I T H E A

International Journal

INFORMATION CONTENT & PROCESSING

2016 Volume 3 Number 3

International Journal
INFORMATION CONTENT & PROCESSING
Volume 3 / 2016, Number 3

IJ ICP is official publisher of the scientific papers of the members of the ITHEA® International Scientific Society

IJ ICP rules for preparing the manuscripts are compulsory.
The **rules for the papers** for ITHEA International Journals as well as the **subscription fees** are given on www.ithea.org .
The papers should be submitted by ITHEA® Submission system http://ij.ithea.org .
Responsibility for papers published in IJ ICP belongs to authors.

International Journal "INFORMATION CONTENT AND PROCESSING" Volume 3, Number 3, 2016

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with
Institute of Mathematics and Informatics, BAS, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politechnika de Madrid, Spain,
Hasselt University, Belgium
Institute of Informatics Problems of the RAS, Russia,
St. Petersburg Institute of Informatics, RAS, Russia
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia.

Publisher: **ITHEA®**

Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: office@ithea.org
Technical editor: Ina Markova

Printed in Bulgaria

ISSN 2367-5128 (printed) **ISSN 2367-5152 (online)**

MODEL OF PROBLEM DOMAIN "MODEL-DRIVEN ARCHITECTURE FORMAL METHODS AND APPROACHES"

Elena Chebanyuk, Krassimir Markov

***Abstract**: Model-Driven Architecture MDA can be regarded as a part of Model-Driven Development (MDD), where the modeling operation and transformation languages are standardized by Object Management Group (OMG).*

This article is devoted to designing of domain model that illustrates interconnections between mathematical foundations used to design formal approaches of software models transformation. Then, review of related researches advantages, according to MDA promising, is represented. A summary of requirements to model to model transformation approach according to MDA promising is outlined. The scope of mathematical foundations for designing model to model transformation techniques is defined. During domain model designing, a controlled vocabulary containing description of basic mathematical foundations that are used for model to model transformations is composed. Proposed domain model will serve as a template for choosing proper mathematical approaches and means for model to model transformation performing with given level of accuracy. It should be considered when new transformation methods are designed.

Means of increasing productiveness for transformation methods designing, involving proposed domain model, are formulated in conclusions.

***Keywords**: software model, model transformation, Model-Driven Architecture (MDA), theory of categories, first-order logic, metalogic, formal language, transformational grammar.*

***ACM Keywords**:*

***- Classification of 2012 year**: Software and its engineering, Software system structure, Software system model, Model driven system engineering.*

***- Classification of 1998 year:** D2 software engineering, D 2.0 Tools*

Introduction

Software models, often represented as Unified Modeling Language (UML) diagrams, are key artifacts in Model-Driven Development (MDD) approach. The key idea behind MDA is to separate the specification of the system functionality from its implementation on specific platforms increasing the degree of automation and achieving interoperability with multiple platforms. Thus, model processing, is a key and the most common activity for different software development lifecycles processes.

Software model transformation is a key activity of Model-Driven Architecture (MDA). Target of model transformation activity is to analyze together information from different software development lifecycle processes. In order to archive this goal the next model transformation activities are implemented: Model to Model (M2M), Model to Text (M2T), and Text to Model (T2M) transformations. Text in M2T and T2M transformations means analytical representation of software model or skeleton of program. In the first case, the role of text is to be subsidiary artifact that saves information about model. In the second case the role of text to be target of transformation [Truyen, 2006].

Other transformation' aspects are horizontal and vertical software model transformations. A horizontal transformation is a transformation where the source and target models reside at the same abstraction level. Typical examples are refactoring (an endogenous transformation) and language migration (an exogenous transformation). A vertical transformation is a transformation where the source and target models reside at different abstraction levels. A typical example is refinement, where a specification is gradually refined into a full-fledged implementation, by means of successive refinement steps that add more concrete information. Also code generation operation are considered as vertical software model transformation [Czarnecki and Helsen, 2006].

The key idea behind MDA is to separate the specification of the system functionality from its implementation on specific platforms increasing the degree of automation and achieving interoperability with multiple platforms.

The promise of Model Driven Architecture is to facilitate the creation of machine-readable models with a goal of long-term flexibility in terms of:

• Technology obsolescence: new implementation infrastructure can be more easily integrated and supported by existing designs;

• Portability: existing functionality can be more rapidly migrated into new environments and platforms as dictated by the business needs;

• Productivity and time-to-market: by automating many tedious development tasks; architects and developers are freed up to focus their attention on the core logic of the system;

• Quality: the formal separation of concerns implied by this approach plus the consistency and reliability of the artifacts produced all contribute to the enhanced quality of the overall system;

• Integration: the production of integration bridges with legacy and/or external systems is greatly facilitated;

• Maintenance: the availability of the design in a machine-readable form gives analysts, developers and testers direct access to the specification of the system, simplifying their maintenance chores;

• Testing and simulation: models can be directly validated against requirements as well as tested against various infrastructures. They can also be used to simulate the behavior of the system under design;

• Return on investment: businesses are able to extract greater value out of their investments in tools [Mens and Van Gorp, 2006].

To realize software model transformation tasks many successful researches are done. These researches are aimed to solve actual software engineering tasks of implementing software model transformation.

Consider papers that make a strong contribution in several MDE promising by means of using tools or means operating with some mathematical solutions.

A review of mathematical foundations for providing realization of model transformation techniques is outlined in [Rabbi et al, 2016].

The MDE promising are achieved solving actual software engineering tasks by means of implementing software model transformation [Favre and Duarte, 2016].

Paper [Greiner et al, 2016] represents a case study dealing with incremental round-trip engineering of UML class models and Java source code.

The MoDisco [Bruneliere et al., 2010] framework is used to parse the Java source code into a model representation. QVT-R is used to formalize a bidirectional model-to-model transformation between the UML model and the Java model. This aspect is an interesting feature of QVT-R, as a transformation

developer may provide a single relational specification which may be executed in both directions, rather than writing two unidirectional transformations separately. Moreover, QVT-R is chosen because of its declarative nature where the developer is supposed to focus on relations and dependencies between the metamodels rather than on single execution steps [Greiner et al, 2016].

Described approach tries to prevent information loss during round-trip engineering by using a so called trace model which is used to synchronize the platform independent and the platform specific models. Furthermore, the source code is updated using a fine grained bidirectional incremental merge. Also, information loss is prevented by using Javadoc tags as annotations. In case model and code are changed simultaneously and the changes are contradicting, one transformation direction has to be chosen, which causes that some changes might get lost [Greiner et al, 2016].

Declarative notation of QVT-R language allows involving predicate logic to express transformation rules.

A review of metamodeling tools is represented in paper [Favre and Duarte, 2016] and several metamodeling frameworks are described. Two points of this review are interesting for us from point of view of providing descriptions of mathematical foundations that consist the basic of metamodeling tools and frameworks:

1. Varró and Pataricza presented a visual and formally precise metamodeling framework that is capable of uniformly handling arbitrary models from engineering and mathematical domains. They propose a multilevel metamodeling technique with precise static and dynamic semantics (based on a refinement calculus and graph transformation) where the structure and operational semantics of mathematical models can be defined in a UML notation [Varro and Pataricza, 2003]. In order to verify metamodel some expressions are composed. First-order logics may be used as mathematical foundations for it. Different semantics (for example static and dynamic semantics) are used to verify the content of metamodels.
2. Also, logics are used to design modeling languages. For example modeling language “Allow” is based on first-order relational logic. It can be a base for creating frameworks for metamodels and model processing by means of analyzing their analytical representation. For example, Boronat and Messeguer describe an algebraic, reflexive and executable framework for metamodeling in MDD [Boronat and Meseguer, 2010]. The framework provides a formal semantic of the notions of metamodel, model and conformance relation between a model and a metamodel.

Authors of [Favre and Duarte, 2016] provided a metamodeling framework based on MOF and the algebraic formalism that focus on automatic proofs and tests. The central components of the proposed approach are the definition of the algebraic language NEREUS and the development of tools for formal metamodeling: the NEREUS analyzer and the NEREUS-to-CASL translator.

Let's consider main possibilities of NEREUS syntax: classes may declare types, attributes, operations and axioms which are formulas of first-order logic. They are structured by different kinds of relations: importing, inheritance, sub-typing, and associations [Favre and Duarte, 2016].

Descriptive notation of NEREUS allows adding new construction to describe variety of connections between objects and operations. First-order logic and algebraic formalisms used to describe relationships between NEREUS components and to compose new expressions for interconnecting NEREUS with other metamodeling tools.

Considering that there exist many formal algebraic languages, NEREUS allows connecting any number of source languages such as different Domain Specific Languages (DSLs) and target languages (different formal languages).

The contribution of paper [Rabbi et al, 2016] is a new web-based metamodeling and model transformation tool called WebDPF based on the Diagram Predicate Framework (DPF). WebDPF has been developed using HTML5 and JavaScript. Any HTML5 and JavaScript enabled web browser can be used for metamodeling with WebDPF. Algorithms, related to model transformation and analysis in WebDPF, are written in JavaScript and therefore executes on the client machine. WebDPF supports multilevel diagrammatic metamodeling and specification of model constraints, and it supports diagrammatic development and analysis of model transformation systems. In WebDPF, one can graphically specify constraints and model transformation rules. Transformation rules have been introduced in WebDPF for two purposes:

i) automatic rewriting of partial (incomplete) models so that they can be made to conform to the underlying metamodel;

ii) modelling the behavior of systems.

The support for model transformation systems in WebDPF can be exploited to (i) support auto-completion of partial models thereby enhancing modeling efficiency, and (ii) provide execution semantics for workflow models.

The WebDPF metamodeling environment supports multilevel metamodeling [Rutle, 2010]. In WebDPF, one can graphically specify constraints and model transformation rules, based on graph transformation rules. The rules are linked to predicates and the standard double-pushout (DPO) approach is used. Attached transformation rules for a predicate p, is given by a set of coupled transformation rules ρ(p) where the meta-models remain unchanged. A rule r ∈ ρ(p) of a predicate p has a matching pattern (L), a gluing condition (K), a replacement pattern (R), and an optional negative application condition. The matching pattern and replacement pattern are also known as left-hand side and right-hand side of a rule, respectively. WebDPF performs termination analysis based on principles adapted from layered graph grammars [Ehrig et al., 2006]. In a layered typed graph grammar, transformation rules are distributed across different layers. The transformation rules of a layer are applied as long as possible before going to the next layer. WebDPF generalizes the layer conditions from [Ehrig et al., 2006] allowing deleting and non deleting rules to reside in the same layer as long as the rules are loop-free [Favre and Duarte, 2016].

Authors of [Zaraket and Noureddine, 2014] proposed a method of checking software with first order logic specification using And-Inverter-Graph (AIG) solvers. AIG can be viewed as a restricted C++ program, specifically a concurrent program in which all variables are either integers, whose range is statically bounded, or Boolean-valued, and dynamic allocation is forbidden. Using first order logics it permits developing and modifying semantics for model checking operations. Using conjunctive normal forms it support defining templates to estimate software models quality.

The results of literature review, matching strong contribution of considered transformation methods and techniques to MDA promising are represented in Table 1.

The first column contains features of the MDA promising. The next four columns contain analysis of transformation methods and techniques outlined in the considered papers:

- Formal MOF Metamodeling and Tool Support [Favre and Duarte, 2016];
- Model Checking Software with First Order Logic Specifications using AIG Solvers [Zaraket and Noureddine, 2014];
- Bidirectional transformations approach with QVT-R [Greiner et al, 2016];
- Web-based metamodeling and model transformation tool called WebDPF [Rabbi et al, 2016].

The last column contains our proposal what features we expect from the full automated method model to model transformations which has to be developed.

Table 1. Matching of considered transformation methods and techniques to MDA promising

Methods and tools / MDA promising	Formal MOF Metamodeling and Tool Support	Model Checking Software with First Order Logic Specifications using AIG Solvers	Bidirectional transformations approach with QVT-R	Web-based metamodeling and model transformation tool called WebDPF	Full automated method model to model transformations (to be developed)
Summary	Approach integrates MOF meta-language with formal specification languages based on the algebraic formalism. More concretely, NEREUS, as a formal metamodeling language, supports processes for reasoning about MOF-like metamodels such as ECORE metamodels.	Synthesis and verification frameworks to validate programs. And-Inverter-Graph (AIG) is a Boolean formula with memory elements, logically complete negated conjunction gates, and a hierarchical structure.	Method uses extensible model for model to model transformations by means of adding relations and elements	WebDPF is based on the Diagram Predicate Framework (DPF). WebDPF supports multilevel diagrammatic metamodeling and specification of model constraints, and it supports diagrammatic development and analysis of model transformation systems.	Provide a both top-down and bottom-up software model transformation from any type of UML diagram to another
Technology obsolescence	Framework is extensible because all of its components are described declaratively.			WebDPF specification allows adding new programming languages	Using existing and new formal approaches and tools

Methods and tools / MDA promising	Formal MOF Metamodeling and Tool Support	Model Checking Software with First Order Logic Specifications using AIG Solvers	Bidirectional transformations approach with QVT-R	Web-based metamodeling and model transformation tool called WebDPF	Full automated method model to model transformations (to be developed)
Portability				Such open standards as javaScript and HTML5 allow using WebDPF for different platforms	To be compatible with open standards, open data specifications, and open model processing environments or tools
Productivity and time-to-market		Effective optimization of model checking operations lets to shrink time for processing large amount of software models	Using tools lets to proceed many software models and modules of code raising effectiveness of software development processes		Providing bidirectional transformations for different types of software models. (productivity). (*) Reusing information from different software models (time to market). (*)

Methods and tools / MDA promising	Formal MOF Metamodeling and Tool Support	Model Checking Software with First Order Logic Specifications using AIG Solvers	Bidirectional transformations approach with QVT-R	Web-based metamodeling and model transformation tool called WebDPF	Full automated method model to model transformations (to be developed)
Quality	Comparability with constraint language allows to design high quality metamodels	Setting model quality characteristics and involving different verification techniques to modeling process allows estimating resulting models		By understanding both SysML semantic and programming languages constructions	Using modularity, restriction languages, and semantic tools
Integration	Framework uses formal language and it is compatible with other model processing tools and plugins, namely CASL, HETS and AST			Web environment allows integrating proposed tool with other metamodeling frameworks	Considering operations on meta-level allows designing integration tools for different platforms
Maintenance:		Quality models are sources for effective software lifecycle development processes	Using open standards simplifies the model maintenance procedure		Maintenance procedure based on matching transformation techniques with visualization ones

Methods and tools / MDA promising	Formal MOF Metamodeling and Tool Support	Model Checking Software with First Order Logic Specifications using AIG Solvers	Bidirectional transformations approach with QVT-R	Web-based metamodeling and model transformation tool called WebDPF	Full automated method model to model transformations (to be developed)
Testing and simulation			Using tools following open standards notations allows verifying all intermediate results in model transformation process		Creating environment for model processing allows combining analytical results with existing plugins and tools for model testing and simulation

Task and challenges

Task: to design model of problem domain "Model-driven architecture formal methods and approaches".

A subsequent task is to represent information for designing model to model transformation automated method, covering all MDA promising.

Explanation, how to follow MDA promising by means of mathematical foundations, is represented in the last grey column of the Table 1.

Challenges to this model:

- Reflect interconnection between mathematical foundations used to design new transformation methods or techniques;
- Serve as a template for defining collaboration between mathematical foundations involved to transformation techniques;
- Simplify the procedure defining compatible data formats for transmitting information about models between different stages of transformation methods.

Domain model design

There is no standard for domain model design, but there are articles containing precise description of this process. According to many recommendations, the first step of domain model designing is to compose a controlled vocabulary. Such one is represented in Table 2.

Table 2. Controlled vocabulary of problem domain "MODEL-DRIVEN ARCHITECTURE FORMAL METHODS AND APPROACHES"

Term	**Definition**
Logic	A particular system or codification of the principles of proof and inference: *Aristotelian logic.* Study of correct reasoning, especially as it involves the drawing of inferences. [Britannica, 2016c] The formal mathematical study of the methods, structure, and validity of mathematical deduction and proof. [MathWorld. 2016g]
Formal logic	Abstract study of propositions, statements, or assertively used sentences and of deductive arguments. From the content of these elements, the discipline abstracts the structures or logical forms that they embody. Alternative titles: mathematical logic; symbolic logic [Britannica, 2016a].
First-order logic	The set of *terms* of first-order logic (also known as first-order predicate calculus) is defined by the following rules: 1. A variable is a *term*. 2. If *f* is an *n*-place function symbol (with n>=0) and $t_1, ..., t_n$ are terms, then $f(t_1,...,t_n)$ is a *term*. If *P* is an *n*-place *predicate* symbol (again with n>=0) and $t_1, ..., t_n$ are *terms*, then $P(t_1,...,t_n)$ is an *atomic statement*. Consider the *sentential formulas* $\forall xB$ and $\exists xB$, where *B* is a *sentential formula*, $\forall$ is the *universal quantifier* ("for all"), and $\exists$ is the *existential quantifier* ("there exists"). *B* is called the scope of the respective quantifier, and any occurrence of variable *x* in the scope of a quantifier is bound by the closest $\forall$x or $\exists$x. The variable *x* is free in the

Term	Definition
	formula *B* if at least one of its occurrences in *B* is not bound by any quantifier within *B*. [MathWorld. 2016e]
First-order predicate calculus	The set of *sentential formulas* of first-order predicate calculus is defined by the following rules: 1. Any *atomic statement* is a *sentential formula*. 2. If *B* and *C* are *sentential formulas*, then ***¬B*** (NOT B), ***B Λ C*** (B AND C), ***B V C*** (B OR C), and ***B⇒C*** (B implies C) are *sentential formulas* (cf. *propositional calculus*). 3. If *B* is a *sentential formula* in which *x* is a *free variable*, then *∀xB* and *∃xB* are *sentential formulas*. In formulas of first-order predicate calculus, all variables are object variables serving as arguments of functions and predicates. The set of axiom schemata of first-order predicate calculus is comprised of the axiom schemata of propositional calculus together with the two following axiom schemata: $\forall x\ F(x) \Rightarrow F(r)$ (1) $F(r) \Rightarrow \exists x\ F(x)$ (2) where *F(x)* is any sentential formula in which *x* occurs free, *r* is a term, *F(r)* is the result of substituting *r* for the free occurrences of *x* in sentential formula *F*, and all occurrences of all variables in *r* are free in *F*. Rules of inference in first-order predicate calculus are the *Modus Ponens* and the two following rules: $\frac{G \Rightarrow F(x)}{G \Rightarrow \forall\, xF(x)}$ (3) $\frac{\mathrm{F(x)} \Rightarrow \mathrm{G}}{\exists xF(x) \Rightarrow G}$ (4) where *F(x)* is any sentential formula in which x occurs as a free variable, *x* does not occur as a free variable in formula *G*, and the notation means that if the formula above the line is a theorem formally deducted from axioms by application of inference rules, then the *sentential formula* below the line is also a *formal theorem* [MathWorld. 2016e].

