X$. This result fails in FSM. The existence of right inverses for surjective functions also fail in FSM.

Theorem 2.2. *1. There exists an infinite invariant set X (particularly the set A of atoms) such that there is no finitely supported surjection $f : X \to X \times X$.*

2. There exists a finitely supported surjective function between two invariant sets which has no finitely supported right inverse.

Schröder-Bernstein theorem which emphasizes the antisymmetry of cardinal ordering remains consistent in FSM. While Schröder-Bernstein theorem is consistent in FSM, its dual is no longer valid in this new framework.

Theorem 2.3. *1. Let B and C be two invariant sets such that there exist a finitely supported one-to-one mapping $f : B \to C$ and a finitely supported one-to-one mapping $g : C \to B$. Then there exists a finitely supported bijective mapping $h : B \to C$.*

2. There are two invariant sets B and C such that there exists a finitely supported onto mapping $f : B \to C$ and a finitely supported onto mapping $g : C \to B$, but there does not exist a finitely supported bijective mapping $h : B \to C$.

3 Dedekind Finiteness in FSM

In ZF, a set is called Dedekind-infinite provided that there exists an injection into one of its proper subsets. Otherwise, it is called Dedekind-

finite. In ZF with choice the concepts 'finite' and 'Dedekind-finite' coincide. In FSM there exist infinite sets that are Dedekind-finite.

Theorem 3.1. *The concepts of finiteness and Dedekind-finiteness are different in FSM, meaning that there exists an infinite invariant set (particularly the set A of atoms) which has no finitely supported injection into any of its finitely supported proper subsets.*

Other examples of infinite invariant sets which are Dedekind-finite in FSM are the set of all finite subsets of A, the set of all finitely supported (i.e. either finite or cofinite) subsets of A, and the set of all finite injective tuples of atoms.

In ZF with the axiom of choice, a set X is finite if and only if every surjective function from X onto itself is one-to-one. Such a characterization of finite sets is not valid in FSM.

Theorem 3.2. *There exists an infinite invariant set X (particularly the set A of atoms) such that every finitely supported surjective function from X onto itself is also one-to-one.*

Corollary 3.3. *Let A be the infinite set of atoms in FSM. A finitely supported function $f : A \to A$ is injective if and only if it is surjective.*

4 Conclusion

The goal of this paper is to enrich the results presented in [1] by mentioning some consistency and inconsistency results regarding cardinalities and Dedekind-finiteness in Finitely Supported Mathematics.

References

[1] A. Alexandru, G. Ciobanu. *Finitely Supported Mathematics: An Introduction.* Springer, 2016.

Andrei Alexandru, Gabriel Ciobanu[1]

[1]Romanian Academy, Institute of Computer Science, Iasi, Romania
Email: gabriel@info.uaic.ro

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

Sequential Polarized Tissue P Systems with Vesicles of Multisets

Artiom Alhazov, Rudolf Freund, Sergiu Ivanov, Sergey Verlan

Abstract

We consider sequential polarized tissue P systems working on vesicles of multisets with the very simple operations of insertion, deletion, and substitution of single objects. With the whole multiset being enclosed in a vesicle, sending it to a target cell can be indicated in those simple rules working on the multiset. Polarizations -1, 0, 1 are sufficient for obtaining computational completeness.

1 Introduction and Preliminaries

For a comprehensive overview of different variants of (tissue) P systems and their expressive power we refer the reader to the handbook [4], and for a state of the art snapshot of the domain to the P systems website [6] as well as to the Bulletin series of the International Membrane Computing Society [5].

Very simple biologically motivated operations on strings are the so-called *point mutations*, i.e., *insertion*, *deletion*, and *substitution*. For example, these point mutations are used in *networks of evolutionary processors (NEPs)*. For an overview on hybrid NEPs and the best results known so far, we refer the reader to [2]. In *NEPs with polarizations*, each symbol has assigned a fixed integer value; the polarization of a string is computed according to a given evaluation function, and in the communication step the obtained string is moved to any of the connected cells having the same polarization, e,g. see [3]. In this extended

abstract, we recall the results from [1] for sequential polarized tissue P systems where a multiset is enclosed in a vesicle, point mutations are working on such a multiset in the evolution step, and the vesicle with the resulting multiset is moved from one cell to another one as a whole in the communication step.

2 Sequential Polarized Tissue P Systems with Vesicles of Multisets

In a polarized tissue P system Π working on vesicles of multisets, each cell gets assigned an elementary polarization from $\{-1, 0, 1\}$; each symbol from the alphabet V also has an integer polarization (again the elementary polarizations $\{-1, 0, 1\}$ are sufficient), but every terminal symbol from the terminal alphabet has polarization 0.

A *sequential polarized tissue P systems working on vesicles of multisets* (a *sequential ptPV* system for short) is a tuple $\Pi = (L, V, T, R, (i_0, w_0), h, \pi_L, \pi_V, \varphi)$ where

- L is a set of labels identifying in a one-to-one manner the $|L|$ cells of Π;
- V is the polarized alphabet of the system;
- T is the terminal alphabet of the system (the terminal symbols have no polarization, i.e., polarization 0);
- R is a set of rules of the form (i, p) where $i \in L$ and p is an insertion, deletion or substitution rule;
- (i_0, w_0) describes the initial vesicle containing the multiset w_0 in cell i_0;
- h is the (label of the) *output cell*;
- π_L is the function assigning an integer polarization(from$\{-1, 0, 1\}$) to each cell;
- π_V is the function assigning an (elementary) integer polarization (from $\{-1, 0, 1\}$) to each symbol in V;
- φ is the evaluation function yielding an integer value for each multiset.

Given a multiset, we need an evaluation function computing the polarization of the whole multiset from the polarizations of the symbols it contains; here we only use the evaluation function φ which computes the value of a multiset as the sum of the values of the symbols contained in it. Given the result m of this evaluation of the multiset in the vesicle, we apply the sign function $sign(m)$, which returns one of the values $+1/0/-1$, provided that m is a positive integer / is 0 / is a negative integer, respectively.

A derivation step in a ptPV system (as in (H)NEPs) consists of two substeps – the *evolution step*, with applying one rule from R, and the *communication step* with sending the vesicle to a cell with the same polarization as the multiset in it. As a special additional feature we require that the vesicle must not stay in the current cell even if its polarization would fit (if there is no other cell with a fitting polarization, the vesicle is eliminated from the system). As the terminal symbols have polarization 0, the output cell itself also has to have polarization 0. The computation of Π starts with a vesicle containing the multiset w_0 in cell i_0 (obviously, the initial multiset w_0 has to have the same polarization as the initial cell i_0), and the computation proceeds using the sequential derivation mode for the evolution steps until a the vesicle has arrived in the output cell h and only contains terminal symbols.

The familiy of sets of vectors of natural numbers generated by sequential ptPV systems is denoted by $Ps(ptPV, sequ)$. With $PsRE$ denoting the family of Parikh sets of recursively enumerable languages we obtain:

Theorem 1. $PsRE = Ps(ptPV, sequ)$.

References

[1] A. Alhazov, R. Freund, S. Ivanov, S. Verlan, (Tissue) P Systems with Vesicles of Multisets. In: *Proceedings AFL 2017,* to appear. 2017.

[2] A. Alhazov, R. Freund, V. Rogozhin, Yu. Rogozhin, Computational completeness of complete, star-like, and linear hybrid networks of evolutionary processors with a small number of processors. *Natural Computing* **15** (2016) 1, 51–68.
http://dx.doi.org/10.1007/s11047-015-9534-1

[3] R. Freund, V. Rogojin, S. Verlan, Computational Completeness of Networks of Evolutionary Processors with Elementary Polarizations and a Small Number of Processors. In: G. Pighizzini, C. Câmpeanu (eds.), *Descriptional Complexity of Formal Systems: 19th IFIP WG 1.02 International Conference, DCFS 2017, Milano, Italy, July 3-5, 2017, Proceedings*. Springer, 2017, 140–151.
https://doi.org/10.1007/978-3-319-60252-3_11

[4] Gh. Păun, G. Rozenberg, A. Salomaa (eds.), *The Oxford Handbook of Membrane Computing*. Oxford University Press, Oxford, England, 2010.

[5] Bulletin of the International Membrane Computing Society (IMCS). http://membranecomputing.net/IMCSBulletin/index.php.

[6] The P Systems Website. http://ppage.psystems.eu/.

Artiom Alhazov[1], Rudolf Freund[2], Sergiu Ivanov[3], Sergey Verlan[3]

[1]Institute of Mathematics and Computer Science
Academy of Sciences of Moldova
Academiei 5, Chişinău, MD-2028, Moldova
Email: artiom@math.md

[2] Faculty of Informatics, TU Wien
Favoritenstraße 9–11, 1040 Vienna, Austria
Email: rudi@emcc.at

[3]Laboratoire d'Algorithmique, Complexité et Logique,
Université Paris Est Créteil,
61 av. du général de Gaulle, 94010 Créteil, France
Email: sergiu.ivanov@u-pec.fr,verlan@u-pec.fr

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

P Systems and the Concept of Fairness

Artiom Alhazov, Rudolf Freund, Sergiu Ivanov

Abstract

We introduce a novel kind of P systems in which the application of rules in each step is controlled by a function on the applicable multisets of rules. Some examples are given to exhibit the power of this general concept. Moreover, for three well-known models of P systems we show how they can be simulated by P systems with a suitable fairness function.

1 Introduction

Membrane computing is a research field originally founded by Gheorghe Păun in 1998, see [6]. Membrane systems (also known as P systems) are a model of computing based on the abstract notion of a membrane and the rules associated to it which control the evolution of the objects inside. In many variants of P systems, the objects are plain symbols from a finite alphabet, but P systems operating on more complex objects (e.g., strings, arrays) have been considered, too, e.g., see [3].

A comprehensive overview of different flavors of membrane systems and their expressive power is given in the handbook, see [7]. For a state of the art snapshot of the domain, we refer the reader to the P systems website [10] as well as to the bulletin series of the International Membrane Computing Society [9].

In this paper we introduce a novel kind of P systems in which the application of rules in each step is controlled by a function on the applicable multisets of rules, possibly also depending on the current configuration; we call this function the *fairness function*. In the standard

A preliminary version of this paper was published as [2].

variant, the fairness function will be used to choose those applicable multisets for which the fairness function yields the minimal value.

After recalling some preliminary notions and definitions in the next section, in Section 3 we will define our new model of *fair P systems* and give some examples to exhibit the power of this general concept. In Section 4, for three well-known models of P systems we will show how they can be simulated by P systems with a suitable fairness function. Future research topics finally are touched in Section 5.

2 Preliminaries

In this paper, the set of positive natural numbers $\{1, 2, \dots\}$ is denoted by $\mathbb{N}_+$, the set of natural numbers also containing 0, i.e., $\{0, 1, 2, \dots\}$, is denoted by $\mathbb{N}$. The set of integers denoted by $\mathbb{Z}$.

An *alphabet* V is a finite set. A (non-empty) *string* s over an alphabet V is defined as a finite ordered sequence of elements of V.

A *multiset* over V is any function $w : V \to \mathbb{N}$; $w(a)$ is the *multiplicity* of a in w. A multiset w is often represented by one of the strings containing exactly $w(a)$ copies of each symbol $a \in V$; the set of all these strings representing the multiset w will be denoted by $str(w)$. The set of all multisets over the alphabet V is denoted by V°. By abusing string notation, the empty multiset is denoted by λ.

The families of sets of Parikh vectors as well as of sets of natural numbers (multiset languages over one-symbol alphabets) obtained from a language family F are denoted by PsF and NF, respectively. The family of recursively enumerable string languages is denoted by RE.

For further introduction to the theory of formal languages and computability, we refer the reader to [7, 8].

2.1 (Hierarchical) P Systems

A *hierarchical P system* (P system, for short) is a construct

$$\Pi = (O, T, \mu, w_1, \dots, w_n, R_1, \dots R_n, h_i, h_o),$$

where O is the alphabet of objects, $T \subseteq O$ is the alphabet of terminal objects, μ is the membrane structure injectively labeled by the numbers from $\{1, \ldots, n\}$ and usually given by a sequence of correctly nested brackets, w_i are the multisets giving the initial contents of each membrane i ($1 \leq i \leq n$), R_i is the finite set of rules associated with membrane i ($1 \leq i \leq n$), and h_i and h_o are the labels of the input and the output membranes, respectively ($1 \leq h_i \leq n$, $1 \leq h_o \leq n$).

In the present work, we will mostly consider the *generative case*, in which Π will be used as a multiset language-generating device. We therefore will systematically omit specifying the input membrane h_i.

Quite often the rules associated with membranes are multiset rewriting rules (or special cases of such rules). Multiset rewriting rules have the form $u \to v$, with $u \in O^o \setminus \{\lambda\}$ and $v \in O^o$. If $|u| = 1$, the rule $u \to v$ is called *non-cooperative*; otherwise it is called *cooperative*. Rules may additionally be allowed to send symbols to the neighboring membranes. In this case, for rules in R_i, $v \in O \times Tar_i$, where Tar_i contains the targets *out* (corresponding to sending the symbol to the parent membrane), *here* (indicating that the symbol should be kept in membrane i), and in_j (indicating that the symbol should be sent into the child membrane j of membrane i).

In P systems, rules are often applied in the maximally parallel way: in any derivation step, a non-extendable multiset of rules has to be applied. The rules are not allowed to consume the same instance of a symbol twice, which creates competition for objects and may lead to the P system choosing non-deterministically between the maximal collections of rules applicable in one step.

A computation of a P system is traditionally considered to be a sequence of configurations it successively can pass through, stopping at the halting configuration. A halting configuration is a configuration in which no rule can be applied any more, in any membrane. The result of a computation of a P system Π as defined above is the contents of the output membrane h_o projected over the terminal alphabet T.

Example 1. *For readability, we will often prefer a graphical representation of P systems; moreover, we will use labels to identify the rules.*

For example, the P system $\Pi_1 = (\{a, b\}, \{b\}, [\,]_1, a, R_1, 1)$ *with the rule set* $R_1 = \{1 : a \to aa, 2 : a \to b\}$ *may be depicted as in Figure 1.*

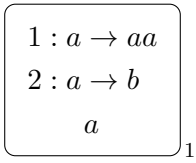


Figure 1. The example P system Π_1

Due to maximal parallelism, at every step Π_1 *may double some of the symbols* a*, while rewriting some other instances into* b*.*

Note that, even though Π_1 *might express the intention of generating the set of numbers of the powers of two, it will actually generate the whole of* $\mathbb{N}_+$ *(due to the halting condition). Indeed, for any* $n \in \mathbb{N}_+$*,* a^n *can be generated in* n *steps by choosing to apply, in the first* $n - 1$ *steps,* $1 : a \to aa$ *to exactly one instance of* a *and* $a \to b$ *to all the other instances, and by applying* $2 : a \to b$ *to every* a *in the last step (in fact, for* $n > 1$*, in each step except the last one, in which* $2 : a \to b$ *is applied twice, both rules are applied exactly once, as exactly two symbols* a *are present, whereas all other symbols are copies of* b*).* □

While maximal parallelism and halting by inapplicability have been standard ingredients from the beginning, various other derivation modes and halting conditions have been considered for P systems, e.g., see [7].

2.2 Flattening

The folklore flattening construction (see [7] for several examples as well as [4] for a general construction) is quite often directly applicable to many variants of P systems. Hence, also for the systems considered in this paper we will not explicitly mention how results are obtained by flattening.

3 P Systems with a Fairness Function

In this section we consider variants of P systems using a so-called *fairness function* for choosing a multiset of rules out of the set of all multisets of rules applicable to a configuration.

3.1 The General Idea of a Fairness Function in P Systems

Take any (standard) variant of P systems and any (standard) derivation mode. The application of a multiset of rules in addition can be guided by a function computed based on specific features of the underlying configuration and of the multisets of rules applicable to this configuration. The choice of the multiset of rules to be applied then depends on the function values computed for all the applicable multisets of rules.

Therefore, in general we extend the model of a hierarchical P system to the model of a *hierarchical P system with fairness function* (*fair P system* for short)

$$\Pi = (O, T, \mu, w_1, \ldots, w_n, R_1, \ldots R_n, h_i, h_o, f),$$

where f is the fairness function defined for any configuration C of Π, the corresponding set $Appl_\delta(\Pi, C)$ of multisets of rules from Π applicable to C in the given derivation mode δ, and any multiset of rules $R \in Appl_\delta(\Pi, C)$. We then use the values $f(C, Appl_\delta(\Pi, C), R)$ for all $R \in Appl_\delta(\Pi, C)$ to choose a multiset $R' \in Appl_\delta(\Pi, C)$ of rules to be applied to the underlying configuration C. A standard option for choosing R' is to require it to yield the minimal value for the fairness function, i.e., we require $f(C, Appl_\delta(\Pi, C), R') \leq f(C, Appl_\delta(\Pi, C)), R)$ for all $R \in Appl_\delta(\Pi, C)$. As usually the derivation mode δ will be obvious from the context, we often shall omit it.

The fairness function may be independent from the underlying configuration, i.e., we may write $f(Appl(\Pi, C), R)$ only; in the simplest case, f is even independent from $Appl(\Pi, C)$, hence, in this case we only write $f(R)$.

Fair or Unfair

One may argue that it is fair to use rules in such a way that each rule should be applied if possible and, moreover, all rules should be applied in a somehow *balanced* way. Hence, a fairness function for applicable multisets should compute the best value for those multisets of rules fulfilling these guidelines.

On the other hand, we may choose the multiset of rules to be applied in such a way that it is the *unfairest* one. In this sense, let us consider the following *unfair example*.

Example 2. *Consider the P system* $\Pi_1 = (\{a, b\}, \{b\}, [\,_1\,]_1, a, R_1, 1)$ *with the rule set* $R_1 = \{1 : a \to aa, 2 : a \to b\}$ *as considered in Example 1 together with the fairness function* f_2 *defined as follows: if a rule is applied* n *times then it contributes to the function value of the fairness function* f_2 *for the multiset of rules with* 2^{-n}*. The total value for* $f_2(R)$ *for a multiset of rules* R *containing* k *copies of rule* $1 : a \to aa$ *and* m *copies of rule* $2 : a \to b$ *then is the sum* $2^{-k} + 2^{-m}$*. The resulting fair P system* $\Pi_2 = (\{a, b\}, \{b\}, [\,_1\,]_1, a, R_1, 1, f_2)$ *is depicted in Figure 2; we observe that it can also be written as* (Π_1, f_2)*.*

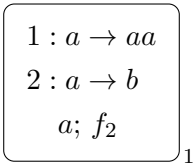


Figure 2. The P system Π_2

In this fair P system (or in this case we might also call it maximally unfair) with one membrane working in the maximally parallel way, we again start with the axiom a *and use the two rules* $1 : a \to aa$ *and* $2 : a \to b$*. If we apply only one of these rules to all* m *objects* a*, then the function value is* 2^{-m} *and is minimal compared to the function values computed for a mixed multiset of rules using both rules at least once.*

Starting with the axiom a we use the rule $1 : a \to aa$ in the maximal way k times thus obtaining 2^k symbols a. Then in the last step, for all a we use the rule $2 : a \to b$ thus obtaining 2^k symbols b. We cannot mix the two rules in one of the derivation steps as only the clean use of exactly one of them yields the minimal value for the fairness function.

We observe that the effect is similar to that of controlling the application of rules by the well-known control mechanism called label selection, e.g., see [5], where either the rule with label 1 or the rule with label 2 has to be chosen. We will return to this model in Section 4.3. □

The following weird example shows that the fairness function should be chosen from a suitable class of (at least recursive) functions, as otherwise the whole computing power comes from the fairness function:

Example 3. *Take the fair P system Π_3 with one membrane working in the maximally parallel way, starting with the axiom a and using the three rules $1 : \ a \to aa$, $2 : \ a \to a$, and $3 : \ a \to b$, see Figure 3.*

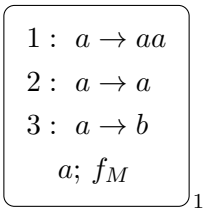


Figure 3. The P system Π_3

Moreover, let $M \subset \mathbb{N}_+$, i.e., an arbitrary set of positive natural numbers. The fairness function f_M on multisets of rules over these three rules and a configuration containing m symbols a is defined as follows: For any multiset of rules R containing copies of the rules $1 : \ a \to aa$, $2 : \ a \to a$, and $3 : \ a \to b$,

- *$f(R) = 0$ if R only contains m copies of rule 3 and $m \in M$,*

- $f(R) = 0$ *if* R *only contains exactly one copy of rule* 1 *and the rest are copies of rule* 2,
- $f(R) = 1$ *for any other applicable multiset of rules.*

Again the choice is made by applying only multisets of rules which yield the minimal value $f(R) = 0$. *If we use rule* $1 : \ a \to aa$ *once and rule* $2 : \ a \to a$ *for the rest, this increases the number of symbols* a *in the skin membrane by one. Thus, in* $m - 1$ *steps we get* m *symbols* a. *If* m *is in* M, *we now may use rule* $3 : \ a \to b$ *for all symbols* a, *thus obtaining* m *symbols* b, *and the system halts. In that way, the system generates exactly* $\{b^m \mid m \in M\}$.

To make this example a little bit less weird, we may only allow computable sets M. *Still, the whole computing power is in the fairness function* f_M *alone, with* f_M *only depending on the multiset of rules.* □

We now again return to Example 2 and illustrate how the same result can be obtained by using another fairness function in the standard *unfair* mode using the multsets of rules with minimal fairness value; on the other hand, we will also show what happens if we try to be *fair* and use the rules in a *balanced* way.

Example 4. *Consider the* P *system* $\Pi_1 = (\{a, b\}, \{b\}, [_1 \,]_1, a, R_1, 1)$ *with the rule set* $R_1 = \{1 : \ a \to aa, 2 : \ a \to b\}$ *as considered in Example 1 together with the fairness function* $f_4(R)$ *for any multiset* R *of rules defined as follows: consider* $f_4(R) = |str(R)|$, *i.e.,* $f_4(R)$ *is the number of different strings representing the multiset* R. *The resulting fair* P *system* $\Pi_4 = (\Pi_1, f_4) = (\{a, b\}, \{b\}, [_1 \,]_1, a, R_1, 1, f_4)$ *is depicted in Figure 4.*

With the standard selection of multisets of rules to be applied by choosing those with the minimal value of the fairness function, we obtain the same result for the set of multisets generated by Π_4, *i.e.,* $\{a^{2^n} \mid n \in \mathbb{N}\}$, *because only the pure multisets of rules* R *containing only copies of rule* 1 *or only copies of rule* 2 *yield* $f(R) = 1$, *whereas any mixed multiset of rules containing both rules at least once yields a bigger value.*

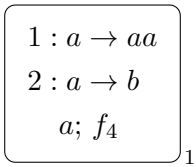


Figure 4. The P system Π_4

On the other hand, if we try to be fair *and use both rules in a balanced way, i.e., by choosing those multisets of rules yielding the maximum values of* f_4*, then the generated set is the singleton* $\{b\}$*, which can be generated in one step from the axiom* a *by using rule* $2 : a \to b$*. Any other derivation starting with using rule* $1 : a \to aa$ *will not yield any result due to running into an infinite computation without any chance to halt: as soon as* aa *has been generated, only once the rule* $1 : a \to aa$ *and once the rule* $2 : a \to b$ *can be used as only this combination of rules yields* $f_4(\langle 1, 2\rangle) = |\{12, 21\}| = 2 > 1 = f_4(\langle 1, 1\rangle) = f_4(\langle 2, 2\rangle)$ *(we here use the brackets* $\langle, \rangle$ *to describe a multiset).* □

The problem with halting observed in the example above when using only non-cooperative rules seems to be an inherent one when using a *fair* (balanced) selection of multisets of rules. These variants may deserve further investigations in the future, but in this paper we will restrict ourselves to the standard (*maximally unfair*) selection of multisets of rules to be applied as in the previous examples.

4 First Results

In this section, we show three general results. The first one describes how priorities can be simulated by a suitable fairness function in P systems of any kind working in the sequential mode. The second one exhibits how P systems with energy control, see [1], can be simulated by suitable fair P systems for any arbitrary derivation mode. Finally

we show how P systems with rule label control, see [5], can be simulated by suitable fair P systems for any arbitrary derivation mode.

4.1 Simulating Priorities in the Sequential Derivation Mode

In the sequential derivation mode, exactly one rule is applied in every derivation step of the P system Π. Given a configuration C and the set of applicable rules $Appl(\Pi, C)$ not taking into account a given priority relation $<$ on the rules, we define the fairness function to yield 0 for each rule in $Appl(\Pi, C)$ for which no rule in $Appl(\Pi, C)$ with higher priority exists, and 1 otherwise. Thus, only a rule with highest priority can be applied. More formally, this result now is proved for any kind of P systems working in the sequential derivation mode:

Theorem 1. *Let $(\Pi, <)$ be a P system of any kind with the priority relation $<$ on its rules and working in the sequential derivation mode. Then there exists a fair P system (Π, f) with fairness function f simulating the computations in $(\Pi, <)$ selecting the multisets of rules with minimal values.*

Proof. First we observe that the main ingredient Π is exactly the same in both $(\Pi, <)$ and (Π, f), i.e., we only replace the priority relation $<$ by the fairness function f. As already outlined above, for any configuration C of Π we now define f for any rule r as follows (we point out that here the fairness function not only depends on $\{r\}$, but also on $Appl(\Pi, C)$):

- $f(Appl(\Pi, C), \{r\}) = 0$ if and only if there exists no rule $r' \in Appl(\Pi, C)$ such that $r < r'$, and
- $f(Appl(\Pi, C), \{r\}) = 1$ if and only if there exists a rule $r' \in Appl(\Pi, C)$ such that $r < r'$.

If we now define the task of f as choosing only those rules with minimal value, i.e., a rule r can be applied to configuration C if and only if $f(Appl(\Pi, C), \{r\}) = 0$, then we obtain the desired result. □

4.2 Simulating Energy Control

Recently we have considered P systems where a specific amount of energy is assigned to each rule, see [1]. There, only those multisets of rules are applied which use the minimal amount of energy. In a similar way the amount of energy coming up with a multiset of rules can be seen as the value of the fairness function. The minimal amount of energy then exactly corresponds with the minimal fairness.

In this paper, from the two variants of energy-controlled P systems we only consider the one where the energy is directly assigned to the rules. This variant of P systems is called a *rule energy-controlled* P system. The multisets or sets of rules to be applied to a given configuration must fulfill the condition of yielding the minimal amount of energy.

Formally, in a rule energy-controlled P system the rules are of the form (p, v) where p is a rule of a specific type like cooperative or non-cooperative and v is an integer energy value. The total energy value of a mutiset of rules can be defined in different ways, but in the following we will assume it to simply be the sum of energy values of the rules in the multiset and denote this function computing the energy value of a multiset of rules in this way by σ.

Theorem 2. *Let (Π, σ) be a rule energy-controlled P system working in any derivation mode, using any kind of rules and using the sum function σ for computing the energy value of a multiset of rules. Then there exists a fair P system (Π', f) with fairness function f simulating the computations in (Π, σ) with f selecting the multisets of rules with minimal values.*

Proof. By definition, in the rule energy-controlled P system (Π, σ) a multiset of rules can be applied to given configuration only if the application of σ yields the minimal value in $\mathbb{Z}$. The fair P system (Π', f) with fairness function f now is constructed from (Π, σ) by replacing any rule with energy (p, v) by the rule p itself, but on the other hand defining the fairness function f for a multiset of rules to take v as the

value assigned to the rule p having been obtained from (p, v). By summing up these values for the whole multiset and selecting only those multisets of rules applicable to a given configuration in the given derivation mode which have minimal values, f fulfills the same task in (Π', f) as σ does in (Π, σ). Hence, in any derivation mode, (Π', f) simulates exactly step by step the derivations in (Π, σ), obviously yielding the same computation results. □

4.3 Simulating Label Selection

In P systems with label selection only rules belonging to one of the predefined subsets of rules can be applied to a given configuration, see [5].

For all the variants of P systems defined in Section 2, we may consider to label all the rules in the sets $R_1, \ldots, R_m$ in a one-to-one manner by labels from a set H and to take a set W containing subsets of H. Then a *P system with label selection* is a construct

$$\Pi^{ls} = (O, T, \mu, w_1, \ldots, w_n, R_1, \ldots R_n, h_i, h_o, H, W),$$

where $\Pi = (O, T, \mu, w_1, \ldots, w_n, R_1, \ldots R_n, h_i, h_o)$ is a P system as in Section 2, H is a set of labels for the rules in the sets $R_1, \ldots, R_m$, and $W \subseteq 2^H$. In any transition step in Π^{ls} we first select a set of labels $U \in W$ and then apply a non-empty multiset R of rules applicable in the given derivation mode restricted to rules with labels in U.

The following proof exhibits how the fairness function can also be used to capture the underlying derivation mode.

Theorem 3. *Let (Π, H, W) be a P system with label selection using any kind of rules in any kind of derivation mode. Then there exists a fair P system (Π', f) with fairness function f simulating the computations in (Π, H, W) with f selecting the multisets of rules with minimal values.*

Proof. By definition, in the P system (Π, H, W) with label selection a multiset of rules can be applied to given configuration only if all the

rules have labels in a selected set of labels $U \in W$. We now consider the set of all multisets of rules applicable to a configuration C, denoted by $Appl_{asyn}(\Pi, C)$, as it corresponds to the asynchronous derivation mode (abbreviated *asyn*); from those we select all R which obey to the label selection criterion, i.e., there exists a $U \in W$ such that the labels of all rules in R belong to U, and then only take those which also fulfill the criteria of the given derivation mode restricted to rules with labels from U.

Hence we define (Π', f) by taking $\Pi' = \Pi$ and, for any derivation mode δ, f_δ for any multiset of rules $R \in Appl_{asyn}(\Pi, C)$ as follows:

- $f_\delta(C, Appl_{asyn}(\Pi, C), R) = 0$ if there exists a $U \in W$ such that the labels of all rules in R belong to U, and, moreover, $R \in Appl_\delta(\Pi_U, C)$, where Π_U is the restricted version of Π only containing rules with labels in U, as well as
- $f_\delta(C, Appl_{asyn}(\Pi, C), R) = 1$ otherwise.

According to our standard selection criterion, we choose only those multisets of rules where the fairness function yields the minimal value 0, i.e., those R such that there exists a $U \in W$ such that the labels of all rules in R belong to U and R is applicable according to the underlying derivation mode with rules restricted to those having a label in U, which exactly mimicks the way of choosing R in (Π, H, W). Therefore, in any derivation mode δ, (Π', f_δ) simulates exactly step by step the derivations in (Π, H, W), obviously yielding the same computation results. □

5 Conclusions and Future Research

In this article, we introduced and partially studied P systems with the application of rules in each step being controlled by a function on the applicable multisets of rules.

We have given several examples exhibiting the power of using suitable fairness functions. Moreover, we have shown how priorities can

be simulated by a suitable fairness function in P systems of any kind working in the sequential mode as well as how P systems with energy control or label selection can be simulated by fair P systems with a suitable fairness function for any derivation mode.

Yet with all these examples and results we have just given a glimpse on what could be investigated in the future for P systems in connection with fairness functions:

- consider other variants of hierarchical P systems working in different derivation modes, e.g., also taking into consideration the set derivation modes;
- extend the notion of *fair* to tissue P systems, i.e., P systems on an arbitrary graph structure;
- extend the notion of fair to P systems with active membranes, there probably also controlling the division of membranes;
- investigate the effect of selecting the multiset of rules to be applied to a given configuration by other criteria than just taking those yielding the minimal values for the fairness function;
- consider other variants of fairness functions, either less powerful or taking into account other features of $Appl(\Pi, C)$ and/or the multiset of rules R;
- investigate the effect of selecting the multiset to be applied to a given configuration by requiring it to contain a balanced (really *fair*) amount of copies of each applicable rule;
- show similar simulation results with suitable fairness functions as in Section 4 for other control mechanisms used in the area of P systems;
- ...

References

[1] Artiom Alhazov, Rudolf Freund, and Sergiu Ivanov. Variants of energy-controlled P systems. In *Proceedings NIT 2016*, 2016.

[2] Artiom Alhazov, Rudolf Freund, and Sergiu Ivanov. Unfair P systems. In *Proceedings BWMC 2017*, 2017.

[3] Rudolf Freund. P systems working in the sequential mode on arrays and strings. In Cristian Calude, Elena Calude, and Michael J. Dinneen, editors, *Developments in Language Theory, 8th International Conference, DLT 2004, Auckland, New Zealand, December 13-17, 2004, Proceedings*, volume 3340 of *Lecture Notes in Computer Science*, pages 188–199. Springer, 2004.

[4] Rudolf Freund, Alberto Leporati, Giancarlo Mauri, Antonio E. Porreca, Sergey Verlan, and Zandron. Flattening in (tissue) P systems. In Artiom Alhazov, Svetlana Cojocaru, Marian Gheorghe, Yurii Rogozhin, Grzegorz Rozenberg, and Arto Salomaa, editors, *Membrane Computing*, volume 8340 of *Lecture Notes in Computer Science*, pages 173–188. Springer, 2014.

[5] Rudolf Freund, Marion Oswald, and Gheorghe Păun. Catalytic and purely catalytic P systems and P automata: Control mechanisms for obtaining computational completeness. *Fundamenta Informaticae*, 136(1-2):59–84, 2015.

[6] Gheorghe Păun. Computing with Membranes. *Journal of Computer and System Sciences*, 61:108–143, 1998.

[7] Gheorghe Păun, Grzegorz Rozenberg, and Arto Salomaa. *The Oxford Handbook of Membrane Computing*. Oxford University Press, Inc., New York, NY, USA, 2010.

[8] Grzegorz Rozenberg and Arto Salomaa, editors. *Handbook of Formal Languages, 3 volumes*. Springer, New York, NY, USA, 1997.

[9] Bulletin of the International Membrane Computing Society (IMCS). http://membranecomputing.net/IMCSBulletin/index.php.

[10] The P Systems Website. http://ppage.psystems.eu/.