Term	Definition
Metalogic: Second-and Higher-order Logic	Study and analysis of the semantics (relations between expressions and meanings) and syntax (relations among expressions) of formal languages and formal systems [Britannica, 2016a]. Metalogic may be second or higher order logic. Second-order logic is an extension of first-order logic where, in addition to quantifiers such as "*for every object* (in the universe of discourse)," one has quantifiers such as "*for every property of objects* (in the universe of discourse)." This augmentation of the language increases its expressive strength, without adding new non-logical symbols, such as new predicate symbols. For classical extensional logic, properties can be identified with sets, so that second-order logic provides us with the quantifier "*for every set of objects*." There are two approaches to the semantics of second-order logic. They differ on the interpretation of the phrase "*for every set of objects*." Does this have some fixed meaning to which we can refer, or do we need to consider the variety of meanings the phrase might have? In the first case (which will be called *standard semantics*), we are taking for granted certain mathematical concepts. In the second case (which will be called *general semantics*), much less is being taken for granted. In this case, to be considered valid, a sentence will need to be true under *all the allowable meanings* of the phrase "*for every set of objects*." [Stanford, 2015]. In second-order predicate calculus, variables may denote predicates, and quantifiers may apply to variables standing for predicates [MathWorld. 2016e]. There is no need to stop at second-order logic; one can keep going. We can add to the language "*super-predicate*" symbols, which take as arguments both *individual symbols* (either variables or constants) and *predicate symbols*. And then we can allow quantification over *super-predicate symbols*. And then we can keep going further. [Stanford, 2015]
Language	Set (finite or infinite) of sentences, each finite in length, and constructed out of a finite set of elements [Chomsky, 1957].
Formal Language	Formal language is normally defined by an alphabet and formation rules. The alphabet of a formal language is a set of symbols on which this language is built. Some of the

Term	Definition
	symbols in an alphabet may have a special meaning. The formation rules specify which strings of symbols count as well-formed. The well-formed strings of symbols are also called words, expressions, formulas, or terms. The formation rules are usually recursive. Some rules postulate that such and such expressions belong to the language in question. Some other rules establish how to build well-formed expressions from other expressions belonging to the language. It is assumed that nothing else is a well-formed expression. [MathWorld. 2016f]
Notation	A series or system of written symbols used to represent numbers, amounts, or elements in something such as music or mathematics [Oxford, 2016].
Grammar	Grammar is best formulated as a self-contained study independent of semantics. We consequently view grammars as having a tripartite structure. A grammar has a sequence of *rules from which phrase structure can be reconstructed* and a sequence of *morphophonemic rules* that convert strings of morphemes into strings of phonemes. Connecting these sequences, there is a sequence of *transformational rules* that carry strings with phrase structure into new strings to which the morphophonemic rules can apply [Chomsky, 1957].
Transfor-mational Grammar (Transforma-tional-gene-rative Grammar)	System of language analysis that recognizes the relationship among the various elements of a sentence and among the possible sentences of a language and uses processes or rules (some of which are called transformations) to express these relationships. Transformational grammar assigns a *"deep structure"* and a *"surface structure"* to show the relationship of such sentences. A type of grammar that describes a language as a system that has a deep structure which changes in particular ways when real sentences are produced [Britannica, 2016e].
Generative grammar	Precisely formulated set of rules whose output is all (and only) the sentences of a language — i.e., of the language that it generates. There are many different kinds of generative grammars, including transformational grammar as developed by Noam Chomsky from the mid-1950s [Britannica, 2016b].

Term	Definition
Algebra	The part of mathematics in which letters and other general symbols are used to represent numbers and quantities in formulae and equations [Corry, 2005]. Term algebra usually denotes various kinds of mathematical ideas and techniques, more or less directly associated with formal manipulation of abstract symbols and/or with finding the solutions of an equation. Examples of algebras include the algebra of real numbers, vectors and matrices, tensors, complex numbers, and quaternions. (Note that linear algebra, which is the study of linear sets of equations and their transformation properties, is not an algebra in the formal sense of the word.) Other more exotic algebras that have been investigated and found to be of interest are usually named after one or more of their investigators. This practice unfortunately leads to entirely unenlightening names which are commonly used by algebraists without further explanation or elaboration [MathWorld. 2016b].
Modern algebra	Branch of mathematics concerned with the general algebraic structure of various sets [Britannica, 2016d]. Modern algebra, also called abstract algebra, is the set of advanced topics of algebra that deal with abstract algebraic structures rather than the usual number systems. The most important of these structures are groups, rings, and fields. Important branches of abstract algebra are commutative algebra, representation theory, and homological algebra. [MathWorld. 2016a].
Category theory	A general mathematical theory of structures and of systems of structures. It is a language, or conceptual framework, allowing us to see the universal components of a family of structures of a given kind, and how structures of different kinds are interrelated Category theory is an alternative to set theory as a foundation for mathematics. As such, it raises many issues about mathematical ontology and epistemology. Categories are algebraic structures with many complementary natures, e.g., geometric, logical, computational, combinatorial, just as groups are many-faceted algebraic structures [Stanford, 2014]. Category theory is a branch of mathematics which formalizes a number of algebraic

Term	Definition
	properties of collections of transformations between mathematical objects (such as binary relations, groups, sets, topological spaces, etc.) of the same type, subject to the constraint that the collections contain the identity mapping and are closed with respect to compositions of mappings. The objects studied in category theory are called *categories*. [MathWorld. 2016c]. A category consists of three things: a collection of objects, for each pair of objects a collection of morphisms (sometimes call "arrows") from one to another, and a binary operation defined on compatible pairs of morphisms called composition [MathWorld. 2016d].

Analyzing Table 2 and defining interconnections between vocabulary entities, domain model of mathematical approaches to be used for model transformation tasks is designed and represented on Figure 1.

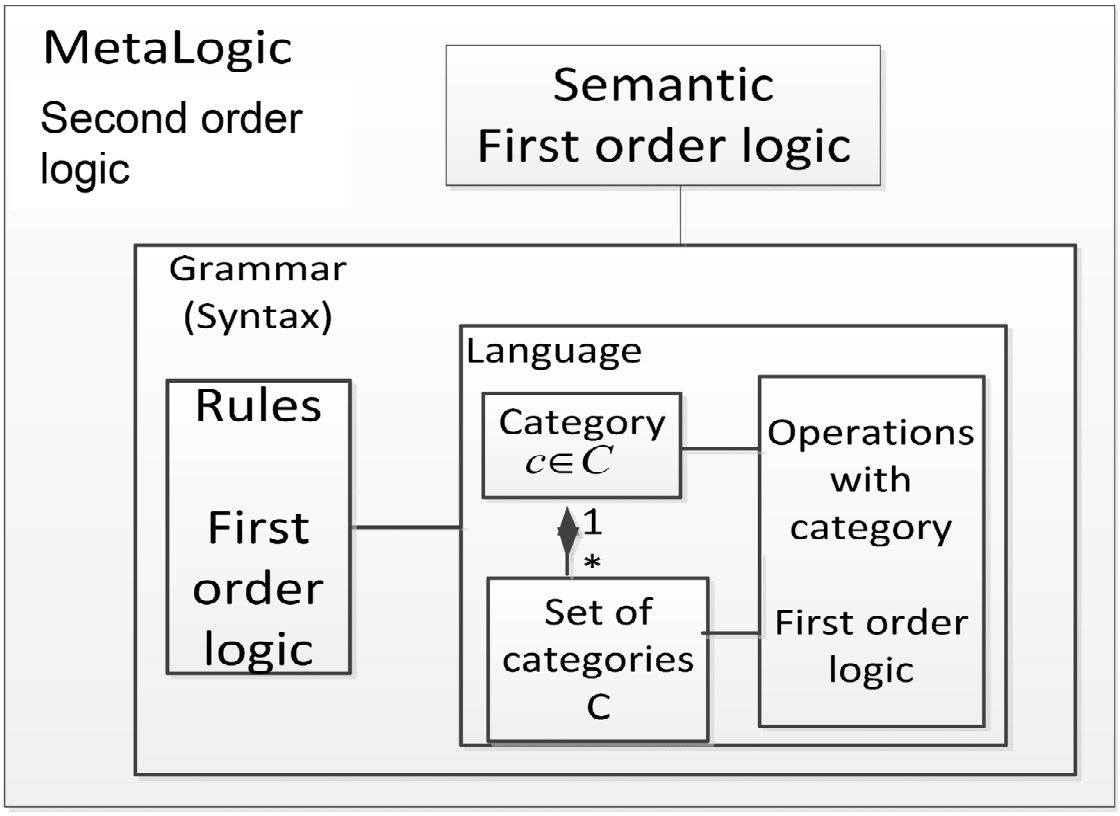

Figure 1. Model of problem domain

Domain model description

Consider mathematical foundations for software model elements description (Table 2).

Languages that are used for designing software models contain a set of elements to be combined for creation of complex software model structures. Considering theory of categories, one may see that each language element corresponds to some category. Combination of language elements allows creating more complex constructions. Also theory of categories is used to describe complex structure from some elements. Thus, it shows complex relations between components of complex structure.

Of course, some rules for designing these constructions should be followed. And first-order logic allows expressing these rules.

From software model elements whole model is designed. Thus, in turn, consider mathematical foundations for whole software model description.

Variety of complex constructions permits design precise software models. Grammar of language, namely its syntax, defines space of rules for checking correctness of software model in a whole. Also, first-order logic allows expressing rules and restrictions for correct model creation.

Semantic rules are used for analysis of software model content. To express semantic rules, first or second order logic is used. First order logic allows composing expressions. Second-order logic is aimed to estimate group of expressions or expressions that works with group of objects. Then chosen logic is used for estimating composed expressions. Also mathematical apparatus for semantic checking should be compatible with tool of software model designing and representation.

Finally, metalogic, or second order logic, as a foundation that studies interconnections between syntax and semantic, should cover all aspects of software model representation and all model processing operations.

Further researches

Further researches are:

1 Analyze model to model transformation techniques in order to define the typical stages of transformation operations.

2 Using controlled vocabulary (Table 2) and domain model (Figure 1), define mathematical foundations for every of these tasks and their interconnections. It has to be done with aim to ground the choice of mathematical foundations for every stage of transformation algorithms.

3 Represent formal description of all model to model transformation operations in terms of chosen mathematical foundations. Then propose techniques, which allow combining different formal solutions of transformational tasks.

Conclusion

Model of problem domain “MDA formal methods and approaches” is proposed in this article. Designed model illustrates relationships between different mathematical foundations that are used in formal software models transformational methods. It is proposed to use domain analysis artifacts, namely controlled vocabulary and domain model, by the following way:

- Decompose transformation techniques into steps;
- Match tasks of every step with possibilities of mathematical foundations, represented in controlled vocabulary (Table 1);
- Define a set of mathematical tools for solving tasks formulated above, analyzing controlled vocabulary and domain model;
- Co-ordinate different mathematical foundations for software model representation and processing.

Performing such operations allows:

- Formulating requirements to data specifications;
- Considering math apparatus for solving transformation tasks more attentively;
- Simplifying software model verification operations;
- Co-ordinate hidden relations between mathematical descriptions.

Bibliography

[Boronat and Meseguer, 2010] Artur Boronat, José Meseguer. An algebraic semantics for MOF. Formal Aspects of Computing. Springer Verlag, 2010, 22 (3), pp.269-296. <10.1007/s00165-009-0140-9>. <hal-00567269> https://hal.archives-ouvertes.fr/hal-00567269/document

[Britannica, 2016a] First-order Logic. Encyclopædia Britannica. 2016. https://www.britannica.com/search?query=first-order%20logic

[Britannica, 2016b] Generative Grammar. Encyclopædia Britannica. 2016. https://www.britannica.com/topic/generative-grammar

[Britannica, 2016c] Logic. Encyclopædia Britannica: 2016. https://www.britannica.com/topic/logic

[Britannica, 2016d] Modern Algebra. Encyclopædia Britannica. 2016.https://www.britannica.com/search?query=Modern%20Algebra

[Britannica, 2016e] Transformational grammar. Encyclopædia Britannica: 2016. https://www.britannica.com/topic/transformational-grammar

[Bruneliere et al., 2010] Hugo Bruneliere, Jordi Cabot, Frederic Jouault, Frederic Madiot. MoDisco: A Generic And Extensible Framework For Model Driven Reverse Engineering. 25th IEEE/ACM International Conference on Automated Software Engineering (ASE 2010), Sep 2010, Anvers, Belgium. pp.173-174, 2010. <hal-00534450> http://hal.univ-nantes.fr/hal-00534450/document

[Chomsky, 1957] Noam Chomsky, Syntactic Structures. Mouton publishers, Eilenberg: Mac Lane The, Hague, 1945 - 1957. ISBN 90 279 3385 5. p.107. http://ewan.website/egg-course-1/readings/syntactic_structures.pdf

[Corry, 2005] Leo Corry. History of algebra. In: Encyclopædia Britannica. 2005. http://www.tau.ac.il/~corry/publications/articles/pdf/algebra%20EB.pdf

[Czarnecki and Helsen, 2006] Krzysztof Czarnecki , Simon Helsen. Feature-based survey of model transformation approaches. IBM Systems Journal Vol. 45 No.:3: 2006. pp. 621 - 645. ISSN :0018-8670, DOI: 10.1147/sj.453.0621. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=5386627

[Ehrig et al., 2006] Ehrig, H., Ehrig, K., Prange, U., and Taentzer, G. (2006). Fundamentals of Algebraic Graph Transformation. Monographs in Theoretical Computer Science. Springer. ISBN 978-3-540-31188-1. http://www.springer.com/us/book/9783540311874

[Favre and Duarte, 2016] Liliana Favre and Daniel Duarte. Formal MOF Metamodeling and Tool Support. In: MODELSWARD 2016, Proceedings of the 4th International Conference on Model-Driven Engineering and Software Development. Edited by S. Hammoudi, L.F. Pires, B. Selic and P. Desfray. SCITEPRESS – Science and Technology Publications, Lda. Portugal, 2016. ISBN: 978-989-758-168-7. pp. 99-110. DOI:10.5220/0005689200990110, http://www.scitepress.org/DigitalLibrary/ProceedingsDetails.aspx?ID=j1i7qrX33Ns=&t=1

[Greiner et al, 2016] Sandra Greiner, Thomas Buchmann, Bernhard Westfechtel. Bidirectional Transformations with QVT-R: A Case Study in Round-trip Engineering UML Class Models and Java Source Code. In: MODELSWARD 2016, Proceedings of the 4th International Conference on Model-Driven Engineering and Software Development. Edited by S. Hammoudi, L.F. Pires, B. Selic and P. Desfray. SCITEPRESS – Science and Technology Publications, Lda. Portugal, 2016. ISBN: 978-989-758-168-7. pp. 15-27. DOI:10.5220/0005644700150027 http://www.scitepress.org/DigitalLibrary/PublicationsDetail.aspx?ID=efZXth7Zbbg=&t=1

[Rabbi et al, 2016] Fazle Rabbi, Yngve Lamo, Ingrid Chieh Yu, Lars Michael Kristensen. WebDPF: A Web-based Metamodelling and Model Transformation Environment. In: MODELSWARD 2016, Proceedings of the 4th International Conference on Model-Driven Engineering and Software Development. Edited by S. Hammoudi, L.F. Pires, B. Selic and P. Desfray. SCITEPRESS – Science and Technology Publications, Lda. Portugal, 2016. ISBN: 978-989-758-168-7. pp. 87-98. DOI:10.5220/0005686900870098, http://www.scitepress.org/DigitalLibrary/PublicationsDetail.aspx?ID=lzjjeczBZuA=&t=1

[MathWorld. 2016a] Abstract Algebra. Wolfram MathWorld. 2016. http://mathworld.wolfram.com/AbstractAlgebra.html

[MathWorld. 2016b] Algebra. Wolfram MathWorld. 2016. http://mathworld.wolfram.com/Algebra.html

[MathWorld. 2016c] Category Theory. Wolfram MathWorld. 2016. http://mathworld.wolfram.com/CategoryTheory.html

[MathWorld. 2016d] Category. Wolfram MathWorld. 2016. http://mathworld.wolfram.com/Category.html

[MathWorld. 2016e] First-Order Logic. Wolfram MathWorld. 2016. http://mathworld.wolfram.com/First-OrderLogic.html

[MathWorld. 2016f] Formal Language. Wolfram MathWorld. 2016. http://mathworld.wolfram.com/FormalLanguage.html

[MathWorld. 2016g] Logic. Wolfram MathWorld. 2016. http://mathworld.wolfram.com/Logic.html

[Mens and Van Gorp, 2006] Tom Mens, Pieter Van Gorp. A Taxonomy of Model Transformation. Electronic Notes in Theoretical Computer Science 152 Elsevier B.V., 2006. pp. 125–142. http://staffwww.dcs.shef.ac.uk/people/A.Simons/remodel/papers/MensVanGorpTaxonomy.pdf

[Oxford, 2016] Notation. Oxford Dictionaries. Oxford University Press, 2016. http://www.oxforddictionaries.com/definition/english/notation

[Rutle, 2010] Rutle, A. Diagram Predicate Framework: A Formal Approach to MDE. PhD thesis, Department of Informatics, University of Bergen, Norway. (2010). http://bora.uib.no/handle/1956/4469

[Stanford, 2014] Category Theory. The Stanford Encyclopedia of Philosophy. The Metaphysics Research Lab, Center for the Study of Language and Information (CSLI), Stanford University. 2015. Library of Congress Catalog Data: ISSN 1095-5054 http://plato.stanford.edu/entries/category-theory/

[Stanford, 2015] Second-order and Higher-order Logic. The Stanford Encyclopedia of Philosophy. The Metaphysics Research Lab, Center for the Study of Language and Information (CSLI), Stanford

University. 2015. Library of Congress Catalog Data: ISSN 1095-5054 http://plato.stanford.edu/entries/logic-higher-order/#4

[Truyen, 2006] Frank Truyen. The Fast Guide to Model Driven Architecture. The Basics of Model Driven Architecture (MDA). Cephas Consulting Corp, 2006. http://www.omg.org/mda/mda_files/Cephas_MDA_Fast_Guide.pdf

[Varro and Pataricza, 2003] Varro, V., Pataricza, A., VPM: A visual, precise and multilevel metamodeling framework for describing mathematical domains and UML. Journal of Software and System Modeling, Vol.2, Issue 3. Springer-Verlag, 2003. pp. 187-210. ISSN: 1619-1366 (print version), ISSN: 1619-1374 (e-version), doi:10.1007/s10270-003-0028-8 http://link.springer.com/article/10.1007/s10270-003-0028-8

[Zaraket and Noureddine, 2014] Fadi A. Zaraket, Mohamad Noureddine. Model Checking Software Programs with First Order Logic Specifications using AIG Solvers. arXiv, Cornell University, Ithaca, New York 14850. arXiv:1409.6825v1 [cs.SE] 24 Sep 2014. https://arxiv.org/pdf/1409.6825v1.pdf

Authors' Information

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

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

e-mail: chebanyuk.elena@ithea.org

Krassimir Markov – *Information Modeling Department, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria*

Major Fields of Scientific Research: *Software Engineering, Cognitive Science, Information Modeling, Multi-dimensional Graph Data Bases, Business informatics, General Information Theory*

e-mail: markov@ithea.org

HEURISTICS-BASED CLASSIFIER IN A FRAMEWORK FOR SENTIMENT ANALYSIS OF NEWS

Filip Andonov, Velina Slavova, Marouane Soula

Abstract: *Contemporary systems for sentiment analysis usually work with texts, that are known to be subjective and discovering the sentiment in them is already well studied and developed topic. When dealing with news articles however the main challenge is that the text itself should not be biased, even though that is not granted. Discovering the sentiments that some subjects in the text have about something in it requires more a priori knowledge about the world, the topic, the language, etc. All this requires a system designed for sentiment analysis in financial news to have flexible architecture and knowledge management capabilities. In addition, the social context of the opinion expressed has to be taken into account. That necessitates the establishment of classification criteria concerning the main social characteristics of the opinion holder.*

Keywords: *Sentiment analysis, Data mining, Heuristics, Text processing, Gender differences*

ACM classification keywords: *I.2 Artificial Intelligence I.2.7: Natural Language Processing, 1.7 Document and text processing*

Introduction

The goal of the project is to build a data-mining system for sentiment analysis of financial news. Sentiment analysis appeared as an NLP task and the commonly adopted definitions and techniques come from text analysis. A number of extended surveys have been published [e.g., Liu 2010, 2012]. After the overview [Slavova and Hinkov, 2014] we discovered that the problem is in capturing the implicit attitude in objective texts such as the texts of financial news. Most contemporary systems are designed to capture sentiments in the domain of marketing via Internet blogs or other sources in which the opinion holder expresses his opinion openly. Our main problem to overcome is to capture the attitude expressed in the news, despite the fact that news articles are meant to be objective. The problem is how to detect the subtle clues via which the sentiment is transmitted. That is why our strategy includes following the social reaction expressed on the Internet with relation to concrete published news, an event in the domain and the attitude of concrete subjects. This requires a semantic description of the domain, its

actors and events as well as the interconnections and influences between all these ingredients. For that reason our system is based on semantics and tracking of social feedback.

Short description of the system

The general architecture of the system includes 4 main stages as shown in Figure 1. The articles are retrieved from the Internet and transformed into a convenient format for further treatment. After extraction of several characteristics of the texts by means of a text processing module and natural language processing procedures, each article is assigned to one or more categories. As mentioned, the categorization to an ontological class is necessary in order to discard the articles that are not from the domain, in order to perform the processing related to the domain knowledge.

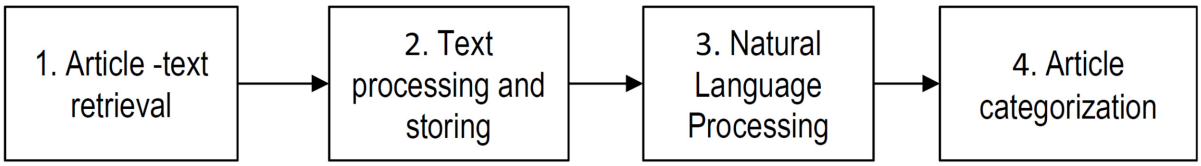

Figure 1. Main architecture of the system

The text processing (Figure 2) is done in the classical way – the text is tokenized in order to separate it into its basic ingredients, i.e. extract the title, paragraphs, sentences and word forms. The content of the article is stored in a database as word-forms (all the strings) and the structure of sentences, paragraphs etc.

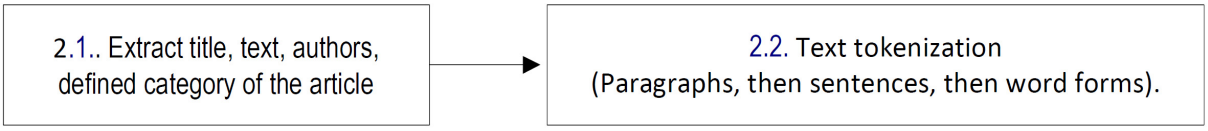

Figure 2. Steps of the text processing module

The Natural Language Processing (NLP) module has two main tasks – to detect the part of speech (POS) for each word form and to identify the Subject of each sentence (figure 3). As the NLP is hard and time consuming task, our strategy is to maintain an Enlarged Dictionary (ED) in the database, which contains the information about the word form itself only once, i.e. what part of speech it is, knowing that a word-form can represent more than one part of speech (table Word-forms in figure 4). In this way the

task of the interconnection between the DB and the texts consists in checking for availability of the word forms in the ED and storage of the extracted characteristics of the sentences and paragraphs only as a structure (word order, order of the sentences and paragraphs). This solution is better from the point of view of memory and processing usage.

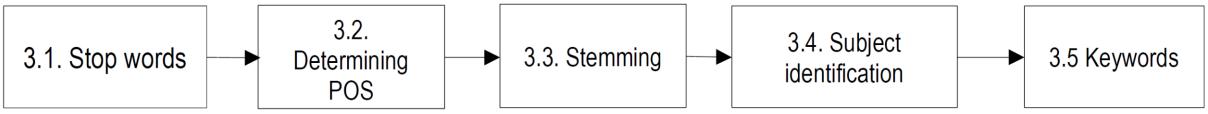

Figure 3. Natural Language Processing module

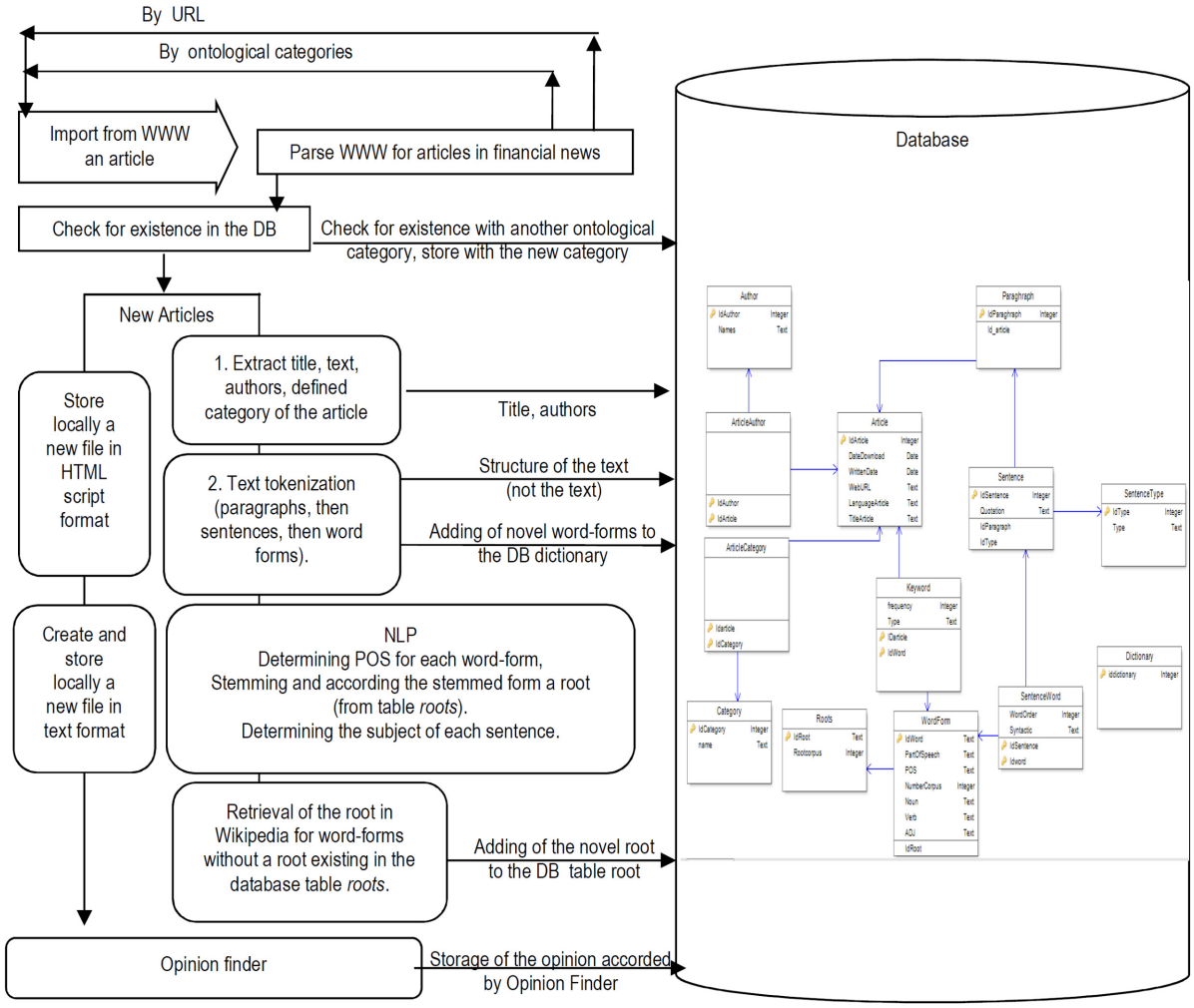

Figure 4. General scheme of the system

The information about Parts of Speech (POS) is necessary for the further sentiment analysis [Nicholls & Song (2009)], namely, as mentioned in the sources, the influence of the different POS on the sentiment detection has to be taken into account.

Concerning the task of Subject identification, it is performed by syntactic tokenization based on the use of Stanford Parser module. We consider the detection of the Subject to be of crucial importance for the further semantic analysis and the detection of the focus, which will be used in combination with the relationships within the ontological structure of the domain.

It is important for the semantic analysis to retrieve the root of the word-form. We consider that stemming gives a result (that we call a ***Stem***) which is similar to the *root* or the basic semantic form of the word. We expect to use the stems to differentiate between different ontological categories. Because of this our further statistical analysis is based on the stems as discussed in the next section.

The more detailed global scheme of the system under development is given in Figure 4. The system is built of several modules (Python is mainly used) and a database (MySQL) for storing the word forms and the structure of the articles. As it is given on the scheme, Internet is parsed for news following a list of URL addresses. The retrieved articles are first kept in a text format for a short time for further classification and longtime storage in the database. As it is shown, the text itself is not stored, but the word forms are identified and stored in table word-forms, which is in fact the ED and contains for the moment 226658 word forms. The structures of the sentences, paragraphs and entire articles are stored in the corresponding tables. Additional information is extracted by the developed modules as follows:

1. The Part of Speech (POS) of each word form (one or more type of POS for the word-form);
2. The subject of the sentence is identified and stored in the table Sentence;
3. The word-form is stemmed and the normal form (stem) is stored in the table Roots which contains also the frequencies of the roots. The procedure of determining these frequencies will be explained in the next section;
4. The article is classified following the ontological categories of the domain under consideration.

One problem on which we concentrate in this paper is related to the classification of the articles to one of the 20 categories that are on the third level of the ontology we have developed (fig. 5).

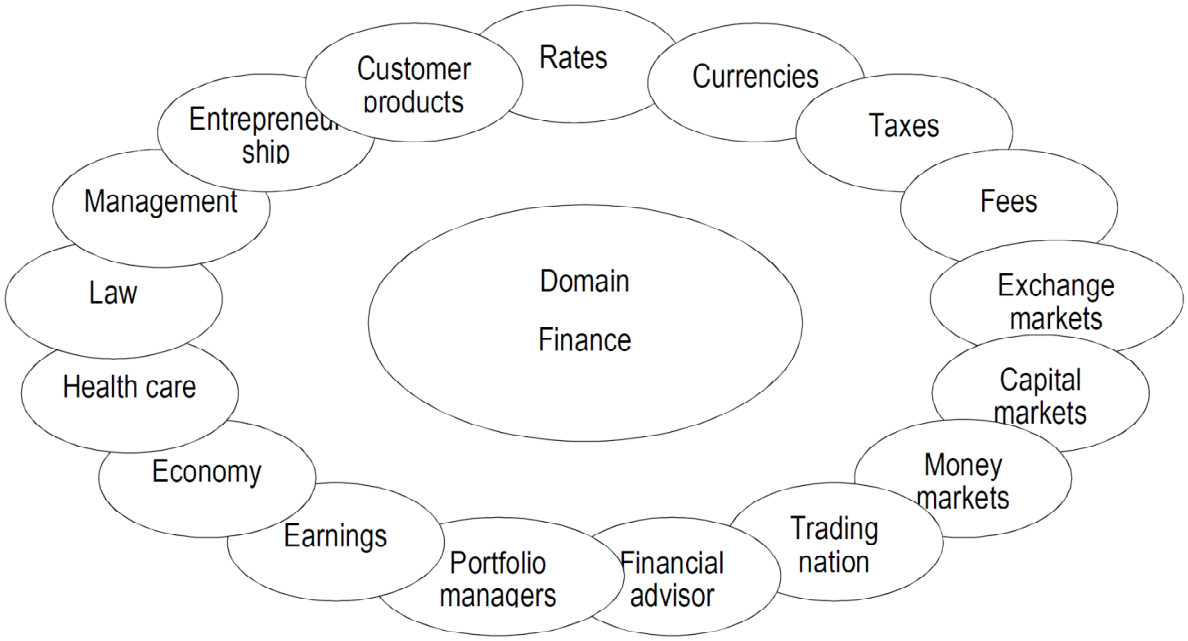

Figure 5. The third level of the Domain ontology developed for the purposes of the system

This problem of classification is related to the lack of explicit information on the Internet about the domain and the subdomain of a given article and, following from this, to the risk to store in the database articles that do not belong to the financial domain. We use this ontology information not only to identify the articles which are in the financial domain, but also because the further development of the system includes a knowledge representation module where the relationships between the events and the actors in the different branches of the ontology will be taken into account.

Using heuristics for classification of articles to the ontological categories

The aim is to build a filter-classifier discarding the articles that do not correspond to the ontology of financial news developed for the system. Each article retrieved from the Internet has to be rejected or classified to one of the 20 third-level classes of the ontology (figure 5). For this purpose, we use statistical methods and heuristics. We assume that the developed tool has to work correctly without being time and memory consuming.

The main strategy is to develop a model of the most frequent word roots in each of the 20 categories and to check each newly retrieved article for belonging to these categories [Andonov & Slavova, 2014]. As a first step we have stored collections of 50 articles per category (see fig. 5), selected manually by human experts. Each category collection C_i is further used as a global text model of the specifics of the language content for the category.

In one category C_i, each word form is stemmed and for each obtained stem S_j the *category representation coefficient* k_{ij} is calculated as follows:

$$k_{ij} = \frac{n_j}{n_{max}} . \log_2 \frac{1}{\frac{n_{jcor}}{n_{maxcor}} 2}$$

Where :

n_j is the frequency of the stem j within the category text C_i

n_{max} is the frequency of the stem with higher frequency in the category text C_i

n_{jcor} is the frequency of stem j in the corpus Subtlex UK of English texts

n_{maxcor} the frequency of the stem with higher frequency in the corpus Subtlex UK

We calculate the frequency n_{jcor} of the stems by summation of the frequencies of all stem derivative word forms (given in Subtlex).

The task of stemming. Some of the extracted stems are for words which do not exist in the used corpus Subtlex UK, so we have enlarged the stems- dictionary (root) taken initially from the corpus. A separate module has been developed for this purpose (Fig. 4), which gets the words for which we cannot find the roots in the database. This is done by retrieving the word from Wikipedia and finding there its root which after that is also saved in the database.

The list of calculated k_{ij} shows which stems are typical for the category C_i and discards the stems which are frequently used in th**e** language in general. The list of first 50 most frequently met stems within a given ontological category C_i represents its model M_i.

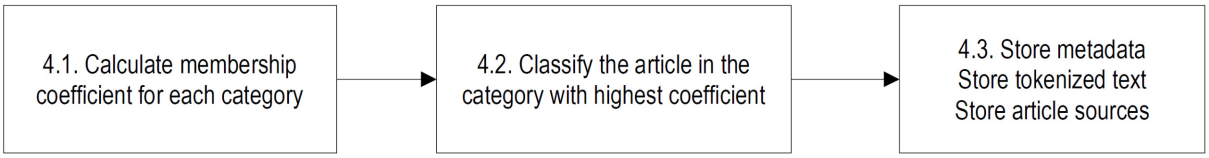

Figure 6. Heuristics classifier

Further the models M_i (i=1 to 20) are used for classifying the new articles to the existing models M_i. Each novel article retrieved on the Internet undergoes a filtering-classification procedure based on the following formula which calculates its coefficient of membership b_i to the category C_i.

$$b_i = \sum_{j=1}^{m} \frac{n_{stem\,j} \cdot k_{i\,stem\,j}}{n_{article}}$$

Where :

$n_{stem\,j}$ is the frequency of stem j (j = 1 to m) existing in the novel article

$k_{i\,stem\,j}$ is the category representation coefficient for the stem j in the model i

$n_{article}$ the number of stems in the novel article

b_i is the membership coefficient of the novel article to the category Ci

As it is seen from the formula, the categorization is based on summation of the “influences” of stems in the newly arriving article multiplied by their corresponding category representation coefficient k_{ij} for each category.