Artiom Alhazov[1], Rudolf Freund[2], Sergiu Ivanov[3]

[1]Institute of Mathematics and Computer Science
Academy of Sciences of Moldova
Academiei 5, Chişinău, MD-2028, Moldova
Email: artiom@math.md

[2] Faculty of Informatics, TU Wien
Favoritenstraße 9–11, 1040 Vienna, Austria
Email: rudi@emcc.at

[3]Laboratoire d'Algorithmique, Complexité et Logique,
Université Paris Est Créteil,
61 av. du général de Gaulle, 94010 Créteil, France
Email: sergiu.ivanov@u-pec.fr

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

P Systems with Random RHS Exchange

Artiom Alhazov, Rudolf Freund, Sergiu Ivanov

Abstract

P systems are a model of hierarchically compartmentalized multiset rewriting. We present a novel kind of P systems in which rules are dynamically constructed in each step by non-deterministic pairing of left-hand and right-hand sides. It turns out that this variant enables non-cooperative P systems to generate exponential (and thus non-semilinear) number languages.

1 Introduction and Preliminaries

For a comprehensive overview of different variants of P systems and their expressive power we refer the reader to the handbook [3], and for a state of the art snapshot of the domain to the P systems website [5] as well as to the Bulletin series of the International Membrane Computing Society [4].

Dynamic evolution of the set of available rules has been considered from the very beginning of membrane computing. Already in 1999, generalized P systems were introduced in [2]. We remark, however, that the previous studies on dynamic rule sets either treated the rules as atomic entities (symport/antiport of rules, operators in generalized P systems), or allowed virtually unlimited possibilities of tampering with their shape (polymorphic P systems). In the present work, we propose a yet different approach which can be seen as an intermediate one.

In *P systems with randomized rule-right-hand sides* (or with randomized RHS, for short), the available left-hand sides and right-hand

sides of rules are fixed, but the associations between them are *re-evaluated in every step*: a left-hand side may pick a right-hand side arbitrarily (randomly).

In this extended abstract, we focus on the expressive power of P systems with randomized RHS, as well as on comparing them to the classical model with or without cooperative rules. One of the central conclusions of the present work is that non-cooperative P systems with randomized RHS can generate *exponential* number languages, thus (partially) surpassing the power of conventional (transitional) P systems. More details about the possible definitions of P systems with randomized RHS as well as about their expressive power can be found in the original article [1].

2 P Systems with Random RHS Exchange

In this variant of transitional P systems, rules randomly exchange right-hand sides at the beginning of every evolution step. This variant was the first to be conceived and is the closest to the classical definition [1].

A *transitional P system with random RHS exchange* is a construct

$$\Pi = (O, T, \mu, w_1, \ldots, w_n, R_1, \ldots R_n, h_o),$$

where the components of the tuple are defined as in the classical model.

As different from conventional transitional P systems, Π does not apply the rules from R_i directly. Instead, Π *non-deterministically permutes* the right-hand sides of rules in each membrane i, and then applies the obtained rules according to the maximally parallel semantics.

The conventional (total) halting condition for P systems can be naturally lifted to randomized RHS: a P system Π with randomized RHS halts on a configuration C if, however it permutes rule right-hand sides, no rule can be applied in C, in any membrane.

Example 1. *Consider the P system* $\Pi_2 = (\{a, b\}, \{b\}, [\,]_1, a, R, 1)$ *with the rule set* $R = \{a \to aa, c \to b\}$. Π_2 *is graphically represented in Figure 1.*

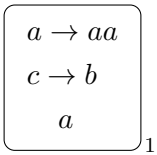


Figure 1. The P system Π_2 with random RHS exchange generating the number language $\{2^n \mid n \in \mathbb{N}\}$.

The number language generated by Π_2 *(the set of numbers of instances of* b *that may appear in the skin after* Π_2 *has halted) is exactly* $\{2^n \mid n \in \mathbb{N}^+\}$. *Indeed, while* Π_2 *applies the identity permutation on the right-hand sides,* $a \to aa$ *will double the number of symbols* a*, while the rule* $c \to b$ *will never be applicable. When* Π_2 *exchanges the right-hand sides of the rules, the rule* $a \to b$ *will rewrite every symbol* a *into a symbol* b*. After this has happened, no rule will ever be applicable any more and* Π_2 *will halt with* 2^n *symbols* b *in the skin, where* $n+1$ *is the number of computation steps taken.*

We will use the notation

$$OP_n(rhsExchange, coo)$$

to denote the family of transitional P systems with random RHS exchange, with at most n membranes, with cooperative rules.

The following statement is one of the central results of the article [1].

Theorem 1.

$$\{2^m \mid m \in \mathbb{N}\} \in NOP_n(rhsExchange, ncoo) \setminus NOP_n(ncoo).$$

Proof. The statement follows (for $n \geq 1$) from the construction given in Example 1 and from the well-known fact that non-cooperative P systems operating under the total halting condition cannot generate non-semilinear number languages (for example, see [3]). □

References

[1] A. Alhazov, R. Freund, S. Ivanov, Transitional P Systems with Randomized Rule Right-hand Sides. In: *Proceedings of CMC 2018, to appear*. 2017.

[2] R. Freund, Generalized P-Systems. In: Gabriel Ciobanu and Gheorghe Păun, editors, *Fundamentals of Computation Theory, 12th International Symposium, August 30–September 3, 1999, Proceedings*. Springer, 1999, 281–292.

[3] Gh. Păun, G. Rozenberg, A. Salomaa (eds.), *The Oxford Handbook of Membrane Computing*. Oxford University Press, Oxford, England, 2010.

[4] Bulletin of the International Membrane Computing Society (IMCS). `http://membranecomputing.net/IMCSBulletin/index.php`.

[5] The P Systems Website. `http://ppage.psystems.eu/`.

Artiom Alhazov[1], Rudolf Freund[2], Sergiu Ivanov[3]

[1]Institute of Mathematics and Computer Science
Academy of Sciences of Moldova
Academiei 5, Chişinău, MD-2028, Moldova
Email: `artiom@math.md`

[2] Faculty of Informatics, TU Wien
Favoritenstraße 9–11, 1040 Vienna, Austria
Email: `rudi@emcc.at`

[3]Laboratoire d'Algorithmique, Complexité et Logique,
Université Paris Est Créteil,
61 av. du général de Gaulle, 94010 Créteil, France
Email: `sergiu.ivanov@u-pec.fr`

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

Sharing Knowledge Network

Bogdan Aman, Gabriel Ciobanu

Abstract

We present the sharing calculus, a calculus of parallel communicating systems in which we can naturally express processes able to add and remove attributes to names. This new calculus can model distributed processes with a shared attribute-based knowledge. We provide its operational semantics, an example and some results.

Keywords: shared attributes, parallel processes, semantics.

1 Introduction

There exist many formal approaches for modelling concurrent systems consisting of parallel processes that interact through communication: e.g, π-calculus [2] and TiMo [1]. However, these approaches abstract away from the fact that these systems use communication to send attributes and values. To be able to model more faithfully the concurrent systems with a shared knowledge we consider important to develop a theoretical foundation that would help in understanding their distinctive attribute-based features.

2 Sharing Calculus

The syntax of the sharing calculus is given in Table 1, where we assume: a set of names (ranged over by $x, y, \ldots$), a set of attributes (ranged over by $a, b, \ldots$), and a set of values (ranged over by $u, v, \ldots$). A pair $\{x.a = u\}$ of the set K is used to assign to each attribute a of name x the value v.

Shared Knowledge	$K :=$	$\emptyset \mid \{x.a = u\} \mid K \uplus K$
Actions	$\alpha ::=$	$\overline{x}(v) \mid \underline{x}(v) \mid x(a)$
Processes	$P ::=$	$\mathbf{0} \mid \alpha.P \mid P + Q \mid P \mid Q \mid (x)P$
		$\mid$ **if** $x \Doteq_S y$ **then** P **else** Q
Knowledge Network	$N :=$	$P \mid\mid K$

Table 1. Syntax of Sharing Calculus

The empty process $\boldsymbol{0}$ cannot perform any action. A prefix $\mu.P$ means that μ is performed and the process continues as P. An output prefix $\overline{x}(v)$ means that a value v is send along name x to be added to the shared knowledge about name x. An output prefix $\underline{x}(v)$ means that a value v is send along name x to be removed from the shared knowledge about name x. An input prefix $x(a)$ means that a name x awaits for an addition of removal of an value of the attribute a from the shared knowledge. A sum $P + Q$ represents a process which can act either as P or as Q, while a parallel composition $P \mid Q$ represents the combined behaviour of P and Q executed in parallel. The pair resulting from a communication between parallel processes is global. A *scope* $(x)P$ defines the scope of x as P such that no communication action have x as its subject. The process **if** $x \Doteq_S y$ **then** P **else** Q executes P if x and y share the same values for the set of shared attributes S, where $S = \{a \mid \{z.a = v\} \in K\}$; otherwise Q is executed. A system N is described by $P \mid\mid K$, namely a shared memory K that process P can read/update.

We first define the structural congruence to equate systems that we never want to distinguish for any reason. The structural congruence $\equiv$ is the least congruence satisfying the abelian monoid laws for summation and parallelism (associativity, commutativity and 0 as identity), the scoping laws for processes: $(x)0 \equiv 0$, $(x)(y)P = (y)(x)P$, $(x)(P + Q) \equiv (x)P + (x)Q$, and also the scope extrusion law $P \mid (x)Q \equiv (x)(P \mid Q)$ if $x \notin fn(P)$.

Transition actions, ranged over by α, consist of actions and pairs. For each process P we denote by $\mathrm{n}(P)$ the *names* of P, $\mathrm{fn}(P)$ the *free names* of P, and by $\mathrm{bn}(P)$ the *bound names* of P.

A COM_ADD and COM_REM rules expresses the synchronous com-

$$COM\_ADD : \frac{P = \overline{x}(v).P' + P'' \quad Q = x(a).Q' + Q''}{(P \mid Q) \mid\mid K \rightarrow (P' \mid Q') \mid\mid K \uplus \{x.a = u\}}$$

$$COM\_REM : \frac{P = \underline{x}(v).P' + P'' \quad Q = x(a).Q' + Q''}{(P \mid Q) \mid\mid K \rightarrow (P' \mid Q') \mid\mid K \backslash \{x.a = u\}}$$

$$PAR : \frac{P \mid\mid K \rightarrow P' \mid\mid K}{(P \mid Q) \mid\mid K \rightarrow (P' \mid Q) \mid\mid K} \qquad SCOPE : \frac{P \mid\mid K \rightarrow P' \mid\mid K}{((z)P) \mid\mid K \xrightarrow{\alpha} ((z)P') \mid\mid K}$$

$$ITE_1 : \frac{P \mid\mid K \rightarrow P' \mid\mid K', x \Bumpeq_S y \ is \ true}{(\textbf{if } x \Bumpeq_S y \textbf{ then } P \textbf{ else } Q) \mid\mid K \xrightarrow{\alpha} P' \mid\mid K'}$$

$$ITE_2 : \frac{Q \mid\mid K \xrightarrow{\alpha} Q' \mid\mid K', x \Bumpeq_S y \ is \ false}{(\textbf{if } x \Bumpeq_S y \textbf{ then } P \textbf{ else } Q) \mid\mid K \xrightarrow{\alpha} Q' \mid\mid K'}$$

Table 2. Operational Semantics of Sharing Calculus.

munication between two processes. If we have a step from P by an action $\overline{x}(a)$, and a step from Q by a corresponding action $x(u)$, then we have a step from the parallel process $(P \mid Q) \mid\mid K$ to $(P' \mid Q') \mid\mid K \uplus \{x.a = u\}$, namely a new pair is created and added as common knowledge into the system, or updates an old knowledge with a new value. If however we have a step from P by an action $\underline{x}(a)$, and a step from Q by a corresponding action $x(u)$, then we have a step from the parallel process $(P \mid Q) \mid\mid K$ to $(P' \mid Q') \mid\mid K \backslash \{x.a = u\}$, namely a pair $\{x.a = u\}$ is removed from common knowledge of the system.

Example 1. *We give a very simple example that demonstrates how easy it is to use our sharing calculus to model access of a client x to a server y using an attribute username u with value name and an attribute password p with value pass. The initial system can be described by: $N = (User \mid Server) \mid\mid K$, where:*

- $K = \{y.u = name\} \uplus \{y.p = pass\}$
- $User = \overline{x}(name').\overline{x}(pass')$
- $Server = x(u).x(p).\textbf{\textit{if}}\ x \Bumpeq_{u,p} y\ \textbf{\textit{then}}\ Access\ \textbf{\textit{else}}\ Deny$

The initial knowledge K illustrate the fact that the server y knows the name name of an user and the necessary password to access its account. The client sends along name x the name $name'$ and $value'$. If these values are equal with the knowledge of y then the user reaches process

Access; otherwise authentication fails and process Deny is reached. It should be noticed that in this case the user x cannot access the information about y in K and that only the server y can check if data is similar and decide if the user grants access or no to its account.

Proposition 1. *$\approx_S$ is an equivalence relation. Consequently, K can be partitioned into disjoint equivalence classes of $K/\approx_S$.*

In what follows we show that any given knowledge network can be obtained starting from a network with empty shared knowledge.

Proposition 2. *If $N' = P' \,||\, K'$ with $K' \neq \emptyset$, then there exists $N = P \,||\, K$ with $K = \emptyset$ such that $N \rightarrow^* N'$.*

The next result claims that a process affects only a part of the general knowledge, while the rest remains unchanged.

Proposition 3. *If $P \,||\, K \rightarrow Q \,||\, K'$, then $P \,||\, (K \uplus K'') \rightarrow Q \,||\, (K' \uplus K'')$ for all K''.*

References

[1] G. Ciobanu, M. Koutny. *Modelling and Verification of Timed Interaction and Migration.* Lecture Notes in Computer Science, vol. 4961 (2008), 215–229.

[2] R. Milner. *Communicating and Mobile Systems: The π-calculus.* Cambridge University Press, 1999.

Bogdan Aman, Gabriel Ciobanu

Romanian Academy, Institute of Computer Science
Blvd. Carol I no.8, 700505 Iaşi, Romania
E–mail: baman@iit.tuiasi.ro, gabriel@info.uaic.ro

On application of P system based algorithms for diachronic text analysis

Lyudmila Burtseva

Abstract

The proposed research relates to modern tendency of application of advanced parallel computing techniques to Big Data problems of computational linguistics. General aspects are studied on particular problem of diachronic analysis of historic texts. The advanced parallel computing techniques are represented here by P system computing. Proposed in this work combination of existing diachronic analysis techniques and our previous developments in P system application is shown to be perspective solution for easy adoption of parallel techniques for analysis of Web-based corpora of historic texts.

Keywords: computational linguistic, diachronic analysis, P system, web-based corpus.

1 Introduction

The proposed research is the contribution to modern tendency of application of advanced parallel computing techniques to Big Data problems of computational linguistics. Building during last decade huge web-based corpora of historic texts inspire development of diachronic analysis frameworks with different underlying technique: word embedding [1], co-occurance graph [2], collocation profiling [3], etc. Increasing of corpora size through scanning new texts in numerous research projects challenges the necessity of modern parallel processing techniques application.

The main idea of proposed research is to support smooth transfer of existing and well tested diachronic analysis techniques to parallel implementation. The support will be achieved by applying natural parallelization strategies of P system computing [4].

2 P system based algorithm advantages

The most actively employed frameworks perform diachronic analysis by methods based on graph built from: (1) word clustering [2] and (2) word embedding [3]. Both cases demonstrate that big corpora result in significant size of analysis graph node sets. Due to need of web access ensuring, the processing of such amount of data requires pre-processing reducing. Another way to provide real-time processing is its parallel implementation.

Advantages of P system application, mentioned in introduction, spring from natural parallelism of algorithms. The parallelism is intrinsic to P system [5] (aka membrane systems) due to its nature: this is bio-inspired paradigm reproducing the membrane structure of the biological cell. Computing is performed by parallel execution of rules of objects movement or transformation.

Employment of P system computing for selected sets of diachronic analysis methods is founded on the following aspect. In both considered cases, regardless of applied processing methods, representation of corpora is vector of vectors. Several already developed applications of P system computing processed such matrix data. In our previous research we dealt with matrix representation of image, applying P system to solving of medical imaging problems [6].

Proposed research intends to test P system based implementation of selected diachronic analysis methods set. Both sets have particular features, which processing can be accelerated by parallelization.

- Methods set (1) uses word embedding to create words co-location graph. There are three methods, which are mostly applied to this set: PPMI (positive point-wise mutual information), SVD (singular value decomposition), SGNS (Skip-gram with negative sampling). Implementation all of these methods usually supposes pre-processing reduction of matrix dimension. Parallel implementation allows

developers to avoid reducing tips-and-tricks because it not afraid of exhaustive search (brute force).

- Methods set (2) uses clustering to obtain graph nodes. There are some P system based solutions [7], [8] of clustering problem but due to matrix data we intend to use image segmentation algorithm [9] that showed acceptable results on both sequential and parallel simulators [6].

On base of test results we will choose suitable methods set to build web-based researchers support toolkit with engine working in parallel.

3 Conclusion

The essence of proposals consists in granting researchers by possibility to build easily, using their previous developments, modern version of diachronic analysis frameworks with parallel processing engine.

As research result, diachronic analysis methods, which will be selected as more suitable for P system based solutions, will be applied to building our own historic texts processing framework.

References

[1] L. W. Hamilton, J. Leskovec, D. Jurafsky. *DiachronicWord Embeddings Reveal Statistical Laws of Semantic Change*, CoRR abs/1605.09096 (2016).

[2] B. Jurish. *DiaCollo: On the trail of diachronic collocations.* CLARIN Annual Conference Proc., Wroclaw, Polen, 14-17 October, (2015), pp.28-31.

[3] A. Kilgarriff, O. Herman, J. Bušta, V. Kovář, M. Jakubíček. *DIACRAN: a framework for diachronic analysis.* Corpus Linguistics 2015 Abst., (2015).

[4] A. Alhazov, L. Burtseva, S. Cojocaru, A. Colesnicov, L. Malahov. *HPC Patterns Based Implementations Of P Systems Based Solutions Of Hard Computational Problems.* Proceedings of the Workshop on Foundations of Informatics FOI-2015, August 24-29, 2015, Chisinau, Republic of Moldova, pp.82-88.

[5] Gh. Păun. *Membrane Computing. An Introduction*, Springer, 2002.

[6] L. Burtseva. *Advantages of application of unconventional computing to image processing and whence these advances come.* Computer Science Journal of Moldova, vol.24, no.1(70), (2016), pp.136-144.

[7] M.-A. Cardona, M. Colomer, A.-J. Zaragoza, M. Perez-Jimenez. *Hierarchical Clustering with Membrane Computing. Computing and Informatics*, vol. 27, no. 3, (2008), pp. 497-513.

[8] Hong Peng, J. Zhang, J. Wang, T. Wang, M. J. Perez-Jimenez, A. Riscos-Nunez. *Membrane Clustering: A Novel Clustering Algorithm under Membrane Computing*, BWMC2014 Proc., (2014), pp. 311-327.

[9] H. A. Christinal, D. Diaz-Pernil, P. Real. *Region-based segmentation of 2D and 3D images with tissue-like P systems*, Pattern Recognition Letters, vol. 32, no. 16, Dec. 2011, pp. 2206–2212.

Lyudmila Burtseva

Institute of Mathematics and Computer Science ASM,
5 Academiei str., Chişinău, MD-2028, Moldova
E–mail: luburtseva@gmail.com

Knowledge level assessment by using Machine learning

Olesea Caftanatov, Tudor Bumbu

Abstract

Knowledge assessment is an important component of learning activity. To lessen teacher's burden we intend to use machine-learning technique to evaluate the students' knowledge level. In this paper, we propose a specific approach for knowledge assessment by analyzing students' behavior in learning activity. Another important aspect in our research is designing methodology for recommending correct answers to improve students' knowledge level. This paper is the first report of a project on the use of machine learning in education, which we recently started.

Keywords: knowledge assessment, machine learning, decision support tool.

1 Introduction

Nowadays there are new approaches in education, based on adaptive learning, personalization or even personal learning paths. However, in each type of learning concepts one of the most important factors is evaluation. Usually, based on evaluation, a teacher or an educational system makes decision what to do next. Daily, teachers are burdened with a lot of work, like preparing educational materials, evaluation, checking pupils' homework, or providing feedback to pupils, to parents on their children learning progress, etc. Because of the big number of pupils teachers really have hard time, so this load of work can be lessen by teaching a learning machine for e.g. how to evaluate pupils knowledge.

Machine learning can provide new solution not only for traditional learning method but also for new learning methods and teaching technologies in education. The biggest advantage of machine learning is

that it is based on principle "the more data is given, the smarter algorithm".

In this paper, we describe a machine learning approach to pupil's knowledge evaluation, which would make it possible to suggest what they should learn next. Thus, our tool can be useful for teachers in their decision-making activities. This paper consists of five main sections. The first one is this "introductory" section. The next one is a short review of what the machine learning means. In the third, we analyzed few related works that gave us some insights of different aspects for our problem. In the fourth section, we present an example of our dataset used for learning activity such as video lessons. In fifth section, we presented our approach to suggest the correct answers.

2 Short review: What is Machine Learning?

Let us recall the main notions of machine learning domain. In [1] machine learning is defined as a "method of teaching computers to synthesize data and use it to make or improve prediction". Another interesting definition one can find in Tomáš Přinda's work [2], where "machine learning is a group of algorithms which are able to learn from your data". Frankly speaking, machine learning is a infrastructure for solving a large number of problems that exist in educational system. There are two main branches of machine learning: supervised learning and unsupervised learning. In our case, we will use both of them – for knowledge assessment approach - we use supervised learning, but for recommendation approach - we use unsupervised learning.

3 Related work

There are different methods in usage of machine learning techniques for educational proposes that include classification and regression algorithms, association rules, sequential pattern analysis, as well as clustering and web mining. Some of them are based on Baker's taxonomy [3]

For instance, S.B. Kotsiantis presented in [4] a case study describing the use of machine learning techniques for educational proposes. Also, based on existing regression techniques he elaborated a prototype version of software support tool for tutors that forecast students' grades and future performance.

Recently, Nazeeh Ghatasheh presented a novel case study [5] describing several classification algorithms that were applied to predict the knowledge level of learners. The experimental results illustrate an overall performance superiority of support vector machine (SVM) model in evaluating the knowledge levels, having 98,6% of correctly classified instances with 0.0069 mean absolute error.

An interesting approach for learning recommendation can be seen in project "OpenEd" [6] that uses machine learning to classify educational resources by learning the goal. In other words, OpenEd uses association rules to prebuild formative assessments where each question is linked with resources, which would automatically surface as recommendation to learn, but it pop-ups only when students give incorrect answers.

Another research was made by Burr Settles team, who developed the *Duolingo* project that helps learn new languages. According to [7] it has millions of students who generate billions of statistics about language learning every day. The main point of their research is that they proposed a new statistical model called *"Half-life regression*", inspired by *logistic regression algorithm*, but using an exponential probability function like the one from Figure 1.

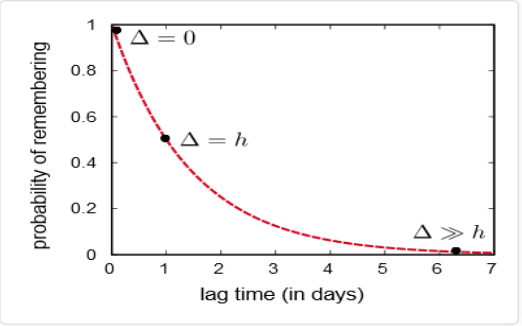


Figure 1. A forgetting curve [8]: the probability of remembering goes down as a function of "lag time" Δ (days since the last practice) and "half-life" *h*.

Technically, they estimated the half-life of a word in our memory using $h=2\Theta\cdot\mathbf{x}$, where Θ denotes the regression model "weights" and $\mathbf{x}$ denotes a bunch of variables that summarize learning history with the word. So, Half-life regression involves finding the "best" model weights

for Θ by minimizing the "loss function" ℓ - across every practice session for every student:

$$\ell(\langle p, \Delta, \mathbf{x}\rangle;\Theta)=(p-2^{-\Delta 2\Theta\cdot\mathbf{x}})2+\alpha(-\Delta\log 2(p)-2\Theta\cdot\mathbf{x})2+\lambda|\Theta|22 \quad (1)$$

In short, they can learn to predict the half-life for each word in our long-term memory, by analyzing the error patterns of millions of languages learners.

4 Dataset of our case

To evaluate knowledge level we need to understand students' behavior in learning activity. The main learning activities are: forum, video lessons, educational games, exercises and quizzes. For instance, in Table 1 are presented the attributes of our dataset for video lessons along with the values of every attributes.

Table 1. Overview of dataset for video lessons interaction

Attributes	Values	Values
Subject	Point and line	Angle
Week	1st	3rd
Video	1 lesson	3 lesson
Length	5 min	3 min
Number of pauses	15	7
Median duration of pauses	50 sec	60 sec
Proportion of skipped video content	0	0
Number of backward seeks	3	0
Replayed video length	00:01:24	00:00:40
Video speed	1.25	1
Drop-out video	-	-

An interesting research on investigating unobtrusive measures of body-language in order to predict student's attention during the class was made in CHILI laboratory, Lausanne [9]. Based on classifier SVM they demonstrated that drops in attention are reflected in decreased intensity of head movements.

In [10] authors used camera or microphone to analyze student's activity over time and extract simple characteristics like the number of:

video watched, post written, post read etc. In the end, by using R and CARPET package due to simplicity and access to the most recent machine learning methods they could discover deeper patterns and to provide more accurate predictions of student's behaviors and outcomes.

For the other type of activities we have the following list of attributes:

Table 2. Overview of dataset for forum activity interaction

Attributes	Values
Post viewed	Point and line
Number of quotes	2
Posting comments	1
Re-reading the post	5 min
Creating post	1
Deleted post	No one
Total time spend on forum per day	2h
Frequency entries on forum per week	20

Table 3. Overview of dataset for game activity interaction

Attributes	Values
Total score	120
Score per level	60/1
Level	1
Total missions	3
Number of attempt per mission	5
Aborting game	No one
Total time	20 min

Table 4. Overview of dataset for exercises activity interaction

Attributes	Values
Total score	10
Correct answer	1
Incorrect answer	0
Time	5 min
Abort	0

Table 5. Overview of dataset for quiz activity interaction

Attributes	Values
Total score	100
Total number of question	10
Total number of correct answers	8
Total numbers of incorrect answers	1
Total numbers of skipped answers	1
Total time	30 min
Attempt for re-doing quiz	0
Abort	0

Note: The level of knowledge assessment will be measured in percentage value and this will represent our classes. For passing one level of knowledge students need to acquire at least 50% of material. In cases where the wrong answers will be found, the system will analyze and recommend the appropriate materials for learning. See more about this in next section.

5 Recommendation Methodology

We intend to use associated rules in recommending correct answers to students. For instance, we have a set of questions and a number of students that gave their personal answers to those questions. Let us consider that one of the students gave incorrect answer to one question. Our approach consist in recommending to that student a list of correct answers for that question that were given by others students. Moreover, it

will also recommend the answers to similar questions. For a better understanding, let's analyze the example presented in Table 6.

Table 6. An example for recommending methodology

	Q1	Q2	***Q3***	Q4	***Q5***
User 1	A1	A2	A3	A4	**?**
User 2	B1	B2	B3	B4	B5
User 3	C1	C2	C3	C4	C5
User 4	D1	D2	D3	D4	D5
User 5	E1	E2	E3	E4	E5

As we can see in Table 6, we have five questions (Q) where Q1, Q3 and Q5 have similar subject, and five users. All of them gave their personal answers for giving set of questions (A1,…,E5). Note, that B1, C5, D3 and E5 are correct answers. So, in our case, the first user doesn't know the correct answer for Q5 and gives an incorrect answer. Our idea is recommending the following objects:

- The correct answers for Q5 given by user 3 and 5;
- For a better understanding of the question 5, user 1 will get as recommendation the correct answers of the similar questions like (Q1 and Q3) that are given by user 2 and user 4.

6 Conclusion and future work

As the modern classroom becomes more and more digitized, we are able to gather big sets of data. Using machine learning we can obtain many new solutions for different problems in education. We intend to bring up a specific approach to student knowledge evaluation. As for the incorrect answers, we present a recommendation method that would analyze and provide necessary information in the next step of learning activity. At the moment we are dealing with training data collection for the future experiments.

References

[1] Harish Agrawal. *Integrating Machine Learning in Education Technology*. August 3, 2017.https://www.getmagicbox.com/integrating-machine-learning-in-education-technology/ .

[2] Tomáš Přinda. *3 Machine Learning Applications that moved products to the next level*. August 7, 2017. https://www.linkedin.com/pulse/3-machine-learning-applications-moved-products-next-level-p%C5%99inda .

[3] Ryan S.J.D. Baker, Kalina Yacef. *The state of educational data mining in 2009: A review and future visions*. Journal of Educational Data Mining, Article 1, Vol1, No 1, Fall 2009.

[4] S.B. Kotsiantis. *Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades*. May 12, 2011 https://link.springer.com/content/pdf/10.1007%2Fs10462-011-9234-x.pdf .

[5] Nazeeh Ghatasheh. *Knowledge level assessment in e-Learning systems using machine learning and user activity analysis*. International Journal of Advanced Computer Science and Applicationa, Vol.6, No.4, 2015. http://thesai.org/Downloads/Volume6No4/Paper_15-Knowledge_Level%20Assessment_in_e-Learning_Systems.pdf .

[6] Adam Blum. *About OpenEd*. Official homepage: https://about.opened.com/.

[7] Burr Settles, Brendan Meeder. *A trainable spaced repetition model for language learning*. Proceedings of the 54th annual meeting of the Association for Computational Linguistics, pages 1848-1858, Berlin, Germany, August 7-12, 2016.http://aclweb.org/anthology/P/P16/P16-1174.pdf.

[8] Hermann Ebbinghaus - a pioneer of memory research. Flashcard Learner. http://www.flashcardlearner.com/articles/hermann-ebbinghaus-a-pioneer-of-memory-research/.

[9] Mirko Raca, Lucasz Kidziński, Pierre Dillenbourg. *Translating Head motion into attention – Towards processing of student's body-language*. http://www.educationaldatamining.org/EDM2015/uploads/papers/paper_83.pdf.

[10] Lucasz Kidziński, Michail Ciannakos, Demetrios G. Sampson and Pierre Dillenbourg. *A tutorial on machine learning in educational science*. https://infoscience.epfl.ch/record/210991/files/la.pdf.

Olesea Caftanatov[1], Tudor Bumbu[2]

Institute of Mathematics and Computer Science
Academy of Sciences of Moldova
Academiei 5, Chişinău, MD-2028, Moldova
E-mail: [1]olesea.caftanatov@math.md, [2]bumbutudor10@gmail.com

Some Techniques to Develop of the Expert Systems

Gheorghe Capatana

Abstract

In the paper are exposed the four techniques for the development of expert systems in the companies' financial diagnosis and comparative analysis of these techniques.

Keywords: the expert systems, companies' financial diagnosis, spiking neural P systems, fuzzy neural P systems

1 Introduction

Several technologies for building expert systems are known nowadays. Airinei D. in the work [1] describes the development of *expert systems for the financial diagnosis* (ESFD) of companies. We will examine the four techniques of designing the ESFD and their comparative.

2 ESFD development with the imperative techniques

Each expert system can be seen in the same time, both as:

(a) a system for the forms of objects recognition, and
(b) a system for automatic classification of the objects.

The objects, the SDFE operates with are companies characterized by the values of the accountants indicators proposed by the International Accountant Study Group and the European Union of Accountants Experts [1].

The knowledge base for the financial diagnosis of companies is presented in Table 1 (adjusted after the [2]). Indicators' values are represented by the hieroglyphs used to build expert systems. The significance of hieroglyphics is not fundamental for interpreting the computer table and concluding with the computer assistance of the expertise results.

Table 1. The values of accounting indicators by classes and groups of companies

The Accounting Indicator	The class „Fragile companies" (f_r)		The class „Strong companies" (r_{ez})	
	The group „In difficulty" ($\hat{i}_d$)	The group „Very vulnerable" (f_v)	The group „Financially stable" (s_f)	The group „Commercially competitive" (c_c)
x_1	s	s	p	p
x_2	r_{id}	r_{id}	m_{od}	m
x_3	s	p	m	m
x_4	r_{id}	r_{id}	m	m
x_5	p	p	p	f_s
x_6	f_p	s	p	m
x_7	r_{ed}	f_{ri}	c	f_{re}
x_8	r_{id}	f_{ri}	r_{id}	f_s
x_9	f_p	a	a	p
x_{10}	l	l	r	r
x_{11}	m_{ed}	λ	r_{id}	s
x_{12}	m_{ed}	c	m_{ed}	s

3 The development of the ESFD with the techniques based on the rules of production

Airinei D., applying the method of the delimitation of the companies, initially in *the categories of companies* and then *in groups of companies*, manages to carry out the expertise of the companies with only 17 rules of production. ESFD is achieved using the instrumental system GURU. The result of the expertise is obtained in less than 12 steps.

4 The development of the SEDF with the spiking neural P systems

The development of the ESFD can be achieved by applying the spiking neural P systems. The author has carried out experiments for the development of the ESFD using spiking neural P systems and SNPS simulator. The work of the expert system has requested 14 neurons. The results of the examination are obtained in 4 steps.

5 The development of the ESFD with the neural fuzzy P systems

The development of the ESDF can be achieved with the neural fuzzy P systems. There was built a neural fuzzy P system that carries out the ESDF on the computer. The expert system has been accomplished in Maple and integrates 48 fuzzy neurons, that communicate telepathically. The end user can adjust the degree of hue of the expert system's neurons. The results of the examination are fuzzy and obtained in 4 steps.

6 Conclusion

In the work have been exposed some of the design techniques of the companies' ESDF 4 techniques for the development of the ESDF and their comparative analysis have been presented as well.

Acknowledgments. The presentation of this article was possible due to:

(a) the project „ SCTU Project 4032 „Power and Efficiency Of Natural Computing: Neural-Like P (Membrane) Systems", Project manager Prof. Iurie Rogojin;

(b) the project „The development of the smart IT systems - oriented on families of decisional problems with implementation in the education and research", keyed 15.817.02.38A.

References

[1] E. Cohen. *Analyse financière.* Les Editions d'Organization, Paris, 1987.

[2] D. Airinei. *Sisteme expert în activitatea financiar-contabilă*. Iaşi, 1997.

Gheorghe Capatana
Moldova State University
E-mail: gh_capatana@yahoo.com

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

Measures of Similarity on Monoids of Strings

Mitrofan M. Cioban, Ivan A. Budanaev

Abstract

In information theory, linguistics and computer science are important distinct string metrics for measuring the difference between two given strings (sequences). In this article we introduce the efficiency, measure of similarity and penalty for given parallel decompositions of two strings. The relations between these characteristics are established.