Development of the system for opinion mining in a social context

As mentioned, the system is conceived to take into account the social opinion expressed about a news article. For that we retrieve posts in the social networks (SN) containing an URL to the article in question, assuming that the post expresses the attitude of its author toward the news. People with different social roles have different understandings and subjectivity level and for that we need to learn more about the person expressing the opinion. Our preliminary choice for the opinion holder’s characteristics are “gender“ and “level of education” as they are of importance in general. This initial set of characteristics will assist us in the methodology and technical approach for detecting more specific categories. In this paper we concentrate on the detection of the gender of the opinion holder.

The final goal is to obtain from the text and metadata of the posts a composite feature vector for classifying the gender categories and to use it in machine learning algorithm.

In order to analyze public reactions to financial (or any other kind of) news we concluded that the best way to harvest that data is to mine twitter. The reasons for that are that the twitter limit of 140 characters is actually beneficent for our task, that twitter is very popular and it has a great API.

Our first task is to identify the **gender of the person** who expresses the opinion. Unfortunately this attribute is completely lacking in the twitter metadata. There are several approaches of deducing the gender from the available information [Burger et al., 2011] – using the user name and the account name, using links in the user profile to social networks having gender attribute in their profiles, analyzing the text itself and using other cues.

Using names. Twitter API gives two attributes: name and screen_name. Screen name, being the account username, has a requirement to not include spaces, but apart from that in both fields the user can write basically anything she desires. Problematic values of those two fields include: symbols that are not characters; nicknames instead of real names; organizational accounts instead of a personal ones; names in any language and ethnicity; names that are not delimited with spaces or any other delimiting symbol.

We developed an algorithm for assigning a probability score of membership to male and female names by comparing substrings of the nickname with lists of known gender-specific personal names. The approach gave satisfactory results.

Another approach to twitter account gender identification is to link the user to one or more accounts in other social network services where gender information is available. Some twitter users use the URL and/or the description fields in their profile to post a link to their profile page in other social networks such as Google plus, blogger, LinkedIn, Facebook, etc. Initially we harvested twitter data for any web URL posted in the above mentioned fields. The observation of 98 974 tweets showed that, as a percentage, only about 3% of the users post any kind of URL in their profiles and those who do are usually merchant or organization accounts. Nevertheless, using users own gender expression is probably the most reliable method. As everywhere in the Internet the truthfulness of the information that users post online about themselves is not easily verifiable. It turned out however that some social networks are used for work and people there tend to use their real names, gender, etc. From all the major SN services, the most suitable for our needs turned out to be Google plus. It allows up to 20 million requests per day, which is more than enough for our needs. The drawback is that twitter users

who post a link to Google plus are very rare – out of 100 000 twitter posts less than 100 have Google plus URL. From those, not all have posted their gender, because the field is not required. As others did before us, we decided that this information can be used to create a training set for machine learning algorithm. The automatic classification should be based on commonly available features. in order to verify the gender identification made by other means. Using Google plus has other advantage though – it contains a lot of structured data about the user which can be helpful on a later stage.

In the application we developed we save all the extracted from different sources information about the user in a database. One record corresponds to one tweet and contains the meta-information about the user. From twitter we get the username, description, date, device, language, location and country and from Google plus - nickname, name, tagline, aboutMe, relationship status, occupation, organization, and places lived.

Using tweet content. The gender differences in the manner of writing were investigated in some recent studies and it has been shown that there is indeed a gender gap that cannot be accounted by external factors such as education, age, etc [Slik et al, 2015]. Others [Wassenburg et al, 2015] have found that girls construct more coherent and vivid mental simulations than boys and rely more heavily on these representations. From this we conclude that it should be possible to get information about the gender of the writer of some text by analysing the text itself. One way to do this is to use the “bag of words” approach. Unfortunately English language contains fewer gender cues than other languages. It still can be done by finding statistical differences between male and female word frequency usage. More complicated approach will be to analyze other text characteristics of the writing. Of course the two can be combined to create a more complex feature vector.

Results

The initial tasks of downloading, classifying, storing and analyzing financial news texts are implemented and functional. The hypothesis that the stems can be used to perform the classification task is confirmed by experiments conducted with the system. The filtering procedure works correctly. The work done up to now encompass all the preliminary steps needed to be performed before the sentiment analysis should begin.

Concerning the social parameters of the opinion, there is no single approach for identifying author’s characteristics by written text that can give satisfactory results. Based on our work on the retrieved from

the SN data, we conclude that in order to identify the gender (and probably other characteristics) of a tweet's author, a complex feature vector is required that takes into account ***both*** the text itself and the available meta-information.

Acknowledgments

This paper is published with partial support by the ITHEA ISS (www.ithea.org) and the Central Fund for Strategic development, New Bulgarian University.

This system has been developed at the New Bulgarian University and made part of a students' project for Internet databases development. We are glad to present our acknowledgments to Hani Akra, a student in computer science, from Syria, who accomplished the task related to the retrieval of new roots by consulting Wiktionary.

References

[Andonov & Slavova, 2014] Andonov F & Slavova V. (2014) Some improvements of the OpenText Summarizer algorithm using heuristics, in proc. of the 10th Annual International Conference on Computer Science and Education in Computer Science.

[Burger et al., 2011] Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (). Discriminating gender on Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, July. (pp. 1301-1309).

[Liu, 2010] Liu, Bing, (2010). "Sentiment Analysis and Subjectivity." Handbook of natural language processing 2 (2010): 568.

[Liu, 2012] Liu, Bing (2012). "Sentiment Analysis and Opinion Mining." Synthesis Lectures on Human Language Technologies 5.1 (2012): 1-167.

[Nicholls & Song, 2009] Nicholls, C. H. R. I. S., and Fei Song. "Improving sentiment analysis with part-of-speech weighting." Machine Learning and Cybernetics, 2009 International Conference on. Vol. 3. IEEE, 2009.

[Slavova and Hinkov, 2014] Slavova V, and B. Hinkov. Multimodal Sentiments Analyses of Financial News – a project outline, in proc. of the 10th Annual International Conference on Computer Science and Education in Computer Science, 2014.

[Slik et al, 2015] Van der Slik, F. W., Van Hout, R. W., Schepens, J. J. (). The Gender Gap in Second Language Acquisition: Gender Differences in the Acquisition of Dutch among Immigrants from 88 Countries with 49 Mother Tongues. PloS one, 10(11), e0142056, 2015. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0142056

[Wassenburg et al, 2015] Wassenburg, S. I., Koning, B. B., Vries, M. H., Boonstra, A. M., & Schoot, M. Gender differences in mental simulation during sentence and word processing. Journal of Research in Reading, 2015.

Authors' Information

Filip Andonov *- New Bulgarian University, department of Computer Science, fandonov@nbu.bg.*

Major Fields of Scientific *Research: multi-criteria optimization, data mining, text processing, Python language*

Velina Slavova *- New Bulgarian University, department of Computer Science, vslavova@nbu.bg*

Major Fields of Scientific Research*: AI, Cognitive Science*

Marouane Soula*, Ph. D Student at the Department of Computer Science, New Bulgarian University*

SEARCH FOR NEIGHBORS AND OUTLIERS VIA SMOOTHED LAYOUT

Elena Kleymenova, Elena Nelyubina, Alexander Vinogradov

Abstract: *A new approach to the problem of quick search of the nearest neighbors and outlying objects in the training sample is presented. The approach is based on a special model for the area of mutual attraction of objects whose shape is consistent with the direction of the principal axes, and the form of the attraction dependency is natural. For this type of the zone a smoothed pre-structuring of the sample can be done that allows one to replace laborious procedure for finding nearest neighbors and outliers by simple bit addressing on the set of structure blocks. All actual distances and summed attraction levels for classes are calculated on the final stage for a limited number of objects*

Keywords: *feature space, logical regularity, attraction zone, even distribution, outlier, nearest neighbor, hyper-parallelepiped, bit addressing, dropout threshold*

Introduction

Accumulation and use only reliable information becomes of special importance in tasks of gathering and analyzing big data [Berman, 2003]. But usually such data are recorded just where the risk of error is raised because of intervention of the "human factor" – in medicine, education, environmental monitoring, social surveys and statistics, etc. Thus, often some parts of medical data records significantly differ from the set of average values of parameters for certain kind of patients. This may occur as a result of hardware malfunction, improper use of measuring methods, errors in the recording of results, etc. Similarly the environmental monitoring data are often taken 'in the field', and it's also associated with increased risk of data corruption while its registration and recording into reports. Mathematical treatment of incomplete, inaccurate and partially contradictory data presupposes revealing erroneous objects in order to provide correct application of precise methods of analysis and forecast to the rest of the data. Let R^N be the feature space of a recognition or prediction problem. For small volume of the training sample $X \subset R^N$ each object $x \in X$ makes essential contribution to formation of significant data clusters. It is usually assumed that the object *x* has its own attraction zone, and it is now known large number of approaches, in which the geometric shape of the attraction zone is modeled in some way - balls, hyper-parallelepipeds, Gaussian "hats", etc. [Tou, 1974]. Such heuristic models allow us to compensate for the deficit of training data at assumption of compactness of classes. In the opposite situation, when the sample has large volume, the use of suitable model helps to optimize the solution, in particular, to

reduce the effects of overfitting. Below we consider the problem of exclusion from the training sample the erroneous objects (outliers), which can occur both in small or large-volume sample. The limitation of attraction zone for outliers will serve here as a tool. We are primarily interested in the case of solving problems of recognition and prediction. We describe the data correction method, which is fast at error-detection stage and simultaneously provides efficient addressing for training sample objects and thus the acceleration of algorithms of the type "nearest neighbor".

Model of attraction zone and the generated density

The approach uses a model of attraction area as uniformly filled hyper-parallelepiped with center *x*, volume $\prod_{n=1}^{N}(2a_n+1)$, and density $1/\prod_{n=1}^{N}(2a_n+1)$, where a_n – half of the smoothing interval along the axis *n, n = 1,2, ..., N*. As a result of this smoothing (or spreading) procedure each central object *x* is evenly represented at all points of the hyper-parallelepiped. Location of a new object in zone of attraction of any training object votes for belonging the former to respective class. It is suggested in the approach to consider this impact only in case when the total generated density at a given point exceeds a predetermined threshold.

We will describe one of the reasons for choosing to rectangular zones of attraction. In case of large amounts the problem of analyzing numerous data highlights the priority of processing speed. Under the new conditions simple and well-researched approaches, in particular linear, get rebirth [Berman, 2003]. The most quick are methods in which all calculations can be reduced to comparisons on special linear scales of a particular type. In this series, one of the highly successful approaches turned out to be the one based on the use of Logical Regularities (LR) [Zhuravlev, 2006], [Ryazanov, 2007]. This approach uses data clusters in the form of hyper-parallelepipeds in R^N, each cluster is described by the conjunction of the form $L = \&R_n, R_n = (A_n \prec x_n \prec B_n)$, and substantially interpreted as recurring joint manifestation of the feature quantities $x = (x_1, x_2, \ldots, x_N)$ on intervals $(A_n, B_n), n = 1, 2, \ldots, N$. The principle of proximity precedents of the same phenomenon to each other here is embodied in the requirement of filling the interior of the cluster by objects of the same class. Same time, the geometric shape of the cluster represented by the parameters A_n, B_n, becomes of particular importance. Multiple joint appearances of feature values inside this shape are regarded as substantive independent phenomenon that is called Elementary Logical Regularity (ELR).

Thus, in our case the calculation of the predicate of finding new object in the zone of attraction of a training object for the given choice of the form of zone is also reduced to calculation just comparisons of numbers on the main axes omitting more complex operations.

Let's iterate the smoothing process, where each descendant of the central object (i.e. point with non-zero generated density) obtained in the previous steps is considered as a new center of attraction.

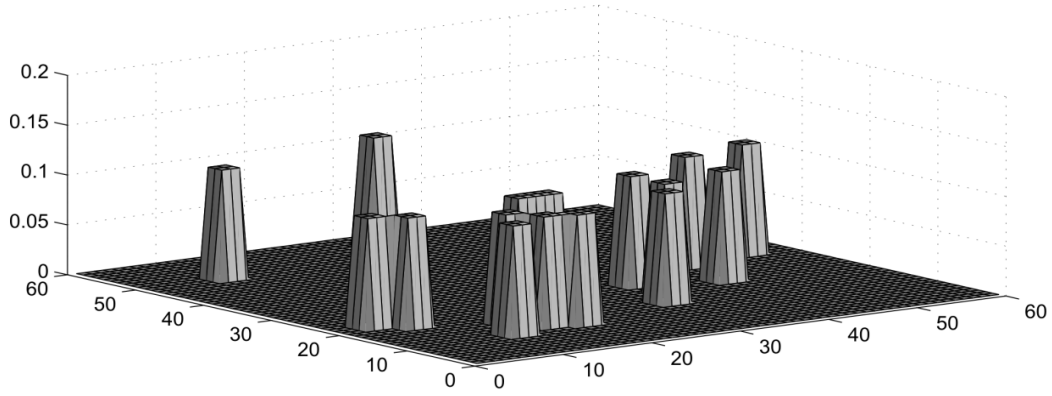

Fig.1. Example of a sample with two outliers after first stages of smoothing.

At *s* steps of smoothing operation, the attraction zone turns out the hyper-parallelepiped of volume $\prod_{n=1}^{N}(2sa_n+1)$ already unevenly filled with generated density. It is easy to show that the distribution within hyper-parallelepiped rapidly normalized with increasing parameter *s*, and already for *s* > *3* approximation of the distribution of descendants of a single point via Gaussian $\mu_i \exp\left(-\frac{1}{2}(x_i-x)^T\sigma^{-1}(x_i-x)\right)$ may in some cases be appropriate to construct numerical estimates for classes. With the expansion of volumes $\prod_{n=1}^{N}(2sa_n+1)$ close training objects are beginning to combine their areas of attraction, and this fact results in summation of estimates from neighbors.

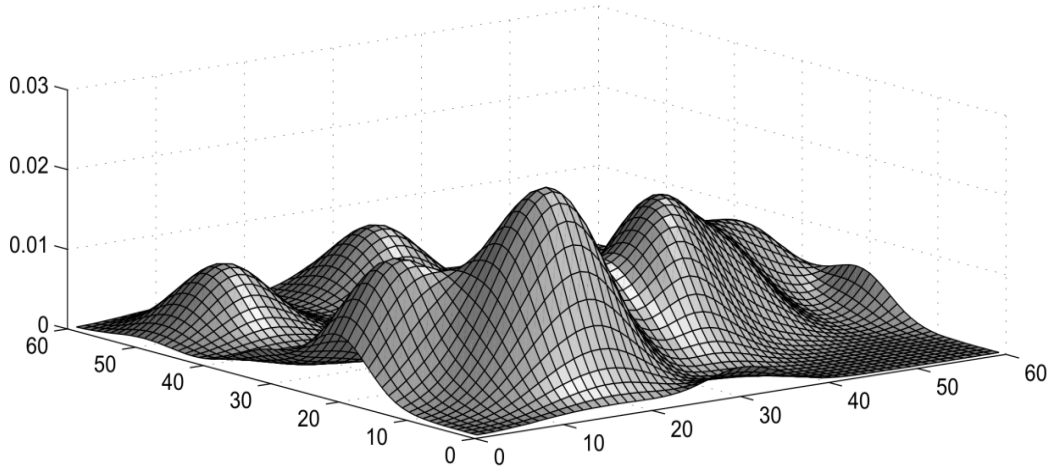

Fig.2. Result of several iterations of the smoothing procedure

The generated density for single objects decreases rapidly, including the maximums in each central point, and at suitable choice of the threshold all isolated objects can be excluded from consideration. The process of normalization generated density for the single point has been well studied, and the desired screening threshold may be calculated in advance with high accuracy. The recognition algorithm proposed in the paper has a structure similar to the standard algorithm of *k* nearest neighbors. It provides advanced possibilities for reconstruction of clusters in generated smoothed densities, and thereby, efficient evaluations for classes. At the same time, controlled expansion of the attraction zones can significantly reduce the amount of search of neighbors at the expense of simple pre-structuring of the sample.

Examples of practical tasks with outliers in data

Below we present two typical practical problems, in which the risk of accidental bias is big in recorded data. The two considered issues are of great importance, and so the quality of such complicated data should be high.

Completion and support of medical registries

In cardiology, neurology, oncology, surgery, including neurosurgery, unified standard forms of reports have been developed on the basis of information from registries. Quality registries are designed for a systematic data acquisition and the application of instruments for improving the quality of healthcare; they can be classified into two categories: disease registries and intervention registries. Quality registries differ from other clinical registries in the existence of special tools that are used in combination with the systematic data acquisition and are aimed at improving the healthcare quality. The tools of support of decision-making analyze the structured data on a patient introduced into the registry and form treatment recommendations on the basis of clinical instructions. During 2013–2014, registries of four directions were introduced with the formation of report templates at the Medical Center of the CB of RF, Moscow. The registry for percutaneous coronary interventions (PCIs, balloon angioplasty and/or stenting of coronary arteries) contains data on 288 patients, of which 138 patients were subjected to planned PCIs and 150 patients were subjected to emergency PCIs. The registry contains 230 indicators, the report on coronary interventions contains 7 sections: implementation of clinical protocols, demographic indicators, characteristics of patients with PCIs, preprocedural state for planned PCIs, preprocedural state for PCIs in case of acute coronary syndrome (ACS) without ST elevation, specific features of the procedure, and postoperative indicators. The registry of the acute cerebrovascular accident (ACVA) is represented by three types: ischemic stroke, hemorrhagic stroke, and transitory ischemic attack (261 patients had ACVA during 2012-2014, of which 197 patients had an ischemic

stroke, 25 a hemorrhagic stroke, and 39 a transitory ischemic attack). The registry of ACVA is formed of 240 indicators, and a report on each type of ACVA consists of six sections. By an example of an ischemic stroke, we can represent the contents of the sections: demographic indicators, indicators at prehospital stage, main risk factors, and the estimate of clinical data and the results of examinations (during the first day) during hospitalization and at discharge from the hospital. The registry of general surgery includes 3 nosological forms by which a decision is made on surgery: cholecystitis, appendicitis and inguinal hernia. The registry contains data on 403 patients operated during 2013-2014: cholecystectomy (214), appendectomy (67), and herniotomy (122). The registry contains 260-240 indicators for each patient, taking into account a specific character of pathology. Reports are formed automatically, separately for each nosology according to the following sections: demographic features, estimate of the condition of a patient before operation, the hospital stage, and the audit of the healthcare quality. Oncological registries include prostate cancer (104 patients), gastric cancer (94), renal cancer (64), and pancreas cancer (11).

The scope of indicators ranges within 316-150. The description of sections is made by an example of CPG: demographic data, regular medical check-up, diagnostics, initial treatment, local recurrence, remote metastases, hormone-resistant CPG, and outcomes. The patient's condition and the general and recurrence-free survival rate are evaluated, and the causes of death and the presence of bone fractures are characterized. Electronic forms are developed for all registries: electronic registration form, protocol of observance of clinical recommendations depending on the stage of a disease and individual risk factors, an electronic form for the audit of the results of treatment and clinical outcomes, and an outpatient form of regular medical check-up for assessing remote results. These numerous and complicated data are further used for gathering statistics, taxonomy and classification, recognition and prediction of events in treatment [Zhuravlev, 2016].

Monitoring of water resources ‘in the field’

At present, all over the world a considerable part of the population consumes contaminated water, of poor quality, because many of the local water intakes on the rivers and lakes have lost the quality of drinking water sources by pollution. At the same time there are many man-made factors of changes in the chemical composition of the water of small rivers and lakes: structural changes in aquatic systems, subtraction of river runoff for local economic needs, direct flows of the domestic wastewater into reservoirs, pollution from fertilizers and pesticides, discharge of industrial waters, and others. For these reasons, there is a great need for regular monitoring of small rivers and a comprehensive analysis of the data. For example, such an application for assessment of the quality of water bodies has been directed

by the local administration to specialists of Kaliningrad Technical University, Russia. During the work sampling of water was carried out at various water bodies in Zelenogradsky, Nesterovsky, Gusevsky, Krasnoznamensky, Ozersky, Chernyakhovsky, Pravdinsky, Slavsky, Guryevsky, Polessky areas of the Kaliningrad region. All recorded samples showed different exceedances of standards for various types of pollutants. Table 1 shows the comprehensive pollution data from 25 water bodies [Velikanov, 2013].

Table 1. Multiplicity of excess regulations and water pollution index

Object No	Multiplicity of excess regulations						Water pollution index
	Oxygen	BOD_5	Permanganate oxidability	Ammonia nitrogen	Phosphate phosphorus	Ferrous iron	
1	0,78	1,68	2,03	7,88	4,88	1,50	3,12
2	0,71	1,86	2,79	1,76	5,44	3,70	2,71
3	1,04	2,20	3,68	1,70	7,44	2,20	3,04
4	5,36	2,93	5,16	21,1	28,0	3,40	10,9
5	0,63	1,78	3,91	1,88	23,6	0,90	5,45
6	66,7	1,93	4,59	67,8	34,4	3,10	29,7
7	0,77	1,28	4,71	1,86	6,36	1,30	2,71
8	0,69	1,26	2,66	1,66	2,32	1,40	1,66
9	0,76	0,96	2,82	1,18	3,10	1,50	1,72
10	0,89	1,14	2,25	2,18	3,28	1,0	1,79
11	0,88	3,02	2,57	5,43	6,30	0,20	3,06
12	0,85	0,45	3,25	1,32	2,32	0,20	1,40
13	0,66	0,96	7,80	1,32	1,54	2,10	2,40
14	0,77	0,59	3,68	1,79	3,06	0,50	1,73
15	0,79	0,44	2,42	1,52	2,24	1,30	1,45

16	1,58	3,48	3,07	20,4	9,60	1,60	6,62
17	0,62	2,14	3,04	4,22	6,38	1,80	3,03
18	0,75	0,68	3,05	7,88	3,54	2,20	3,02
19	1,48	2,87	7,57	1,76	0,80	3,10	2,93
20	0,82	2,04	4,03	1,70	1,34	2,10	2,01
21	0,86	0,87	5,79	21,1	0,58	1,0	5,03
22	0,75	0,44	2,93	1,88	0,58	0,80	1,23
23	0,96	0,31	2,91	67,8	1,18	0,90	12,3
24	0,96	2,70	10,2	1,86	2,76	3,20	3,61
25	0,81	2,28	2,50	1,66	0,74	0,80	1,46

Samples collected at 3 sites (4, 6, and 23 in the Table 1) were assigned to the 5th class of water quality (extremely dirty). If one is interested in analyses of data for the just ordinary water bodies, then these three precedents should be excluded from consideration or, at least, analyzed separately as representatives of other taxons. The table represents only the most important integrated indicators of pollution, as well as some specific fixed concentration of harmful substances. In fact, for environmental monitoring of water bodies several groups of symptoms is used including organoleptic and sanitary characteristics of the water, indicators of presence of suspensions and emulsions of various substances, objects of micro-flora and other components of biological origin, concentrations of dissolved chemical compounds and individual elements. In total, this list can unite many tens of numerical parameters, and significant part of them has subjectivized expert origin. Of course, for single act of monitoring it is difficult to talk about the use of quite exact methods and techniques of pollutants sampling and recording of the results of their research. But the environmental safety gradually comes to the fore in many different aspects of human life and activity. It should be noted that the North-West Russia is bordered by several EU countries, and for this reason, EU environmental services are very interested in cooperation in matters of protection of water resources and the improvement techniques of monitoring, data storage and analysis. Improving representation of monitoring data recorded 'in the field' can also be a useful factor in ensuring such cooperation.

We will not show here examples of erroneous records found in complex data of this kind, and continue to consider the model example of a sample with outliers from the previous section as an illustration for explaining the algorithm usage in various applications such as the two shown above.

Marking training sample by zones of attraction

In what follows we show the use of controlled expansion of attraction zones of specified kind for arrangement of efficient addressing to the training data. The latter is especially important in the case of large dimensions, as occurs in two practical problems mentioned above.

Let $x_n^m, n = 1,\dots,N, m = 1,\dots,M$, be a training table and $K^l, l = 1,\dots,L$, be its marking by classes. The characteristic function $k(m) = \{l,\ if\ m \in K^l\}$ yields the number of a class by the number *m* of an object in the table. For the object x^0 to be recognized, we will seek a set $T = \{x^p\},\ p = 1,2,\dots,P$ of close points of the sample (i.e., nearest neighbors) of the vector x^0, that are located within the hyper-parallelepiped $[x_n^0 - sa_n, x_n^0 + sa_n],\ n = 1,\dots,N$ with volume $\prod_{n=1}^{N}(2sa_n + 1)$, which arises as a result of application of *s* smoothing operations. One should just find all the points of the sample that fall within the hyper-parallelepiped. The fact whether the point falls within the hyper-parallelepiped can be checked independently with respect to each of the axes *n, n=1,2,…,N*, for all the points of the sample $x_n^m, n = 1,\dots,N, m = 1,\dots,M$. For these reasons, one can start the test from any axis, say, from the first, *n=1*, and, on each subsequent axis, check only those points that withstood the closeness test on the previous axes. Having constructed the set $T = \{x^p\},\ p = 1,2,\dots,P$, we find all the points of the sample that extend at least minimal attraction to the object x^0.

On each of the main axes $n = 1,\dots,N$ one and the same test on detection sample points $x_n^m,\ n = 1,\dots,N, m = 1,\dots,M$ inside the limits of interval $[x_n^0 - sa_n, x_n^0 + sa_n]$ has to be performed. If the parameters $s, a_n,\ n = 1,\dots,N$ are fixed in advance, this detection is possible only for the points of a certain restricted subset of the training sample. Thus, when structuring the sample into blocks of size $2sa_n + 1$ for each axis *n = 1, ..., N*, the test should only be done to 3^N blocks that are the nearest to the point x^0 in R^N.

We will go further along this path, and choose blocks such that their boundaries are aligned with the binary bit grid. (Without loss of generality, we assume that all of the data are recorded in the fixed-point numbers) Namely, let q_n – the minimum bit such that $2sa_n + 1 \leq 2^{q_n}$, and the sample is structured into blocks with edge length 2^{q_n} for each axis $n = 1,\ldots,N$. Let $w = (w_1, w_2, \ldots, w_N)$ be block indices. Then q_n-th bit of the coordinate x_n^0 of the new object x^0 serves as immediate address w_n of the interval of length 2^{q_n} on the n-th axis, within which we can find the nearest neighbors. Of course, the two other adjacent intervals also should be taken into account, thus 3 intervals in total for each axis.

So, we have replaced the search for the nearest neighbors in the space R^N by the search for the nearest blocks in the space of block indices $W = \{w\}$. Again, such a search can be performed independently on each axis, and we get a subset W^0 of 3^N blocks as a result. After that each selected block is replaced with training objects it contains, and the final direct search is performed to create the set of neighbors $T = \{x^p\}$, $p = 1, 2, \ldots, P$.

A further reduction of the entire volume of the search is possible via organization of hierarchical structuring, when binary (or another but in concordance with the binary) bit mesh is used to construct index, similar to $w = (w_1, w_2, \ldots, w_N)$, as for the entire grid of blocks $W = \{w\}$, as well as within each of the blocks.

Search for the nearest classes and screening outliers

Let $\mu^l(x)$ the total probability density at position x that is generated by attraction zones of points of a class l, $l = 1, \ldots, L$. Using points x^p of the set $T = \{x^p\}$, $p = 1, 2, \ldots, P$, one can construct at the point x^0 a vector of estimates $\mu = (\mu^1, \mu^2, \ldots, \mu^L)$ for summed densities of all classes, $l = 1, \ldots, L$, and use them as votes for respective classes at decision making. In contrast to the ordinary method of k nearest neighbors, one needn't calculate here the distances immediately during the search. Factually, only carrying out the fast hit test for intervals $[x_n^0 - sa_n, x_n^0 + sa_n]$ is enough during the whole search for subset $T \subseteq X$ on the structured sample. Moreover, the first stage of the search consists in simple collecting indices $W^0 \subseteq W$. Distances and exact contributions of neighbors to the total generated densities of classes may be calculated at the final stage after detection of all points $x^p \in T$.

As is known, the evolution of the distribution of multiple smoothing for a single point falls in conditions of the central limit theorem. The deviation of this distribution $F_s(x)$ from the multivariate normal

$$N(x) - \frac{1}{(2\pi)^{\frac{N}{2}}|\sigma|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(x_1-x)^T\sigma^{-1}(x_1-x)\right)}$$

is described by the Berry-Esseen inequality [Berry, 1941], [Esseen, 1942]:

$$|F_s(x) - N(x)| \leq Const\frac{\rho}{\sigma^3\sqrt{s}},$$

where functions of the second and third absolute moments of the distribution of single smoothing (and, thus, the values of all variables a_n, $n = 1, \ldots, N$) are included as multipliers, and $Const \approx 0.4784$ (the exact value of this constant poses big challenge and continues to be refined in math statistics so far [Shvetsova, 2010]).

Therefore, at strict adherence to the proposed approach the parameter s should always be taken into account, especially in the case of small values. In other cases for larger values of the parameter s, a good estimate can be also obtained easily without reference to this parameter, for example, when used in computing tabulated Gaussian function, which in these cases already is a good approximation for the distribution of multiple smoothing. Same time, the dropout threshold for outliers should be adjusted accordingly, and the reduction of domain of the Gaussian function within boundaries of the block and the necessary renormalization of final distribution should also be considered as additional cost.

One can continue to work with the vector $\mu = (\mu^1, \mu^2, \ldots, \mu^L)$ in order to decide on the assignment of the object x^0 to one of the classes $l, l = 1, \ldots, L$, such as by using a maximum likelihood criterion, etc.

Thus, in the case of solving problems of recognition or classification the proposed approach leaves a significant range of possibilities. On the contrary, in the case of the problem of sifting alien objects there is an obvious simple way. You can pre-select the parameters s and a_n, $n = 1, \ldots, N$, so that the spread of the region of attraction for s iterations will be consistent with substantive expert views on the dropout

of outliers. Then, for strictly single objects of the training sample the maximum of the total generated density $K_s(x^0)$ at the center of the hyper-parallelepiped x^0 can serve as the screening threshold.

Fig.3. presents levels of the total generated density for the training sample of Fig.1. The lowest level serves as a drop-out threshold for two outliers located in the upper part of the figure. At solving the problem of recognition, classification or prediction, the screening of emissions can be carried out in parallel and in coordination with the decision of the main task. The rest of the sample with reliable part of the data can be used with large bases for decisions, identification natural regularities, creation of forecasts, and assessment of various risks.

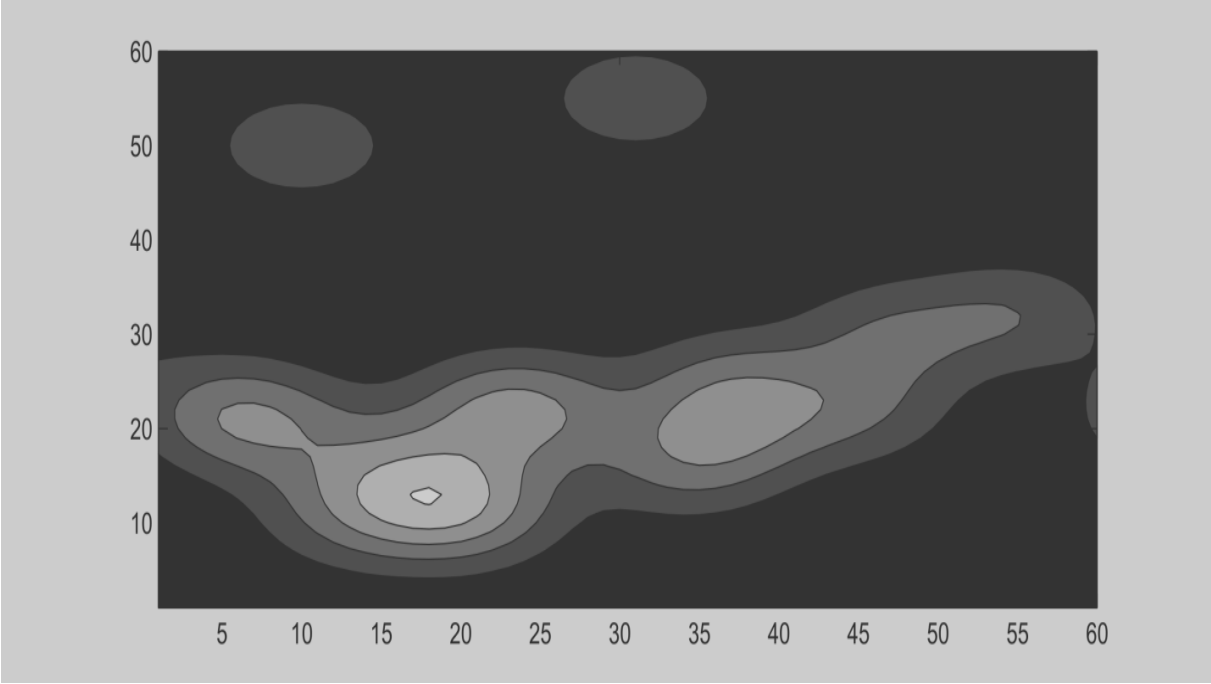

Fig.3. Levels of total generated density built for Fig.1. Senior levels can be used as reliable at construction of a decision rule that cuts off the impact of isolated objects.

Conclusion

This paper presents a fast method of finding the nearest neighbors and separating outlying objects in the training sample. The approach is based on a special model for the zone of mutual attraction of objects and on the pre-structuring of the sample into bit blocks, coordinated with the local geometry of zones. This makes it possible to reduce the search for the set of nearest neighbors to simple choice of a set of relevant blocks, the addresses of which are directly elder bits in the values of parameters of the

new object and construction of the final set of neighbors using only comparisons of numbers. The possibilities of the approach analyzed in application to recognition methods of type '*k* nearest neighbors'. In contrast to conventional methods of such kind, there is no need to calculate the distance between the objects immediately. All calculations of distances and local densities of probability distributions for the classes can be made in the final stages and for limited number of objects. Examples are shown of important practical problems in which the decision-making is especially critical with respect to the reliability of the data used in training. In the model example a procedure was presented for removal from the training sample outlying objects that can be executed in parallel and in accordance with the main data processing procedure. The main proposed innovations are related to the model of attraction zone that is compliant to the main axes and uses a natural form of the dependence of attraction on the distance. This model provides efficient bit addressing and fast data processing. In fact, this scheme can be further developed using a variety of other species of such dependence in models of attraction zone but still immersed in hyper-parallelepiped. The approach can be applied in various tasks of information processing and decision making where it is important to monitor the quality of data and perform operational steps to improvement.

Bibliography

[Berry, 1941] Berry, Andrew C. (1941). "The Accuracy of the Gaussian Approximation to the Sum of Independent Variates" // Transactions of the American Mathematical Society 49 (1), pp. 122–136.

[Esseen, 1942] Esseen, Carl-Gustav (1942). "On the Liapunoff limit of error in the theory of probability" // Arkiv för matematik, astronomi och fysik A28: 1–19.

[Tou, 1974] J.T.Tou, R.C.Gonzales Pattern recognition principles, Addison-Wesley, 1977, 390 p.

[Berman, 2003] Berman J. Principles of Big Data, 2003, pp. 1–14.

[Zhuravlev, 2006] Zhuravlev Yu. I., Ryazanov V.V., Senko O.V. "RASPOZNAVANIE. Matematicheskie metody. Programmnaya sistema. Prakticheskie primeneniya", Moscow: Izdatelstvo ”FAZIS” 2006, 168 p. (Russian).

[Ryazanov, 2007] Ryazanov V.V.: Logicheskie zakonomernosti v zadachakh raspoznavaniya (parametricheskiy podkhod) // Zhurnal vychislitelnoy matematiki i matematicheskoy fiziki, T. 47/10 (2007) str.1793 -1809 (Russian).

[Shvetsova, 2010] Shevtsova, I. G. (2010). "An Improvement of Convergence Rate Estimates in the Lyapunov Theorem". Doklady Mathematics 82 (3): 862–864

[Velikanov, 2013] N.L. Velikanov, V.A. Naumov, L.V. Markova, A.A. Smirnova "Results of Natural Researches of Small Water Flows on Reclaimed Soils" // Water: chemistry and ecology No 7 2013 pp.18-26 (Russian).

[Zhuravlev, 2016] Yu. I. Zhuravlev et al. "Methods for discrete analysis of medical information based on recognition theory and some of their applications" // Pattern Recognition and Image Analysis, 2016 (to appear).