Keywords: invariant distance, measure of similarity, penalty, Levenshtein distance, Hamming distance.

1 Introduction

Let G be a semigroup and d be a metric on G. The metric d is called:

- *left* (respectively, *right*) *invariant* if $d(xa, xb) \leq d(a, b)$ (respectively, $d(ax, bx) \leq d(a, b)$) for all $x, a, b \in G$;

- *invariant* if it is both left and right invariant.

A monoid is a semigroup with an identity element. Fix a non-empty set A. The set A is called an alphabet. Let $L(A)$ be the set of all finite strings $a_1a_2 \dots a_n$ with $a_1, a_2, \dots, a_n \in A$. Let ε be the empty string. Consider the strings $a_1a_2 \dots a_n$ for which $a_i = \varepsilon$ for some $i \leq n$. If $a_i \neq \varepsilon$, for any $i \leq n$ or $n = 1$ and $a_1 = \varepsilon$, the string $a_1a_2 \dots a_n$ is called a *canonical string*. The number $l(a_1a_2 \dots a_n) = |\{i \leq n : a_i \neq \varepsilon\}|$ is the length of the string $a_1a_2 \dots a_n$. For two strings $a_1 \dots a_n$ and $b_1 \dots b_m$, their product (concatenation) is $a_1 \dots a_n b_1 \dots b_m$. If $n \geq 2$, $i < n$ and $a_i = \varepsilon$, then the strings $a_1 \dots a_n$ and $a_1 \dots a_{i-1}a_{i+1} \dots a_n$ are considered equivalent. In this case any string is equivalent to one

unique canonical string. We identify the equivalent strings. In this case $L(A)$ becomes a monoid with identity ε.

Fix an alphabet A and let $\bar{A} = A \cup \{\varepsilon\}$. We assume that $\varepsilon \in \bar{A} \subseteq L(A)$. Let a, b be two strings. For any two representations $a = a_1a_2\cdots a_n$ and $b = b_1b_2\cdots b_m$ we put

$d_H(a_1a_2\cdots a_n, b_1b_2\cdots b_m) = |\{i : a_i \neq bi, i \leq min\{n,m\}\}| + |\{i : n < i \leq m, b_i \neq e_i\}| + |\{j : m < j \leq n, a_j \neq e_i\}|$. The function d_H is called the Hamming distance on the space of strings [7].

Now we put $d_G(a,b) = inf\{d_H(a_1a_2\cdots a_n, b_1b_2\cdots b_n) : a = a_1a_2\cdots a_n \ , \ b = b_1b_2\cdots b_n\}$.

The distance d_G for free group on A was defined by M. I. Graev [5, 9, 3, 4].

Levenshtein distance d_L between two strings $a = a_1a_2\cdots a_n$ and $b = b_1b_2\cdots b_m$ is defined as the minimum number of insertions, deletions, and substitutions required to transform one string to the other [8, 3, 4].

Theorem 1. *The metric d_G on a monoid $L(A)$ has the following properties:*

1. *d_G is an invariant metric on $L(A)$ and $d_G(x,y) = d_G(x,y)$ for all $x, y \in \bar{A}$.*

2. *If ρ is an invariant metric on $L(A)$ and $\rho(x,y) \leq d_G(x,y))$ for all $x, y \in \bar{A}$, then $\rho(a,b) \leq d_G(a,b)$ for all $a, b \in L(A)$.*

3. *For any $a, b \in L(A)$ there exist $n \in \mathbb{N}$, $x_1, x_2, \ldots, x_n \in Sup(a,a)$ and $y_1, y_2, \ldots, y_n \in Sup(b,b)$ with $a = x_1x_2\cdots x_n$, $b = y_1y_2\cdots y_n$, such that $n \leq l(a) + l(b)$ and $d_G(a,b) = |\{i : i \leq n, a_i \neq b_i\}| = d_H(x_1x_2....x_n, y_1y_2....y_n)$.*

4. *$d_G(a,b) = d_L(a,b) \leq d_H(a,b)$ for all $a, b \in L(A)$.*

Remark 1. *The method of extensions of distances for free groups, used by us, was proposed by A. A. Markov [9] and M. I. Graev [5]. For free universal algebras it was extended in [2].*

2 Parallel decompositions of two strings

The longest common substring and pattern matching in two or more strings is a well known class of problems. For any two strings $a, b \in L(A)$ we find the decompositions of the form $a = v_1u_1v_2u_2\cdots v_ku_kv_{k+1}$ and $b = w_1u_1w_2u_2\cdots w_ku_kw_{k+1}$, which can be represented as $a = a_1a_2\cdots a_n$, $b = b_1b_2\cdots b_n$ with the following properties:

- some a_i and b_j may be empty strings, i.e. $a_i = \varepsilon$, $b_j = \varepsilon$;

- if $a_i = \varepsilon$, then $b_i \neq \varepsilon$, and if $b_j = \varepsilon$, then $a_j \neq \varepsilon$; - if $u_1 = \varepsilon$, then $a = v_1$ and $b = w_1$; - if $u_1 \neq \varepsilon$, then there exists a sequence $1 \leq i_1 \leq j_1 < i_2 \leq j_2 < \cdots < i_k \leq j_k \leq n$ such that:

$u_1 = a_{i_1}\cdots a_{j_1} = b_{i_1}\cdots b_{j_1}$, $u_2 = a_{i_2}\cdots a_{j_2} = b_{i_2}\cdots b_{j_2}$, $u_k = a_{i_k}\cdots a_{j_k} = b_{i_k}\cdots b_{j_k}$;

if $v_1 = w_1 = \varepsilon$, then $i_1 = 1$;

if $v_{k+1} = w_{k+1} = \varepsilon$, then $j_k = n$;

if $k \geq 2$, then for any $i \in \{2, \cdots, k\}$ we have $v_i \neq \varepsilon$ or $w_i \neq \varepsilon$.

In this case $l(u_1) + l(u_2) + \cdots l(u_k) = |\{i : a_i = b_i\}|$.

The above decompositions forms are called *parallel decompositions* of strings a and b [3, 4]. For any parallel decompositions $a = v_1u_1\cdots v_ku_kv_{k+1}$ and $b = w_1u_1\cdots w_ku_kw_{k+1}$ the number

$$E(v_1u_1\cdots v_ku_kv_{k+1}, w_1u_1\cdots w_ku_kw_{k+1}) = \sum_{i\leq k+1}\{\max\{l(v_i), l(w_i)\}\} = d_H(x_1x_2...x_n, y_1y_2...y_n)$$

is called the efficiency of the given parallel decompositions. The number $E(a, b)$ is equal to the minimum of efficiency values of all parallel decompositions of the strings a, b and is called the *common efficiency of the strings a,b*. It is obvious that $E(a, b)$ is well determined and $E(a, b) = d_G(a, b)$. We say that the parallel decompositions $a = v_1u_1v_2u_2\cdots v_ku_kv_{k+1}$ and $b = w_1u_1w_2u_2\cdots w_ku_kw_{k+1}$ are optimal if the following equality holds:

$$E(v_1u_1v_2u_2\cdots v_ku_kv_{k+1}, w_1u_1w_2u_2\cdots w_ku_kw_{k+1}) = E(a, b).$$

This type of decompositions are associated with the problem of approximate string matching [10]. If the decompositions $a = v_1u_1 \cdots v_ku_kv_{k+1}$ and $b = w_1u_1 \cdots w_ku_kw_{k+1}$ are optimal and $k \geq 2$, then we may consider that $u_i \neq \varepsilon$ for any $i \leq k$.

Any parallel decompositions $a = a_1a_2 \cdots a_n = v_1u_1v_2u_2 \cdots v_ku_kv_{k+1}$ and $b = b_1b_2 \cdots b_n = w_1u_1w_2u_2 \cdots w_ku_kw_{k+1}$ generate a common subsequence $u_1u_2 \cdots u_k$. The number $m(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) = l(u_1) + l(u_2) + \cdots + l(u_k)$ is the *measure of similarity* of the decompositions [1, 11]. There exist parallel decompositions $a = v_1u_1v_2u_2 \cdots v_ku_kv_{k+1}$ and $b = w_1u_1w_2u_2 \cdots w_ku_kw_{k+1}$ for which the measure of similarity is maximal. The maximum value of the measure of similarity of all decompositions is denoted by $m^*(a, b)$. The maximum value of the measure of similarity of all optimal decompositions is denoted by $m^\omega(a, b)$. We can note that $m^\omega(a, b) \leq m^*(a, b)$. For any two parallel decompositions $a = a_1a_2 \cdots a_n$ and $b = b_1b_2 \cdots b_n$ as in [4], we define the *penalty factors* as

$$p_r(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) = |\{i \leq n : a_i = \varepsilon\}|,$$

$$p_l(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) = |\{j \leq n : b_j = \varepsilon\}|,$$

$$\begin{aligned} p(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) &= |\{i \leq n : a_i = \varepsilon\}| + |\{j \leq n : b_j = \varepsilon\}| \\ &= p_r(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) + p_l(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) \end{aligned}$$

and

$$\begin{aligned} &M_r(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) \\ &\quad = m(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) - p_r(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) \\ &M_l(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) \\ &\quad = m(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) - p_l(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) \\ &M(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) \\ &\quad = m(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) - p(a_1a_2 \cdots a_n, b_1b_2 \cdots b_n) \end{aligned}$$

as the *measures of proper similarity.*

The number $d_H(a_1a_2\cdots a_n, b_1b_2\cdots b_n) = |\{i \leq n : a_i \neq b_i\}|$ is the Hamming distance between decompositions and it is another type of penalty: we have that $p(a_1\cdots a_n, b_1\cdots b_n) \leq d_H(a_1\cdots a_n, b_1\cdots b_n)$.

The assertions from the following theorem establish the main results.

Theorem 2. Let a and b be two non-empty strings, $a = a_1a_2\cdots a_n$ and $b = b_1b_2\cdots b_n$ be the initial optimal decompositions, and $a = a'_1a'_2\cdots a'_q$ and $b = b'_1b'_2\cdots b'_q$ be the second decompositions, which are arbitrary. Denote by

$$
\begin{aligned}
&m = m(a_1a_2\cdots a_n, b_1b_2\cdots b_n), && m' = m(a'_1a'_2\cdots a'_n, b'_1b'_2\cdots b'_q),\\
&p = p(a_1a_2\cdots a_n, b_1b_2\cdots b_n), && p' = p(a'_1a'_2\cdots a'_n, b'_1b'_2\cdots b'_q),\\
&p_l = p_l(a_1a_2\cdots a_n, b_1b_2\cdots b_n), && p'_l = p_l(a'_1a'_2\cdots a'_n, b'_1b'_2\cdots b'_q),\\
&p_r = p_r(a_1a_2\cdots a_n, b_1b_2\cdots b_n), && p'_r = p_r(a'_1a'_2\cdots a'_n, b'_1b'_2\cdots b'_q),\\
&r = d_H(a_1a_2\cdots a_n, b_1b_2\cdots b_n), && r' = d_H(a'_1a'_2\cdots a'_n, b'_1b'_2\cdots b'_q),\\
&M = m - p, M' = m' - p', && M_l = m - p_l, M'_l = m' - p'_l,\\
&M_r = m - p_r, M'_r = m' - p'_r.
\end{aligned}
$$

The following assertions are true:

1. $p' - p = 2(m' - m) + 2(r' - r)$.

2. If the second decompositions are non optimal, then $M_l > M'_l$ and $M_r > M'_r$.

3. If the second decompositions are optimal, then $M_l = M'_l$ and $M_r = M'_r$ and the measures M_l and M_r are constant on the set of optimal parallel decompositions.

4. If $m' \geq m$ and the second decompositions are non optimal, then $p' > p$, $p_l\prime > p_l$, $p'_r > p_r$ and $M > M'$.

5. If $m' = m$ and the second decompositions are optimal, then $p' = p$, $p_l\prime = p_l$, $p'_r = p_r$ and $M' = M$.

6. If $m' \leq m$ and the second decompositions are non optimal, then $m' - r' < m - r$.

For any distinct $x, y \in L(A)$ there exists algorithm for computing the distance $d_G(x, y)$ (see [3, 4]).

3 Conclusions

From Assertions 1 and 3 of Theorem 2 it follows that on the class of all optimal decompositions of given two strings:

- the maximal measure of proper similarity is attained on the optimal parallel decomposition with minimal penalties (minimal measure of similarity);

- the minimal measure of proper similarity is attained on the optimal parallel decomposition with maximal penalties (maximal measure of similarity).

For any two non-empty strings there exist parallel decompositions with maximal measure of similarity and optimal decompositions on which measure of similarity is minimal.

Decompositions with minimal penalty and maximal proper similarity are of significant interest. Moreover, if we consider the problem of text editing and correction, the optimal decompositions are more favorable. Therefore, optimal decompositions are the best parallel decompositions and we may solve the string match problems only on class of optimal decompositions.

To summarize the results, we established that optimal decompositions:

- describe the proper similarity of two strings;
- permit to obtain long common sub-sequences;
- permit to calculate the distance between strings;
- permit to appreciate changeability of information over time.

References

[1] V. B. Barahnin, V. A. Nehaeva, A. M. Fedotov. *O zadanii mery shodstva dlja klasterizacii tekstovyh dokumentov,* Vestnik NGU. Ser.: Informacionnye tehnologii, vol. 1 (2008), pp. 3–9. (in Russian)

[2] M. M. Choban. *The theory of stable metrics,* Math. Balkanica, vol. 2 (1988), pp. 357–373.

[3] M. M. Choban, I. A. Budanaev. *Distances on Monoids of Strings and Their Applications,* Proceedings of the Conference on Mathematical Foundations of Informatics MFOI2016, July 25-29, 2016, Chisinau, Republic of Moldova, (20016), pp. 144–159.

[4] M. M. Choban, I. A. Budanaev. *About Applications of Distances on Monoids of Strings*, Computer Science Journal of Moldova, vol. 24, no 3 (2016), pp. 335–356.

[5] M. I. Graev. *Free topological groups* Trans. Moscow Math. Soc., vol. 8 (1962), 303–364 (Russian original: Izvestia Akad. Nauk SSSR, vol. 12 (1948), 279–323).

[6] D. Gusfield, R. W. Irving, *The Stable Marriage Problem: Structure and Algorithms,* in: Foundations of Computing Series, Cambridge, MA, USA: MIT Press, 1989.

[7] R. W. Hamming. *Error Detecting and Error Correcting Codes,* Bell System Technical Journal, vol. 29, no 2 (1952), pp. 147–160.

[8] V. I. Levenshtein. *Binary codes capable of correcting deletions, insertions, and reversals,* DAN SSSR, vol. 163, no 4 (1965), pp. 845–848 (in Russian) (English translation: Soviet Physics – Doklady, vol. 10, no. 8 (1966), pp. 707–710).

[9] A. A. Markov. *On free topological groups,* Trans. Moscow Math. Soc., vol. 8 (1962) pp. 195–272.

[10] G. Navarro. *A guided tour to approximate string matching,* ACM Computing Surveys, vol. 33, no 1 (2001), pp. 31–88.

[11] S. B. Needleman, C. D. Wunsch. *A general method applicable to the search for similarities in the amino acid sequence of two proteins,* Journal of Molecular Biology, vol. 48, no 3 (1970), pp. 443–453.

Mitrofan M. Cioban[1], Ivan A. Budanaev[2]

[1]Professor, Doctor of Science,
Academician of the Academy of Science of Moldova
Tiraspol State University, Republic of Moldova
E–mail: mmchoban@gmail.com

[2]Doctoral School of Mathematics and Information Science
Institute of Mathematics and Computer Sciences of ASM
Tiraspol State University, Republic of Moldova
E–mail: ivan.budanaev@gmail.com

Data preparation in the process of prognostic model STROKE.MD creation

Svetlana Cojocaru, Constantin Gaindric,
Galina Magariu, Tatiana Verlan

Abstract

Process of data preparation for prognostic model creation is the most durational one which can take up to 80% of the total project. In the article the process of data preparation for construction of the prognostic classification model for stroke patients in the conditions of Moldova hospitals.

Keywords: prognostic model, stroke patients, data preparation, feature selection, training set, test set.

1 Introduction

Process of data preparation for prognostic model creation is the most durational one which can take up to 80% of the total project [1,2].

Data preparation is not only information collection, selecting attributes for the model, but also a subsequent processing of the collected data. This process includes such stages as: selection of the information relevant to the considered problem; data cleaning; filling in of missed values of some features; work with noise and outliers – non-typical observations which are not in line with general data regularity; data filtration, i.e., their rescaling; the most significant features selection (or the insignificant ones sifting); bringing the features to certain formats (e.g. numerical, categorial, etc).

2 Data selection

The construction of models that could help a doctor in assessing the severity of the disease and the possible course of the disease with these

indicators of the patient's condition is very complex and requires a deep understanding of the most important parameters. The ambulance hospital is accumulating a database of patients who had a stroke. Not all indicators have the same effect on the course of treatment. And to help the doctor it is necessary a relatively simple tool, which we see as a mathematical model convenient to use.

Our aim is to construct the prognostic model of classification of stroke patients in the conditions of Moldova hospitals. Data for such model were taken from the data base STROKE.MD [3], in which the information about patients admitted to the urgency hospital with stroke or stroke suspicion was input.

The initial number of parameters in the data base is 172. The current number of patients is 32. For model construction only some parameters were selected, which at the initial stage are considered as relevant to the problem.

It is necessary to mention that such information about a patient as his first name, last name, personal address and phone initially are not taken into consideration, because it does not influence the problem. There are such parameters in the data base as "Deep vein thrombosis (DVT)", "Rheumatism" which we would like to include into our model, but they are not taken, because these data in the information about patients mostly is absent.

As the result, the following 39 parameters were taken as the starting point: 1. physical activity; 2. coffee consumption; 3. age; 4. type of home place; 5. sex; 6. smoking; 7. migraine; 8. valvulopathy; 9. dyslipidemia; 10. myocardial infarction; 11. hypertension; 12. transient cerebral attack; 13. diabetes; 14. septic endocarditis; 15. stroke; 16. heart failure; 17. degree of obesity; 18. type of obesity; 19. main artery stenosis; 20. prosthetic heart valves; 21. atrial fibrillation; 22. ischemic heart disease; 23. Rankin scale; 24. NIHSS- volume of eyeball movements; 25. NIHSS-field of view; 26. NIHSS- muscle strength in the left leg; 27. NIHSS-muscle strength in the right leg; 28. NIHSS-ataxia; 29. NIHSS-reaction to stimuli; 30. NIHSS-dysarthria; 31. NIHSS-facial paralysis; 32. NIHSS-muscle strength in the right hand; 33. NIHSS-muscle strength in the left hand; 34. NIHSS-sensitivity; 35. NIHSS-orders fulfilling; 36. NIHSS-

level of consciousness; 37. NIHSS-Coherent answer to two questions; 38. NIHSS-aphasia; 39. gravity.

3 Feature selection

At the next stage significant features selection has been performed. The means of WEKA platform [4] were used for this purpose. The "Select Attributes" tools provides 8 Attribute Evaluators for significant features ranking and selection.

Table 1. Feature selection on existing data set

Attribute Evaluator + Search Method	Selected Attributes
weka.attributeSelection.CorrelationAttributeEval weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1	**28,29,38,30,24,25,23,32,34,31,33,27,26,35,37,7,13,11,17,19,9,10,3,4,21,18,16,1,6,22,36,8,15,2,5,20,14,12 : 38**
weka.attributeSelection.CfsSubsetEval -P 1 -E 1 weka.attributeSelection.GreedyStepwise -T -1.7976931348623157E308 -N -1 -num-slots 1	**6,11,23,24,25,28,30,33 : 8**
weka.attributeSelection.GainRatioAttributeEval weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1	**24,28,30,23,25,29,38,34,31,32,33,26,27,35,11,37,19,21,3,17,13,18,6,8,36,4,22,10,16,7,1,9,2,15,5,12,20,14 : 38**
weka.attributeSelection.InfoGainAttributeEval weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1	**23,30,28,34,31,29,38,32,24,25,11,33,26,27,3,17,6,19,22,18,35,13,37,36,1,10,8,4,16,7,21,9,2,15,5,12,14,20 : 38**
weka.attributeSelection.OneRAttributeEval -S 1 -F 10 -B 6 + weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1	**30,23,33,31,29,28,38,34,26,4,32,3,36,27,11,6,9,19,21,24,1,22,13,7,35,18,16,5,25,20,15,14,12,37,2,8,17,10 : 38**
weka.attributeSelection.PrincipalComponents -R 0.95 -A 5 +weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1	**1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 : 16 - these are formulae**
weka.attributeSelection.ReliefFAttributeEval -M -1 -D 1 -K 10 +weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1	**30,23,28,34,24,29,31,38,32,25,33,26,27,35,37,11,19,13,16,7,4,17,3,20,14,12,9,18,21,10,1,6,36,8,22,5,15,2 : 38**
weka.attributeSelection.SymmetricalUncertAttributeEval+weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1	**23,30,28,24,34,29,31,38,32,25,33,26,11,27,3,17,19,6,35,37,22,18,13,36,8,10,1,4,16,21,7,9,2,15,5,12,14,20 : 38**

Having analyzed ranked chains in different Evaluators, we initially had picked out candidates for elimination according to the first applied by us Attribute Evaluator «weka.attributeSelection.CorrelationAttribute Eval»: 9,10,3,4,21,18,16,1,6,22,36,8,15,2,5,20,14,12. In the rest of Attribute Evaluators almost all these candidates are the least significant too (at different degree). But some of these candidates (3,4,36,6,21)

according to other Attribute Evaluators fall into the more significant ones. That is why we decided to leave them among the significant parameters.

So, the following parameters were taken for the model creation: 28,29,38,30,24,25,23,32,34,31,33,27,26,35,37,7,13,11,17,19,3,4,36,6,21

The parameters type is categorial. The parameter "gravity" was defined as class with the values: Light, Moderate, Moderate- Severe, Severe.

All the data sets (32 observations objects, i.e. patients) are divided in 2 parts: training set (21 of patients) and test set (11 patients).

4 Classification models construction and their performance comparison

Further on the model of patients classification into 4 classes is constructed (according to the values of class Gravity).

Table 2. Correctly Classified Instances on training and test sets (methods of classification in WEKA)

Method	**Correctly Classified Instances %**			
	training set		**Test set**	
	25 attributes	**20 attributes**	**25 attributes**	**20 attributes**
rules.ZeroR	23.8 under CV 28.5 whole set	28.5 whole set		
trees.J48	80.9 under CV 95.2 whole set	95.2 whole set	81.8	54.5
functions.Logistic -R	100	100	90.9	81.8
bayes.NaiveBayes	100	95.2	81.8	81.8
lazy.IBk	100	100	81.8	90.9
trees.REPTree	76.19	76.19	81.8	81.8
functions.SMO	100	100	81.8	90.9

Since the number of parameters as compared to the number of patients remains large enough, we decided to make the next iteration decreasing the number of parameters.

Additional candidates for parameters exclusion are: 37,7,13,11,17,19. But we do not eliminate parameter 11 because in some evaluators it falls in the first half of the significant parameters. So, we eliminated: 37,7,13,17,19.

As a result, we eliminate the following 18 parameters from the initially selected list: 9,10, 18,16,1, 22, 8,15,2,5,20,14,12, 37,7,13,17,19.

The following attributes have left (with rank indication according to Attribute Evaluator «weka.attributeSelection.CorrelationAttributeEval»):

Ranked attributes:

0.5099 28 NIHSS-ataxia
0.447 29 NIHSS-reaction to stimuli
0.4378 38 NIHSS-aphasia
0.4332 30 NIHSS-dysarthria
0.4303 24 NIHSS- volume of eyeball movements
0.3954 25 NIHSS- field of view
0.3912 23 Rankin scale
0.3799 32 NIHSS-muscle strength in the right hand
0.3579 34 NIHSS-sensitivity
0.3465 31 NIHSS-facial paralysis
0.2922 33 NIHSS-muscle strength in the left hand
0.289 27 NIHSS-muscle strength in the right leg
0.2577 26 NIHSS- muscle strength in the left leg
0.2461 35 NIHSS-orders fulfilling
0.2219 11 hypertension
0.2036 10 Myocardial infarction
0.1994 3 age
0.1962 4 type of home place
0.1945 21 atrial fibrillation
0.1664 6 smoking
0.1239 36 NIHSS-level of consciousness

Based the remained 20 attributes we construct classification models using WEKA platform. Table 2 shows the comparison results. We see that the number of Correctly Classified Instances (in %) for 25 and 20

attributes is comparable, and for some methods (lazy.Ibk and SMO) the result is even better.

The next stage in data preparation and significant features selection was the attempt to exclude parameters with Rankin and NIHSS scales. From the remained (without Rankin and NIHSS scales) there were eliminated 6 parameters (following the procedure of feature selection). But, the work with the classification methods showed the results worse than at the previous iteration.

5 Conclusion

Further on we intend to continue consultations with physicians – specialists in the domain of stroke diagnostics – for specification from their point of view the order of parameters exclusion. Also the data improvement and the increase of number of observation objects is supposed.

Acknowledgments. A part of the research for this paper is supported by the project "Mathematical modeling of risk factors and clustering of patients for preventive management of stroke", 16.00418.80.07A.

References

[1] I.A. Chiubukova. *Data Mining*, (1chubukova_i_a_data_mining.pdf). (in Russian)

[2] V.Kitov. *Practical aspects of Machine learning*, Open systems. DBMS, vol.1, 2016, https://www.osp.ru/os/2016/01/13048648/. (in Russian)

[3] E. Zamsa. *Medical Software User Interfaces, Stroke MD application design*. Proceedings of the 5th IEEE International Conference on e-Health and Bioengineering, 2015 IEEE, 4 p. (format electronic).

[4] https://www.cs.waikato.ac.nz/ml/weka/downloading.html?__s=mayfbc9vfaq7kwudf2nx

S. Cojocaru[1], C. Gaindric[2], G. Magariu[3], T. Verlan[4]

Institute of Mathematics and Computer Science, Academy of Sciences of Moldova
E-mails: svetlana.cojocaru@math.md[1], constantin.gaindric@math.md[2], galina.magariu@math.md[3], tatiana.verlan@math.md[4]

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

Towards an Algebraic Explication of Quantity

Ioachim Drugus

Abstract

There are two aspects of quantity, which in natural languages manifest as "mass-count distinction" – one treated by measure theory, and the other treated by arithmetic. However, a mathematical structure explicating the notion of quantity under both these aspects sounds to be missing in literature. To fill this gap, in current paper a class of universal algebras called "metrologic algebras" are introduced, which explicate both aspects of quantity, and treat them on a common basis.

Keywords: generalized Boolean algebra, measure theory, mass-count distinction, metrology.

1 Introduction

Quantity is a key notion of mathematics, invoked in practically all definitions of this discipline. A concrete manifestation of quantity is called "magnitude". We split the class of magnitudes into two classes - those, which are said to be "measured", an aspect explicated by measure theory, and those, which are obtained as a result of counting, an aspect explicated by arithmetic. It sounds natural to treat the process of counting as a partial kind of the process of measuring, but it would sound unacceptable to treat measuring as a partial kind of counting. Therefore, from among two terms, "measure" and "count", the term "measure" was selected to serve here as a general term for the process of determining a magnitude (one *cannot* say "determine the *value* of a magnitude", since the meaning of Latin "magnitude" is "largeness" or "smallness", which can be treated as "value").

One needs a term for magnitudes said to be "measured", but which cannot be said to be "counted", like length, volume, duration, and similar ones. Each such magnitude is usually said to be the measure of an "extent" or "extension" in space (like length or volume), in time (like duration) or within any other "domain" or "range" (like luminous intensity or thermodynamic temperature). The term "extent" is chosen here to refer to such a magnitude. In addition to magnitudes called "extents", there are magnitudes which are obtained by counting. These are referenced here as "counts" (like in English expressions "a count of" or "head count", i.e. count of heads). To sum up, the assumption is made here that there are exactly two types of magnitudes – "extents" and "counts".

The grammars of European languages make a sharp distinction between two kinds of quantity – a distinction referenced as "mass-count distinction" [1], and this distinction must reflect an objective phenomenon external to languages. In this paper, to cast some light upon this phenomenon, we will discuss about an algebra of extents, which is called "extent algebra", an algebra of "counts" which is a mono-unary algebra (or a "unar"), and an algebra of both the extents and the counts, which is called "metrologic algebra". The last term comes from "metrology" which, according to International Bureau of Weights and Measures, is "the science of measurement, embracing both experimental and theoretical determinations at any level of uncertainty in any field of science and technology".

The term "metrologic algebra" is descriptive, i.e. it refers to many kinds of algebras, including those which might appear in the future, and not only those defined in this paper. This choice of a term was made to emphasize that the algebra introduced here is not claimed to be the only possible algebra which explicates the notion of quantity.

2 The metrologic algebra and its two reducts

The metrologic algebras are universal algebras, the signature of which consists of two symbols "+", "–" for two binary operations, and a sym-

bol "′" (prime) for a unary operation, called "successor operation". The axioms of metrologic algebras state that with respect to the operation "+", these are commutative idempotent monoids with the following additional axioms:

$$(a + b) - c = (a - c) + (b - c), \tag{1}$$
$$a - (b + c) = (a - b) - c, \tag{2}$$
$$a + (b - a) = a + b, \tag{3}$$
$$a + (a - b) = a, \tag{4}$$
$$(a - b) - c = (a - c) - (b - c), \tag{5}$$
$$a - (b - c) = (a - b) + (a - (a - c)), \tag{6}$$

In order that these axioms make sense, consider that the symbol "+" denotes the set-theoretic union usually denoted as "∪", and the symbol "–" denotes the set-theoretic difference usually denoted as "\".

The axioms of "metrologic algebra" defined in this paper are exactly the same as the axioms of an algebra defined in [2]. However, the metrologic algebra has one extra operation, a unary operation "′" and this changes its type of a universal algebra. To get an intutive idea of the operation "′", treat it as the "singleton operation" defined in the universe of discourse of any set theory, which results in the one-element set $\{x\}$ (singleton) when applied to any object x, be x an atom or a set. Complying with this intuition is a metrologic algebra with the invertible operation, i.e. with the following additional axiom:

$$a' = b' \to a = b. \tag{7}$$

However, there might exist alternative axiomatic set theories, where the statement (7) does not hold, and for generality sake, this statement was not postulated as an axiom of metrologic algebra. Thus, no axioms mentioning the successor operation are postulated – a situation similar to the ZF-algebras described below.

The reduct of metrologic algebra with only the two symbols "+" and "–" in the signature is called "extent algebra" because this algebra

is intended as domain of definition of the function of measure in most general case. The section 4 gives more detail on this. The reduct of metrologic algebra with only one symbol "′" in the signature is what the algebraists mono-unary algebra (or a "unar").

3 The correlations of metrologic algebras with the ZF-algebras

The Zermelo-Fraenkel algebras ("ZF-algebras") were introduced to serve as models of the ZF set theory in the widely accepted algebraization of set theory [3]. A ZF-algebra is a complete sup-lattice (that is a complete lattice regarded as a complete upper semilattice in homomorphisms), equipped with an additional unary operation called "successor operation". In [3], it was proved that for a ZF-algebra A:

(a) A is a model of ZF set theory, iff A is a free ZF algebra;

(b) If A is a free algebra then A is a "big algebra" - i.e. its support is necessarily a proper class.

Given the free ZF-algebra V, one can "recover" the membership relation between sets from the ZF-algebra structure by setting

$$a \in b \leftrightarrow a \leq b,$$

where "$\leq$" is the relation of partial order in the sup-lattice.

The fact that ZF-algebras can be big algebras manifest in some peculiar features of these algebras. In particular, without enough attention to details one can get to wrong conclusions and even to contraditions in reasoning about these algebras. For example, one can judge this manner: since a ZF-algebra A is a complete lattice, the union of all its elements U must be an element of A, and thus all ZF-algebras have a top element. But, ZF set theory does not allow such a set to exist and, thus, no ZF-algebra can be a model of ZF set theory - a wrong statement, given that these algebras were specifically designed to serve as models for ZF set theory. The mistake in the reasoning which brought to the wrong statement is that one can take the union of a set, but

not of a proper class, and the object U mentioned above cannot be obtained. Generally, the notion of completeness should be defined for big algebras differently than for the regular universal algebras.

One cannot treat a complete sup-lattice as a universal algebra in the customary treatment of this notion. Namely, a universal algebra is defined as an algebraic structure with a certain signature, which is a set of symbols of n-ary operations, where n is a non-negative integer – one says that the signature of universal algebras has only "finitary operations". The operation of supremum is an infinitary operation and, thus, a ZF-algebra is not a universal algebra. On the other hand, the metrologic algebras are universal algebras by definition, and using them, one is not same prone to mistakes like the mistake with the object U. A correlation between metrologic algebras and the ZF-algebras is determined by the correlation between lattices and complete lattices – the metrologic algebras are a generalization of ZF-algebras.

4 Extent algebras for measure theory

The work on laying down the foundations of the measure theory was motivated by the discovery of unmeasurable figures on plane or unmeasurible bodies in space. In [4], the notion of measure is defined as an additive function with a boolean algebra as its domain of definition. This treatment of measure complies with the intuition of measure as an additive function, but due to the fact that a boolean algebra has its top element, is limited to objects of a limited measure. Such a definition of measure is good for probability theory where the probability of an event is limited by 1 from above.

Stone [5] introduced the "generalized Boolean algebras" (GBA) and gave as most representative example of this notion the algebra of Lebesgue or Borel measurable sets. The main result of [2] is the introduction of an alternative form of GBA - the form of a monoid (one binary operation with its inverse), called "extension algebra". Since the word "extension" can incorrectly invoke the meaning of "extension of (something)" instead of the meaning intended here, in this paper

the name of this algebra was changed to "extent algebra". From the perspective of abstract algebra, this new form of GBA can be referenced as "Boolean magma" (or "Boolean groupoid" using an older term "groupoid" used for what is currently preferably called "magma").

References

[1] Nicolas, D. (2008). Mass Nouns and Plural Logic. Linguistics and Philosophy 31.2, (2008), pp. 211-244.

[2] Drugus, I. *Generalized Boolean Algebras as Single Composition Systems for Measure Theory.* Proceedings of the 4th Conference of Mathematical Society of Moldova (CMSM42017), June 28 – July 2, 2017, Chisinau, Republic of Moldova, (2017), pp. 75-78.

[3] A. Joyal and I. Moerdijk, Algebraic Set Theory. Cambridge University Press, Cambridge, 1995.

[4] A. Horn, A. Tarski, *Measures in Boolean algebras*, Trans. Amer. Math. Soc. 64 (1948), 467–497.

[5] M. H. Stone. *Postulates for Boolean Algebras and Generalized Boolean Algebras.* American Journal of Mathematics, Vol. 57, no 4 (1935), pp. 703-732.

Ioachim Drugus

Institute of Mathematics and Computer Science, Academy of Sciences of Moldova
E-mail: ioachim.drugus@math.md

Automation, Computer Supported Decision-Making, and The New Enabling Information and Communication Technologies

Florin Gheorghe Filip

In recent years, important progresses have been noticed in information and communication technologies (I&CT) and platforms. They have had seriously influenced the industry business models and human's skills, knowledge, and ways of behavior. The scope of automation has been enlarged and its fundamental concepts have been evolved. The traditional position of the human agent in control (Fitts, 1951; Bainbridge, 1963; Dekker, Woods, 2002) and decision-making processes (Drucker, 1967) and his/her interaction and "division of labor" with the computer are apparently radically changed (Dewhurst, Willmott, 2014). Collaboration engineering (Nunamaker et al 2015), a specific methodology supported by a toolset of modern information and communication technologies, got ever more traction.

This survey paper aims at reviewing the above mentioned developments with a particular emphasis on the collaborative decision-making activities and supporting tools in control and management (C&M) settings. The attributes of the modern organization and collaborative networks (Camarinha-Matos, Afsarmanesh, 2005), which are characterized by an ever increased degree of intra-and inter-enterprise collaboration are reviewed first. Then the transition of the C&M schemes from genuine multilevel rigid hierarchical structures (Mestrovic et al 1970) to more and more cooperative schemes (Nof et al, 2015) are described. The significant developments in the Decision Support Systems (DSS) domain are highlighted together with the results of a survey of the

published papers (Filip et al 2014). Particular emphasis is put on the description of specific requirements and practical solutions proposed over the time in the domain of Group DSS. Several advanced key I&CT such as Big Data (Chen et al, 2012; Shi, 2016; Boncea et al, 2017), BI&A (business intelligence and analytics), web technology, social networks, mobile and cloud computing that enable collaboration (Filip et al, 2017) are reviewed and their impact on collaborative management and control activities is discussed. The criteria of choosing the appropriate I&CT tools and platforms and the usage of various Multi Attribute Decision Models-MADM (Zavadskas et al, 2014) to support the designer decisions is described together with the possible mistakes to be avoided. Several practical applications and platforms designed to enable collaborative C&M activities (Bitterman, et al,2014; Brandas et al ,2016; Candea, Filip, 2016) are eventually briefly described.

References

[1] Bainbridge L. (1983). *Ironies of automation*. IFAC J. Automatica 19(6), pp.775-779.

[2] Bitterman T., et al. (2014). *Simulation as a service (SMaaS): a cloud-based framework to support the educational use of scientific software.* International Journal of Cloud Computing 1, 3(2), pp.177-190.

[3] Boncea R., Petre I., Smada D.M., and Zmarandoiu A.Z. (2017). *A Maturity Analysis of Big Data Technologies*. Informatica Economica, 21(1), pp.60.

[4] Brandas C., Panzaru C., Filip F. G. (2016). *Data driven decision support systems: an application case in labour market analysis*. ROMJIST, 19(1-2), pp. 65–77.

[5] Camarinha-Matos L. M., Afsarmanesh H. (2005). *Collaborative networks: a new scientific discipline.* Journal of Intelligent Manufacturing, 16(4-5), pp. 439-452.

[6] Candea C., Filip F.G. (2016). *Towards intelligent collaborative decision support platforms*. Studies in Informatics and Control, 25(2), p.143-152.

[7] Chen H., Chiang R. H. L., Storey V. C. (2012). *Business Intelligence and Analytics: from Big Data to Big Impact*. MIS Quarterly, 36(4), pp. 1-24.

[8] Dekker S.W., Woods D.D. (2002). *MABA-MABA or abracadabra? Progress on human–automation co-ordination*. Cognition, Technology & Work, 4(4), pp.240-244.

[9] Dewhurst M., Willmott P. (2014). *Manager and machine: The new leadership equation.* McKinsey Quarterly.

[10] Drucker P. (1967). *The manager and the moron*. In: Drucker P, Technology, Management and Society: Essays by Peter F. Drucker, Harper & Row, New York: pp. 166-177.

[11] Filip F.G., Suduc A.M., Bîzoi M. (2014). *DSS in numbers*. Technological and Economic Development of Economy, 20(1), pp. 154-164.

[12] Filip F.G., Zamfirescu C.B., and Ciurea C. (2016). *Computer Supported Collaborative Decision Making*. Springer.

[13] Fitts P. M. (1951). *Human engineering for an effective air navigation and traffic control system*. Washington, DC: National Research Council.

[14] Mesarovic M D, Macko D Takahara I (1970). *Theory of Hierarchical Multilevel Systems*. Academic Press, New York.

[15] Nof S. Y., Ceroni J., Jeong W., Moghaddam M. (2015). *Revolutionizing Collaboration through e-Work*, e-Business, and e-Service. Springer.

[16] Nunamaker Jr., J. F, Romero Jr., N C, Briggs R. O. (2015). *Collaboration Systems. Part II: Foundations*. In: Nunamaker J. F. et al (eds). Collaboration Systems: Concept, Value and Use. Routledge, pp. 9-23.

[17] Shi Y.(2016). *Challenges to Engineering Management in the Big Data Era*. Frontiers of Engineering Management, pp.293-303.

[18] Syberfeldt, A., et al (2013). *A web-based platform for the simulation–optimization of industrial problems*. Computers & Industrial Engineering, 64(4), pp.987-998.

[19] Zavadskas, E.K., Turskis Z, Kildiene S. (2014). *State of art surveys of overviews on MCDM/MADM methods*. Technological and Economic Development of Economy, 20(1), pp.165-179.

Florin Gheorghe Filip

Member of the Romanian Academy
E-mail: ffilip@acad.ro

Malaria Detection System

Daniela Gîfu

Abstract: Natural Language Processing (NLP) is a research area aimed at exploiting rich knowledge resources, in order to understand and identify concepts in standardized formats. In this paper, a prompt and reliable clinical-concept extraction procedure from a corpus of full-text academic articles, using a new text mining tool, is described. The goal of this study is to develop a framework for detecting malaria concept. This method can be useful to the direct beneficiaries (health professionals), but, also, researchers in the fields of BioNLP and NLP, etc.

Keywords: academic corpora, malaria concept, annotated medical corpora, text mining, clinical-concept extraction.

1 Introduction

This study is based on clinical-concept extraction procedure from a corpus of full-text academic articles in order to implement a tool for detecting automatically the malaria concept using text mining or knowledge discovery from text [Fledman *et al.*, 1995]. The malaria disease remains a major public health problem [Hemingway and Bates, 2003], especially, in the underdeveloped countries.

Actually, this work is a continuation of a previous one [Amarandei *et al.*, 2017; Onofrei *et al.*, 2017; Curea & Gifu, 2017].

The present research is based on the question: *Text mining techniques could be applied successfully to extract concepts from biomedical text?*

The paper is structured as follows: Section 2 presents briefly relevant mining biomedical literature that reveals a large interest for clinical-concept extraction. Section 3 describes shortly the reliable clinical-concept extraction procedure from a corpus of full-text academic articles, using a keyword-generator algorithms, Section 4 presents the statistics and results interpretation. At last, section 5 highlights the conclusions and

future work focused by the developing the framework for describing biomedical concepts and clinical information systems.

2 Mining Biomedical Literature

Until now, biomedical text mining (BioNLP) uses sophisticated predictive models to understand, identify and obtain concepts from a large corpus of scientific texts in medicine, biology, biophysics, chemistry, etc. in order to reveal innovator knowledge that can increase value in biomedical research. For these aims, all language resources include complex lexicons, thesauri and ontology that cover the entire range of clinical concepts. Keizer [Keizer and Abu-Hanna, 2000; Keizer *et al.*, 2000] and Cornet [Cornet *et al.*, 2006] described a terminological and typology system to provide a uniform conceptual understanding.

3 Clinical-Concept Extraction Procedure

This section describes briefly the modules of architectural design study based on an academic corpus used for malaria concept extraction. This work is based on TF-IDF (T*erm Frequency-Inverse Document Frequency*) matrix from a bag-of-words model and LDA (*Latent Dirichlet Allocation)* topic model, discovery by [Blei *et al.*, 2003].

3.1 Corpus

Starting with a manual semantic annotation of an important lexical resource, the Colorado Richly Annotated Full Text (CRAFT) corpus[1] (97 Open Access journal articles ~ 800k tokens) [Verspoor *et al.*, 2012], this study describes an important experience made within the EUROLAN Summer School[2].

3.2 Project Architecture Diagram

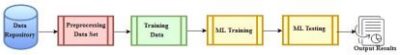


Figure 1. Study Phases Diagram

[1] http://bionlp-corpora.sourceforge.net/CRAFT/index.shtml
[2] http://eurolan.info.uaic.ro/2017

```
INPUT -> Analysis -> Pre-processing (XML to text, tokenization, lemmatization) -> Implementation of the algorithm for key concepts extraction -> OUTPUT- Set of key words/concepts -> Rating Algorithm -> Final Output -> Testing
```

For the current tool, the corpus is auto-generated by the keyword-generator algorithm from the given input.

3.3. Analysis Phases

In this study, the clinical-concept extraction procedure consists on the following phases: (1) extracting the first word from the first paragraph; (2) identifying the titles of sections; (3) using citations; (4) searching title of papers from References; (5) automatic preprocessing chain: segmentation, tokenization, part-of-speech tagging, lemmatization, number of unique words.

This research is concerned with extracting a k number of key concept from a collection of sample texts provided by the user, namely an extractor based on TF-IDF method and LDA topic modeling.

A. `keywords_tfidf.py - k` (number of keywords to be generated) – `d` (doc will be split into parts containing the specified number of words);

B. `keywords_lda.py – t` (number of topics) – `w` (number of words per topic) – `k` (number of keywords) – `d` (doc will be split into parts containing the specified number of words)

C. `keywords_tfidf.py – k 20 – d 1000`

4 Statistics and interpretation

For this study, only a quarter of the corpus has been used. The results will be improved using the rest of the corpus.

The scores obtained running the TF-IDF algorithm of the total number of concepts show that the number of false results is only with a third part bigger that the right ones.

Using LDA model to discover the topics, the results are almost similar with those obtaining running the other algorithm. However, to concentrate on medical concepts (e.g. malaria), training an LDA model, the results seem to be more precise in identifying the key concepts.

5 Conclusions and discussions

The described methodology offers a friendly and rapid malaria detection system, being language independent. Also, it offers a basis for future large-scale studies, having an important impact on reducing the amount of human effort required by semantic analysis of clinical language.

For the moment, this tool includes two predictive algorithms that show a tendency towards better performance, delimitated so more complex models do not necessarily lead to improving the results. In the future, each model would be improved and then they will be united in order to obtain better results together with a shorter running time.

Acknowledgments. I would like to thank to my master students in the 1st year of Computational Linguistics at the Faculty of Computer Science of the "Alexandru Ioan Cuza" of Iaşi that have participated in the initial phases of the acquisition and annotation of the biomedical corpus.

References

[1] Amarandei, S., Fleşcan, A., Ioniță, G., Turcu, R., Trandabat, D., Gifu, D. (2017). *Key Biomedical Concepts Extraction*, at the *Workshop on Curative Power of Medical Data*, MEDA, a satellite event of the 13th EUROLAN Summer School on Biomedical Text Processing, 10-17 September 2017, Constanța, Romania.

[2] Blei, D. M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet Allocation. In *Journal of Machine Learning Research*, 3, pp. 993-1022.

[3] Cornet R, De Keizer NF, Abu-Hanna A. (2006). *A framework for characterizing terminological systems*. In: Methods Inf Med., vol. 45, pp. 253-266.

[4] Curea, E. and Gifu, D. (2017). A *Framework for Medical Data Retrieval*, at the *Workshop on Curative Power of Medical Data*, MEDA, a satellite event of the 13th EUROLAN Summer School on Biomedical Text Processing, 10-17 September 2017, Constanța, Romania.

[5] Feldman, R. and Dagan, I. (1995). *Knowledge Discovery in Textual Databases* (KDT). In: KDD, vol. 95, pp. 112-117.

[6] Hemingway, J. and Bates, I. (2003). *Malaria: past problems and future prospects*. In: EMBO Rep., 4 (Suppl. 1), pp. 29-31, doi: 10.1038/sj.embor.embor841.

[7] de Keizer NF, Abu-Hanna A. (2000). *Understanding terminological systems II: terminology and typology*. In: Methods Inf Med., vol. 39, pp. 22-29.

[8] de Keizer NF, Abu-Hanna A, Zwetsloot-Schonl JHM. (2000). *Understanding terminological systems I: terminology and typology*. In: Methods Inf Med., vol. 39, pp. 16-21.

[9] Onofrei, M., Hulub, I., Hriscu, A., Alexa, L., Trandabat, D., Gifu, D. (2017). *Developing a Technology Allowing Automatic Identification the Author's Confidence in Biomedical Papers*, at the *Workshop on Curative Power of Medical Data*, MEDA, a satellite event of the 13th EUROLAN Summer School on Biomedical Text Processing, 10-17 September 2017, Constanța, Romania.

[10] Verspoor, K., Cohen, K.B., Lanfranchi, A., Warner, C., Johnson, H.L., Roeder, C., Choi, J.D., Funk, C., Malenkiy, Y., Eckert, M., Xue, N., Baumgartner Jr., W.A., Bada, M., Palmer, M. and Hunter, L.E. (2012). *A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools*. In: BMC Bioinformatics, 13:207.

Daniela Gîfu

Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași
Institute of Computer Sicence, Romanian Academy – Iasi Branch
e-mail: daniela.gifu@info.uaic.ro

What is Statistical Stability: Mathematical Regularity or Physical Phenomenon?

Igor Gorban

Abstract

Mathematical and physical interpretations of statistical stability of mass phenomena are analyzed. It is compared the features of the mathematical probability theory interpreting the statistical stability as a manifestation of the mathematical law of large numbers and the physical-mathematical theory of hyper-random phenomena considering the statistical stability as a physical phenomenon. Attention is focused on the necessity of a detailed study of the physical properties of the statistical stability phenomenon and the development of physical-mathematical, mathematical, and applied theories that take into account its features.

Keywords: statistical stability, violation of statistical stability, probability theory, theory of hyper-random phenomena.

1 Introduction

Statistical stability of mass phenomena, manifested in the stability of relative frequencies of events, sample averages, and other functions of samples (statistics) with a large amount of data, is known for more than 350 years. For more than 300 years, the well-known law of large numbers discovered by J. Bernoulli is known. Starting from those ancient times to this day, scientists are arguing about whether statistical stability is a mathematical regularity or a physical phenomenon.

At first glance, the question under discussion is only a theoretical interest. However, in reality this is not so. The interpretation of the

concept of statistical stability determines the strategy for the long-term development of a number of disciplines.

If statistical stability is only a manifestation of the mathematical regularity, then the widely known probability theory, which describes statistical stability, is a mathematical discipline. Then, for the development of the fundamental part of this theory and other theories associated with it, it is not necessary to carry out any experimental studies of real physical phenomena (events, quantities, processes, and fields), and the main attention should be paid to the development of probability theory as mathematical discipline.

Another strategy should be if statistical stability is a physical phenomenon. Then for its study, it is necessary to carry out experimental and theoretical studies typical for physical phenomena.

The aim of the paper is to compare mathematical and physical interpretations of the statistical stability, as well as the mathematical probability theory that treats statistical stability as a manifestation of the mathematical law of large numbers, and the physical-mathematical theory of hyper-random phenomena that treats it as a physical phenomenon.

2 Mathematical Interpretation of the Statistical Stability

Among many scientists, primarily mathematicians, the widespread belief is that statistical stability is a manifestation of the mathematical law of large numbers, and the probability theory describing it is a mathematical discipline.

This understanding of statistical stability and the probability theory has largely evolved as a result of the universal recognition and dissemination of the mathematical approach to the axiomatization of probability theory, proposed in the late 20s of the last century by A.N. Kolmogorov.

The Kolmogorov's probability theory operates not with real physical objects, but with their abstract mathematical models that are random events, variables, and functions.

The subject matter of this theory is an abstract probabilistic space, and the scope of study is links between abstract random models.

Using of the probability theory in practice is based on the hypothesis of perfect statistical stability assuming the convergence of any statistics.

3 Physical Interpretation of the Statistical Stability

Physicists and engineers, as a rule, perceive statistical stability as a physical phenomenon. Some well-known mathematicians agree with them. Among them for instance are G. Korn and T. Korn, the authors of the world-famous fundamental English reference book on mathematics [1]. Representing statistical stability (regularity), they characterize it as follows: "Statistical regularity, in each individual situation, is an empirical physical law which, like the law of gravity or the induction law, is ultimately derived from experience and not from mathematics."

Note, the authors of this handbook do not simply emphasize the physical nature of the phenomenon of statistical stability, but place it in a number of fundamental phenomena of nature, such as the phenomenon of gravity and the phenomenon of induction.

4 Research of the Statistical Stability Phenomenon

Paying tribute to the tremendous positive role in mathematical axiomatization of the probability theory, we must admit that the widespread dissemination in the mathematical environment of ignoring the physical foundations of probability theory has in fact interrupted for decades the systematic study of the physical properties of the statistical stability phenomenon.

The need for serious research into this phenomenon became apparent at the turn of the 1970s and 1980s, as a result of the discrepancy between certain assumptions of the probability theory and experimental data. In particular, this concerns the potential accuracy of measurements.

According to the probability theory, the random error in measuring of physical quantities tends to zero when the data volume infinitely increases (Cramer-Rao estimates). But the actual measurement accuracy is always limited. It is impossible to overcome the existing accuracy limit by statistical processing of the data.

Clarifying the reasons for the discrepancy between theory and practice led to the understanding that the problem is related to the unreasonable idealization of the phenomenon of statistical stability.

Experimental investigations of various processes of different physical kinds over broad observation intervals have shown [2] that the hypothesis of perfect statistical stability is not confirmed experimentally.

For relatively short temporal, spatial, or spatio-temporal observation intervals, an increase in data volume usually reduces the level of fluctuation in the statistics. However, when the volumes become very large, this tendency is no longer visible, and once a certain level is reached, the fluctuations remain practically unchanged or even grow. This indicates a lack of convergence for real statistics (their inconsistency).

5 The Theory of Hyper-random Phenomenon

Investigation of the physical properties of the statistical stability phenomenon and the development of effective methods for describing of it, taking into account the violation of the convergence of statistics, led to the formation of a new physical-mathematical theory of hyper-random phenomena.

This theory is based on the Kolmogorov's axiom system and two physical hypotheses – the hypothesis of limited statistical stability and the hypothesis of an adequate description of real physical phenomena by so called hyper-random models that are the sets of random models.

The subject matter of this theory is the statistical stability phenomenon, and the scope of study is representation of this phenomenon by hyper-random models.

6 Conclusion

The study points the need for careful study of the physical properties of the statistical stability phenomenon and development of physical-mathematical, mathematical, and applied theories, taking into account its features.

References

[1] G.A. Korn, T.M. Korn. *Mathematical Handbook for Scientists and Engineers*, 2000.

[2] I.I. Gorban. *The Statistical Stability Phenomenon*, 2016.

Igor Gorban

The Institute of Mathematical Machines and Systems Problems of National Academy of Sciences of Ukraine
E-mail: igor.gorban@yahoo.com

Analysis of data selection criteria for a given visualization method

Vadim Grinshpun

Abstract

The author considers how a set of data selection criteria affects the dataset response to the variations in quantity and quality of attributes, describing an observation, for the purposes of minimizing the complexity of the dataset while maintaining the veracity of the outcome for a desired tasking.

Keywords: computer science, big data, data visualization, multi-dimensional data, data-selection criteria, dataset reduction.

1 Introduction

The informational characteristics of a given composition of the data selection criteria in an implementation depend on the size and type of the observations under consideration. The set of selection criteria for the multi-dimensional data can be categorized into two subsets:

- The "territory" of the observations, when the characteristic indicator is universal for the whole subset and the information action between the bases is realized based on the algorithms for determining indicator's average;
- The "thematic" focus, when any private base of the lowest level can have the maximum detailed information flow and a specific rubricator of characteristics and transfers to the general database only some summary indicators;

2 Model Definition

The model of the observations dataset response to the variations in the degree of multiple parameter influences and the quantitative analysis of

the normative indices of the cause-effect factors' influences can generally be modeled as building and subsequent analysis of a system of equations:

$$\begin{aligned} \mathrm{H}\,(x,y,t) &= \mathrm{F1}(\mathrm{h0},\mathrm{h1},\mathrm{h2},\mathrm{h3}\,\ldots.x,y,t) \\ \mathrm{X}(x,y,t) &= \mathrm{F2}(\mathrm{x0},\mathrm{x1},\mathrm{x2},\mathrm{x3}\,\ldots.x,y,t) \end{aligned} \quad (1)$$

Where $\mathrm{H}\,(x,y,t)$ – is an indicator of quality of the observations integrated over the time-period t, calculated for a given (x,y) point;

$(h0,h1,h2,h3\,\ldots.x,y,t)$ and $(\mathrm{x0},\mathrm{x1},\mathrm{x2},\mathrm{x3}\,\ldots.x,y,t)$ – sets of characteristics, describing the corresponding states of the original observations and factors influencing the objects at the same time-space coordinates;

F1– is a function describing the strength of influences and the final state of the objects being observed;

F2– is a function describing the relationship between the quality of the object under observation and its state;

The quality of the observed object H often is unidimensional, and can take very few pre-determined values. Having said this, it is still reasonably difficult to offer a clear line of demarcation between the characteristics sets H and X, so any indicators evaluating the object's response to the external disturbances are called the "response" functions of the objects under observation [1].

To solve the problem the database is being loaded with various characteristics that indicate the overall state of all observations within the territory being researched, to be use as a basis for deriving the response indicators; additionally, subgroups of "typical" representatives of observable objects identified based on the structural or functional similarities, are being loaded including the lower-level details.

Besides the characteristics derived directly from the observable objects and their groupings, the database is also being loaded or automatically generated various indices traditionally used for indicator processing [2], as shown in the (Fig. 1):

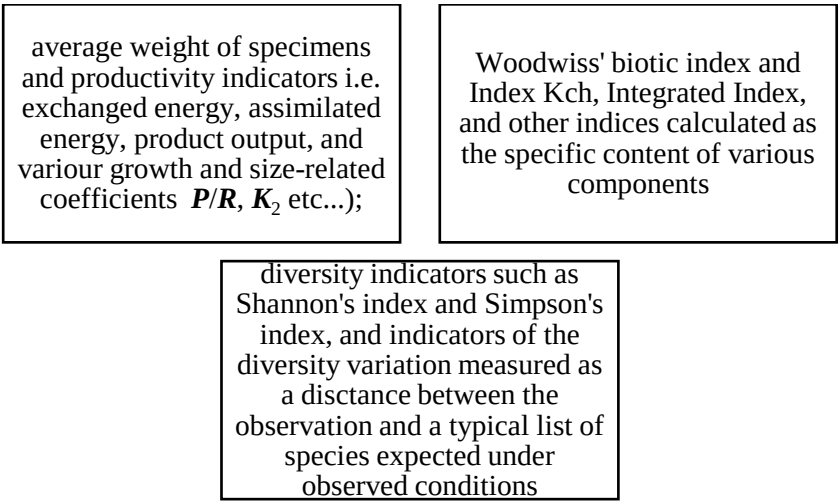


Figure 1. An example of characteristics derived directly from the observation.

The main limitation for the practical application of tables of factor indicators for an adequate description of the observable object is the insufficient reliability, associated with stationary measurement errors and the influence of subjective factors, the uncertainty of the formation algorithms, temporal and spatial inhomogeneity. Indicators that are loaded into databases can have the most diverse units of measure, scale, reference points and periods of variation.

Empirical series of modifications adapt to the laws of distributions that are very far from theoretically normal or measured, and the graphs of dependencies often have the form of apparent fluctuations approaching "white noise". Thus, for these characteristics to become not only an essential, but also an establishing link in the information characteristics of the model of the observed territory, it is necessary to develop effective procedures for preprocessing the initial data to determine robust estimates of the observed indicators.

3 Processing of Individual Characteristics

Let’s assume that the application used for the data analysis can perform the functions listed in Fig 2 [2].

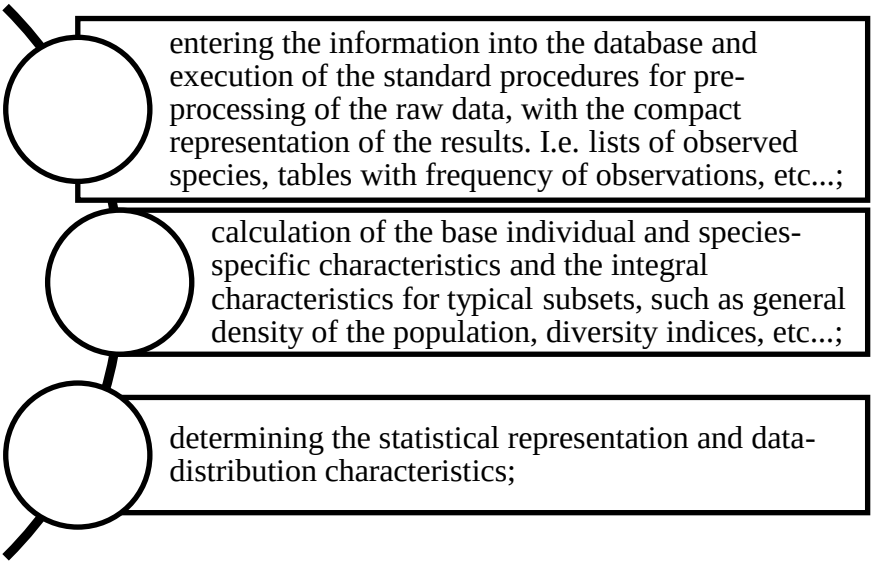


Figure 2. Implemented Functions available for data analysis.

The initial stage of mathematical processing is the estimation of the level of spatial and periodic inhomogeneity of the investigated factors and the quantitative characteristics of the observation states, i.e. for each analyzed indicator, the relevance of the statistical hypothesis on the effect of the measurement site (**x-y** coordinates) or the existence of interval dynamic $X = f(t)$ [3].

The verification of the assumptions about the homogeneous structure of the series' variances is accomplished through the multifactor dispersion analysis utilizing the Fisher's criterion and as well as Friedman's and Page's criteria [4].

If spatial inhomogeneity of the considered set of observed factors and system characteristics is represented, then in this case the further formation of the task of territorial division into regions, the construction of dependencies of the response from significant factors is authorized. In the opposite case, the object of observation is treated as a single and homogeneous, and the available data on it are applied to the highest degree for the construction of geographically-objective models. If the hypothesis about the heterogeneity of economic objects is accepted, then

the typical problem of analyzing transient series, constructing multifactorial trend autonomies and economic forecasting is solved.

Exploratory Data Research assumes a special set of service mechanisms - the "sifting" of information to determine in it the features and artifacts that are specified by the image of templates or thresholds. Along with the model data on the identification of "emissions", filtering the incorrect and recreating the missed indicators, which are carried out by traditional methods, one can notice the author's algorithms of functional transformations, rationing and translation of the characteristics of the initial data into interval series [3].

The proximity of the separation of empirical samples to normal law by the functional transformation method is an effective measure of the growth of the validity of a typical statistical analysis. To do this the X line developed through application methods undergoes a series of transformations:

$$Xn(n = \pm 3,\ \pm 2, -1,\ \pm 0.5,\ \pm 0.33), \ln(X), \mathrm{ek}X \qquad (2)$$

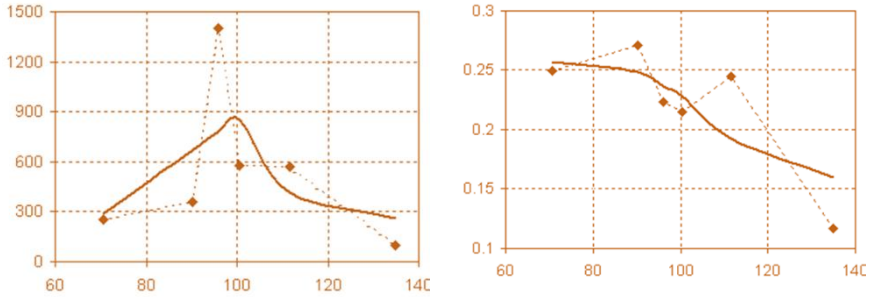
And for any of the observed samples, the asymmetry and kurtosis. Of the whole transformation formulas, there is one that brings a minimum of l-indicator, and, often, the hypothesis of the normality of the modified sample is made reliable. For such purposes, the use of Fisher's angular transformation $arcsin\ (2X - 1)$ [5].

The presentation of contiguous characteristics as normalized and interval-based graphs is also widely used for the formation of a uniform space of characteristics and the generation of complex indicators. The general principle of this transformation involves achieving large amounts of data in encoded data, which can be if the uniform distribution of the variant is formed over data intervals.

In a case like that, the following algorithm can be applied for the migration of the data from the quantitative to the ordinal scale: the universal range of the admissible characteristics of the indicator is broken by the number of classes per n segments with a length proportional to the number of measurements of each class in the starting sample, that is, $\Delta xk = pk/p$, where "pk" - the number of measurements of class "k", and "p" - the total number of measurements [6].

The most common improved Shannon’s entropy algorithm can be applied for the optimal partitioning and to establishing both the range

limits and the best number of gradations n. The developed interval and binary data structures are, to a lesser extent than the objects of observation themselves, sensitive to systematic and random errors, since the following mathematical processing is subjected not to variations, but to their normalizing frequencies of hits into the cell intervals. This circumstance is used in the execution of many the following ordinal methods.



$\boldsymbol{N_h}$ (F = 2,55, p = 0,029) $\boldsymbol{X_h/X_{ch}}$ (F = 1,83, p = 0,11)

Figure 3. Modification of characteristic factors of the observations.

The execution of the algorithm for determining the optimality of the decomposition δ of the domain of the exponent x_q at the gradation boundaries, which optimally emphasizes the discriminant basis of some external a priori systematization of the measurements of K_1, K_2, …, K_n can also be of a interest. This method of formalization was developed by AA Genkin in 1999 [7].

Note the frequency of hits of the values $p_j(x_q\,|K_s)$ of the exponent x_q from subsets $\{x_q\}_{Ds}$ in the j-th range (j = 1,2, ..., k). In this case, for the 2 classes of K_S and K_l, as the best partitioning of the ranges into k segments, we choose one that enhances the information space of the Kulbak divergence, which has the significance of a generalized estimate of the difference of 2 empirical separations:

$$\left[U\left(\frac{K_S}{K_l}\right)=\sum_{j=1}^{k}\left(p_j\left(\frac{x_q}{K}\right)-p_j\left(\frac{x_q}{K_l}\right)\right)*ln\frac{p_j\left(\frac{x_q}{K}\right)}{p_j\left(\frac{x_q}{K_l}\right)}\rightarrow max,\right. \qquad (3)$$

The boundary characteristics of the intervals are easily located both in the half-sum of the adjacent sorted indicators x_{q_i} of training sample data relating to various ranges. In the general case of n classes, the indicator is maximized:

$$J = \sum_{s=1}^{n} \sum_{l=1}^{s} J(\frac{K_S}{K_l}; x_q), \quad (4)$$

Thus, the results of sample observations that can be presented in databases are usually related to a specific date, and the totality of characteristics of the indicator in question for successive time periods creates a dynamic series. Any value of a time series is created under the influence of many obvious or latent determining factors that can be conditionally divided into three classes:

- factors forming the universal long-term tendency of the series;
- factors that form cyclical fluctuations;
- unidentifiable factors that cause stochastic fluctuations in multidimensional data.

4 Conclusion

The actual process of formation of the time series takes different forms in a variety of combinations of these factors, so the main task of modeling is to quantify each of the above characteristics in order to apply the obtained information data to predict future indicators or when modeling the relationships of multidimensional data.

If we analyze many practical applications of the work of researchers, especially when analyzing the seasonal components, it turns out to be sufficient to use traditional smoothing mechanisms, to calculate the autocorrelation relationship or to construct a stochastic model that includes the lagged indicators of the effective trait as independent characteristics.

References

[1] O.G. Berestneva, Ia.S. Pekker, K.A. Sharopin, V.A. Volovodenko. *Disovering hidden patterns in the medical and socio-psychological research*, in 2nd Int-l Conference on Applied Information Systems, Moscow, 2010.

[2] O.V. Marukhina, O.G. Berestneva, V.A. Volovodenko, K.A. Sharopin. *Technologies for visualization of the experimental research results.*

/Informaiton and Mathematical Technologies in science and management. Records of XVI Bajkal Conference, Irkutsk 2010, Part 3, pp 165–171.

[3] K.A. Sharoshin, O.G. Berestneva, G.I. Shkatova. *Visualization of the results of the experimental research.* Tomsk Politechnical University News 2010-T.136, pp. 172-176.

[4] Lu L.F., Huang M.L., Huang T.H. *A New Axes Re-Ordering Method in Parallel Coordinates Visualization*. // Proc. of the 11th International Conference on Machine Learning and Applications (ICMLA), 2012, 12-15 December, Boca Raton, USA. Boca Raton, 2012. pp. 252-257. DOI:10.1109/ICMLA.2012.148.

[5] Kendall, M. Stuart, A. *The advanced theory of statistics. Vol.2: Inference and relationship.* - London: Griffin, 1979, 4th ed.

[6] A. U. Zinoviev, A.A. Pitenk. *Mapping of an arbitrary datasets*. // "Students and the Scientific Progress": Information technology. Proceeds of the XXXVIII International Scientific Student Conference. Novosibirsk: NSU. - 2000. - pp. 38.

[7] Zhigirev N.N., Korzh V.V., Onykii B.N. *Use of asymmetry of frequency properties of information signs for constructing automated systems for classification of text documents* // Proceedings of the All-Russian Scientific Conference "Neuroinformatics". Moscow, 1999. Part 3. pp. 83-91.