Authors' Information

Elena Kleymenova *– Professor, Medical Center of the Bank of Russia, Sevastoposkii pr. 66, 117593 Russia; e-mail: e.kleymenova@gmail.com*

Elena Nelyubina *– Assistant Professor, Kaliningrad State Technical University, Sovietsky prospect 1, 236022 Kaliningrad, Russia; e-mail: e.nelubina@gmail.com*

Alexander Vinogradov *– Senior Researcher, Dorodnicyn Computing Centre, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences; Vavilova 40, 119333 Moscow, Russia; e-mail: vngrccas@mail.ru*

SYSTEMIC APPROACH TO ESTIMATION OF FINANCIAL RISKS

Petro Bidyuk, Svitlana Trukhan

***Abstract**: A computer based decision support system is proposed the basic tasks of which are adaptive model constructing and forecasting of financial risks. The DSS development is based on the system analysis principles, i.e. the possibility for taking into consideration of some stochastic and information uncertainties, forming alternatives for models and forecasts, and tracking of the computing procedures correctness during all stages of data processing. A modular architecture is implemented that provides a possibility for the further enhancement and modification of the system functional possibilities with new forecasting and parameter estimation techniques. A high quality of final result is achieved thanks to appropriate tracking of the computing procedures at all stages of data processing during computational experiments: preliminary data processing, model constructing, and forecasts estimation. The tracking is performed with appropriate set of statistical quality parameters. Examples are given for estimation of financial credit. The examples solved show that the system developed has good perspectives for the practical use. It is supposed that the system will find its applications as an extra tool for decision making when developing the strategies for financial companies and enterprises of various types.*

***Keywords**: mathematical model, system analysis principles, adaptive forecasting, decision support system, risk estimation.*

***ACM Classification Keywords**: CCS - Information systems - Information systems applications - Decision support systems - Data analytics.*

Introduction

Financial risk analysis and management is an urgent problem not only for the active financial organizations and companies but for all industrial enterprises, small and medium business, investment and insurance companies etc. Adequate models of multidimensional risks and the loss forecasts based upon them help to take into consideration a set of various influencing risk factors and make objective quality managerial decisions. There are many types of financial risks that could be described with mathematical models in the form of appropriately constructed equations or probability distributions. The market and some other types of risks are estimated with different modifications of VaR methodology that

provides a possibility to reach practically acceptable quality of risk estimates [MCNeil, 2005], [Basel Committee on Banking Supervision, 2006]. One of the widely spread type of risks is credit risk that arises due to failures of clients to return loans to banks. To analyze credit risks in banks the following models are used as of today: linear and nonlinear regression (logit and probit), Bayesian networks, decision trees, fuzzy logic, factor analysis, support vector machine (SVM), neural networks and neuro-fuzzy techniques, and combinations of the approaches mentioned [Mays, 2001], [Neil, 2005], [Shakhov, 2002].

All types of mathematical modeling usually need to cope with various kinds of uncertainties related to data, structure of the process under study and its model, parameter uncertainty, and uncertainties relevant to the models and forecasts quality. To avoid or to take into consideration the uncertainties and improve this way the quality of final result (risk values forecasts and decisions based on them) it is necessary to construct appropriate computer based systems for solving specific problems.

Selection and application of a specific model for process description and forecasts estimation depends on application area, availability of statistical data, qualification of personnel, who work on the financial analysis problems, and availability of appropriate applied software. Better results for estimation of financial processes risks is usually achieved with application of ideologically different techniques combined in the frames of one computer based system. Such approach to solving the problems of quality risk forecasts estimation can be implemented in the frames of modern decision support systems (DSS). DSS is a powerful instrument for supporting user's (managerial) decision making as far as it combines a set of appropriately selected data and expert estimates processing procedures aiming to reach final result of high quality – objective high quality alternatives for a decision making person (DMP). Development of a DSS is based on modern theories and techniques of system analysis, information processing systems, estimation and optimization theories, mathematical and statistical modeling and forecasting, decision making theory as well as many other results of theory and practice of processing data and expert estimates [Burstein, 2008], [Hollsapple, 1996], [Bidyuk, 2012].

The paper considers the problem of DSS constructing for solving the problems of modeling and estimating selected types of financial risks with the possibility for application of alternative data processing techniques, modeling and estimation of parameters and states for the processes under study.

Problem formulation

The purpose of the study is as follows: 1) analysis and development of requirements to the modern decision support systems; 2) development of the system architecture for financial risk evaluation; 3) selection of mathematical modeling and forecasting techniques for selected financial risks; 4) illustration of the system application to solving selected problem of financial risk estimation using statistical data.

Requirements to modern Decision support system

Modern DSS are rather complex multifunctional (possibly distributed) highly developed computing systems of informational type with hierarchical architecture that corresponds to the nature of decision making by a human. To make their performance maximum useful and convenient for users of different levels (like engineering and managerial staff) they should satisfy some general requirements. Define DSS formally as follows:

$$DSS = \{DKB, PDP, ST, MSE, MPE, RGP, DQ, MQ, REQ, AQ\},$$

where DKB is data and knowledge base; PDP is a set of procedures for preliminary data processing; ST is a set of statistical tests for determining possible effects contained in data (like integration or heteroskedasticity); MSE is a set of procedures for estimation of mathematical model structure; MPE is a set of procedures for estimation of mathematical model parameters; RGP are generating procedures for the risk estimates; DQ, MQ, REQ, AQ are the sets of statistical quality criteria for estimating quality of data, models, risk estimates, and decision alternatives, accordingly.

Such systems should satisfy the following general requirements that follow from the system analysis principles: 1) – contain highly developed bases of data and knowledge with mathematical models, quality criteria for each type of computing, and model selection rules, as well as necessary computational procedures; 2) – to achieve high quality of the final result the hierarchy of the system functioning should correspond to the hierarchic process of making decision by a human; 3) – their interface should be based on the human factors principles: user friendly, convenient and simple for use, as well as adaptive to users of various levels (e.g., engineering and managerial staff); 4) – the system should possess an ability for learning in the process of its functioning, i.e. accumulate appropriate knowledge regarding possibilities of solving the problems of definite (selected) class; 5) an active use of artificial intelligence data processing techniques, helping to gradually transform the DSS into intelligent

one; 6) – the organization aspects and techniques for computing procedures should provide for appropriate rate of computing that corresponds to the human requirements with regard to the rate of alternatives generating and reaching the final result; 7) – precision (quality) of computing should satisfy preliminary established requirements by a user and developer; 8) – intermediate and final results of computations should be controlled with appropriate sets of analytic quality criteria, what will allow to enhance significantly quality and reliability of the final result (decision alternatives); 9) – DSS should generate all necessary for a user formats and types of intermediate and final results representations with taking into consideration the users of various levels; 10) – the system should contain the means for exchanging with data and knowledge with other information processing systems via local and/or global computer nets; 11) – to make the system functionality complete and flexible DSS should be easily expandable with new functions regarding data processing, results representation and control with appropriate statistical criteria, model constructing and alternatives generating.

Satisfaction of all the requirements mentioned above provides a possibility for effective practical application of the system developed and enhancing general behavioral effect of the DSS as a whole for a specific company or an enterprise within long periods of time [Hollsapple, 1996].

Basic mathematical tools for DSS

All mathematical methods and techniques that are hired for development and implementation of DSS could be divided in the two following groups: 1 – general purpose methods that provide for implementation of system functions; and 2 – special purpose methods and techniques that are necessary for solving specific problems regarding preliminary and basic data processing, model constructing, alternatives generating, selecting the best alternative for further implementation and forecasting of the implementation consequences.

The group of the general purpose methods includes the following ones: – data and knowledge collecting and editing procedures; – preliminary data processing techniques such as digital filtering, normalization, imputation of missing values, detecting special effects (such as regime switching, seasonal effects, spikes, nonstationarity etc); – the methods for accumulating information regarding previous applications of DSS to problem solving for the retrospective analysis and repetitive use; – computer graphics techniques; – techniques for syntactic analysis to be used in a command interpreter (language system of DSS)); – methods for setting up necessary communications with other information processing

systems via local and global nets; – logical rules to control the system functioning. The set of the methods mentioned could be modified or expanded depending on a specific practical application.

Selection of the application defined mathematical methods for a DSS depends on the specific system application area, possible specific problem statements regarding data processing, model building, processes forecasting, and alternatives generation. However, it is possible to state that in most cases of DSS development it is necessary to use the following mathematical methods: – methods and methodologies for mathematical (statistical and probabilistic) modeling using statistical/experimental data; – risk estimating and forecasting techniques on the basis of the models constructed with possibilities for combining the forecasts computed with different techniques; – operations research optimization techniques and dynamic optimization (optimal control) methods; – the methods for forecasting/foresight of decision implementation consequences; – the sets of special analytic (statistical) criteria to control the processes of computations performed at each stage of data processing, model constructing and alternatives generation aiming to reach high quality of a final result.

All the methods and methodologies mentioned are described with necessary completeness in special modern literature. For example, time series modeling and forecasting are presented in many references, more particularly in [Tsay, 2010], [Bidyuk, Menyailenko, 2008], and financial risks modeling, evaluation and management is considered in a vast literature, say in [MCNeil, 2005], [Basel Committee on Banking Supervision, 2006], [Mays, 2001], [Neil, 2005], [Shakhov, 2002], [Bidyuk, Matros, 2008], [Jong, 2008]. The task for a DSS developer is in appropriate selection of model classes, modeling and optimization techniques, quality criteria as well as relevant methodologies for appropriate organization of all computational procedures.

Mathematical models used in DSS

Generalized linear models (GLM). GLM can be considered as further extension of multiple linear regression (MLR) model. It is distinguished from MLR with the following features: – distribution of dependent variable can be non-Gaussian and not necessarily continuous, say binomial; – predicted values of dependent variable are computed as linear combination of predictors that are linked to dependent variable via selected link function. Generally, GLM create a class of statistical models that includes linear regression, variance analysis relations, nonlinear models like logit and probit, Poisson regression and some others [Jong, 2008]. In a general linear model independent variable is supposed to

be normally distributed and the link function is called identity function, i.e. linear combination of independent variables is not subjected to any transform. Thus, GLM is a model of the following type:

$$y = g^{-1}\left(\sum_{i=1}^{m} \mathbf{b}_i \, g_i(x) \right),$$

where m is a number of independent (explaining) variables; $g(\cdot)$ is a link function. It is usually supposed that dependent variable y belongs to the class of exponential distributions. Thus, characteristics of GLM suppose the knowledge of dependent variable distribution, characteristics and parameters of the link function $g(\cdot)$, and of linear predictor $\mathbf{X}\,\mathbf{b}$, where $\mathbf{X}$ is a measurement matrix for independent variables; $\mathbf{b}$ is parameter vector. The class of exponential distributions includes the following distribution types: normal, gamma, and beta, and the discrete families: binomial, Poisson, and negative binomial. General representation of probability density functions (PDFs) for them is as follows:

$$f(x \mid \theta) = h(x)\, c(\theta) \exp\left(\sum_{i=1}^{k} w_i(\theta) l_i(x) \right),$$

where $h(x) \geq 0$ and $l_1(x),...,l_k(x)$ are real-valued functions of the observation x (they cannot depend on θ); $c(\theta) \geq 0$ and $w_1(x),...,w_k(x)$ are real-valued functions of the possibly vector-valued parameter θ (they cannot depend on x).

Nonlinear models logit and probit. To solve the problem of classifying credit borrowers into two groups it is quite logically to use appropriately transformed cumulative distribution function (CDF). CDF belongs to the class of monotonous functions that monotonously decrease or increase on some interval. Suppose that for determining probability of crediting a client p_c it is chosen a normal distribution:

$$p_c = \Phi(\mathbf{b}^T \mathbf{x}) = \int_{-\infty}^{u} \varphi(z)\, dz,$$

where $\varphi(z)$ is a density for standard normal distribution; $u = \mathbf{b}^T \mathbf{x}$ is upper integration limit. This way so called probit model is constructed.

If the probability for successful crediting is determined with logistic distribution function, then logit model is constructed. In this case we have:

$$p_c = \Phi(\mathbf{b}^T\mathbf{x}) = \int_{-\infty}^{u} \varphi(z)dz = \frac{1}{1+\exp(-\mathbf{b}^T\mathbf{x})},$$

or

$$p_c = \frac{\exp(b_1x_1 + ... + b_mx_m)}{1+\exp(b_1x_1 + ... + b_mx_m)}.$$

In contrast to the normal distribution logistic function has so called closed form that provides a possibility for simplified computations in comparison to probit. Parameter estimates for both models can be found with maximum likelihood technique. An alternative possibility is Markov chain Monte Carlo (MCMC) approach that is based on correct generation of pseudorandom sequences that satisfy certain conditions. Due to availability of multiple alternative techniques for generating pseudorandom sequences MCMC has found wide applications [Gilks, 2000]. The classification results achieved with logit and probit are usually acceptable in most cases of application.

Generation and implementation of alternatives with DSS

Decision making process includes rather sophisticated procedures that could be partially or completely iterative, i.e. executed repeatedly when the alternative found is not satisfactory for a decision making person (DMP). DSS could return automatically (or on DMP initiative) to the previous stages of available data and knowledge analysis.

The whole process of making and implementing decision could be considered as consisting of the stages given below.

1 – A thorough analysis of the decision problem using all available sources of information, collection of data and knowledge relevant to the problem. At this stage it is also important to consider and use former solutions to the problem if such are available. The information regarding former solutions of similar problem could be helpful for correct problem statement, to select appropriate techniques for data analysis, to speed up alternatives generation, and to decline the alternatives that turned out to be ineffective in the past.

2 – Selection of a class (classes) of mathematical models for the problem description, and analysis of the possibility for the use of available (previously developed) models. The models could belong to different classes as far as they can be formulated in continuous or discrete time, be linear or nonlinear, they could be developed according to the structural or functional approach etc. In some cases, it is necessary to construct complex simulative model that would include a set of simpler models of different classes.

3 – Development of new models for the problem (process, object, system) under study what includes structure and parameter estimation for candidate models using available data (and possibly expert estimates) and knowledge of various types. The alternative structures of candidate models provide a possibility for selecting the best one of them for generating alternative decisions (loss estimates, forecasts, probability of risk estimates, control actions etc.) on their bases.

4 – Analysis of the candidate models constructed and selecting of the best one of them with application of a set of statistical quality criteria and expert estimates. At this stage again more than one model could be selected for the further use as far as the best model (for a particular application) can be found only after application of the candidates for solving particular problem, i.e. after alternatives generating and estimating possible consequences of their implementation.

5 – Application of the model (models) selected for solving the problem of risk estimation and/or control (or management) problem (when necessary). If the forecasts or controls computed are not satisfactory we should return back to the stage one or stage three, and repeat the process of model constructing. At this stage another set of statistical quality criteria should be applied to the analysis of risk estimates, forecasts or controls.

6 – Generating of a set of alternatives with the use of the model (models) constructed and various admissible initial conditions and constraints on variables. In a case of controls generating the alternatives could be built with different optimality criteria, utility functions or other criteria.

7 – Analysis of the alternatives generated with the experts of an enterprise or a company, and final selection of the best one for practical implementation. In a case when no alternative is acceptable we should return back to the model constructing or alternative generating stages. New knowledge or data could be required for the next iteration of computing new decision alternatives.

8 – Planning of actions and estimation of financial, material and human resources that are necessary for implementation of the alternative selected. Determining of the time horizon (horizon of control) necessary for implementing the decision made.

9 – Implementation of the decision made: current monitoring of availability and spending the necessary resources, estimation of necessary time frames, registering and quality estimation of intermediate and final results.

10 – Application of possible analytic and expert quality criteria to estimation of final results.

11 – Analysis of the final results by the company experts, and final elucidation of advantages and disadvantages of the alternative implemented; analysis of the decision making and implementing process, and forming forecasts (foresights) for the future.

12 – Writing the final report on the tasks performed.

Processing uncertainties

As it was mentioned above all types of mathematical modeling usually need to cope with various kinds of uncertainties linked with data, structure of the process under study and its model, parameter uncertainty, and uncertainties relevant to the models and forecasts quality. In many cases a researcher has to cope with the following types of uncertainties: structural, statistical and parametric. Structural uncertainties are encountered in the cases when structure of the process under study (and respectively its model) is unknown or not clearly enough defined (known partially). For example, when the functional approach to model constructing is applied usually we do not know object (or a process) structure, it is estimated with appropriate model structure estimation techniques: correlation analysis, estimation of mutual probabilities, lags estimation, testing for nonlinearities and nonstationarity, identification of external disturbances type etc. The sequence of actions necessary for identification, processing and taking into consideration of uncertainties is given in Figure 1. All the tasks mentioned in the figure are solved successfully with appropriately designed and implemented DSS.

We consider uncertainties as the factors that influence negatively the whole process of mathematical model constructing, risk forecasts estimating and alternative decisions generating. They are inherent to the process due to incompleteness or inexactness of our knowledge regarding the objects (systems)

under study, incorrect selection or application of computational procedures etc. The uncertainties very often appear due to incompleteness of data, noisy measurements or they are invoked by stochastic external disturbances with unknown probability distribution, poor estimates of model structure or by a wrong selection of parameter estimation procedure. The problem of uncertainties identification is solved with special statistical tests and visual data studying.