Vadim Grinshpun

Institute of Mathematics and Informatics, Academy of Science of Moldova
E-mail:vgrinshpun@hotmail.com

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

On Implementation of the Composition-nominative Approach to Program Formalization in Mizar System

Ievgen Ivanov, Artur Korniłowicz, Mykola Nikitchenko

Abstract

In this talk we describe an ongoing work on implementation of the composition-nominative approach to program formalization in Mizar proof assistant based on the first-order logic and axiomatic set theory. The further aim of this work is development of a formal verification tool for software which processes and communicates with complex forms of data.

Keywords: Formal methods, program semantics, semistructured data, formalization, proof assistant.

1 Introduction

Formal verification of software systems has been a topic of interest of researches in computer science for more than fifty years. During this period many formal software verification tools based on different theoretical frameworks (automata theory, first-order logic, dynamic logics, program logics, etc.) were developed, but most of the existing tools are still in research stage and their usage in software industry is negligible. Some reasons include:

- they do not integrate well into typical software development cycles;
- successful practical application of such tools requires specialized knowledge, is labour intensive, time consuming, and not cost effective for most software projects.

However, in industries related to development of safety-critical systems such as aerospace, automotive, health technology formal verification of software plays a more significant, but still limited role.

The well known tools that support or aid formal verification of software for safety-critical systems include:

- model checkers such as Simulink Design Verifier, Systerel Smart Solver;

- verified translators and compilers such as CompCert that generate code in a low-level language or machine code from the source code in a high-level language that is proven to be equivalent to the source code under the assumptions of the formalized source and target language semantics;

- microkernels and hypervisors such as seL4 and CertiKOS that are formally verified with respect to formal specifications of their application programming interfaces and formal models of microprocessor instruction set architectures.

These tools allow one to eliminate some sources of deviations of software implementation from its specification and the implied safety problems, however, they and their underlying theoretical frameworks have serious limitations – they focus on verification of

- system software;

- software which performs basic logical operations or numerical computations (e.g. software controllers);

- communication protocols which involve simple types of data;

but lack sufficiently easily usable methods of verification of

- software which performs which complex processing of partially structured (semistructured) data;

- communication protocols which involve complex types of data.

Besides, their application requires specialized knowledge, is labour intensive, and time consuming.

These limitations are a factor that may prevent expansion of the mentioned tools and theoretical frameworks outside of traditional safety-critical systems domains. Emerging high tech areas like the Internet of Things (IoT) rely on the idea of combining software systems

and hardware devices and physical objects which involve complex interaction protocols, large-scale interaction, and processing of large volumes of semistructured data e.g. in home automation, smart buildings, smart cities, etc. Errors in IoT software can impact the real world and lead to cyber security breaches or direct hazards to humans, but due to the nature of IoT systems in each particular case the potential impact of software errors is difficult to asses. Moreover, such errors are difficult to eliminate through testing, because IoT systems have to be able to function under variety of circumstances which are costly to model or reproduce. Thus IoT and other relevant high-tech areas could benefit from introduction of formal software verification approaches to their systems development processes.

Software for the Internet of Things (IoT) and other emerging high-tech areas has some differences from traditional safety-critical software which make application of the state-of-the-art verification tools to it difficult. One of them is processing of complex, usually semistructured types of data, instead of simple types of data such as logical values or numbers. Usually such data are encoded in data formats like JSON and XML which have tree-like, hierarchical nature.

2 Main result

The work described in this talk aims to implement a formal verification tool which may overcome some limitations of the existing verification tools which prevent them to deal with software which processes complex, semistructured types of data.

The implementation of this tool is based on the composition-nominative approach to program formalization [11, 12] which is a development of composition programming [13], and the Mizar system [1], a software for formalizing mathematical theories (proof assistant) based on first-order logic and Tarski-Grothendieck set theory and a library of already formalized theories (Mizar Mathematical Library).

Composition-nominative approach provides the means of formalization of data – the notion of nominative data which is able to uniformly

represent common forms of data used in programming (e.g. lists, trees, tables, multidimensional arrays, etc.), mathematical models of software which operates on such data based on generalization of Glushkov algorithmic algebras [14], and a logic for reasoning about properties and correctness of such software – a generalized Floyd-Hoare logic [2, 3] with partial pre- and post-conditions for programs which operate on nominative data [9, 10]. The Mizar system provides an environment where the notions, models and logics of the composition-nominative approach can be formalized and implemented. In more detail the plan of their implementation in Mizar is described in [6].

A benefit of usage of Mizar as such an environment is that the Mizar Mathematical Library includes a large amount of notions and facts about continuous mathematics which allow formalization of mathematical models of physical aspects of IoT systems. More details on the link between the composition-nominative approach and the mathematical systems theory can be found in [4, 5].

The notion of a simple-named complex-valued nominative data and basic operations on such data were formalized in Mizar in [7]. This formalization is already included in Mizar Mathematical Library under the name `NOMIN_1`. The notion of a *binominative function* which represents semantics of a program which operates on complex forms of data, and *nominative predicate* and the main *compositions* of such functions and predicates (sequential composition, branching, cycle, etc.) [14] were formalized in [8]. The next steps include formalization of the extension of the Floyd-Hoare logic proposed in [9] for reasoning about properties of programs which operate on nominative data.

References

[1] Bancerek, G., Bylinski, C., Grabowski, A., Kornilowicz, A., Matuszewski, R., Naumowicz, A., P.K., Urban, J. *Mizar: State-of-the-art and beyond*. pp. 261–279. In Intelligent Computer Mathematics International Conference, CICM 2015, Washington, DC, USA, Proceedings (2015).

[2] Floyd, R. *Assigning meanings to programs*. Mathematical aspects of computer science 19(19–32) (1967).

[3] Hoare, C. *An axiomatic basis for computer programming*. Commun. ACM 12(10), 576–580 (1969).

[4] Ivanov, I. *On Representations of Abstract Systems with Partial Inputs and Outputs*, pp. 104–123. Springer International Publishing, Cham (2014), https://doi.org/10.1007/978-3-319-06089-7$_{8.}$

[5] Ivanov, I. *On Local Characterization of Global Timed Bisimulation for Abstract Continuous-Time Systems*, pp. 216–234. Springer International Publishing, Cham (2016), https://doi.org/10.1007/978-3-319-40370-0$_1$3.

[6] Korniłowicz, A., Kryvolap, A., Nikitchenko, M., Ivanov, I. *An approach to formalization of an extension of Floyd-Hoare logic*. In: Proceedings of the 13th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer, Kyiv, Ukraine, May 15-18, 2017. pp. 504–523 (2017), http://ceur-ws.org/Vol-1844/10000504.pdf.

[7] Kornilowicz, A., Kryvolap, A., Nikitchenko, M., Ivanov, I. *Formalization of the algebra of nominative data in Mizar*. In: Ganzha, M., Maciaszek, L.A., Paprzycki, M. (eds.) Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017, Prague, Czech Republic, September 3-6, 2017. pp. 237–244 (2017), https://doi.org/10.15439/2017F301.

[8] Korniłowicz, A., Kryvolap, A., Nikitchenko, M., Ivanov, I.: Formalization of the nominative algorithmic algebra in Mizar. In: Świątek, J., Borzemski, L., Wilimowska, Z. (eds.) Information Systems Architecture and Technology: Proceedings of 38th International Conference on Information Systems Architecture and Technology – ISAT 2017: Part II.

pp. 176–186. Springer International Publishing, Cham (2018), https://doi.org/10.1007/978-3-319-67229-8$_1$6.

[9] Kryvolap, A., Nikitchenko, M., Schreiner, W. *Extending Floyd-Hoare Logic for Partial Pre- and Postconditions*, pp. 355–378. Springer International Publishing, Cham (2013), https://doi.org/10.1007/978-3-319-03998-5$_1$8.

[10] Nikitchenko, M., Kryvolap, A. *Properties of inference systems for Floyd-Hoare logic with partial predicates*. Acta Electrotechnica et Informatica 13(4), 70–78 (2013), https://doi.org/10.15546/aeei-2013-0052.

[11] Nikitchenko, N.S. *A composition nominative approach to program semantics*. Tech. Rep. IT-TR 1998-020, Department of In-formation Technology, Technical University of Denmark (1998).

[12] Nikitchenko, N. *Abstract computability of non-deterministic programs over various data structures*. In: Bjorner D., Broy M., Z.A. (ed.) Perspectives of System Informatics: 4th International Andrei Ershov Memorial Conference, PSI 2001. Lecture Notes in Computer Science, vol. 2244, pp. 468–481. Springer, Berlin, Heidelberg (2001), https://doi.org/10.1007/3-540-45575-2$_4$5.

[13] Red'ko, V. *Backgrounds of compositional programming*. Programming [in Russian] (3), 3–13 (1979).

[14] Skobelev, V.G., Nikitchenko, M., Ivanov, I. *On Algebraic Properties of Nominative Data and Functions*, pp. 117–138. Springer International Publishing, Cham (2014), https://doi.org/10.1007/978-3-319-13206-8$_6$.

Ievgen Ivanov[1], Artur Korniłowicz[2], Mykola Nikitchenko[3]

[1]Taras Shevchenko National University of Kyiv, Ukraine
Email: ivanov.eugen@gmail.com

[2] University of Białystok, Poland
Email: arturk@math.uwb.edu.pl

[3] Taras Shevchenko National University of Kyiv, Ukraine
Email: nikitchenko@unicyb.kiev.ua

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

EA-Style Mathematical Text Processing in English SAD System

Alexander Lyaletski

Abstract

The paper contains a short description of some of the features and peculiarities of the English SAD system intended for theorem proving/text verification in the Evidence Algorithm-style in an environment of self-contained mathematical texts.

Keywords: EA-style proof search, self-contained text, Evidence Algorithm, English SAD system, Russian SAD system.

1 Introduction

In the early 1960s, Academician V.M. Glushkov initiated a research on the problem of automated theorem-proving in mathematics. In 1970, his paper [1] appeared, in which it was described his vision of this problem and announced the so-called Evidence Algorithm (EA) programme, in the framework of which Glushkov suggested to focus attention firstly on the construction of formal natural-like languages for writing mathematical texts and on evolutionarily development of an evidence of a machine-made proof step.

The first steps in the direction of the realization of EA were made in 1962, but a serious progress in this direction began to be observed after the publication of the paper [1]. As a result, in 1978, the Russian-language SAD system [2] with the input natural-like formal language TL was designed and implemented. In 2002, the English-language SAD system [3] appeared with the input language ForTheL [4] being an English modification and extended version of TL. (For all details, see

[5]). That is why below we pay attention to theorem proving and proof verification doing by the English SAD system operating within a so-called self-contained ForTheL-text environment.

2 Language processing in English SAD

The TL language was proposed in 1974 as the first representative of natural-like mathematical languages satisfying the EA requirements. Later, in 2000, its English ForTheL version [4] admitting writing texts containing both theorems and theorems with their proofs appeared.

In English SAD, a ForTheL-text is an ordered set of first-order sentences satisfying the ForTheL-grammar. In this connection, for applying the deductive tools for finding/verifying theorem proofs, in English SAD an original ForTheL-text is converted into a so-called ForTheL1-text, sentences of which, being a kind of first-order formulas, preserve the signature of an original ForTheL-text, its syntax and structure (i.e. its division into sections: definitions, theorems, lemmas, etc.). Note that in English SAD, the transformation of a ForTheL-text into a ForTheL1-text is performed immediately after syntactical analysis and then the obtained ForTheL1-text undergos a deductive processing.

3 Deductive processing in English SAD

Evidence of a proof step is largely determined by the deductive tools used for proving theorems/verifying ForTheL-texts. According to EA, deductive processing should satisfy the following requirements: the syntactical form of an initial problem should be preserved, deduction should be done in the signature of an initial theory, in particular, preliminary skolemization should be non-obligatory, proof search should be goal-oriented, equality handling should be separated from a deductive processes. For this, a sequent approach was selected as basic for the construction of the native English SAD prover Moses. Note that the sequent approach exploits the original notion of an admissible substitution allowing to preserve the initial signature of a task under con-

sideration so that accumulated equations can be sent to a specialized solver or external (w.r.t. English SAD) computer algebra system. Also note that English SAD was implemented in such a way that its external first-order prover can be Vampire, SPASS, Otter, or E Prover.

4 EA-style text processing in English SAD

An EA-style deductive processing is made by the English SAD system in an information environment constructed from a self-contained structured mathematical ForTheL-text to be processed. (A self-contained ForTheL-text is defined as a text that should contain such syntactical ForTheL-units as sentences, sentences with proofs, cases, and top-level sections being sufficient for proving a theorem under consideration.)

The general scheme of theorem proving/proof verification is as follows. A proved theorem/verified proof is considered in a context of a self-contained mathematical text, that is, as part of a set of interconnected definitions, theorems, statements, etc. After this text appears in ForTheL, English SAD translates it into a ForTheL1-text preserving the structure of the (original) ForTheL-text, which permits to construct a ForTheL1 information environment for the English SAD deductive engine for finding/verifying a proof, which can use the naive prover Moses (based on a sequential formalism) or one of the above-mentioned resolution provers (being external w.r.t. English SAD).

The correctness and completeness of this scheme of an EA-style text processing in English SAD is provided by the following proposition.

Theorem. Let Txt be a self-contained noncontradictory ForTheL-text for a theorem T (for a theorem T with its proof Pr) and $\overline{Txt}$ and $\overline{T}$ are results of the translation of Txt and T respectively into ForTheL1-sentences ($\overline{Txt}$, $\overline{T}$, and $\overline{Pr}$ are results of the translation of Txt, T, and Pr respectively into ForTheL1-sentences). Then $\overline{T}$ is a logical consequence of $\overline{Txt}$ ($\overline{Pr}$ is a correct proof for $\overline{T}$ in the self-contained environment $\overline{Txt}$) if, and only if, the English SAD system can establish this fact using its deductive engine in the assumption that it has an infinite memory to store data and an infinite time to operate.

5 Conclusion

The above-said demonstrates that the EA ideas realized in the English SAD system (see the site "nevidal.org") can serve as a background for the creation of an info-structure for the presentation and complex processing of mathematical knowledge that will be useful in both teaching and academical daily activity of a man in "doing mathematics".

References

[1] V. M. Glushkov. *Some problems in the theories of automata and artificial intelligence*. Cybernetics and System Analysis, vol. 6, no. 2 (1970), pp. 17–27.

[2] Yu. V. Kapitonova, K. P. Vershinin, A. I. Degtyarev, A. P. Zhezherun, and A. V. Lyaletski. *System for processing mathematical texts*. Cybernetics and System Analysis, vol. 15, no. 2 (1079), pp. 209–210.

[3] A. Lyaletski, K. Verchinine, A. Degtyarev, and A. Paskevich. *System for automated deduction (SAD): linguistic and deductive peculiarities*. Advances in Soft Computing: Proc. of the IIS'2002 Symposium, Siedlce, Poland (2002), pp. 413-422.

[4] K. Vershinin and A. Paskevich. *ForTheL — the language of formal theories*. International Journal of Information Theories and Applications, vol. 7, no. 3 (2000), pp. 120-126.

[5] A. Lyaletski, M. Morokhovets, and A. Paskevich. *Kyiv School of Automated Theorem Proving: a Historical Chronicle*. In book: "Logic in Central and Eastern Europe: History, Science, and Discourse", University Press of America, USA, 2012, pp. 431-469.

Alexander Lyaletski

Institute of Mathematics and Computer Science, Chisinau
Email: forlav@mail.ru

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

On Correct Computations on Fuzzy Data

Alexandre Lyaletsky

Abstract

The problem of doing correct computations and actions with fuzzy data for solving fuzzy tasks in such areas as economics, trade, finances, planning, etc. is considered and solved.

Keywords: fuzzy set, fuzzy computation, fuzzy task, correctness of a solution of a fuzzy task.

The problem of finding correct actions with fuzzy data for solving fuzzy tasks is considered and solved. Its brief description is given below.

Let T be a parameterized problem with parameters $p_1, \ldots, p_n$, the solution of which is found by computing an "exact" value of a known (perhaps, algorithmically determined) function $f : A_1 \times \ldots A_n \mapsto B$ of its arguments $p_1, \ldots, p_n$, where each A_i is a set of possible values for the corresponding argument p_i and B is a set of possible solutions. That is, with selected values of parameters of T, f maps these possible values $\nu_1, \ldots, \nu_n$ from $A_1, \ldots, A_n$ to an exact calculated value of f being a solution of the problem T. Suppose that values $\nu_1, \ldots, \nu_n$ of the parameters $p_1, \ldots, p_n$ are given as "fuzzy", with some "degree of confidence". Then it is natural to assume that each of $A_1 \ldots A_n$ is a set of pairwise incompatible elementary events, which informally means that, first, at a given instant of time, exactly one event should occur and, second, that a fuzzy value of each parameter p_i is presented in the form of a corresponding fuzzy set $c_i \in [0,1]^{A_i}$ $(i = 1, \ldots, n)$.

According to such a fuzzy approach, the following questions arise: (i) What should be understood as a solution of a "fuzzy version" of the task T? (ii) What information is sufficient for finding such a solution and can it be calculated with the help of certain fuzzy operations?

(iii) Is it possible to substantiate the correctness of a found solution in the case of the existence of such operations?

In order to answer the first two questions, the notions of so-called fuzzy contexts of the first and second orders are introduced and using them, the definition of a fuzzy task is given and operations over fuzzy data are determined leading to solutions of such fuzzy tasks. At that, the positive answer to the third question is achieved.

Thus, we get that knowing how to perform some actions on an arbitrary set of "exact events", one can correctly perform similar actions on fuzzy sets representing forecasts about these events.

The results presented here will be useful in constructing expert systems using the fuzzy set paradigm as well as in solving fuzzy tasks in economics, trade, financial sector, planning, and so on.

Finally note that the papers [1, 2] reflect the first steps in this research. Additionally note that they also contain a brief description of a software system based on the ideas presented in them and this system has significant advantages over the software package from [3].

References

[1] A.A. Lyaletsky and O.M. Yaremchuk. *On a method for solving finitely-parametrzed tasks and its soft implementation.* Proc. of the X-th Int. Conf. KDS'2003, Bulgaria, 2003, pp. 86–92. (In Russian.)

[2] A.A. Lyaletsky and O.M. Yaremchuk. *On prediction-based method in fuzzy logic and its soft implementation.* Proc. of the 19th Int. Conf. on Art. Int. (AI'19), Siedlce, Poland, 2004, pp. 43–48.

[3] V.P. Bocharnikov, S.V. Sveshnikov, S.N. Voznyak. *Business forecast calculations and risk analysis on Fuzzy for Excel.* K: Ineks, 2000, 159 pp. (In Russian.)

Alexandre Lyaletsky

Taras Shevchenko National University of Kyiv
Email: foraal@mail.ru

On Recognition of Manuscripts in the Romanian Cyrillic Script

Ludmila Malahov Svetlana Cojocaru
Alexandru Colesnicov Tudor Bumbu

Abstract

The paper discusses approaches and problems at the reciognition of old Romanian handwritten texts. The Transkribus engine was used at the first attempts.

Keywords: Romanian language, handwritten text recognition, old Romanian Cyrillic alphabets, cultural heritage.

1 Introduction

Several years our research group at the Institute of Mathematics and Computer Science in Chisinau performs studies in the recognition of old Romanian texts printed in the Cyrillic script. Starting from the Moldavian Cyrillic texts of the 20th century from the Moldavian SSR, we drew our attention back in time and covered sequentially the transitional alphabets of the 19th century, and then old Romanian Cyrillic scripts up to the first printed book in Romanian in 1561 [1,2].

Researchers of the old printings in the Latin script got the opinion that ABBYY Finereader (AFR) isn't a suitable OCR engine in their case because of great variety of fonts and orthographic irregularities of old texts [2]. We found that the situation with Romanian Cyrillic printings is much simpler and that the mentioned variety does exist but don't hinder usage of AFR. We found the dependencies of fonts from locations and typographies and proposed to use AFR with specific set

of trained recognition patterns in each case. For example, the printed Romanian texts of the 17th century needs only two sets of patterns.

We collected such pattern sets that made in total approx. 5,000 patterns, other information for AFR, and several useful utilities (selector of AFR data depending of epoch and location, transliteration to the modern Latin script and reverse, and virtual keyboard) in a Tool Pack. This collection of tools substantially helps to process old Romanian books with AFR. With our tools and AFR, we get for Romanian Cyrillic printings WER (Word Error Rate) of 3%, that corresponds to CER (Character Error Rate) less than 1%.

An important part of data for AFR is presented by word lists that should correspond the epoch, script, and reflect peculiarities of the old orthography. We collected these lists from all available sources including the recognized text itself. Transliteration of modern dictionaries to the old script was also useful. Our dictionaries cover four epochs and contain more than 5,000 words in total.

Further step is to continue research over handwritten Romanian documents in the Cyrillic script. Libraries and archives propose big collections of manuscripts of great importance for research in cultural heritage. For example, Strempel's catalog of Romanian manuscripts [3] lists approx. 6,000 items. Many of them are described as being manual copies of older books and documents that would be lost otherwise.

A small part of older manuscripts is written in the uncial script that was imitated later in typographies. Nevertheless, the prevailing cursive handwriting is totally different [4] and needs a specific approach to recognition.

2 Transkribus: a platform for automated recognition

For the recognition of Romanian Cyrillic manuscripts, we tried Transkribus.[1]

[1] https://transkribus.eu/Transkribus/

Transkribus is the core part of a EU project *Recognition and Enrichment of Archival Documents* (READ).[2] It combines research, services, and network. Research are carried out, for example, in Handwritten Text Recognition (HTR) and Document Layout Analysis that are necessary for recognition and are included as services that we can use.

Transkribus is freely downloadable and provides through its interface access to the whole platform that proposes document management, user management, layout analysis, HTR, OCR, manual transcription, and correction.

HTR and OCR are different in approach to the recognition. They start analyzing document layout and then segmenting the image. OCR engines segment them up to the separate letters, while HTR ones stop segmentation on word and even line level. Then OCR engines recognize separate letters but HTR engines recognize words or lines at once.

Transkribus provides access to the OCR engine (AFR version 11), but the selection (HTR or OCR) is on the user.

Transkribus was successfully used in several big projects. One of well-known projects gaining from use of Transkribus is *Transcribe Bentham*[3], a major online initiative to transcribe the manuscripts of the English philosopher Jeremy Bentham (1748–1832). The project is managed by the Bentham Project at University College London (UCL).

Bentham's collection contains more than 43,000 handwritten pages. To transcribe them, the crowdsourcing was started. Until present more than 19,000 pages (44%) were processed.

Each volunteer registers and gets access to the TSX, the Web interface for transcription. There are at least three possibilities: transcription may be performed manually; transcription may be automated with the HTR module (Transkribus) and then corrected; transcription may be performed manually with the help (suggestions) from the HTR module at any place. The final checking is performing by full-time staff at UCL. Number of necessary correction is slightly less than in 1% of words.

[2] https://read.transkribus.eu/

[3] http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham

Output of transcription process is a free accessible XML file in the TEI (Text Encoding Initiative[4]) standard.

3 Applying Transkribus to a handwritten Romanian text in the Cyrillic script

We took [5] as an example of manuscript in the Romanian Cyrillic script. The text is a manuscript of the very first novel written in Romanian (*Polydor and Charity*, 1843), completed with a collection (*Miscellania*) of short moral stories (1848).

The novel is a compilation of some lost Greek book. The manuscript author (Dmitrie Balica) noted that his knowledge of Greek is poor and that he wrote in fact a new text.

This edition [5] contains images of the whole manuscript. The images of the novel are accompanied with their manual transcription into the modern Romanian Latin script. *Miscellania* was transcribed both into printed uncial Cyrillic and modern Latin scripts. We started our experiments with *Miscellania* because of the possibility to compare results with the existing transcription.

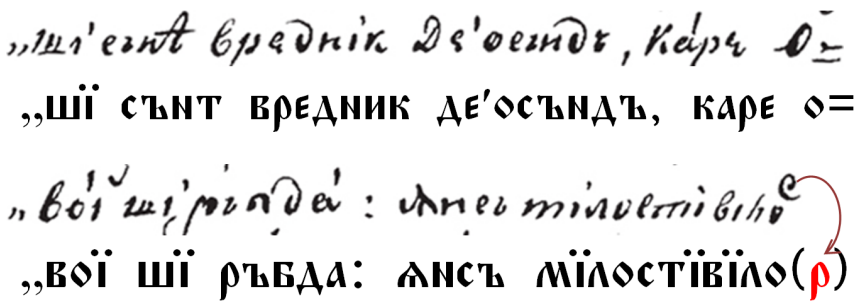


Figure 1. Data to create a model for HTR engine (fragment)

[4] http://www.tei-c.org/index.xml

We followed the usual procedure when using Transkribus. We prepared manual Cyrillic transcription and images of 50 pages (Fig. 1). The pages were used by the Transkribus staff to provide a model for HTR engine. Then the training of the HTR engine is continuing by introducing new images and editing the result. Our first experiments with HTR engine from Transkribus show that the quantity of introduced data is not enough to obtain good results so we are to introduce more pages. In addition, we should prepare and submit dictionary (word list).

To enter the Romanian Cyrillic text, we created the corresponding keyboard. Transkribus provides a very simple way to create additional keyboard layout: it should be enlisted and encoded in the file `virtualKeyboards.xml`. To encode the keyboard, it is necessary to write out Unicode points of characters, or characters themselves. We implemented Romanian Cyrillic alphabet of 44 characters[5] from *The Romanian Grammar* of 1797 [6]. Fig. 2 shows Transkribus with its virtual keyboard during work with a fragment of *Miscellania*.

4 Conclusion

We began the work over recognition of old Romanian Cyrillic handwritten texts. The Transkribus transcription platform, constructed by the University of Innsbruck, facilitates transcription and addition of metadata via tags, whilst also ensuring the transcripts export in several formats. In addition, Transkribus can be used to apply HTR technology to a set of documents.

Transkribus seems suitable for our tasks, and we would feed it with the information necessary for good recognition. This include introduction of transcribed pages (images and text), and dictionaries.

[5] Author included in the alphabet the Cyrillic letter *Yery* (Ы) that is not used in Romanian.

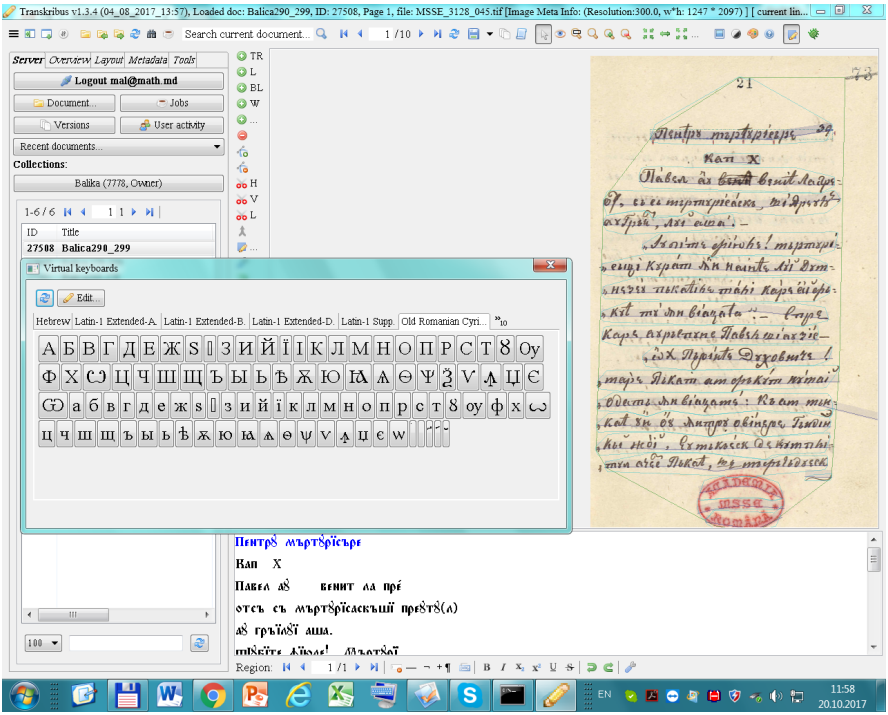


Figure 2. Working with Transkribus HTR

References

[1] S. Cojocaru, A. Colesnicov, L. Malahov, T. Bumbu. *Optical Character Recognition Applied to Romanian Printed Texts of the 18th-20th Century.* Computer Science Journal of Moldova, vol. 24, no. 1(70) (2016), pp. 106–117.

[2] S. Cojocaru, A. Colesnicov, L. Malahov. *Digitization of Old Romanian Texts Printed in the Cyrillic Script.* Second International Conference on Digital Access to Textual Cultural Heritage DATeCH-2017, Goettingen, June 1-2, 2017, pp. 143–148.

[3] G. Strempel. *Catalog of Romanian manuscripts.* Bucuresti, vol. 1

(1978), vol. 2 (1983), vol. 3 (1987), vol. 4 (1992). – In Romanian.
[4] E. Virtosu. *The Romanian Cyrillic Palaeography*. Bucuresti (1968). – In Romanian.
[5] T.-T. Marsalcovschi et al. *Manuscripts of* comisul *Dimitrie Balica*, Chisinau, 2013. – In Romanian.
[6] R. Tempea. *The Romanian Grammar*. Sibiu, 1797. – In Romanian.

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

[1]Institute of Mathematics and Computer Science
5 Academiei str., MD-2028, Chisinau
Republic of Moldova
[2]Email: lmalahov@gmail.com
[3]Email: svetlana.cojocaru@math.md
[4]Email: acolesnicov@gmx.com
[5]Email: tudorbumbu10@gmail.com

Non Standard Treebank Romania – Republic of Moldova in the Universal Dependencies

Cătălina Mărănduc, Victoria Bobicev

Abstract

Morphological and syntactic annotated corpora are very important for language computerization. A Dependency Treebank was created in 2007 at the Al. I. Cuza University. 4,500 sentences of it were introduced into the Universal Dependencies (UD), consisting of over 80 treebanks for 50 languages annotated in unique conventions. Added to other 5,000 sentences annotated at the Artificial Intelligence Institute in Bucharest, they form a Romanian Contemporary Standard Treebank, affiliated with UD. But our Treebank has other non-standard language sentences. We intend to affiliate with UD a Non-Standard Romanian Dependency Treebank, having 15,000 sentences, part of them collected and annotated in the Republic of Moldova.

Keywords: treebank corpus, dependency grammar, syntactic parser, non-standard language, language specific peculiarities.

1 Introduction

The Al. I. Cuza University Romanian Diachronic Treebank, UAIC-RoDia, (ISLRN 156-635-615-024-0) has now 16,187 sentences, with 322,404 words, punctuation included. The 4,500 sentences in Contemporary Standard Romanian, which were affiliated with UD, come from the research conducted by Augusto Perez within the doctoral thesis and are mostly translations from English [12].

Initially, the UD project aimed to create a universal syntactic parser, and for this purpose, not very complex sentences in Contemporary

Standard languages are requested. The general features were highlighted and the language specific ones were admitted only as sub-classifications.

However, as the group grew up, other uses of affiliated treebanks emerged, comparative language study, old language study, or automated translations. All types of treebanks were allowed, both consisting of a single type of text and balanced ones, like ours, with texts from social media communication to old texts.

In fact, there are few occasions when standard language is used in communication, official relationships, scientific communications, books for publication, exams. We cannot study and process the natural language only on the basis of the simplified standard examples; especially they do not give information about linguistic creativity. That's why we've annotated more and more non-standard text types: oral regional fiction, social media communication, poetry.

2 Related Work

At the moment, the international community is interested in preserving and digitizing the cultural heritage, i.e. the processing of old texts. One of the projects affiliated with UD is PROIEL (Pragmatic Resources in Old Indo-European Languages) It contains the New Testament in Latin, Ancient Greek, Ancient Slavonic, Aramaic languages. We choose to introduce in the UAIC Treebank the Alba Iulia New Testament (1648), the first printed in Romanian, with the intention to compare it with the other Old New Testaments in this project **[**8]. We have already annotated the four Gospels and we are going to annotate the Acts of the Apostles.

The text with Cyrillic letters was obtained using an Optical Character Recognizer (OCR) made at the Institute of Mathematics and Computer Science in Chisinau [4]. We provided data from our corpus for the old language lexicon required for this OCR program.

Participating at the DATeCH (Digital Access of Textual Cultural Heritage) conference in Gottingen on June 1-2, 2017 (https://www.digitisation.eu/datech-international-conference), we noticed that our OCR program for old Romanian Cyrillic letters and Part of Speech (POS) tagger for Old Romanian are compatible with similar programs presented, for example, OCR for the Old Gothic letters [6].

More UD-affiliated treebanks have another format outside the UD one, for research on the language specific peculiarities, semantics, pragmatics, text annotation, which the international format does not favors. For example, the Tectogrammatic layer of Prague Treebank [2], or the Head-Driven Phrase Structure Grammar (HPSG) format of the Bulgarian Treebank [11].

3 The Regional Folk Poetry

Oral folk creation is also part of the cultural heritage and must be preserved and protected, especially as a phenomenon of extinction as the written culture spreads in the villages.

Computer scientists who prefer simple texts to get a good accuracy of processing tools have always avoided the annotation of the lyrics. But these have also to be taken into account as a creative phenomenon of natural language. Recently, a project lead by Cristina Vertan has begun in Hamburg with the purpose to annotate poetry.

Another topic of our research is to compare the language spoken in Romania to the one spoken in the Republic of Moldova. Lexical studies were recently conducted on journals published in the two countries in the nineteenth century and at the beginning of the twentieth century. Their results show that the differences between them are minimal [7].

However, the differences appear at the topic and syntactic level, as well as in the non-standard language used by villagers who do not have access to normative Romanian texts.

A recent study on social media communication in the Republic of Moldova shows that many Russian words are used in non-standard language. This is a peculiarity of the bilingualism. The same study shows that speakers control these linguistic interferences and exclude them when a person who does not speak Russian participates in the conversation [5].

Both folklore and lyrics with rhymes are poorly processed and annotated by Natural Language Processing (NLP) specialists around the world. For all these reasons, we decided to make a comparative study of popular texts with rhymes in Romania and the Republic of Moldova. We intend to organize this annotation on the UD platform for the international visibility of this study. In Romania, we started to annotate local texts from

Muntenia and Oltenia. At the same time, an annotation of a collection of Moldovan Ballads began in the Republic of Moldova [1].

By now, we have been annotating in XML format and in UAIC treebank conventions, because we do not have any processing tools for another format. A large gold corpus in the new format is necessary in order to build the necessary tools. A tool called Treeops was created to transfer XML from one format to another desired format [3]. Using this tool, we plan to transform the texts in the UAIC format into the Universal Dependencies one.

It should be pointed out that these transformations need to be corrected by human annotators. We have two work interfaces to do it [9]. We also have converters from the XML format to the CONLLU format used by UD. This transformation does not require supervision and is performed in automate mode.

4 Brief Introduction to Universal Dependencies

Universal Dependencies (UD) is a framework for cross-linguistically consistent treebank annotation that has so far been applied to over 50 languages [11].

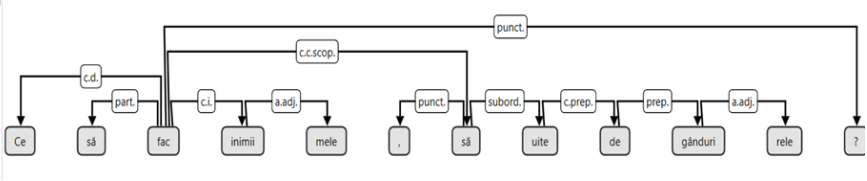


Figure 1. An example of an annotated sentence in UAIC format.

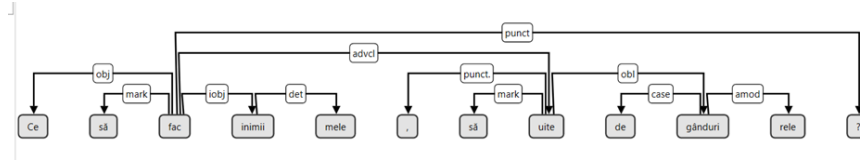


Figure 2. An example of an annotated sentence in UD format.

These are treebanks built into the dependency grammar system: the words and the punctuation are nodes, and the arcs of the graph are inscribed with the relationships between them; no equality is allowed, but only subordination; the subordination from more than one head is no

allowed [13]. This is an economical and flexible system similar to finite state automata.

The UD annotation convention highlights words with full meaning and the relational words are subordinated to them (see Figure 2). In the UAIC convention, the related words are heads (see Figure 1). In the UD system, it is easier to compare texts in very different languages and to emphasize the semantic structure. In the UAIC system, the logical structure consisting of semantic units and connectors are more visible.

5 Conclusion

The paper presents an ongoing work of syntactically annotated corpora creation. Several efforts were made to enrich the standard corpora with non-standard and more difficult annotated examples such as folklore and lyrics. Sub-corpora from different regions where Romanian language is spoken are in the process of creation and annotation. The necessary volume of annotated and manually corrected texts would serve as a training corpus for the statistical parser.

A resource is the more useful as while it is enriched and has a flexible form, easy to adapt to international formats. Our affiliation to UD will create a great visibility of our common efforts and perspectives to participate in international projects.

References

[1] V. Bobicev, T. Bumbu, V. Lazu, V. Maxim, 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 (2016), pp. 39-50.

[2] A. Bohmova, J. Hajic, E. Hajicova, B. Hladka. *The Prague Dependency Treebank: A Three-Level Annotation Scenario*. Text, Speech and Language Technology. Springer Publisher, Prague. (2003).

[3] M. Colhon, C. Mărănduc, C. Mititelu, *A Multiform Balanced Dependency Treebank for Romanian.* Proceedings of Knowledge Resources for the Socio-Economic Sciences and Humanities, (KnowRSH), pp. 9-18,

[4] S. Cojocaru, A. Colesnicov, L. Malahov. *Digitization of Old Romanian Texts Printed in the Cyrillic Script.* Proceedings of DATeCH (2017), pp. 143-148. https://www.digitisation.eu/datech-international-conference/

[5] V. Cojocaru. *Discourse markers in Romanian spoken in the Republic of Moldova: pragmatic and sociolinguistic aspects*. PhD Thesis, Faculty of Letters, University of Bucharest. (2016).

[6] F. Fink, K. U. Schulz, U. Springmann *Profiling of OCR'ed Historical Texts Revisited.* Proceedings of DATeCH (2017), pp. 61-66.

[7] D. Gîfu. *The Analysis of Diachronic Variation in Romanian Print Press.* In: Proceedings of the First PhD Symposium on Sustainable Ultrascale Computing Systems, NESSUS PhD (2016), pp. 49-53.

[8] D. T. T. Haug. *The PROIEL corpus: annotation of morphology, syntax and information structure*, Perspective Project kick-off meeting, University of Nijmegen, (2014).

[9] C. Mărănduc, F. Hociung, V. Bobicev, *Treebank Annotator for multiple formats and conventions.* Proceedings of The 4th Conference of Mathematical and Computer Science Society of the Republic of Moldova, (2017), pp. 529-534.

[10] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, D. Zeman. *Universal Dependencies v1: A Multilingual Treebank Collection*. Proceedings of LREC (2016). http://universaldependencies.org/

[11] P. Osenova, K. Simov. *Syntactic-Semantic Treebank for Domain Ontology Creation.* Cognitive Studies. SOW Publishing House, Warsaw, Poland, (2011), pp 213-225.

[12] C. A. Perez. *Linguistic Resources for Natural Language Processing*. PhD thesis. Faculty of Computer Science, Al. I. Cuza University, Iasi, (2014).

[13] P. Tapanainen, T. Jarvinen. *Towards an implementable dependency grammar*. CoLing-ACL98 workshop Processing of Dependency-based Grammars. (1998).

Cătălina Mărănduc[1,2], Victoria Bobicev[3]

[1]Faculty of Computer Science, Al. I. Cuza University, Iași
catalinamaranduc@gmail.com:

[2]Iorgu Iordan – Al. Rosetti Academic Institute of Linguistics, Bucharest

[3]Tehnical University of Moldova
victoria.bobicev@ia.utm.md:

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

Non-commutative finite associative algebras of 2-dimension vectors

Alexander Moldovyan, Nicolay Moldovyan, Victor Shcherbacov

Abstract

In this paper properties of the non-commutative finite associative algebra of two-dimension vectors are presented. An interesting features of the algebra is mutual associativity of all modifications of the defined parameterized multiplication operation and existing of a large set of the single-side unit elements. In the ordinary case one unique two-side unit element is connected with each element of the algebra, except the elements that are square roots from zero element.

Keywords: finite algebra; ring; Galois field; vector; associative multiplication; parameterized multiplication; cryptoscheme

AMS: 16U60, 11G20, 11T71

1 Introduction

Finite non-commutative associative algebras (FNAA) are interesting for applications in the desin of the public-key cryptoschemes characterized in using the hidden conjugacy search problem (called also discrete logarithm problem in hidden commutative subgroup)[2, 3, 1]. In the literature there are considered different FNAA defined over the finite vector spaces with dimensions $m = 4, 6$, and 8. The main attention was paid to the case $m = 4$ that provides lower computational difficulty of the multiplication operation in the FNAA, while defining the vector spaces over the same finite field $GF(p)$.

In present paper it is shown that the FNAA can be defined over the vector spaces of the dimensions less than 4. There are introduced

the FNAA of two-dimension vectors and investigated some properties of such FNAA.

Suppose $\mathbf{e}$ and $\mathbf{i}$ be some formal basis vectors and $a, b \in GF(p)$, where prime $p \geq 3$, be coordinates. The two-dimension vectors are denoted as $a\mathbf{e} + b\mathbf{i}$ or as (a, b). The terms $\tau\mathbf{v}$, where $\tau \in GF(p^d)$ and $\mathbf{v} \in \{\mathbf{e}, \mathbf{i}$ are called components of the vector.

The addition of two vectors (a, b) and (x, y) is defined as addition of the corresponding coordinates, i.e. with thefollowing formula $(a, b) + (x, y) = (a + x, b + y)$.

The multiplication of two vectors $a\mathbf{e}+b\mathbf{i}$ and $x\mathbf{e}+y\mathbf{i}$ is defined with the following formula

$$(a\mathbf{e} + b\mathbf{i}) \circ (x\mathbf{e} + y\mathbf{i}) = ax\mathbf{e} \circ \mathbf{e} + bx\mathbf{i} \circ \mathbf{e} + ay\mathbf{e} \circ \mathbf{i} + by\mathbf{ii},$$

where $\circ$ denotes the vector multiplication operation and each product of two basis vectors is to be replaced by some basis vector or by a one-component vector in accordance with so called basis-vector multiplication table (BVMT) which defines associative and non-commutative multiplication of the two dimension vectors. In this section the are considered two variants of the BVMT presented in Table 1 (Section 2) and Table 2 (Section 3).

2 Algebra with unique local right-side unit elements

The multiplication of two-dimension vectors defined with Table 1, where $\mu \neq 0$ and $\tau \neq 0$, is a parametrized operation different modifications of which correspond to different pairs of the values of so called structural coefficients μ and τ. As compared with the case of the commutative finite algebra of the 2-dimension vectors [4] the defined non-commutative multiplication operation is characterized in the mutual associativity of all its modifications.

Statement 1. Suppose $\circ$ and $\star$ are two arbitrary modifications of the vector multiplication operation, which correspond to different

Table 1. The basis-vector multiplication table for the case $m = 2$

$\circ$	$\overrightarrow{e}$	$\overrightarrow{i}$
$\overrightarrow{e}$	$\mu\mathbf{e}$	$\mu\mathbf{i}$
$\overrightarrow{i}$	$\tau\mathbf{e}$	$\tau\mathbf{i}$

pair of structural coefficients (μ_1, τ_1) and $(\mu_2, \tau_2) \neq (\mu_1, \tau_1)$. Then for arbitrary three vectors A, B, and C it holds the following formula $(A \circ B) \star C = A \circ (B \star C)$.

Proof of this statement consists in straightforward using the definition of the multiplication operation and Table 1.

To find the right unit element of the considered FNAA one can solve the following vector equation

$$(a\mathbf{e} + b\mathbf{i} \circ (x\mathbf{e} + y\mathbf{i}) = (a\mathbf{e} + b\mathbf{i}), \tag{1}$$

where $V = (a\mathbf{e} + b\mathbf{i})$ is an arbitrary vector and $X = (x\mathbf{e} + y\mathbf{i})$ is the unknown.

Equation (1) can be reduced to solving the following system of two linear equations in $GF(p)$:

$$\begin{cases} (a\mu + b\tau)x = a \\ (a\mu + b\tau)y = b. \end{cases} \tag{2}$$

In the case $a\mu + b\tau \neq 0$ This system has u nique solution

$$\begin{cases} x = \dfrac{a}{a\mu + b\tau} \\ y = \dfrac{b}{a\mu + b\tau}. \end{cases} \tag{3}$$

All vectors (a, b) such that $a\mu + b\tau \neq 0$ have only one right unit element. In general case the right unit elements corresponding to different vectors are different, therefore these unite elements can be called local, since they acts only in frame of some sufficiently restricted subset of

the two-dimension vectors. There exists no global right unit element, i.e. right unit acting over the whole two-dimension vector space. The following is evident:

Statement 2. Suppose $V = (a, b)$ be a vector such that $a\mu + b\tau \neq 0$. Then the vector

$$E_r = \left(\frac{a}{a\mu + b\tau}, \quad \frac{b}{a\mu + b\tau} \right) \tag{4}$$

acts as local right unit in the following subset of two-dimension vectors $V, V^2, ..., V^i, ...$, where i is an arbitrary integer.

Let us consider the sequence $V, V^2, ..., V^i$ (for $i = 1, 2, 3, ...$). If the vector V is not a zero-divisor relatively some its power (zero-divisors are considered below and it is shown that vectors satisfying condition $a\mu + b\tau \neq 0$ are not zero-divisors), then for some two integers h and $k > h$ we have $V^k = V^h$ and $V^k = V^{k-h} \circ V^h = V^h \circ V^{k-h} = V^{k-h} \circ V^h$, i.e. the mentioned sequence is periodic and for some integer ω (that can be called order of the vector V) it holds $V^\omega = E'$, where E' is bi-side local unit such that $V^i \circ E' = E' \circ V^i = V^i$ holds for all integers i.Thus, taking into account that the local right unit element corresponding to the vector V is unique one can conclude the following:

Statement 3. Suppose $V = (a, b)$ be a vector such that $a\mu + b\tau \neq 0$.Then the vector E_r described with formula (4) acts as unique bi-side local unit element E' in the subset $\{V, V^2, ..., V^i, ..., \}$ and the value E' can be computed as some power of V.

The following computational example illustrates this fact: for $p = 16832914260232697023$ and $\mu = 276474637$; $\tau = 948576254546$ we have

$$\begin{aligned} &N = (a, b) = \\ &(17235252752952, 29124252511124) \end{aligned} \tag{5}$$

computation of the value E' as $E' = N^{p-1}$ and with using formula (3) gives the same result

$$\begin{aligned} &E' = \\ &(12597150130467515608, 9876457378547066970). \end{aligned} \tag{6}$$

To find the left unit elements of the considered FNAA one can solve the following vector equation

$$(x\mathbf{e} + y\mathbf{i}) \circ (a\mathbf{e} + b\mathbf{i} = (a\mathbf{e} + b\mathbf{i}). \tag{7}$$

Equation (7) can be reduced to solving the following system of two linear equations in $GF(p)$:

$$\begin{cases} a\mu x + a\tau y = a \\ b\mu x + b\tau y = b. \end{cases} \tag{8}$$

The last system defines the following set of the left unit elements:

$$E_l = (x,\ y) = \left(x,\ \tau^{-1}(1-x)\right), \tag{9}$$

where x takes on all possible values in $GF(p)$. Each element of the last set acts on all elements of the considered FNAA as the left unit, i.e. elements of set (9) are global left unit elements. Substituting the value $x = a(a\mu + b\tau)^{-1}$ in (9) one can show that all local right unites are contained in the set of the (global) left unit elements. This is in compliance with Statement 3.

Let us consider the question of the existence of the right and left zero-divisors. The first case case is connected with solving the vector equation

$$(a\mathbf{e} + b\mathbf{i} \circ (x\mathbf{e} + y\mathbf{i}) = (0,0), \tag{10}$$

where $V = (a\mathbf{e} + b\mathbf{i})$ is an arbitrary vector different from (0,0) and $X = (x\mathbf{e} + y\mathbf{i})$ is the unknown.

Equation (10) can be reduced to solving the following system of two linear equations in $GF(p)$:

$$\begin{cases} (a\mu + b\tau)x = 0 \\ (a\mu + b\tau)y = 0. \end{cases} \tag{11}$$

In the case of the vectors V coordinates of which satisfies condition $a\mu + b\tau \neq 0$ this system has u nique solution $(x, y) = (0, 0)$ that represents

zero of the considered FNAA. Each two-dimension vector acts on the vectors V such that $a\mu + b\tau = 0$ as the right zero-divisor.

Consideration of the case of the left zero-divisors is connected with solving the vector equation

$$(x\mathbf{e} + y\mathbf{i}) \circ (a\mathbf{e} + b\mathbf{i} = (0,0), \tag{12}$$

that can be reduced to the following system of two linear equations in $GF(p)$:

$$\begin{cases} a\mu x + a\tau y = 0 \\ b\mu x + b\tau y = 0. \end{cases} \tag{13}$$

One can see that each of the vectors

$$D_l = \left(x, -\tau^{-1}\mu x\right),$$

where x takes on all values in $GF(p)$, acts on each element of the considered FNAA as the left zero-divisor.

Some zero-divisor D satisfying equation

$$D^2 = D \circ D = (0,0)$$

can be called sqware root from zero of the FNAA. Finding such elements is connected with solving the vector equation

$$(x\mathbf{e} + y\mathbf{i}) \circ (x\mathbf{e} + y\mathbf{i} = (0,0),$$

connected with the following system of two linear equations in $GF(p)$

$$\begin{cases} \mu x^2 + \tau xy = 0 \\ \mu xy + \tau y^2 = 0. \end{cases} \tag{14}$$

For the last system we have the following solutions that define the set of the sqware roots from zero element $(0,0)$:

$$D = (x,\ y) = \left(x,\ \ -\mu\tau^{-1}x\right), \tag{15}$$

where $x = 0, 1, ..., p-1$. Taking into account the condition of Statement 2 one can conclude that elements to which no right unit element corresponds are square roots from the zero vector $(0,0)$.

3 Algebra with unique local left-side unit elements

The FNAA of two-dimension vectors with the multiplication operation defined with Table 2 , where $\mu \neq 0$ and $\tau \neq 0$, has properties analogous to the properties of the FNAA described in Subsection 2.1, for example Statement 1 is valid.

$\circ$	$\overrightarrow{e}$	$\overrightarrow{i}$
$\overrightarrow{e}$	$\mu\mathbf{e}$	$\tau\mathbf{e}$
$\overrightarrow{i}$	$\mu\mathbf{i}$	$\tau\mathbf{i}$

Consideration of the vector equations defining the right and left unit elements, the right and left zero divisors, and square roots from zero $(0,0)$ have given the following statements.

Statement 4. Each two-dimension vector from the set

$$E_r = (x,\ y) = \left(x,\ \tau^{-1}(1-x)\right), \tag{16}$$

where x takes on all possible values in $GF(p)$, represents a global right-side unit element.

Statement 5. Suppose $V = (a, b)$ be a vector such that $a\mu + b\tau \neq 0$. Then the vector

$$E_l = \left(\frac{a}{a\mu + b\tau},\ \ \frac{b}{a\mu + b\tau}\right) \tag{17}$$

is unique local left-side unit for all vectors from the following set $\left\{V, V^2, ..., V^i, ...\right\}$, where i is an arbitrary integer.

Statement 6. A unique local bi-side unit element $E' = E_l$ acts in the set $\left\{V, V^2, ..., V^i, ...\right\}$, where i is an arbitrary integer and vector $V = (a, b)$ is such that $a\mu + b\tau \neq 0$. The value E' can be computed as $E' = V^\omega$ for some integer ω.

Statement 7. Each two-dimension vector acts on the vectors $V = (a, b)$ such that $a\mu + b\tau = 0$ as the left zero-divisor.

Table 2. Alternative BVMT for the case $m = 2$

$\circ$	$\overrightarrow{e}$	$\overrightarrow{i}$
$\overrightarrow{e}$	$\mu\mathbf{e}$	$\mu\mathbf{i}$
$\overrightarrow{i}$	$\tau\mathbf{e}$	$\tau\mathbf{i}$

Statement 8. Each of the vectors

$$D_r = \left(x, -\tau^{-1}\mu x\right),$$

where x takes on all values in $GF(p)$, acts on each element of the considered FNAA as the right-side zero-divisor.

Statement 9. The set of the square roots from zero element $(0, 0)$ is described by formula (15).

4 Conclusion

It has been introduced associative FNA of the two-dimension vectors defined over the field $GF(p)$. One of the interesting properties of the investigated FNA is mutual associativity of all modifications of the parameterized non-commutative multiplication operation. The known in the literature parameterized commutative multiplication operation for the case $m = 2$ [4] do not possess such property.

Future research in frame of the concerned topic is connected with investigation properties of the associative FNAs of m-dimension vectors for cases $m = 3$ and $m = 5$.

References

[1] E. Sakalauskas and P. Tvarijonas and A. Raulynaitis. Key Agreement Protocol (KAP) Using Conjugacy and Discrete Logarithm Problems in Group Representation Level, *Informatica*, **18**, 1, (2007), p. 115–124.

[2] D.N. Moldovyan. Non-Commutative Finite Groups as Primitive of Public-Key Cryptoschemes, *Quasigroups and Related Systems*, **18**, 2, (2010), p. 165–176.

[3] D.N. Moldovyan, N.A. Moldovyan. Cryptoschemes over hidden conjugacy search problem and attacks using homomorphisms, *Quasigroups Related Systems*, **18**, 2, (2010), p. 177–186.

[4] Moldovyan N.A., Moldovyanu P.A. Vector Form of the Finite Fields $GF(p^m)$. Bul. Acad. Ştiinţe Repub. Mold. Mat. 2009. No 3 (61). P. 1-7.

Alexander Moldovyan[1], Nicolai Moldovyan[2], Victor Shcherbacov[3]

[1]Professor/St. Petersburg ITMO University Email: maa1305@yandex.ru

[2]Head of the laboratory/St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences Email: nmold@mail.ru

[2]Principal Researcher/Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova Email: victor.scerbacov@math.md

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

Oracle as Modality: Two Examples

Alexei Muravitsky

Abstract

We propose to use modality in conditional contexts where some of the elements of conclusion must satisfy certain constraints in order for the conclusion to be fulfilled. Thus, modality acts as an oracle in relative algorithms, if we understand deducibility in a base logic as recursiveness and deducibility in a modal conservative consistent extension of the base logic as relative recursiveness.

Keywords: relative recursiveness, intuitionistic propositional logic, classical tautology, modality, Heyting algebra.

1 Introduction

The concept of *oracle* was introduced into logical-computational context by Alan Turing [4, 5] to extend the notion of computability to that of computability by o-machine (or *oracle machine*). This leads to the notions of *relative algorithm* and *relative recursiveness*; see, e.g., [3], § 9.2. Applying his approach to systems of formal arithmetic, Turing [4, 5] defined "'non-constructive' systems of logic with which not all the steps in a proof are mechanical, some being intuitive," which are conducted by the oracle.

The implicit presence of oracles can however be spotted in areas far beyond foundational issues. Take, for instance, *inductive definitions*; see [2], § 53. In defining a concept, an inductive definition starts with a *basic clause* which demonstrates *direct determination* of the primitive elements of the concept, and then goes through an *inductive clause*

which demonstrates *conditional determination* of the compound elements of the concept. (The *extremal clause* guarantees that there are no classes besides the classes of primitive and compound elements.) We can require that both determinations be effectively decidable, or recursive. Then, if we define the notion of a formula of first order logic, we are forced to limit ourselves with a countable formal language. To overcome this inconvenience, we can resolve the first definition to be recursive and second relative recursive. This ensures that the concept of a formula is a solvable concept, although the formulas need not necessarily be pairwise effectively distinguishable.

The example of inductive definition shows that an oracle can be employed when we have a conditional context and are not concerned, whether the premise can be evaluated constructively. In other words, the oracle can be asked whether we are allowed to continue to evaluate the conclusion when the condition is met, which may require that the task be more complex than simply verifying its truth. For instance, assume that we have a basis logic L and want to investigate some property $\mathcal{P}$ which designates a class of formulas $\mathcal{F}$. Also, assume the class $\mathcal{F}$ as a whole cannot be supported by L, that is $L \nvdash A$, for some $A \in \mathcal{F}$. Further, suppose that $\mathcal{P}(A)$ is fulfilled if some parts of A, $E(A)$, satisfy some constrains which cannot be expressed in L. We ask the oracle whether the elements of $E(A)$ satisfy the constrains. If the answer is 'yes', we turn to the evaluation of $\mathcal{P}(A)$. If the answer is 'no', we can stop and leave the task as inconclusive. Let us informally interpret $\Box E(A)$ (more exactly $\Box E(A) := \wedge\{\Box x : x \in E(A)\}$) that an oracle ensures that $E(A)$ satisfies the desired conditions. One can imagine that in the language expanding the original language by modality $\Box$ there is a system M such that $M \vdash \Box E(A) \to A$, for any $A \in \mathcal{F}$. Thus M does what L cannot do. However, it should be no more than that. Taking this into account, we require that M be a conservative extension of L. Although semantics of L may be well known, semantics of M will not be discussed here at all. It suffices to have as M a consistent conservative extension of L.

It should be noted that in this approach the use of modality as

an oracle is assumed only for conditional sentences. Also, $\Box$ cannot be iterated. This approach will be illustrated by the following two examples.

2 Motivation for the first example

It is well known that any two classical tautologies are equivalent in **Cl**. Will it stay true for a weaker logic than **Cl**? The following example shows that the answer is negative. Indeed, let **S** be the logic of a 3-element Heyting algebra $\mathbf{G}_3$ whose elements are arranged as follows: $\mathbf{0} < \tau < \mathbf{1}$. We claim that $\mathbf{S} \nvdash (p \vee \neg p) \leftrightarrow (\neg p \vee \neg\neg p)$. Defining the assignment $v : p \mapsto \tau$, we obtain that $v(p \vee \neg p) = \tau$ and $v(\neg p \vee \neg\neg p) = \mathbf{1}$. Since the set $\{\mathbf{0}, \mathbf{1}\}$ is closed under the Heyting operations, the algebra $\mathbf{G}_3$ validates $\neg p \vee \neg\neg p$ and hence this algebra refutes the equivalence $(p \vee \neg p) \leftrightarrow (\neg p \vee \neg\neg p)$. Since any intermediate logic weaker than **Cl** is included in **S** (see [1], theorem 1), only **Cl** can support the equivalence $(p \vee \neg p) \leftrightarrow (\neg p \vee \neg\neg p)$. The question arises: Can a modal extension of **Int** support all equivalences of classical tautologies? Thus our goal is to formulate a logic L in modal language with modality $\Box$ such that $\mathbf{Int} \subseteq L$ and for any assertoric formulas A and B,

$$\mathbf{Cl} \vdash A \wedge B \Longrightarrow L \vdash (\Box p \wedge \ldots \wedge \Box q) \rightarrow (A \leftrightarrow B), \tag{1}$$

where $\{p, \ldots, q\} = \boldsymbol{Var}(A) \cup \boldsymbol{Var}(B)$. Also, we must ensure that not for all classical tautologies A and B, $L \vdash A \leftrightarrow B$. The intended interpretation of the last formula of (1) is the following. If the range of the truth values of each variable in the given formulas A and B is the set that consists of *true* and *false*, then A and B are equivalent. Thus modality $\Box$ acts as an oracle. If the oracle answers 'no' or remains silent, logic L perhaps is not powerful enough to assert the equivalence of A and B.

We give an example of a modal consistent extension $\mathbf{M}_0$ of **Int** such that (1) with $L = \mathbf{M}_0$ holds. Also, we prove that $\mathbf{M}_0$ is a conservative over **Int** and, hence, is consistent.

3 Motivation for the second example

It is well known that **Int** derives only classical tautologies, but there is no means to distinguish by **Int** alone those classical tautologies which are derivable in **Int** from those which are not. Suppose the intended interpretation of $\Box A$ is that the oracle confirms that A is proved by finitary means and that of $\neg\Box A$ is that the oracle fails to confirm that. We seek to find a modal system L such that if A is a classical tautology, then $L \vdash \Box A \to A$ implies $\mathbf{Int} \vdash A$ and for any formula A, if $L \vdash \neg\Box A$, then $\mathbf{Int} \nvdash A$.

We define a conservative modal extension $\mathbf{M}_1$ of **Int** such that

(a) $\mathbf{Cl} \vdash A \Longrightarrow (\mathbf{M}_1 \vdash \Box A \to A \Leftrightarrow \mathbf{Int} \vdash A)$;
(b) $\mathbf{M}_1 \vdash \neg\Box A \Longrightarrow \mathbf{Int} \nvdash A$.

We note that the conservativity of $\mathbf{M}_1$ implies its consistency.

References

[1] V. A. Jankov. *On certain superconstructive propositional calculi.* Dokl. Akad. Nauk SSSR, vol. 151 (1963), pp. 796–798.

[2] S. C. Kleene. *Introduction to metamathematics.* D. Van Nostrand Co., Inc., New York, N. Y., 1952.

[3] H. Rogers, Jr. *Theory of recursive functions and effective computability.* MIT Press, Cambridge, MA, second edition, 1987.

[4] A. M. Turing. *Systems of logic based on ordinals.* ProQuest LLC, Ann Arbor, MI, 1938. Thesis (Ph.D.)–Princeton University.

[5] A. M. Turing. *Systems of logic based on ordinals.* Proc. London Math. Soc. (2), vol. 45, no 1 (1939), pp. 161–228.

Alexei Muravitsky

Louisiana Scholars' College, Northwestern State University
Email: alexeim@nsula.edu

Graphical representation of statistical data as an alternative method of ecodopplerographic score

Ana Nastasiu

Abstract

The author hypothesized that a graphical method of displaying multivariate data in the form of a five-dimensional chart represented on axes starting from the same point - a spider chart can be used as an alternative method of the ecodopplerographic score.

Keywords: medical scoring, medical statistics, spider chart, portal hypertension, Doppler ultrasound imaging.

1 Introduction

In the last decades, both theoreticians and practitioners have widely recognized that the models, procedures and rational methods of applied statistics are effective means in solving various technical, social and economic problems. Such a statistical method can also be applied in case of a problem of an efficient graphical representation of statistical data as an alternative ecodoplerographic score for portal hypertension (PHT) assessment [1, 2].

2 Representation of portal hemodynamics disorders severity using spider chart

Numerous data representation models have been used over the time to display the information. Depending on the field of application, graphical methods are often more suitable than numerical methods for visual identification of the experimental data trend.

This paper describes the experience of the use of the spider (radar) charts in order to assess the degree of liver disorders of a cohort of patients diagnosed with liver cirrhosis as an alternative ecodoplerographic score.

Starting from the fact that all cases/precedents are described by 5 parameters (congestion index, splenoportal index, vascular portal index, PHT index, spleen area), we can conclude that each case can be described by a 5-node polygon , using the "radar" representation. And the area of the obtained polygon can be used as a comparative value between two or more precedents.

Subsequently, the partial and general areas of all 97 polygons (using the tools available in Microsoft Excel) were obtained and calculated, representing 97 cases analyzed in the ecodopplerographic score development procces.

Of course, the value of the areas also depends on how the parameters are arranged on the chart, but for the moment being a random arrangement, identical for all cases, was chosen to see if there really is a perspective in this approach.

Analyzing the obtained multicriterial charts, it was found that there is no clear delimitation between the groups of cases: expressed, moderate and mild. In order to make the result more efficient, we have proceed to the segmentation of the area value domain, so that we can develop a new score.

3 Results and Conclusion

The ecodopplerograph score, developed in previous researches, where each parameter had its weight on the final result, showed a validation rate of 94 cases out of 97. In our case, using the graphical method of spider chart, the weight of the parameters is merged into one - the total area of the polygon, that shows a 87 validated cases from 97.

Although the "radar" representation of patient cohort data seems to be a good alternative for the ecodropplerographic score, however, following the validation of the final results, it was found that it does not provide an accuracy as good as the score.

The gap of 87 validated cases compared to 94 in the case of the use of the ecodoprographic score is only a preliminary stage of the respective research. The result can be affected for the following reasons:

- The small number of cases in the cohort of patients;
- In the graphical representation, the poligon area can be affected by the parameter order on the spider chart.

This research will continue by manipulating the parameters on the spider chart, and if a better result is obtained, a larger cohort of patients will be required, which will increase the accuracy of the developed score.

Acknowledgments. The Technology Transfer Project 17.80015.5007.213T has supported the research for this paper.

References

[1] Tambala, C.; Secrieru, Iu. *Portal hemodynamics disorders severity in liver cirrhosis assessment by duplex ultrasound*. In: Scientific medical journal "Curierul medical", Vol.59, No. 1, 2016, Chisinau, Moldova, pp. 37-40, ISSN 1857-0666

[2] Secrieru, IU.; Tambala, C.; Macari, D.; Nastasiu, A. *Improvement of scoring interpretation in liver cirrhosis assessment by duplex ultrasound*. In: Book of abstracts of the 3rd International Conference Health Technology Management (ICHTM-2016). Chisinau, Moldova, October 6-7, 2016, p. 67., ISBN 978-9975-51-774-4

Ana Nastasiu

Affiliation/Institution: Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova
E-mail: ana.nastasiu19@gmail.com

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9–11, 2017, Chisinau, Republic of Moldova

Towards Representation of Classical Logic as Logic of Partial Quasiary Predicates

Mykola Nikitchenko, Stepan Shkilniak

Abstract

Quasiary predicates are partial predicates over partial states (partial assignments) of variables. Such predicates do not have a fixed arity. They are used to represent program conditions and requirements, therefore logics of quasiary predicates are program-oriented logics. Classical predicate logic is based on total n-ary predicates. We demonstrate how classical logic can be represented as logic of quasiary predicates. Properties of such representation are investigated. Obtained results can be used for program verification.

Keywords: classical predicate logic, first-order logic, partial predicate, quasiary predicate, quasiary logic.

1 Introduction

Classical logic is widely used in computer science. Still, some of its features restrict its usage. In particular, classical logic is based upon total n-ary predicates which represent fixed and static properties of subject domains. Therefore this logic is not convenient for representing such dynamical properties as partiality and varying arity of predicates. In our previous works [1–3] we proposed to construct program-oriented logics based on quasiary predicates. Such predicates are partial predicates which do not have fixed arity. Conventional n-ary predicates can be considered as a special case of quasiary predicates. Therefore

logic of quasiary predicates (quasiary logic) can be considered as a natural generalization of classical predicate logic. Still, the relationship between two logics is not trivial and calls for further investigation.

In this paper we demonstrate how classical logic can be represented as logic of quasiary predicates. Here we simplify formal definitions of logic, in particular, we do not consider an inference relation, also we consider only pure logics (without function symbols); therefore we treat a *logic* L as a triple $(Fr, INT, \models)$ where Fr is a set of *formulas* (language of a logic), INT is a class of *logic interpretations* (L-interpretations), $\models: INT \xrightarrow{t} \mathcal{B}(\mathcal{B}(Fr) \times \mathcal{B}(Fr))$ is a *consequence relation* for sets of formulas ($\mathcal{B}(S)$ is the set of all subsets of S). In the sequel, $\models (J)$ is denoted as ${}_J\models$ for any $J \in INT$; for any formula Φ from Fr its interpretation in J is denoted by Φ_J. A set of formulas Δ is a *logical consequence* of a set of formulas Γ (denoted $\Gamma_L \models \Delta$) if $\Gamma_J \models \Delta$ for all $J \in INT$.

Let $L = (Fr, INT, \models)$ and $L' = (Fr', INT', \models')$ be two logics. We say that L is a *sublogic* of L' if $Fr \subseteq Fr'$, $INT \subseteq INT'$, and for any $J \in INT$ we have that $({}_J\models) \supseteq ({}_J\models')$. We say that L is a *strict logical consequence sublogic* of L' if $({}_L\models) \supset ({}_{L'}\models')$. We say that sublogic L is *consequence isomorphic* with L' if there are bijections $b_{Fr} : Fr \xrightarrow{t} Fr'$ and $b_{INT} : INT \xrightarrow{t} INT'$ such that for any sets of formulas $\Gamma, \Delta \in \mathcal{B}(Fr)$ and L-interpretation J we have that $(\Gamma_J \models \Delta)$ iff $b_{Fr}(\Gamma)\,{}_{b_{INT}(J)}\models' b_{Fr}(\Delta)$ where $b_{Fr}(\Gamma) = \{b_{Fr}(\Phi) | \Phi \in \Gamma\}$ and $b_{Fr}(\Delta) = \{b_{Fr}(\Psi) | \Psi \in \Delta\}$.