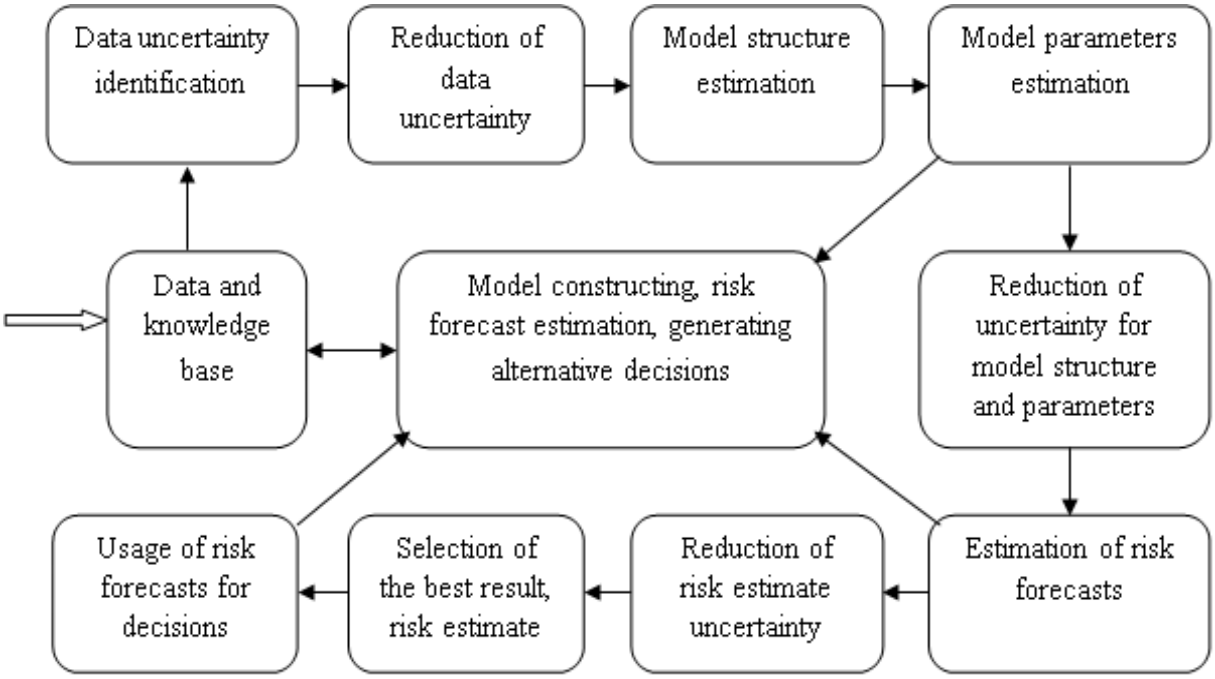

Figure 1. Sequence of actions directed towards identification of uncertainties, processing and taking them into consideration

As far as we usually work with stochastic data, application of statistical techniques mentioned provides a possibility for approximate estimation of an object (and its model) structure. To find “the best” model structure it is recommended to apply adaptive estimation schemes that provide automatic search in a wide range of model structure parameters (model order, time lags, and nonlinearities).

Usually the search is performed in the class of regression type models with the use of integrated criterion of the following type [Bidyuk, Menyailenko, 2008]:

$$V_N(\theta, D_N) = e^{|1-R^2|} + e^{|2-DW|} + \alpha \ln\left(1+\frac{SSE}{N}\right) + \beta\{\ln(1+MSE)+\ln(MAPE)\} \qquad (1)$$

where θ is a vector of model parameters; D_N data in the form of time series (N is a power of time series used); R^2 is a determination coefficient; DW is Durbin-Watson statistic; MSE is mean square error; MAPE is mean absolute percentage error for forecasts; α, β are adjustment coefficients.

There are several possibilities for adaptive model structure estimation: (1) automatic analysis of partial autocorrelation for determining autoregression order; (2) automatic estimation of the exogeneous variable lag (detection of leading indicators); (3) automatic analysis of residual properties; (4) analysis of data distribution type and its use for selecting correct model estimation method; (5) adaptive model parameter estimation with hiring extra data; (7) optimal selection of weighting coefficients for exponential smoothing, nearest neighbor and some other techniques; (6) the use of adaptive approach to model type (linear, nonlinear) selection. The use of a specific adaptation scheme depends on volume and quality of data, specific problem statement, requirements to forecast estimates, etc.

The adaptive estimation schemes also help to cope with the model parameters uncertainties. New data are used to compute model parameter estimates that correspond to possible changes in the object under study. In the cases when model can be nonlinear alternative parameter estimation techniques can be hired to compute alternative (though admissible) sets of parameters and to select the most suitable of them using statistical quality criteria.

Processing some types of stochastic uncertainties. While performing practical modeling very often we don't know statistical characteristics (covariance matrix) of stochastic external disturbances and measurement noise (errors). To eliminate this uncertainty optimal filtering algorithms are usually applied that provide for a possibility of simultaneous estimation of object (system) states and the covariance matrices. One of the possibilities to be hired is optimal Kalman filter. Kalman filter is used to find optimal estimates of system states on the bases of the system model represented in a convenient state space form as follows:

$$\mathbf{x}(k) = \mathbf{A}(k, k-1)\mathbf{x}(k-1) + \mathbf{B}(k, k-1)\mathbf{u}(k-1) + \mathbf{w}(k) \qquad (2)$$

where $\mathbf{x}(k)$ is n-dimensional vector of system states; $k = 0,1,2,...$ is discrete time; $\mathbf{u}(k-1)$ is m-dimensional vector of deterministic control variables; $\mathbf{w}(k)$ is n- dimensional vector of external random disturbances; $\mathbf{A}(k, k-1)$ is $(n \times n)$ -matrix of system dynamics; $\mathbf{B}(k, k-1)$ is $(n \times m)$ matrix of control coefficients. The double argument $(k, k-1)$ means that the variable or parameter is used at the moment k, but its value is based on the former (earlier) data processing including moment $(k-1)$. Usually the

matrices **A** and **B** are written with one argument like **A**(k), and **B**(k), to simplify the text. Obviously stationary system model is described with constant parameters like **A**, and **B**. As far as matrix **A** is a link between two consequent system states, it is also called state transition matrix. Discrete time k and continuous time t are linked to each other via data sampling time T_s: $t = kT_s$. In the classic problem statement for optimal filtering the vector sequence of external disturbances $\mathbf{w}(k)$ is supposed to be zero mean white Gaussian noise with covariance matrix **Q**, i.e. the noise statistics are as follows:

$$\begin{aligned} &E[\mathbf{w}(k)] = 0, \quad \forall k, \\ &E[\mathbf{w}(k)\mathbf{w}^T(j)] = \mathbf{Q}(k)\,\delta_{kj}, \end{aligned} \tag{3}$$

where δ_{kj} is Kronecker delta-function: $\delta_{kj} = \begin{cases} 0, & k \neq j \\ 1, & k = j \end{cases}$; $\mathbf{Q}(k)$ is positively defined covariance $(n \times n)$ matrix. The diagonal elements of the matrix are variances for the components of disturbance vector $\mathbf{w}(k)$. Initial system state $\mathbf{x}_0$ is supposed to be known with the following statistics:

$$E[\mathbf{x}_0] = \bar{\mathbf{x}}_0; \quad E[\mathbf{x}_0\mathbf{x}_0^T] = \mathbf{M}; \quad E[\mathbf{w}(k)\,\mathbf{x}_0^T] = 0, \ \forall k$$

The measurement equation for vector $\mathbf{z}(k)$ of output variables is described by the equation:

$$\mathbf{z}(k) = \mathbf{H}(k)\mathbf{x}(k) + \mathbf{v}(k) \tag{4}$$

where **H**(k) is $(r \times n)$ observation (coefficients) matrix; **v**(k) is r-dimensional vector of measurement noise with statistics:

$$E[\mathbf{v}(k)] = 0, \quad E[\mathbf{v}(k)\,\mathbf{v}^T(j)] = \mathbf{R}(k)\,\delta_{kj}, \tag{5}$$

where $R(k)$ is $(r \times r)$ positively defined measurement noise covariance matrix, the diagonal elements of which represent variances of additive noise for each measurable variable. The noise of

measurements is also supposed to be zero mean white noise sequence that is not correlated with external disturbance $\mathbf{w}(k)$ and initial system state:

$$E[\mathbf{v}(k)\mathbf{w}^T(j)] = 0, \quad \forall\ k, j;$$
$$E[\mathbf{v}(k)\mathbf{x}_0^T] = 0, \quad \forall\ k. \tag{6}$$

For the system (2) – (6) with state vector $\mathbf{x}(k)$ it is necessary to find optimal state estimate $\hat{\mathbf{x}}(k)$ at arbitrary moment k as a linear combination of estimate $\hat{\mathbf{x}}(k-1)$ at the previous moment $(k-1)$ and the last measurement available, $\mathbf{z}(k)$. The estimate of state vector $\hat{\mathbf{x}}(k)$ is computed as optimal one with minimizing the expectation of the sum of squared errors, i.e.:

$$E[(\hat{\mathbf{x}}(k) - \mathbf{x}(k))^T(\hat{\mathbf{x}}(k) - \mathbf{x}(k))] = \min_K, \tag{7}$$

where $\mathbf{x}(k)$ is an exact value of state vector that can be found as deterministic part of the state equation (2); $\mathbf{K}$ is optimal matrix gain that is determined as a result of minimizing quadratic criterion (7).

Thus, the filter is constructed to compute optimal state vector $\hat{\mathbf{x}}(k)$ in conditions of influence of random external system disturbances and measurement noise. Here uncertainty arises when we don't know estimates of covariance matrices $\mathbf{Q}$ and $\mathbf{R}$ in (3) and (4), respectively. To solve the problem an adaptive Kalman filter is constructed that allows to find estimates of $\hat{\mathbf{Q}}$ and $\hat{\mathbf{R}}$ simultaneously with the state vector $\hat{\mathbf{x}}(k)$. Another choice is in constructing separate algorithm for computing the values of $\hat{\mathbf{Q}}$ and $\hat{\mathbf{R}}$.

Other appropriate instruments for fighting uncertainties are fuzzy logic, neuro-fuzzy models, Bayesian networks, appropriate types of distributions etc. Some of statistical data uncertainties such as missing measurements, extreme values and high level jumps of stochastic origin could be processed with appropriately selected statistical procedures. There exists a number of data imputation procedures that help to complete the sets of data collected. For example, very often missing measurements for time series could be generated with appropriately selected distributions or in the form of short term forecasts.

Appropriate processing of jumps and extreme values helps with adjusting data non-stationarity and to estimate correctly the probability distribution for the stochastic processes under study.

Processing data with missing observations (data in the form of time series). For the data in the time series form the most suitable imputation techniques are: – simple averaging when it is possible (when only a few values are missing); – generation of forecast estimates with the model constructed using available measurements; – generation of missing (lost) estimates from distributions the form and parameters of which are again determined using available part of data; – the use of optimization techniques, say appropriate forms of EM-algorithms (expectation maximization); – exponential smoothing etc. The simplest model that could be hired for generating forecasts is AR(1):

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k),$$

where a_0, a_1 are model parameters; random process $\varepsilon(k)$ takes into consideration model structure uncertainties (say lack of appropriate regressors, external random disturbances, errors of parameter computing etc.), and possible measurement errors. If parameters a_0, a_1 are known, we could compute one step-ahead prediction as conditional expectation of the form:

$$\hat{y}(k+1) = E_k[y(k+1)] = E_k[y(k+1)\,|\,(k), y(k-1), ..., \varepsilon(k), \varepsilon(k-1), ...] = a_0 + a_1 E_k[y(k)] = a_0 + a_1 y(k),$$

as far as $y(k)$ at the moment k takes known value. Iteratively it is possible to derive the forecasting function for s steps-ahead [Bidyuk, Menyailenko, 2008]:

$$\hat{y}(k+s) = E_S[y(k+s)] = a_0\left(\sum_{i=0}^{S-1} a_1{}^i\right) + a_1{}^S y(k) = a_0 \sum_{i=0}^{S-1} a_1{}^i + a_1{}^S y(k).$$

The sequence of forecast estimates $\{\hat{y}(k+i)\}$, $i = 1, ..., s$ is convergent if $|a_1| < 1$:

$$\lim_{s \to \infty} E_k[y(k+s)] = \frac{a_0}{1 - a_1}.$$

The last expression means that for stationary AR or ARMA processes the estimates of conditional forecasts asymptotically ($s \to \infty$) converge to unconditional mean (long-term forecast). It should also be mentioned here that optimal filter can also be used for missing data imputation because it contains "internal" forecasting function that provides a possibility for generating quality short-term forecasts.

Further reduction of the uncertainty is possible thanks to application of several forecasting techniques to the same problem with subsequent combining of separate forecasts using appropriate weighting coefficients. The best results of combining the forecasts is achieved when variances of forecasting errors for different forecasting techniques do not differ substantially.

Coping with uncertainties of model parameters estimates. Usually uncertainties of model parameter estimates such as bias and inconsistency result from low informative data, or data do not correspond to normal distribution, what is required in a case of application LS for parameter estimation. This situation may also take place in a case of regressors multicollinearity and substantial influence of process nonlinearity that for some reason has not been taken into account when model was constructed. When power of data sample is not satisfactory for model construction it could be expanded by applying special techniques, or simulation is hired, or special model building techniques, such as GMDH, are applied. Very often GMDH produces results of acceptable quality with short samples. If data do not correspond to normal distribution, then ML technique could be used or appropriate Monte Carlo procedures for Markov Chains (MCMC) [Bidyuk, 2012]. The last techniques could be applied with quite modest computational expenses when the number of parameters is not large.

Dealing with model structure uncertainties. When considering mathematical models, it is convenient to use proposed here a unified notion of a model structure which we define as follows:

$$S = \{r, p, m, n, d, w, l\},$$

where r` is model dimensionality (number of equations); p is model order (maximum order of differential or difference equation in a model); m is a number of independent variables in the right hand side of a model; n is a nonlinearity and its type; d is a lag or output reaction delay time; w is stochastic external disturbance and its type; l are possible restrictions for the variables and/or parameters. When using DSS, model structure should practically always be estimated using data. It means that elements of model structure accept almost always only approximate values. When a model

is constructed for forecasting we build several candidates and select the best of them with set model quality statistics.

Generally we could define the following techniques to fight structural uncertainties: gradual improvement of model order (AR(p) or ARMA(p, q)) applying adaptive approach to modeling and automatic search for the "best" structure using complex statistical quality criteria; adaptive estimation (improvement) of input delay time (lag) and data distribution type with its parameters; describing detected process nonlinearities with alternative analytical forms with subsequent estimation of model adequacy and forecast quality. An example of complex model and forecast criterion may look as follows:

$$J = \left|1 - R^2\right| + \alpha \ln\left[\sum_{k=1}^{N} e^2(k)\right] + \left|2 - DW\right| + \beta \ln(MAPE) + U \to \min_{\hat{\theta}_i},$$

where R^2 is determination coefficient; $\sum_{k=1}^{N} e^2(k) = \sum_{k=1}^{N} [y(k) - \hat{y}(k)]^2$ is sum of squared model errors; DW is Durbin-Watson statistic; $MAPE$ is mean absolute percentage error for one step-ahead forecasts; U is Theil coefficient that measures forecasting characteristic of a model; α, β are appropriately selected weighting coefficients; $\hat{\theta}_i$ is parameter vector for $i-th$ candidate model. A criterion of this type is used for automatic selection of the best candidate model. The criterion also allows operation of DSS in adaptive mode. Certainly, other forms of the complex criteria are possible. While constructing the criterion it is important not to overweigh separate members in right hand side.

Coping with uncertainties of a level (amplitude) type. The use of random (i.e. with random amplitude or a level) and/or non-measurable variables leads to necessity of hiring fuzzy sets for describing such situations. The variable with random amplitude can be described with some probability distribution if the measurements are available or they come for analysis in acceptable time span. However, some variables cannot be measured (registered) in principle, say amount of shadow capital that "disappears" every month in offshore, or amount of shadow salaries paid at some company, or a technology parameter that cannot be measures on-line due to absence of appropriate gauge. In such situations we could assign to the variable a set of possible values in linguistic form as follows: capital amount = {very low, low, medium, high, very high}. There is a necessary set of mathematical operations to be applied to such fuzzy variables. Finally, fuzzy value could be transformed into exact form using known techniques.

Processing probabilistic uncertainties. To fight probabilistic uncertainties, it is possible to use Bayesian approach that helps to construct models in the form of conditional distributions for the sets of random variables. Usually such models represent the process (under study) variables themselves, stochastic disturbances and measurement errors or noise. Certainly the problem of distribution type identification also arises in regression modeling. Each probability distribution is characterized by a set of specific values that random variable could take and the probabilities for these values. The problem is in the distribution type identification and estimating its parameters. The probabilistic uncertainty (will some event happen or not) could be solved with various models of Bayesian type. This approach is known as Bayesian programming or paradigm. The generalized structure of the Bayesian program includes the following steps: (1) problem description and statement with putting the question regarding estimation of conditional probability in the form: $p(X_i \mid D, Kn)$, where X_i – is the main (goal) variable or event; the probability p should be found as a result of application of some probabilistic inference procedure; (2) statistical (experimental) data D and knowledge Kn are to be used for estimating model and parameters of specific type; (3) selected and applied probabilistic inference technique should give an answer to the question put above; (4) analysis of quality of the final result. The steps given above are to some extent “standard” regarding model constructing and computing probabilistic inference using statistical data available. This sequence of actions is naturally consistent with the methods of cyclic structural and parametric model adaptation to the new data and operating modes (and possibly expert estimates).

One of the most popular Bayesian approaches today is created by the models in the form of static and dynamic Bayesian networks (BN). Bayesian networks are probabilistic and statistical models represented in the form of directed acyclic graphs (DAG) with vertices as variables of an object (system) under study, and the arcs showing existing causal relations between the variables. Each variable of BN is characterized with complete finite set of mutually excluding states. Formally BN could be represented with the four following components: $\mathbf{N}=<\mathbf{V}, \mathbf{G}, \mathbf{P}, \mathbf{T}>$, where $\mathbf{V}$ stands for the set of model variables; $\mathbf{G}$ represents directed acyclic graph; $\mathbf{P}$ is joint distribution of probabilities for the graph variables (vertices), $\mathbf{V}=\{X_1,...,X_n\}$; and $\mathbf{T}$ denotes conditional and unconditional probability tables for the graphical model variables. The relations between the variables are established via expert estimates or applying special statistical and probabilistic tests to statistical data (when available) characterizing dynamics of the variables.

The process of constructing BN is generally the same as for models of other types, say regression models. For example, as model parameters for BN are unconditional and conditional probabilities for specific values of variables, that are stored in respective tables. For parent variables these are unconditional probabilities and for daughter variables – conditional probability tables (CPT). Unconditional and conditional probabilities are determined by experts (in simpler cases), and by special computational procedures when appropriate sets of statistical (or experimental) data are available. Thus to each node of DAG is assigned CPT that is used for computing probabilistic inference over the BN [Jensen, 2007], [Zgurovsky, 2015].

The set of the model variables should satisfy the Markov condition that each variable of the network does not depend on all other variables but for the variable's parents. In the process of BN constructing first the problem is solved of computing mutual information values between all variables of the net. Then an optimal BN structure is searched using acceptable quality criterion, say well-known minimum description length (MDL) that allows analyzing and improving the graph (model) structure at each iteration of the learning algorithm applied. Bayesian networks provide the following advantages for modeling: the model may include qualitative and quantitative variables simultaneously as well as discrete and continuous ones; number of the variables could be very large (thousands); the values for conditional probability tables could be computed with the use of statistical data and expert estimates; the methodology of BN constructing is directed towards identification of actual causal relations between the variables hired what results in high adequacy of the model; the model is also operable in conditions of missing data.