Notations not defined here are treated in the sense of [3].

2 Basic logic of partial quasiary predicates

To define a basic logic of partial quasiary predicates L^{QB} we should define its language (based on logic signature), its class of interpretations, and its consequence relation.

2.1 Quasiary logic signature and formulas

Let V be a nonempty set of *variables*. Basic *logical operations* (*compositions*) [3] for quasiary predicates are disjunction $\vee$, negation $\neg$, renomination $R^{v_1,...,v_n}_{x_1,...,x_n}$, and existential quantification $\exists x$ (x, x_1,..., x_n, v_1,...,$v_n \in V$).

Let $CBs(V) = \{\vee, \neg, R^{v_1,...,v_n}_{x_1,...,x_n}, \exists x\}$ be a set of *basic composition symbols*, Ps be a set of *partial quasiary predicate symbols*. A tuple $\Sigma^{QB} = (V, CBs(V), Ps)$ is called a *signature* of L^{QB}.

Given Σ^{QB}, we define inductively the language of L^{QB} – *the set of formulas* $Fr(\Sigma^{QB})$. Formulas of the form P are *atomic* ($P \in Ps$); *composite formulas* are of the form $\Phi \vee \Psi$, $\neg\Phi$, $R^{v_1,...,v_n}_{x_1,...,x_n}\Phi$, and $\exists x\Phi$ where Φ and Ψ are formulas. Formulas of the form $R^{v_1,...,v_n}_{x_1,...,x_n}P$ ($P \in Ps$) are called *primitive*.

2.2 Quasiary logic interpretations

Let A be a nonempty set called a set of *values*. Given V and A, the class ${}^VA$ of *partial assignments* (*partial data*, *nominative sets*) is defined as the class of all partial mappings from V to A, thus, ${}^VA = V \xrightarrow{p} A$.

Let $Pr^V_A = {}^VA \xrightarrow{p} Bool$ be the set of all *partial quasiary predicates* over ${}^VA$.

Formal definitions of basic quasiary compositions from the set $CB(V) = \{\vee, \neg, R^{v_1,...,v_n}_{x_1,...,x_n}, \exists x\}$ on Pr^V_A can be found in [3] (we use the same notation for compositions and their symbols). Please note that the compositions are similar to *strong Kleene's operations*.

For $p \in Pr^V_A$ the *truth* and *falsity domains* of p are denoted $T(p)$ and $F(p)$ respectively.

A pair $AQB(V, A) =< Pr^V_A; CB(V) >$ is called *a first-order pure basic algebra of partial quasiary predicates*. Such algebras (for various A) form a semantic base for L^{QB}.

Composition symbols have fixed interpretation. We also need interpretation $I^{Ps}_Q : Ps \xrightarrow{t} Pr^V_A$ of predicate symbols. An L^{QB}-*interpretation* J^Q is a tuple $(\Sigma^{QB}, AQB(V, A), I^{Ps}_Q)$. A class of such interpretations for various A and I^{Ps}_Q is denoted INT^{QB}.

2.3 Quasiary logic consequence relation

Usually, for logics of quasiary predicates an *irrefutability consequence relation* is defined [1–3]: a set of formulas Δ is a consequence of a set of formulas Γ in an interpretation J^Q (denoted $\Gamma_{J^Q} \models_{IR} \Delta$), if

$$\bigcap_{\Phi\in\Gamma} T(\Phi_{J^Q}) \cap \bigcap_{\Psi\in\Delta} F(\Psi_{J^Q}) = \emptyset.$$

Please note that we can characterize L^{QB} is a logic of *double partiality*: its predicates are *partial* predicates over *partial* assignments.

3 Classical predicate logic

For classical logic L^{Cl} [4] we use notations of the previous section.

3.1 Classical logic signature and formulas

A *signature* of L^{Cl} is $\Sigma^{Cl} = (V, CCs(V), Ps, ar_n)$ where $CCs(V) = \{\vee, \neg, \exists x\}$ and $ar_n : Ps \xrightarrow{t} Nat$ is an arity mapping.

A *set of formulas* $Fr(\Sigma^{Cl})$ is defined in a usual way [4].

3.2 Classical logic interpretations

A set of *total assignments* (*total data*) is $A^V = V \xrightarrow{t} A$. A set of total predicates over total data is $PrTT_A^V = A^V \xrightarrow{t} Bool$.

A pair $AC(V, A) =< PrTT_A^V; CC(V) >$ is called *a first-order pure classical algebra of n-ary predicates* where $CC(V) = \{\vee, \neg, \exists x\}$. Such algebras (for various A) form a semantic base for L^{Cl}.

We also need interpretation $I_{Cl}^{Ps} : Ps \xrightarrow{t} \cup_{i\in Nat}(A^i \xrightarrow{t} Bool)$; this mapping should agree with the arity mapping is such a way that $I_{Cl}^{Ps}(P) \in (A^n \xrightarrow{t} Bool)$ if $ar(P) = n$ $(P \in Ps)$.

An L^{Cl}*-interpretation* J^C is a tuple $(\Sigma^{Cl}, AC(V, A), I_{Cl}^{Ps})$. A class of such interpretations for various A and I_{Cl}^{Ps} is denoted INT^{Cl}.

3.3 Classical logic consequence relation

Usually, a *classical (related to always-true predicate) consequence relation* is defined: a set of formulas Δ is classical consequence of a set of formulas Γ in an interpretation J^C (denoted $\Gamma_{J^C} \models_{Cl} \Delta$), if

$$\bigcup_{\Phi\in\Gamma} F(\Phi_{J^C}) \cup \bigcup_{\Psi\in\Delta} T(\Psi_{J^C}) = V^A.$$

Please note that L^{Cl} can be characterized as a logic of *double totallity*: its predicates are *total* predicates over *total* assignments.

4 Defining classical logic as quasiary logic

It is not possible to represent directly classical logic as quasiary logic because of differences in their languages, interpretations, and consequence relations. Therefore we first construct a special quasiary sublogic L^{QF}, called *finitary logic*, and then prove its consequence isomorphism with classical logic.

A *signature* of L^{QF} is $\Sigma^{QF} = (V, CBs(V), Ps, ar_f)$ where $ar_f : Ps \xrightarrow{t} V^*$ is a finite arity mapping such that for any $P \in Ps$ if $ar_f(P) = (v_1, ..., v_n)$ then $v_1, ..., v_n$ are different variables.

Given Σ^{QF}, we define inductively the language of L^{QF} – *the set of formulas* $Fr(\Sigma^{QF})$. Primitive formulas of the form $R^{v_1,...,v_n}_{x_1,...,x_n} P$ are *atomic* ($ar_f(P) = (v_1, ..., v_n)$, $P \in Ps$); *composite formulas* are of the form $\Phi \vee \Psi$, $\neg\Phi$, and $\exists x\Phi$ where Φ and Ψ are formulas. Note that renomination is used only in atomic formulas; such formulas in classical logic are presented as atomic formulas of the form $P(x_1, ..., x_n)$.

For L^{QF} we define a *special interpretation* of predicate symbols $I^{Ps}_{QF} : Ps \xrightarrow{t} Pr^V_A$ in such a way that predicate $I^{Ps}_{QF}(P)$ is defined on all data where $v_1, ..., v_n$ are assigned while other variables from $V \setminus \{v_1, ..., v_n\}$ are unessential for it ($P \in Ps$, $ar_f(P) = (v_1, ..., v_n)$).

An L^{QF}*-interpretation* J^{QF} is a tuple $(\Sigma^{QF}, AQB(V, A), I^{Ps}_{QF})$. A class of such interpretations for various A and I^{Ps}_{QF} is denoted INT^{QF}.

The main results of the paper are the following.

Theorem 1. *Logic L^{QF} is a strict logical consequence sublogic of L^{QB}.*

Theorem 2. *Logic L^{QF} is consequence isomorphic with L^{Cl}.*

5 Conclusion

In this paper we have investigated how to represent classical predicate logic as a logic of partial quasiary predicates. We have formulated two theorems saying that classical logic is strictly richer with respect to logical consequence relation than logic of partial quasiary predicates but there is a sublogic of quasiary logic that is consequence isomorphic with classical logic. In other words, some laws of classical logic fail in quasiary logic but within the latter we can define a sublogic which has the same properties as classical logic.

References

[1] M. Nikitchenko, S. Shkilniak. *Applied Logic*, Publishing house of Taras Shevchenko National University of Kyiv, Kyiv, 2013, 278 p. (in Ukrainian).

[2] M. Nikitchenko, V. Tymofieiev. *Satisfiability in composition-nominative logics*, Central European Journal of Computer Science, vol. 2, no. 3 (2012), pp. 194–213.

[3] M. Nikitchenko, S. Shkilniak. *Algebras and Logics of Partial Quasiary Predicates*, Algebra and Discrete Mathematics, vol. 23, no. 2 (2017), pp. 263–278.

[4] E. Mendelson. *Introduction to Mathematical Logic (4th ed.)*, London: Chapman & Hall, 1997, 440 p.

Mykola Nikitchenko[1], Stepan Shkilniak[2]

[1]Taras Shevchenko National University of Kyiv, Ukraine
E-mail: nikitchenko@unicyb.kiev.ua

[2]Taras Shevchenko National University of Kyiv, Ukraine
E-mail: sssh@unicyb.kiev.ua

Software Application for Interconnecting PACS systems

Lucian Nita, Sinica Alboaie, Paul Herghelegiu

Abstract

The paper presents a software application that interconnects PACS [1] systems from various medical institutions making it possible to transfer medical imaging files between these institutions. By using this system, a doctor from an institution can access all medical data available in the other institutions for a given patient. With more data available, the doctor can give a better diagnosis and suggest appropriate treatment.

Keywords: PACS interconnection, software application

1 Introduction

During his life, a patient acquires many medical data, which are stored in different places, databases and different formats (Fig.1). The problem is that these data are not available in critical cases when the doctor needs them to study the patient's history and give a correct diagnosis. All the doctor can access is the data of that patient stored in that institution and possibly the information given by the patient from his memory. However, this information can be poor in content and subjective, especially in emergencies when the ambulance brings the patient from another institution or locality.

The proposed software application interconnects the medical databases, making possible to transfer automatically data for a given patient when the doctor needs that information. In this way, the patient is not ask to store any more data, no data is lost and the doctor obtains a better context for giving his diagnosis.

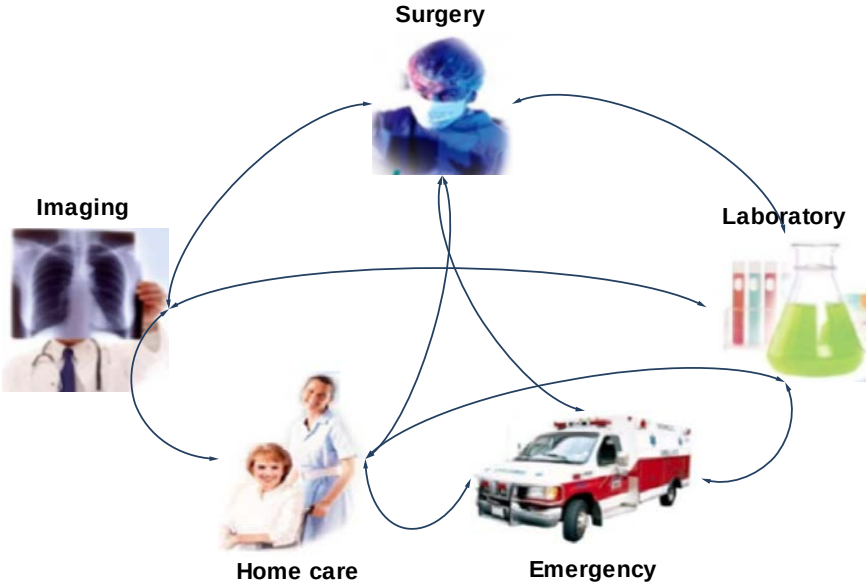


Fig. 1. Actors who need or produce medical data

2 System Architecture

Each institution included into this unique system for medical images (USMED) has a local Picture Archiving and Communications System (PACS) which collects all medical imaging data produced in that institution. If the institution has many laboratories or sections, which produce or consume data images, then the laboratory receives a PACS and operates as an independent institution into the system.

In each local node, the system installs an USMED agent, which verifies periodically the PACS state. Every update produced into local PACS is transmitted by the agent to the USMED server (Fig.2). The server maintains a unique database which stores all information regarding the system state:

- The institutions included into USMED
- Accounts for patients and doctors
- The place for each patient data

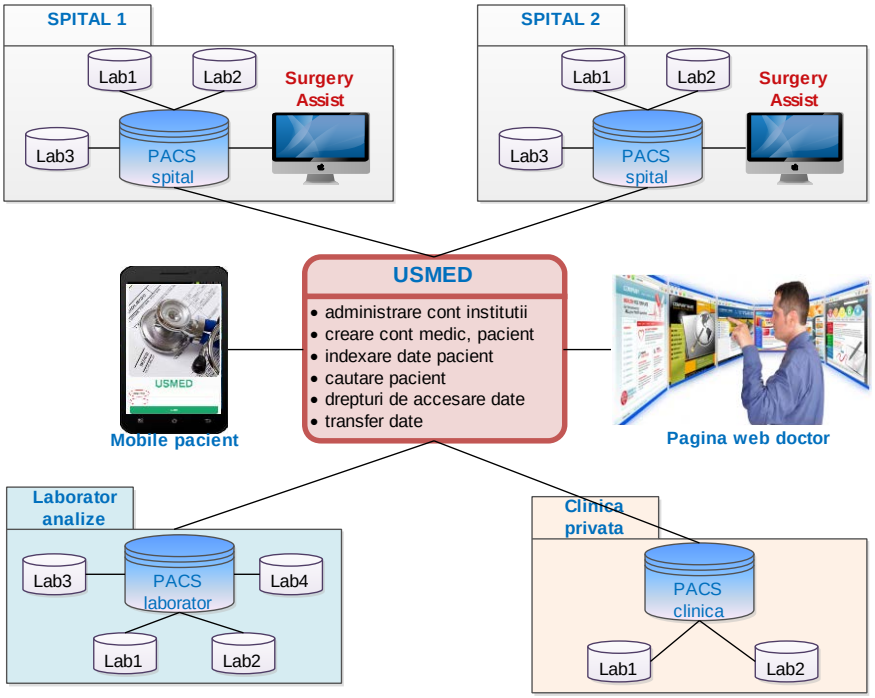


Fig. 2. The system architecture

The USMED server does not store patient data, but only the place where this data can be found. When a doctor requires data for a given patient, then the server sends this demand to all nodes that have information for that patient and then, the data is sent directly to the PACS where the doctor raised the request.

3 The user Interface

3.1 Creating accounts for hospitals and doctors

The application is developed as a web page accessible anytime and anywhere. The system administrator has a specific web page where can add new institutions to USMED (Fig.3). Then, the administrator creates

accounts for the doctors working on that hospital and attach them to the institution account.

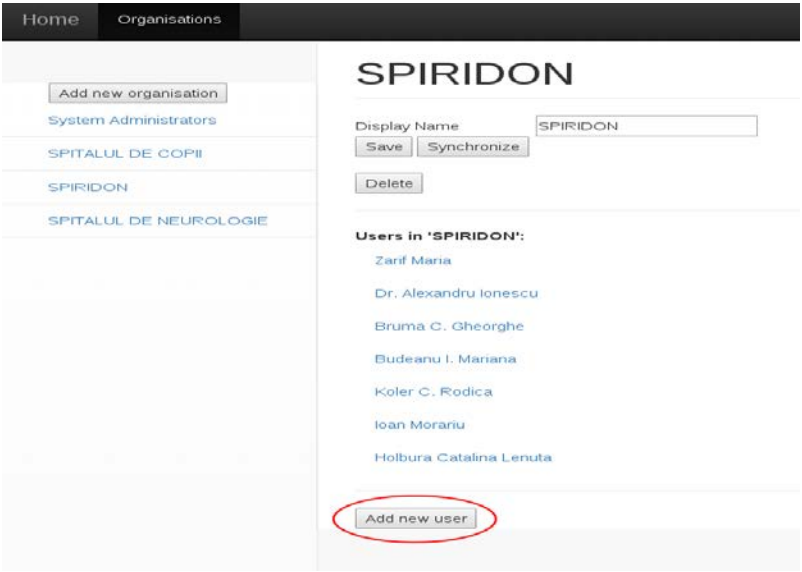
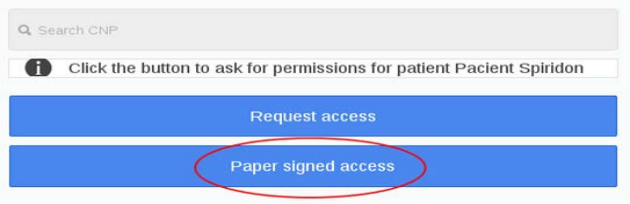


Fig. 3. Adding new institutions and doctors to USMED

3.2 Data transferring workflow

In order to ask data for a given patient, the doctor should follow the steps listed bellow:

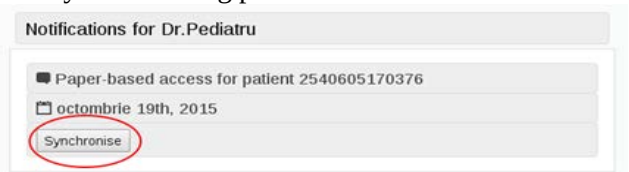
- Login into USMED system using an active account and password.
- Search the medical data for a given patient identified by the Personal Numeric Code (CNP), which is unique in the country.
- The system displays if that patient is found or not into database.
- The doctor asks the permission for data accessing:



- Sends the request to Patient Application.
- Makes use of the written consent given by the patient during the hospital admission.

- The patient grants the rights to access the data:

- The doctor receives the grant notification and raises the synchronization request between the local PACS and the other PACS systems having patient data:



3.3 Data displaying

Once the medical imaging files are downloaded to the local PACS, the doctor can open and visualize the data. The USMED system includes a DICOM [2] viewer named Surgery Assist, which displays medical images and also implements some voice commands with which the doctor can manipulate the image during surgery (Fig.4).

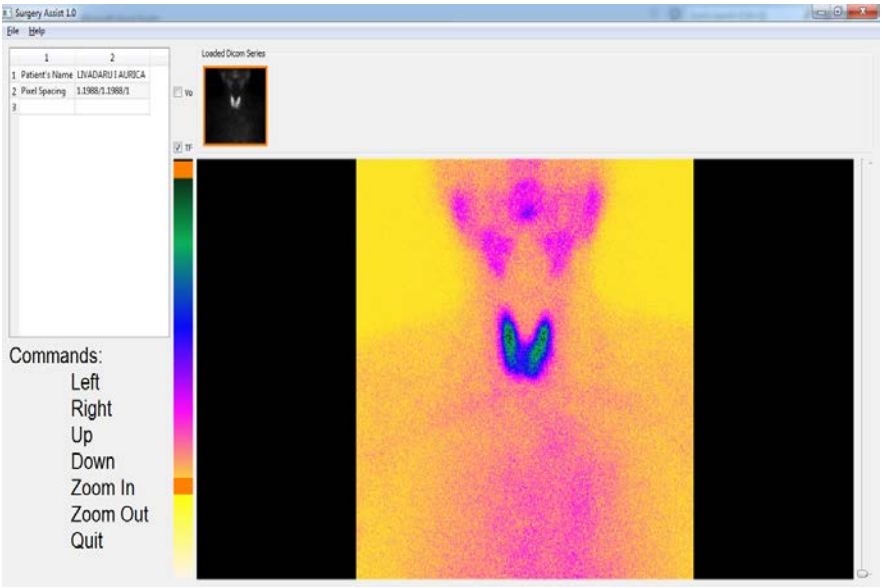


Fig. 4. The Surgery Assist application

4 Conclusion

The paper presents a system that indexes and automatically transfers imaging data files between medical institutions. The doctor from an institution asks for access and receives all the data found in the other institutions for a given patient. The system also includes an image viewer that allows voice commands that helps the doctor during surgery.

Acknowledgments. Part of the research for this paper was supported by the AM POS CCE project “Consolidarea institutionala si cresterea vizibilitatii Clusterului Regional Inovativ de Imagistica Moleculara si Structurala Nord-Est, cadru-suport pentru cresterea capacitatii de Cercetare Dezvoltare Inovare a membrilor si a competitivitatii IMM-urilor in domeniu din Romania – IMAGO MOL”, SMIS code: 49820.

References

[1] PACS system: https://en.wikipedia.org/wiki/Picture_archiving_and_communication_system

[2] DICOM files: http://dicom.nema.org/

Lucian Nita[1], Sinica Alboaie[2], Paul Herghelegiu[3]

[1]Technical University of Iasi, RomSoft SRL
E-mail:luc@rms.ro

[2]RomSoft SRL, http://romsoft.eu
E-mail: sinica.alboaie@romsoft.eu

[3]Technical University of Iasi
E-mail: paulhergh@gmail.com

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

Determining emotional classifiers for social disasters text clustering

Mircea Petic, Victor Cozlov

Abstract

In this study, we describe a tool set that includes methods for extracting relevant texts from the networks. The article describes a research of a methodology of Web crawler development. The experiment is done on a news site noi.md which is both Romanian and Russian language. The way of enriching existent classifier list is described.

Keywords: computational linguistic resources, linguistic classifier.

1 Introduction

This research is made within a project whose goal is related to the development of information systems oriented to ensure the security of citizens in extreme situations (natural calamities, technogenic catastrophes, etc.). The main source of information for the means of preventing and mitigating the consequences of social disasters are large volumes of unstructured data accessible to global information networks: mass media, social networks, blogs, and so on.

One of the first steps of the project our aim is the finding texts containing any signals about something that has occurred or is about to occur somewhere. In selecting the relevant information, we can point out relatively distinct approaches for different sources of information: news sites and social networks.

Another step is the enriching existent classifier list by means of internal language mechanisms that can be automatized.

2 Collecting texts

In order the processing phase be more rapid we elaborated a Crawler-based application service. Its work consists of page downloading, revision for new reference and page indexing. Indexing is used for the quicker search of pages inquired by a user [1]. It search through web news articles, downloads and extracts the text of this news, and stores them in the database. As every news site has its own structure we should take into account its particularity.

The developed Web Crawler works in the noi.md site area. For the data storage a database is used and all the information is written in files. We can find each news in two main languages: Romanian and Russian, so for more complete data each news is downloaded in both languages. We need about 15.4 seconds for a page download and extraction of information from it. This is a relative calculation based on 3000 downloaded pages [2].

We performed the following experiments on a sub-collection of texts (100 news articles, consisting of 35257 words, referring to railway, air and car accidents). The same procedure was applied: annotation at sentence and word levels, providing morphological information using UAIC Romanian Part of Speech Tagger [3]. Based on the obtained results, we got 7453 unique lemmas. In addition, extracting only those which have the frequency more than one, and part of speech noun, verb, adjective and adverb, we obtained 3456 different lemmas. So, the procedure showed how to reduce the number of susceptible words for markers and to optimize the processing time.

3 Enriching classifier list

As the Romanian language belongs to the class of inflectional ones, the process of word forming or derivation of a number of vowel or consonant alternations may occur, generating new stems. For every of these words, the context, where it is present, is highlighted and rules of inflection and derivation with a high degree of accuracy are applied

to generate more semantically related words, thus enriching the set of classifiers [4].

Therefore, for each word it is necessary to have all the possible stems. We use in-house elaborated tool to inflect the selected keywords (it has a general purpose, also applicable in our case). The tool is based on grammar rules with scattered contexts, and word-forms generation is reduced to the corresponding grammar rules interpretation [4].

As we described in [5] there are four algorithms: affix substitution, derivatives projection, formal derivation rules and derivational constraints. A few affixes form the overwhelming majority of derivatives: 12 of 41 prefixes formed 88.2% of all derivatives with prefixes, analogously, 52 of the 420 suffixes formed 87.7% of all derivatives with suffixes.

Even if we apply these four algorithms a step of validation is needed. Our approach was based on the Internet filtration and manual validation of the generated words. The method shows that we can increase the vocabulary by approximately 15%, with the accuracy of 89% [5].

4 Conclusion

The developed tools are useful in computational linguistic resources creation, which is of great importance in natural language processing applications. Building both large and good quality text corpora is the challenge we face nowadays. The results of this article can become a starting point for data processing of the downloaded text and the creation of text corpora of different domains.

Acknowledgments. The research is performed in the frame of the project Development of a toolkit for modeling strategies to mitigate social disasters caused by catastrophes and terrorism (17.80013.5007.01/Ua).

References

[1] V. Shkapenyuk, T. Suel. *Design and implementation of a high performance distributed web crawler*. In Proceedings of the 18th International Conference on Data Engineering (ICDE), San Jose, California. IEEE CS Press, 2002, pp. 357–368.

[2] V. Cozlov, M. Petic. *Downloading useful information from Web with the help of a Crawler*. In: Proceedings CMSM4 The Fourth Conference of Mathematical Society of the Republic of Moldova edicated to the centenary of Vladimir Andrunachievici (1917-1997), June 28 July 2, 2017, Chisinau, pp. 501–504.

[3] R. Simionescu. *Hybrid POS Tagger*, in Proceedings of Language Resources and Tools with Industrial Applications Workshop (Eurolan 2011 Summer School), Cluj-Napoca, Romania, 2011, pp. 21–28.

[4] Sv. Cojocaru, M. Petic, Gr. Horoş. *Tools for Texts Monitoring and Analysis Aimed at the Field of Social Disasters, Catastrophes, and Terrorism*. In: Computer Science Journal of Moldova, v.24, n.2 (71), 2016, pp. 157–171.

[5] M. Petic, V. Gîsca, O. Palade. *Multilingual mechanisms in computational derivational morphology*, in Proceedings of Workshop on Language Resources and Tools with Industrial Applications LRTIA-2011, Cluj-Napoca, Romania, pp. 29–38.

Mircea Petic [1,2], Victor Cozlov [1]

[1] Departament of Mathematics and Computer Science,
Alecu Russo Balti State University
Email: petic.mircea@gmail.com, viteak91@mail.ru

[2] Laboratory Programming Systems,
Institute of Mathematics and Computer Science

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

Controlling directed protein interaction networks, an overview

Vladimir Rogojin, Krishna Kanhaiya, Wu Kai Chiu,
Cristian Gratie, Keivan Kazemi, Eugen Czeizler, Ion Petre

Abstract

Network controllability research aims at discovering sets of external interventions that can drive a biological system to a targeted state. From the practical point of view, this means finding a (minimal) set of drugs that in combination are able induce a desired response from a cell. Hereby, network controllability may lead to the development of novel therapeutic approaches for systemic diseases like cancer. We summarize here our recent results related to this research direction. In particular, we overview the algorithm of structured network controllability, its implementation as a web application and a number of case studies.

Keywords: network controllability, software pipeline, web service, protein interaction networks, multi-drug therapy, bioinformatics.

1 Introduction

In recent years, high-throughput experimental technologies such as microarray gene expression profiling, gene and RNA sequencing, proteomics, etc. have generated a large set of biomedical data and became the core of biomedical research [2]. The overwhelming amount of already collected experimental data allow researchers to study functions and significance of various proteins, RNAs and genes as well as interactions between them in the context of some cellular processes and complex

diseases like cancer. The set of intracellular protein interactions serves as the base for signaling [14] and metabolic [6] pathways, and a number of essential intracellular processes for normal functioning of a living organism [9, 16]. In recent decade, analysis of protein interaction networks became significant for and provided novel insights into modern molecular biology from the network perspective [1].

A number of software tools have been recently developed for the analysis of interaction patterns and topological properties of protein interactions networks [5] and mostly focused at finding structurally important disease-specific protein interactions. On the other hand, the authors are not aware of existence of software solutions identifying strategies for controlling biochemical interaction networks. A number of algorithms have been developed recently to perform network structural analysis in order to suggest optimal sets of *input* (also called, *driven*) nodes through which one can control a network [12, 4]. We say that a network is *controllable* if there exists a time-dependent sequence of input signals such that while being delivered through the selected set of input nodes, it can drive the network from any initial state to any targeted final state within finite time [12, 11]. In [12], there was presented an efficient algorithm that selects a minimal set of input nodes through which one can control the whole directed protein interaction network. However, it was shown in [12] through several computer-based experimental tests that in real-life use cases it may be necessary to use as mush as 80% of all the nodes in a network for input in order to control the whole network. Of course, this fact makes the full network controllability approach for biomedical applications highly impractical. In practice, in majority cases, one needs to control only a tiny fraction of all the nodes in a network (for instance, a disease-specific set of essential genes) in order to reach the targeted cellular response [7, 4]. This kind of approach is called *target controllability*. It may be used, for instance, for development of combined personalized multi-drug therapy for a particular disease.

Here, we will review briefly our work in the domain of structural network controllability. In particular, we consider here the algorithm for

structural controllability of linear directed networks [4] (single source and single head directed edges, no parallel edges and self-loops), implementation of this algorithm and its use in our web-based application [15] and a number of case studies related to breast, pancreatic and ovarian cancer diseases [8].

2 Target Controllability Algorithms

From now on, we focus on the following controllability problem. We assume having a directed linear network and a set of target nodes that we want no control. We want to find a minimal (or almost minimal) set of input nodes through which we can control the target nodes (induce a finite sequence of signals on the input nodes to propagate to the target nodes, so that target nodes reach the desired state). Our goal is to find as small set of the input nodes as possible, which, hopefully, is much smaller than the set of target nodes. In practice, we can imagine the task of looking for a combinatorial drug therapy using few drug targets (genes or proteins serving as input nodes) through which we can control in parallel several dozens of disease specific and significant genes or proteins. We note that in our current research we try to answer questions like "what input nodes we need to control the targets" while not focusing yet on how to control those input nodes (i.e., should we up or down regulate a particular input node and how much).

We have shown in [4] that the original structural control greedy algorithm for finding set of input nodes for given targets from [7] is NP-hard. Moreover, we have improved that algorithm to cover a number of particular cases where the algorithm from [7] failed. In order to make the algorithm applicable in practice, we have introduced a number of heuristics, reducing in this way to 10-fold the average running time and significant reduction of the size of the average minimal solution. We have shown through a number of simulations that our approximation algorithm in [4] can efficiently find almost minimal sets of input nodes to control given target nodes in real-life networks containing thousands of nodes and edges. However, the efficiency of the algorithm is very

case-specific and depends on such characteristics of a network as its topology, density and other structural characteristics.

3 Pipeline for automated discovery of multi-drug therapy

In [15] we have developed a bioinformatics data analysis pipeline with web-based GUI front-end for automated generation of multi-drug therapy suggestions through the analysis of directed biochemical interaction networks. We have deployed this pipeline as a web service [3] that we host on our servers, as a stand-alone application in docker containers and as a source code under open-source license.

The pipeline performs three main steps:

1. **Preprocessing:** Generates biochemical interaction network basing on the user query (user identifies set of genes/proteins to be included into the network as well as the set of target nodes). The pipeline combines pathway data from multiple public sources such as KEGG, WikiPathways, Pathway Commons and generates comprehensive network that is to be analyzed further.

2. **Analysis:** Identifying an (almost) minimal set of input nodes through which one can control the target genes/proteins. This step is implemented by using our network controllability algorithm from [4].

3. **Postprocessing:** Mapping set of known drug targets (from the public databases of FDA-approved drugs) onto the set of discovered input nodes. Generating outputs: visual representation of the network with marked input, drug-targetable input nodes, and target nodes; network in XML format compatible with other bioinformatics and network software (like, Cytoscape); etc.

The pipeline is implemented within Anduril workflow framework [13], and the automated network generation in the first step is performed by means of Moksiskaan generic database [10].

4 Controlling Directed Protein Interaction Networks in Cancer

In [8] we have performed an analysis of protein interaction networks associated to breast, pancreatic and ovarian cancers. We have chosen as targets the survivability-essential proteins specific for each of the considered cancer types. We have shown that these targets are efficiently controllable from a relatively small computable set of input nodes. Also, we have adjusted our method to search for the input nodes among those for which FDA-approved drugs exist. We have discovered that, while majority of the drugs associated on the found input nodes are parts of known anti-cancer therapies, some of the input nodes with the associated drug targets were not known to be used in any of anti-cancer therapies.

5 Conclusion

We have demonstrated that a better understanding of control mechanisms in cancer-associated networks can drive us towards new more efficient therapies and personalized medicine.

Acknowledgments. The authors are partially supported by the Academy of Finland through grant 272451, by the Finnish Funding Agency for Innovation through grant 1758/31/2016, and by the Romanian National Authority for Scientific Research and Innovation, through the POC grant P_37_257.

References

[1] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12(1):56–68, jan 2011.

[2] Hamid Bolouri. Modeling genomic regulatory networks with big data. *Trends in Genetics*, 30(5):182–191, may 2014.

[3] COMBIO. Netcontrol4biomed: Network controllability for biomedicine. http://combio.abo.fi/software/netcontrol/, 2017.

[4] Eugen Czeizler, Cristian Gratie, Wu Kai Chiu, Krishna Kanhaiya, and Ion Petre. Target controllability of linear networks. In *Computational Methods in Systems Biology*, pages 67–81. Springer Nature, 2016.

[5] Nadezhda T Doncheva, Yassen Assenov, Francisco S Domingues, and Mario Albrecht. Topological analysis and interactive visualization of biological networks and protein structures. *Nat Protoc*, 7(4):670–685, mar 2012.

[6] Pawel Durek and Dirk Walther. The integrated analysis of metabolic and protein interaction networks reveals novel molecular organizing principles. *BMC Systems Biology*, 2(1):100, 2008.

[7] Jianxi Gao, Yang-Yu Liu, Raissa M. D'Souza, and Albert-László Barabási. Target control of complex networks. *Nature Communications*, 5:5415, nov 2014.

[8] Krishna Kanhaiya, Eugen Czeizler, Cristian Gratie, and Ion Petre. Controlling directed protein interaction networks in cancer. *Scientific Reports*, 7(1):10327, September 2017.

[9] Walter Kolch, Melinda Halasz, Marina Granovskaya, and Boris N. K Holodenko. The dynamic control of signal transduction networks in cancer cells. *Nature Reviews Cancer*, 15(9):515–527, aug 2015.

[10] Marko Laakso and Sampsa Hautaniemi. Integrative platform to translate gene sets to networks. *Bioinformatics*, 26:1802–1803, 7 2010.

[11] Ching-Tai Lin. Structural controllability. *IEEE Transactions on Automatic Control*, 19(3):201–208, 1974.

[12] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Controllability of complex networks. *Nature*, 473(7346):167–173, May 2011.

[13] Kristian Ovaska, Marko Laakso, Saija Haapa-Paananen, Riku Louhimo, Ping Chen, Viljami Aittomäki, and Erkka Valo. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome medicine*, 2(9):65+, September 2010.

[14] T. Pawson and P. Nash. Protein-protein interactions define specificity in signal transduction. *Genes&Development*, 14(9):1027–47, may 2000.

[15] Vladimir Rogojin, Keivan Kazemi, Krishna Kanhaiya, Eugen Czeizler, and Ion Petre. Netcontrol4biomed automated discovery of combined drug therapy. techreport 1162, Turku Centre for Computer Science, 2016.

[16] Takuji Yamada and Peer Bork. Evolution of biomolecular networks — lessons from metabolic and protein interactions. *Nature Reviews Molecular Cell Biology*, 10(11):791–803, nov 2009.

Vladimir Rogojin[1], Krishna Kanhaiya[1], Wu Kai Chiu[1], Christian Gratie[1], Keivan Kazemi[1], Eugen Czeizler[1], Ion Petre[1]

[1]Computational Biomodelling Laboratory, Turku Centre for Computer Science, and Department of Computer Science, Åbo Akademi University
E–mail: vrogojin@abo.fi, kkanhaiy@abo.fi, eugen.czeizler@abo.fi, ipetre@abo.fi

Determination of thresholds for liver disorders severity in cirrhotic PHT

Victoria Rotaru, Iulian Secrieru, Carolina Ţâmbală

Abstract

Differential and quantitative diagnostics of liver cirrhosis remains a major problem for both clinicians and developers of medical information systems. Determination of liver disorders severity in cirrhotic portal hypertension and its classification into Low, Middle, High would allow prescribing a personalized treatment and lifestyle. This objective corresponds to the latest trends in medical diagnostics.

Keywords: threshold, cirrhosis assessment, medical statistics, spider chart, PHT-portal hypertension, Doppler ultrasound imaging

1 Introduction

Nowadays, statistics have a variety of methods and its knowledge power is verified in various fields and constantly developed general theory. The statistics culture becomes a component of the general culture of contemporary mankind, and the statistics thinking – an indispensable scientific way for analysis and interpretation of very broad classes of phenomena. Currently, decision-making based on statistics is a very important factor.

Information statistics has found practical applications in various fields, including medicine.

In this paper we propose a method for determination of delimitation values of haemodynamic portal disorders severity in patients diagnosed with liver cirrhosis, where we used radar (spider) charts as comparison

mode. These values allow the classification of these disorders as: low severity, middle severity and high severity.

2 Selection criteria, material and methods

The study group consisted of 97 cases (41% men and 59% women aged 20-70 years), who were patients diagnosed with liver cirrhosis of various etiologies (especially of viral origin). Conventional ultrasound (2D) in Doppler duplex color mode was used as the diagnostic method. It has a number of advantages, being an accessible, non-irradiating, repeatable method that can be done even for the patient's bed.

For the description and evaluation of liver disorders severity in cirrhotic PHT the system of 5 parameters was proposed in [1,2] - congestion index, splenoportal index, vascular portal index, index PHT and spleen area.

Subsequently, all 97 of cases, described by these 5 parameters, were grouped into 3 categories: expressed (high), moderate (middle) and mild (low), depending on the general conclusions of the clinicians.

As comparison mode of 97 of cases, in order to obtain delimitation thresholds, the area of the figure obtained by representing the values of the 5 parameters in form of radar chart was used.

3 Use of Radar Charts

Radar chart is a graphical method of displaying multivariate data in form of a 2-dimensional or 3-dimensional scheme or with more quantitative variables represented on axes starting from the same point. One of radar charts application is quality enhancement control to display the performance values of any ongoing program [3].

Each of the 5 parameters, used to describe the 97 cases, were represented as individual axes that were radially arranged around a point. The value of each parameter is represented by the node (anchor) on the axis. After that the lines connecting the 5 nodes were drawn. So, we have obtained the radar graphic - polygon, as in the following example (see Fig.1).

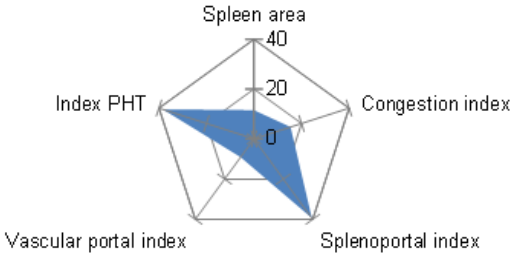


Fig. 1 Radar chart for patient nr.14 (with "Low" liver disorders severity)

Based on the obtained chart, the area of each section is calculated according to the formula $S=1/2*\sin(\alpha)$ and it is summed. The obtained sum, representing the polygon area that will be used as the numerical value of the liver disorder severity in cirrhotic PHT of each particular case. We repeat this procedure for each patient in our study group.

At the next stage, the precedents were arranged in ascending order (separately for each of the three groups), taking the value of the radar chart area as index (see Fig.2).

					category	area "spider" chart
0,073	38%	12	1,3	47	Low	10,7518
0,072	36%	11	1,4	49	Low	11,61419
0,06	41%	12	1,3	53	Low	11,72803
0,073	24%	13	1,5	45	Low	11,85641
0,068	36%	13	1,3	53	Low	11,94347
0,067	35%	13	1,6	52	Low	14,2214
0,09	36%	12	1,5	64	Low	15,75131
0,07	43%	14	1,4	70	Low	16,31451
0,1	38%	9	1,5	70	Low	16,36528
0,1	48%	11	1,6	68	Low	17,51039
0,09	51%	12	1,6	80	Low	20,37769
0,085	49%	13	1,4	94	Low	20,8511
0,1	55%	12	1,4	95	Low	21,06163
0,077	38%	13	1,6	85	Low	21,36105
0,073	47%	12	1,4	101	Low	21,73328
0,06	60%	13	1,4	112	Low	24,05687
0,077	51%	14	1,6	97	Low	24,39855
0,078	56%	13	1,6	98	Low	24,43909

Fig. 2 Sequence from the table of ascending sorted data, mild category

Finally, with the aim of thresholds for liver disorders severity in cirrhotic PHT determination, the cohort of 97 patients was segmented, using the radar chart area as parameter and the method proposed in [1]. Below the graphical representation of the segmentation result according to a single characteristic - the area of the "spider" polygon - is given. The minimum and maximum area values for each category (expressed, moderate and mild) were highlighted (see Fig.3).

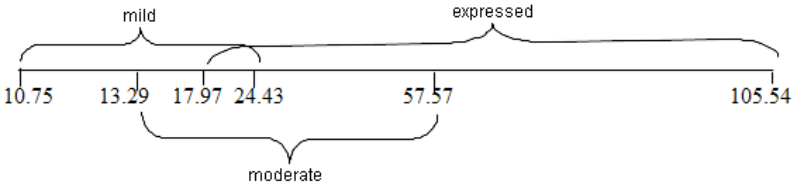


Fig. 3 Graphical representation of the segmentation result

4 Results and Conclusion

As a result of this research, the following delimitation of thresholds were established:

- threshold value 13,29 - if the area of the 5 calculated indices is less than 13,29 we can state that the liver disorders severity is "Low".
- threshold value 57,57 - if the area is greater than 57,57 we can state that the liver disorders severity is "High".
- Threshold value 17,97 - if the area is less than 17,97, we can state that the liver disorders severity is not "High".
- threshold 24,43 - if the area is higher than 24,43, we can state that the the liver disorders severity is not "Low".

The proposed method of using "spider" charts with subsequent segmentation can be successfully used to determine the threshold values for severity of hepatic disorder in chirotic HTP.

Acknowledgments. The Technology Transfer Project 17.80015.5007.213T has supported the research for this paper.

References

[1] Tambala, C.; Secrieru, Iu. *Portal hemodynamics disorders severity in liver cirrhosis assessment by duplex ultrasound*. In: Scientific medical journal "Curierul medical", Vol.59, No. 1, 2016, Chisinau, Moldova, pp. 37-40, ISSN 1857-0666

[2] Tambala, C. *Dopplerographic haemodynamic predictive parameters for portal hypertension associated with hepatic cirrhosis*. Curierul medical 58(4) (2015), pp. 20-24.

[3] Basu, R. *Implementing Quality: A Practical Guide to Tools and Techniques : Enabling the Power of Operational Excellence*. Thomson Learning, 2004, p.311, ISBN 1844800571, 9781844800575

Victoria Rotaru[1], Iulian Secrieru[2], Carolina Ţâmbală[3]

[1,2]Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova
[1]E-mail: gurdisvaleriu@gmail.com
[2]E-mail: iulian.secrieru@math.md
[3]Nicolae Testemitanu State University of Medicine and Pharmacy, Chisinau, Republic of Moldova
[3]E-mail: caroli@bk.ru

Modeling Disasters Using Networks of Morphological Tables

Illia Savchenko

Abstract

In this paper the concept of using networks of morphological tables is introduced for studying problems related to complex, unstructured systems, objects or situations. The proposed approach is applied to the problem of preventing and mitigating social disasters. The developed technique is adapted to be used in one of three modes: evaluating preparedness, monitoring and reaction regarding a disaster.

Keywords: morphological analysis method; disasters; system analysis; foresight; morphological tables.

1 Introduction

In recent years, people have become concerned with one of the most common types of disasters – social. Examples could be humanitarian catastrophes resulting from natural disasters or acts of terrorism.

These disasters are different by nature, they all have various reasons and all have different courses. They could manifest themselves separately from each other or in a chain with other disasters. Social disasters as processes have specific features, such as diversity and diverse nature of the causes and factors, as well as actions that lead to their appearance; spatial distribution of the conditions of uncertainty in time and space dynamics of the regions and their impact; unsteady properties and the uncertainty of their characteristics.

Analysis shows that at the present time various types of social processes, their causes, occurrence, effects and scope are investigated separately, excluding relationships, interdependencies, interaction. These approaches do not take into account some important factors that influence

these processes, the severity of undesirable effects, the ability and effectiveness of prevention.

2 Modified morphological analysis method

One of the highly productive ways of dealing with complex, unstructured problems is applying the morphological analysis method (MAM) [1]. The essence of this method is in analyzing objects, processes or phenomena by picking several their characteristics which can have alternate values. By combining different values for each characteristic, a very large multitude of object configurations is produced.

A modification of this method developed in IASA [2, 3] quantifies the main MAM elements, assigning values to the alternatives, representing either probabilities (for non-controlled parameters), or the expected efficiencies (for a decision or strategy that has to be constructed). Alternatives are interconnected by a consistency matrix that defines a relation between the correspondent values. This approach allows to easily estimate and analyze large amounts of possible configurations.

The procedure of two-stage morphological analysis, described in [2], is suitable for dealing with relatively small or poorly detailed problems, up to 8–10 parameters. To deal with more complex systems and problems, a network approach is proposed, which assigns separate morphological tables to different aspects of the problem, or different objects in a system. The connections between the tables are described by a consistency matrix, and the dependence can be either bidirectional, or unidirectional (causal).

This technique is especially useful when considering scenarios for a problem [4]. This model identifies the key factors and parameters of the scenario and provides the means to evaluate their probabilities (for non-controlled parameters) or efficiencies (for controlled parameters), taking into account the connections and interdependencies between them.

3 Constructing a network of morphological tables for disaster situations

A network of morphological tables is introduced to describe the decision to be made for the social disaster (Fig. 1). The network consists of three stages represented by morphological tables or groups of tables. The first group of tables consists of the table with common disaster parameters

applicable to all disasters, and a table with specific disaster parameters relevant for a specific type of disaster. This group as a whole describes either a potential multitude of disaster situations with probability estimates, or a specific situation which has to be responded. These tables can be prepared in advance for a number of relevant disaster types.

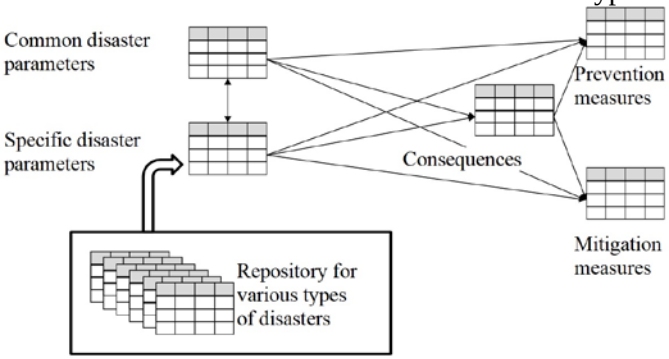


Figure 1. A network of morphological tables for disaster situations

The second group consists of a single table that describes the consequences of the disaster, which are crucial for determining the adequate response. This table can include parameters for threats to integrity of communication, transport and control, type and level of threat to people, degree of irreversibility of damage etc. The third group of tables contains the measures of preventing and/or mitigating the considered disasters. These tables are formed on the expert panel sessions, as they are very specific for the problem.

This network of morphological tables can be exploited in several different modes, depending on the problem at hand.

Evaluating preparedness. The probabilities of a certain disaster parameters are evaluated, and the efficiency of measures is estimated for a hypothetical variety of disaster situations. This shows the most critical problems that may cause larger damage if they happen. A 'what-if' model can also be constructed, making one or several of the parameters fixed to observe the changes in probability estimates of the other parameters.

Monitoring. This mode is relevant when a threat of a disaster is always present, e.g. a social disturbance that can potentially transform into social unrest. The input data for the MAM in this case can be extracted from the results of categorization and sentiment analysis for a collection

of text fragments from the available data sources (media, Internet, social networks etc.) [5]. The result for the disaster is constantly recalculated, and an early warning may emerge if the estimates for some critical parameters reach certain values.

Reaction. In this case the disaster has already happened, and its parameter alternatives are known, so the measures are evaluated according to a fixed configuration of tables for non-controlled parameters. This mode is useful for a quick reaction and choosing the best methods of mitigation for a disaster.

4 Conclusion

The proposed solution allows to improve the control and management during disasters, and provide decision-making support for social disaster situations. The developed technique may be used in one of three modes: evaluating preparedness, monitoring and reaction regarding a disaster.

References

[1] T. Ritchey. *Modeling Alternative Futures with General Morphological Analysis*, World Future Review (2011), pp. 83–94.

[2] N.D. Pankratova, I.O. Savchenko. *Morphological analysis. Theory, problems, application.* Naukova Dumka, Kyiv, Ukraine (2015).

[3] I.O. Savchenko. *Methodological and mathematical support for solving foresight problems using modified morphological analysis method*, System research and information technologies, Vol. 3 (2011), pp. 18−28.

[4] N.D. Pankratova, P.I. Bidyuk, Y.M. Selin, I.O. Savchenko, L.Y. Malafeeva, M.P. Makukha, V.V. Savastiyanov. *Foresight and Forecast for Prevention, Mitigation and Recovering after Social, Technical and Environmental Disasters*, Improving Disasters Resilience and Mitigation, IT Means and Tools, Springer (2014), pp. 119–134.

[5] I. Savchenko. *Estimation of Morphological Tables Using Text Analysis Results*, Computer Science Journal of Moldova, vol.24, no.2(71) (2016), pp. 148–156.

Illia Savchenko

Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine
E-mail: savil.ua@gmail.com

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

Quantification and Assessment of Diffuse Liver Diseases using Deep Data Analysis

Iulian Secrieru, Olga Popcova, Elena Gutuleac

Abstract

Diffuse chronic liver diseases play an important role in morbidity and mortality of the population in both economically developed and developing countries. Correct and early assessment of liver diseases combined with their appropriate management can certainly increase the patient's quality of life and its duration. The development of deep data analysis methods for quantification and assessment of diffuse liver diseases based on non-invasive measurements and laboratory tests is the basic research objective.

Keywords: Medical Informatics, Diffuse Liver Diseases.

We have witnessed the wide spread of diffuse liver diseases, which predominantly affect people of working age, having a significant negative impact on social and economic development of the country. Utilization of the latest technologies, in particular advanced information technologies, offers new opportunities and perspectives in diagnostics and treatment of liver diseases. These opportunities are mainly related to analysis of large data sets, representing both positive and negative diagnostic practices in domain.

The main data collected by the physicians is related to the following arias: patient personal data, anamnesis, clinical disease characteristics (including pain localization), laboratory tests, non-invasive measurements (ultrasound, MRI, elastography). The total number of parameters which clinicians have to analyse is around 120. Based of

them gastroenterologists establish diagnostics and make decisions on treatment.

Creating a data warehouse for diffuse liver diseases, structured based on a single protocol, unified and widely accepted by the medical community, would be useful. However, the collected data volumes are rapidly and inevitably increasing, that implies a need to use some specialized computing resources, capacities and algorithms.

At the moment, the research methodology for quantifying portal hemodynamics disorders severity assessment in liver cirrosis (based on duplex ultrasound scoring) have been developed [1]. This result was validated and appreciated by the specialized medical community.

Extending this result to the whole range of diffuse liver diseases by development of scoring for assessment of the severity of liver disorders (both in the early and advanced stages), as well as application of deep data analysis, will allow to enhance the decision making process in the field of gastroenterology.

We will determine a suitable subset of 5-30 parameters from 120 that will allow to quantify and then to assess a particular disease, subclass of diseases and all diffuse liver diseases.

Acknowledgments. The Technology Transfer Project 17.80015.5007.213T has supported the research for this paper.

References

[1] C. Tambala, Iu. Secrieru. *Portal hemodynamics disorders severity in liver cirrhosis assessment by duplex ultrasound.* Scientific medical journal "Curierul medical", vol. 59, no. 1 (2016), Chisinau, Moldova, pp. 37–40.

Iulian Secrieru[1], Olga Popcova[2], Elena Gutuleac[3]

Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova
[1]Email: iulian.secrieru@math.md
[2]Email: olea.popcova@yahoo.com
[3]Email: elena.gutuleac@math.md

Some efficient Game Development Techniques for Mobile Devices

Victor Seiciuc

Abstract

Development of some efficient IT Methodology in creation of Games for eLearning education on mobile devices and PC, are elaborated. Described are the Artificial Intelligence elements that allow the categorization of the proposed Games as Intelligent Support Systems.

Keywords: Games for eLearning on mobile devices

1 IT Methodology, Goals and Stages of Game Development

The developed *IT Methodology* it is a set of Programs and Computer Techniques organized in such a way, to allow can create on your PC or on your mobile device the desired Game, using minimum time and IT resources.

The basic *goals* of the proposed *IT Methodology* are to ensure (see [1]): -The *simplicity* and *comfort* of the building process of needed applications in Games elaboration; -Minimal *time* and *IT resources* in creation a certain type of Game requested by the customer.

The basic *stages* of Game development are:

(1) *Writing scripts*;

(2) *Integrating sound objects*;

(3) *Integrating graphic objects*;

(4) *Creating the actions and the functional elements for their management*;

(5) *Compiling and exporting to mobile platforms*;

(6) *Testing the application;*

(7) *Final on-line realization.*

Concurrently with Computer Techniques for create of Games, are exposed adjustment and delivery Tools of completed Games on widely known on-line Platforms: *Google play*; *App Store; BlackBerry App World; Windows Phone.*

2 Implementation in development of Games

The proposed *IT Methodology* can be applied to the development different types of Games, for example: 1) *Arcade*; 2) *Simulators*; 3) *Action*; 4) *Platforms*; 5) *Speed*; 6) *Training*; 7) *eLearning*; 8) *Gambling*, etc.

Games are organized on several levels. Each level contains its characteristic *elements* (buttons) that are responsible for the realization of certain concrete *events*.

Remark 1. *The result of the actions made by activating the Game buttons are still called* ***events****, and the respective* ***buttons*** *are called* ***functional elements*** *of this Game.*

Any *functional element* of the Game disappears when it is activated and the resulting *events* occur. *Events* may be one or more.

Remark 2. *We call the* ***purpose event****, the one who achieves the main result in the Game level. Otherwise, the* ***event*** *is called an* ***attendant event.***

3 Examples

I. Action Game - *Izolda Ninja Girl*: It's a fun game, developed for mobile platform Google Play. The game contains ten levels of play in which the player passes various obstacles and faces different opponents. In the play levels of the Game the *purpose event* is to destroy the Monsters and to move to the next level.

The *main character* as well as auxiliaries (*Monsters*) possesses elements possesses elements of *Artifcial Intelligence*, allowing them to interact by themselves. It is composed of the following parameters: 1) moving and stopping the movement; 2) recognition to the target object and distance calculation to him; 3) change of animation and behavior; 4) combat mode; 5) regeneration.

II. Speed Game - *Space Runner*: It's a Game of speed and reaction,

developed for mobile platform Google Play. The *purpose event* in the *Space Runner* is to jump over obstacles and accumulate *maximum points*. The *main character* of speed Game possesses elements of *Artifcial Intelligence* - the variation of the speed of movement according to the distance traveled.

III. E-learning Game - *Test-Game TP.T1*: The *Test-Game TP.T1* is developed for testing the knowledge in the distance learning. The test contains a set of evaluation questions and variants of answers to the course Theory of Probability and Mathematical Statistics, placed on the Moodle platform at Trade Co-operative University of Moldova (TCUM), HYPERLINK "http://www.uccm.md/" www.uccm.md/. This *Test-Game TP.T1* was developed for mobile devices and PC and runs on Windows and / or Mac OS. The mobile version can be accessed on Google Play, PC version - accessed on the site TCUM. The *purpose event* in test levels is to get a *maximum of points*. Elements of *Artifcial Intelligence* in *Test-Game TP.T1*are possibilities to fnish the Game, depending on the *time spent* on the test.

Conclusion. The presence of *Artifcial Intelligence* elements categorizes the described Games as *Intelligent Support Systems* (see [2]).

4 IT Resources in Creating the Games

The presented Games were created using three tools: *Construct 2* - the game modeling platform at the Programming level and Visual Design level. *Photoshop* - tools for creating Graphics. *Audacity* - Tools for creating and editing Sounds.

References

[1] V.V. Seichiuc. *Game Development Techniques for Mobile Devices.* Transactions of XVII International Symposium, Discrete singularities methods in mathematical physics – DSMMPh'15 Proc. LNCS, Kharkov-Sumy, Ukraine (2015), pp.227-231.

[2] F.G. Filip. *Sisteme suport pentru decizii.* Bucureşti:Ed. Tehnică, 2007, 363 p.

Victor Seiciuc

Moldova State University
E-mail: runsmaric@gmail.com

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI'2017, November 9-11, 2017, Chisinau, Republic of Moldova

Development of Expert System

Elena Socolov

Abstract

The difficulty in managing manufactory with custom and individual production forces development and evolution in expert system technology. The uniqueness of each order, resource constraint and existence of different technical ways to fulfill an order make this task intresting. This paper considers the use of expert systems techniques in the field of printing house manufacturing and gives as an example a prototype developed by author of this paper. The prototype addresses the problems of front end scheduling and the allocation of critical resources for maximum efficiency. The special value of the prototype is computer compilation of all technical and production documents based on order description from the manager.

Keywords: expert system, database, data mining.

1 Introduction

The individual production management problem is defined as the specification of the number of each type of resources (raw materials, equipment, technological chain) to use in a manufacturing system for a given planning horizon. Mainly, the approaches that dealt with this problem are complicated production route for every order, which overlaps with other orders in the struggle for resources. Also is important, an initial assessment of the cost. Filling in the order form, manager get technological map, where is indicated optimal path (all the necessary processes, optimal equipment and employee for each, astronomical time, workers time, raw materials, other production costs, cost total).

So, the remaining of this paper is organized as follows. First, Section 2 introduces the Expert System Simulation Approach (ESSA) that constitutes the utilization scope of ESMRS. Then, Section 3 describes the

Expert System static and dynamic knowledge representation, whereas Section 4 presents the basic Expert System features as well as its development using a commercial system shell. Finally, Section is dedicated to conclusion and future work perspectives.

2 Knowledge representation

Two main kinds of knowledge constitutes the core of any Expert System, also called knowledge based system: the static and the dynamic knowledge. The first is the group of concepts describing the expertise domain, where as the second is the reasoning mechanism. Both are exploited by the Expert System inference engine either to inductively answer a question or to deductively generate new facts.

In the job selection production route problem using the system, the main concepts are: optimization objective, performance measures, machine departments, performance limits and optimization history. Theses concepts are modeled using an bid data base of experts knowledge, in which each object is defined by a set of private data called attributes and a set of intrinsic functions called methods. Objects are also categorized into classes and sub-classes and both attributes and methods are inheritable.

The dynamic knowledge, also called ‘know-how’ can be schematized by a general resolution procedure in which each function or step is realized by a set of production rules.

Olso all problems are diagnosed, the Exper System tries to establish the list of corresponding feasible recommendations. Such recommendations should ensure the feasibility of all orders without deterioration in indicators a previously obtained orders. Then, the system ranks the feasible recommendations according to the severity of the related problems. Thus, a ‘lack of resource’ problem is generating order for purchase department, using data of all providers, purchases and experience of wirk with this material. Other related problem is manufacturing employment. Expert System evaluate possible production plans and checks order date of issue, considering other orders and raw material delivery time. In addition, problem solutions are ranked by decreasing in order of their difficulty of eliminating, with offered solutions.

3 Results

The Expert System for manufactory with custom and individual production was developed and implemented. User-friendly interface is supported this product too.

As a result of implementation is increase in the capacity of the enterprise by optimizing process planning. Reduction of stock and in-depth analysis of offers from suppliers allow you to save on raw materials. Increase in accuracy of calculation of the preliminary cost price reduces the risk of losses and raises competitiveness. Accuracy and transparency of production plans allowed to transfer workers to piece-rate wages. Master Data administration make the system flexible and viable. Order system is making possible all users see at what stage the order is located and what parameters it has. There are electronic job ticketing and shipping labels, what reduces the percentage of errors in order execution.

After the order is maked the final calculation (verification) is calculated. If there were large deviations between the plan and the fact, the system detects the points of deviation and looks for the reasons. Human experts are join this detection, they write reasons and all is introduced into system knowledge of the industry.

The System collects and provides all needed statistical and financila analysis for different type of users.

4 Conclusion

Once a expert system is developed and the data based is designed, it must be implemented, a task often more easily discussed then carried out. Expert system, about which there is a discourse, is implemented and works. Nonetheless, there will inevitably be changes needed. There is a desire to contribute to the system more and more intelligence. Besides, artificial intelligence techniques such as expert system, will be used in each industry, ousting human experts.

Acknowledgments.

The research described in this publication was made possible in part by - Casa Editorial Poligrafică “Bons Offices”.

References

[1] Jayaraman, V., Srivastava, R. (1996). *Expert systems, in production and operations management*. International Journal of Operations and Production Management, 16(12).

[2] Kusiak, A. (1990). *Intelligent manufacturing systems*. Englewood Cliffs, NJ: Prentice-Hall.

[3] Kusiak, A. (1997). *Artificial intelligence and operation research in flexible manufacturing systems*. Information processing and operation research, 25(1), pp. 2–12.

[4] He´di Chtouroua, Wassim Masmoudia, Aref Maalej. *An expert system for manufacturing systems machine selection*. Journal Expert Systems with Applications 28 (2005), pp. 461-467.

[5] Dyson, Robert G. Strategy, *Performance and Operational Research*. Journal of the Operational Research Society. January 2000.

Elena Socolov

Moldova State University
E-mail: lenocikasokolov@gmail.com

Proceedings of the Conference on Mathematical Foundations of Informatics
MFOI2017, November 9-11, 2017, Chisinau, Republic of Moldova

Approaches to the modeling of the movement of human flows by HLPNs in case of disaster

Inga Titchiev

Abstract

Humanity are facing with different kind of disasters daily. These can generate new accidents and catastrophes. The aim of these article is to propose an approaches to the modeling of the movement of human flows during evacuation from multi-storey buildings by using High Level Petri Nets (HLPNs) in order to analysis and mitigate the consequences of disaster.

Two issues can be solved by High Level Petri Nets: modeling and analysis. Modeling concerns the abstracting and representing the systems under consideration using HLPNs, and analysis deals with effective ways to study the behaviors and properties of the resulting HLPN models.

Keywords: the simulation of human flows, HLPN, disaster

1 Introduction

The problem of the successfully evacuation of people from buildings is the actual one, since people expose their life to the dangerous impact of the environment and the economic factors everyday.

In[5, 6, 7] was shown the scenario analysis for people evacuation from one storey buildings, but in many cases people working in multi-storey buildings. The essential difference in the movement of human flows into multi-storey buildings is the dominance of movement time in the staircase during evacuation time, a characteristic feature of which is the formation of flows with maximum density at the outlets of the floors.

Therefore, there is an actual problem of determining the structure and size of the evacuation tracks of human flows through the staircases, ensuring freely movement of people.

2 Evacuation rules

In the Republic of Moldova The Normative Supervision Section of Buildings and Fire Department exist, which performs the activity in the field of normative supervision in construction. In [4] it is specified the maximum flux density on the staircases, and the width of the staircases, which ensure freely movement of people.

For successfully evacuation of inhabitants the following features must be taken into account:

- $(t_{j(r+1)} - t_{jr})c_{jr}$, the amount of people arrived at j staircase, $j = 1, ..., N$.

  where

  N - number of floors.

  t_{jr} - the time of evacuation of all persons from j floor.

  c_{jr} - throughput of r evacuation exit from j floor.

- the maximum flux density of the evacuees not exceed the unobstructed value for movement D_{max}

- the contribution of the j^{th} floor to the density of the stationary flow at the zero time moment:

  $H_{jr} = h_j + t_{jr} * v$ floor height, $D_{jr} = \frac{c_{jr}}{w*v}$

  where

  H_{jr} - floor height, v speed of evacuation, w width of the staircase.

- $w_{max} = \frac{max\{D_j\}}{D_{max}}$ the maximum width of the staircase.

  where

  D_j flux density from j^{th} floor.

3 High-Level Petri Nets

Creating large, complex nets can be a hard task, as in particular modeling of the movement of human flows into multi-storey buildings. But High Level Petri[1, 2, 3] nets provide a good framework for the design, specification, validation, and verification of such systems. They support data and functionality definitions, such as using complex structured data as tokens and algebraic expressions as transition formulas. HLPNs have a wide range of application areas, like in the areas of communication protocols, operating systems, hardware design, business process re-engineering.

Modeling and simulation are the tools and methods that are effective, efficient, and can be used in study of the mitigation of consequences of disaster. In order to monitor and study these processes in a realistic and interactive environment, animation and gaming are the other two rapidly growing fields associated with HLPNs.

A *High-level Petri Nets* is a structure
$HLPN = (P; T; D; Type; Pre; Post; M_0)$ where

- P is a finite set of elements called Places.
- T is a finite set of elements called Transitions disjoint from $P(P \bigcap T = \emptyset;)$.
- D is a non-empty finite set of non-empty domains where each element of D is called a *type*.
- $Type : P \bigcup T \rightarrow D$ is a function used to assign types to places and to determine transition modes.
- $Pre; Post : TRANS \rightarrow \mu PLACE$ are the pre and post mappings with
  $TRANS = \{(t; m) | t \in T; m \in Type(t)\}$
  $PLACE = \{(p; g) | p \in P; g \in Type(p)\}$
- $M_0 \in \mu PLACE$ is a multiset called the initial marking of the net.

A *Marking* of the HLPN is a multiset, $M \in \mu PLACE$

A transition is enabled with respect to a *net marking* or in a particular *transition mode*. A transition mode is an assignment of values to the transitions variables, that satisfies the transition condition (i.e., the transition condition is true). The transitions variables are all those variables that occur in the expressions associated with the transition. These are the transition condition, and the annotations of arcs involving the transition.

A finite multiset of transition modes, $T \in \mu TRANS$, is enabled at a marking M iff $Pre(T_\mu) \leq M$

A step may occur resulting in a new marking M' given by $M' = M - Pre(T_\mu) + Post(T_\mu)$

Based on the above approach, we can simulate the movement of people modeling a steady flow by means of HLPNs and indicating the distance between floors, the number of people on each of the floors, the time of arrival of the first people from the floors to the staircase. We can obtain full evacuation time.

4 Conclusion

For the problem of mitigation of some consequences of disasters, a method of HLPNs was proposed. They allow to model and simulate of the system that represent emergency evacuation of people in case of disaster from multi-storey buildings. By this method it is possible to model large systems in a manageable and modular way. In particular, the problem of determining the structure and size of the evacuation routes of human flows through the staircases, ensuring freely movement of people, can be done.

Acknowledgments. 17.80013.5007.01/Ua, bilateral project *Development of a toolkit for modeling strategies to mitigate social disasters caused by catastrophes and terrorism* has supported part of the research for this paper.

References

[1] He X., Murata T. *High-Level Petri Nets Extensions, Analysis, and Applications*, Electrical Engineering Handbook (ed. Wai-Kai Chen), Elsevier Academic Press, pp. 459–476, 2005.

[2] Jensen K. *An Introduction to High-level Petri Nets*. Proceedings of the 1985 International Symposium on Circuits and Systems: Kyoto 85, pp. 723–726, Kyoto, Japan, 1985.

[3] Jensen K., Rozenberg G. *High-level Petri Nets: Theory and Applications*. Springer-Verlag Eds., pp. 724, London, UK, 1991.

[4] Komyak V., Danilin A. *Approaches to the simulation of the motion of human flows in the building and their comparison*. Proceedings of the Problems of fire safety. Edition 35, pp. 110–115, 2014.

http://nuczu.edu.ua/sciencearchive/ProblemsOfFireSafety/vol35

[5] Titchiev I. *Petri nets to model disaster prevention*, Proceedings of the Workshop on Foundations of Informatics, August 24-29, Chisinau, Republic of Moldova, pp. 445–449, 2015.

[6] Titchiev I. *Modelling and verification of evacuation system using Time Petri nets in case of disaster*, Proceedings of the 18-th International Conference System Analysis and Information Technology (SAIT 2016), May 30 June 2, Kyiv, Ukraine, pp. 46–47, 2016.

[7] Cojocaru S., Petic M., Titchiev I. *Adapting Tools for Text Monitoring and for Scenario Analysis Related to the Field of Social Disasters*, In the proceedings of The 18th International Conference on Computer Science and Electrical Engineering (ICCSEE 2016), October 6-7, Prague, Czech Republic, pp. 886–892, 2016.

Inga Titchiev

Institute of Mathematics and Computer Science
E–mail: inga.titchiev@math.md

Table of contents