The process of constructing the model in the form of BN could be represented with the following steps: 1) a thorough analysis of the process (object) under study aiming to detecting of its special functioning features and identification of parent and daughter variables; 2) search and analysis of existing process models and determining the possibility of their usage in DSS; 3) determining degree of relations between the process variables using special tests and expert estimates; 4) reduction of the process dimensionality whenever this is possible; 5) scaling and discretization of the data available when necessary; 6) determining semantic restrictions on the future model; 7) estimation of candidate model (directed acyclic graphs) structures using appropriate optimization procedures and score functions; 8) candidate models analysis and selection of the best one using model quality criteria (including values of score functions); 9) application of the model(s) constructed to solve the problem stated; 10) computing inference with the model(s) constructed with regards to the variables selected, quality analysis of the result. In our case the final result of the model application is computing of client default probability with

the conditions established by other model variables. According to alternative problem statement BM could be constructed for estimation of operational or other type of financial risks.

The model of this type could be integrated with other model types such as regression, neural networks, the models constructed using soft computing paradigm, fuzzy sets etc. To perform computing probabilistic inference for BN today there exists rather wide set of techniques the selection of which depends on specific problem statement and requirements to accuracy of the final result. Thus, it is possible to state that BN is a powerful probabilistic and statistical instrument for modeling processes (systems) of arbitrary nature in conditions of availability of various uncertainties types of statistical and structural nature.

To reduce an influence of probabilistic and statistical uncertainties on models quality and the forecasts based upon them it is also possible to use the models in the form of Bayesian regression based on analysis of actual distributions of model variables and parameters. Consider a simple two variables regression

$$y(k) \mid x(k) = \beta_1 + \beta_2\, x(k) + u(k), \quad k = 0,1,..., n\,.$$

It is supposed that of random values $u_1,...,u_n$ are independent and could belong, for example, to normal distribution, $\{u(k)\} \sim N(0,\sigma_u^2)$; here vector of unknown parameters includes three elements, $\theta = (\beta_1, \beta_2, \sigma_u^2)^T$. The likelihood function for dependent variable $\mathbf{y} = (y_1,..., y_n)^T$ and predictor $\mathbf{x} = (x_1,..., x_n)^T$ without proportion coefficient is determined as follows:

$$L(\mathbf{y} \mid \mathbf{x}, \beta_1, \beta_2, \sigma_u) = \frac{1}{\sigma_u^N} \exp\left\{ -\frac{1}{2\sigma_u^2} \sum_{k=1}^{N} [y(k) - \beta_1 - \beta_2 x(k)]^2 \right\}.$$

Using simplified (non-informative) distributions for the model parameters:

$$g(\beta_1, \beta_2, \sigma_u) = g_1(\beta_1)\, g_2(\beta_2)\, g_3(\sigma_u),$$
$$g_1(\beta_1) \propto const,$$
$$g_2(\beta_{21}) \propto const,$$
$$g_3(\sigma_u) \propto 1/\sigma_u$$

and Bayes theorem it is possible to find joint posterior distribution for the parameters in the form [Bernardo, 2000]:

$$h(\beta_1,\beta_2,\sigma_u \mid x,y) \propto \frac{1}{\sigma}\frac{1}{\sigma^N}\exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{N}\left(y(k)-\beta_1-\beta_2 x(k)\right)^2\right], -\infty < \beta_1,\beta_2 < +\infty, \quad 0<\sigma_u<\infty.$$

Maximum likelihood estimates for the model parameters are determined as follows:

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}; \hat{\beta}_2 = \frac{\sum_{k=1}^{N}[x(k)-\bar{x}][y(k)-\bar{y}]}{\sum_{k=1}^{N}[x(k)-\bar{x}]\sum_{k=1}^{N}[y(k)-\bar{y}]},$$

where $\bar{x}=N^{-1}\sum_{k=1}^{N}x(k), \quad \bar{y}=N^{-1}\sum_{k=1}^{N}y(k)$, with unbiased sample estimate of variance:

$$\hat{\sigma}_u^2 = s^2 = \frac{1}{N-2}\sum_{k=1}^{N}[y(k) - \hat{\beta}_1 - \hat{\beta}_2 x(k)].$$

Joint posterior density for the model parameters corresponds to two dimensional Student distribution:

$$h_1(\beta_1,\beta_2 \mid \mathbf{y},\mathbf{x}) \propto \left\{(N-2)s^2 + N(\beta_1-\hat{\beta}_1)^2 + \right.$$
$$\left. + (\beta_2-\hat{\beta}_2)^2\sum_{k=1}^{N}x(k)^2 + 2(\beta_1-\hat{\beta}_1)(\beta_2-\hat{\beta}_2)\sum_{k=1}^{N}x(k)\right\}^{-0.5N}$$

This way we get a possibility for using more exact distributions of models variables and parameters what helps to enhance model quality. Using new observation x^* and prior information regarding particular model it is possible to determine the forecast interval for the dependent variable, y^*:

$$p(y^* \mid x^*) = \iiint L(y^* \mid x^*,\beta_1,\beta_2,\sigma)h(\beta_1,\beta_2,\sigma) \mid \mathbf{x},\mathbf{y})d\beta_1,d\beta_2,d\sigma.$$

Another useful Bayesian approach is in hierarchical modeling that is based on a set of simple conditional distributions comprising one model. The approach is naturally combined with the theory of computing Bayesian probabilistic inference using modern computational procedures [Bolstad, 2010]. The hierarchical models belong to the class of marginal models where the final result is provided in the form of a distribution $P(\mathbf{y})$, where $\mathbf{y}$ is available data vector. The models are formed from the sequence of conditional distributions for selected variables including the hidden ones. The hierarchical representation of parameters supposes usually supposes that data, $\mathbf{y}$, is situated at the lower (first) level, model parameters (second level) $\theta = (\theta_i,\ i=1, 2,..., n)$, $\theta_i \sim N(\mu, \tau^2)$, determine distributions of dependent variables $y_i \sim N(\theta_i, \sigma^2)$, $i=1,2,..., n$, and parameters $\{\theta_i\}$ are determined by the pair, (μ, τ^2), of the third level. Supposing the parameters σ^2 and τ^2 accept known finite values, and parameter μ is unknown with the prior π_μ, then joint prior density for (θ, μ) could be presented in the form: $\pi_\mu(\mu)\prod_i \pi_\theta(\theta_i \mid \mu)$, and the prior for parameter vector θ will be defined by the integral: $p(\theta) = \int \pi_\mu(\mu) \prod_i \pi_\theta(\theta_i \mid \mu)\, d\mu$.

Architecture and functional layout of DSS for estimation of financial risks

DSS architecture is a generalized large-scale representation of basic system elements with links between them. The architecture gives a notion for the general purpose of system constructing and its basic functions (Figure 2).

DSS functionality is controlled by user commands, correctness of which is monitored by the command interpreter which constitutes a part of the user interface. The user commands are implemented by the central control unit that coordinates functioning of all system elements.

Specific commands and actions are as follows: expanding and modification of bases available in the system; initiation and starting of data and knowledge processing procedures; model constructing, risks and forecasts estimation, alternatives generating; viewing intermediate and final results of computing; retrospective analysis of previous results of decision making; comparing of current results with the previous ones.

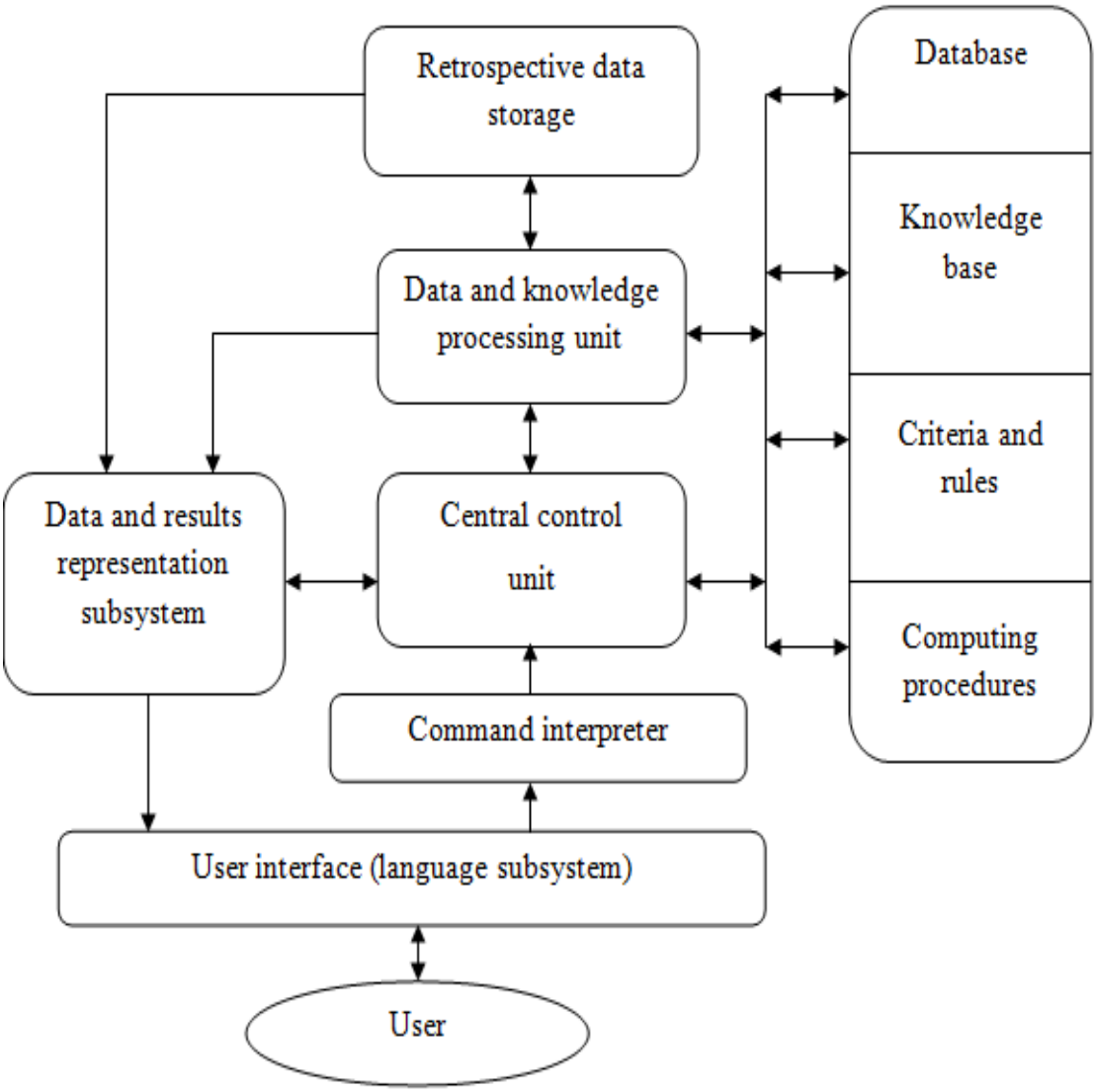

Figure 2. DSS architecture for estimation of financial risks

Specific commands and actions are as follows: expanding and modification of bases available in the system; initiation and starting of data and knowledge processing procedures; model constructing, risks and forecasts estimation, alternatives generating; viewing intermediate and final results of computing; retrospective analysis of previous results of decision making; comparing of current results with the previous ones.

The functional layout of the DSS is shown in Figure 3. It shows some details regarding the set of computational functions implemented in the system. It can be seen from the Figure 2 that the system has three levels of hierarchy: preliminary data processing, model structure and parameters estimation, and estimation (forecasting) of risk for a specific problem statement (say, market, credit or operational risk).

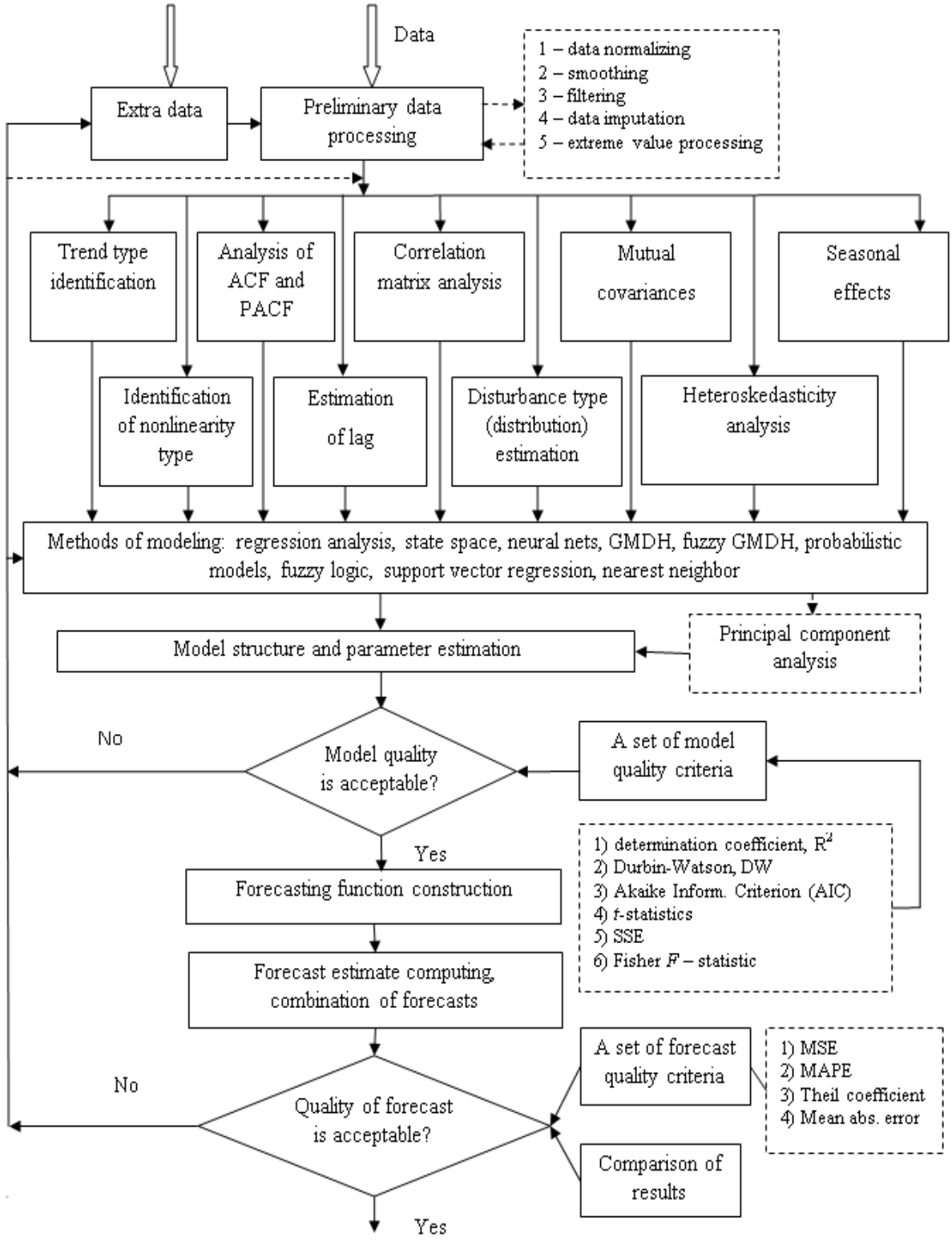

Figure 3. Functional layout of the DSS

Data, model and forecasts quality criteria

To achieve reliable high quality final result of risk estimation and forecasting at each stage of computational hierarchy separate sets of statistical quality criteria have been used. Data quality control is performed with the following criteria:

– database analysis for missing values using developed logical rules, and imputation of missed values with appropriately selected techniques;

– analysis of data for availability of outliers with special statistical tests, and processing of outliers to reduce their negative influence on statistical properties of the data available;

– normalizing of data in a case of necessity;

– application of low-order digital filters (usually that's low-pass filters) for separation of observations from measurement noise;

application of optimal (usually Kalman) filters for optimal state estimation and fighting stochastic uncertainties;

– application of principal component method to achieve desirable level of orthogonalization between the variables selected;

– computing of extra indicators for the use in regression and other models (say, moving average processes based upon measurements of dependent variables).

It is also useful to test how informative is the data collected. Very formal indicator for the data being informative is its sample variance. It is considered formally that the higher is the variance the richer is the data with information. Another criterion is based on computing derivatives with a polynomial that describes data in the form of a time series. For example, the polynomial given below may describe rather complex process trend:

$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + c_1 k + c_2 k^2 + ... + c_m k^m + \varepsilon(k), \qquad (9)$$

where $y(k)$ is basic dependent variable; a_i, c_i are model parameters; $k = 0,1,2,...$ is discrete time; $\varepsilon(k)$ is a random process that integrates the influence of external disturbances to the process being modeled as well as model structure and parameters errors. Autoregressive part of model (1) describes

the deviations that are imposed on a trend, and the trend itself is described with the m-th order polynomial of discrete time k. In this case maximum number of derivatives could be m, though in practice actual number of derivatives is defined by the largest number i of parameter c_i, that is statistically significant. To select the best model constructed the following statistical criteria are used: determination coefficient (R^2); Durbin-Watson statistic (DW); Fisher F-statistic; Akaike information criterion (AIC), and residual sum of squares (SSE). The forecasts quality is estimated with hiring the following criteria: mean squared error (MSE); mean absolute percentage error (MAPE); and Theil inequality coefficient (U). To perform automatic model selection, the above mentioned combined criterion (1) could be hired. The power of the criterion was tested experimentally and proved with a wide set of models and statistical data. Thus, the three sets of quality criteria are used to insure high quality of final result.

One of types of financial risks that are analyzed by the DSS is the credit risk. To analyze the quality of credit borrower's classification, model the following quality criteria were used: common accuracy, errors of type I and type II, ROC-curve, and Gini index. Common accuracy is computed as follows [Mays, 2001]:

$$CA = \frac{Correct\ Forecast}{N},$$

where $Correct\ Forecast$ is a number of correctly forecasted (classified) cases; N is a general number of cases (clients) considered. To some extent this criteria is subjective because it depends on a number of defaults as well as on the cut-off threshold value. The ROC-curve (Receiver Operation Characteristic) shows relation between the number of correctly classified positive cases (positives) and the number of incorrectly classified negative cases (negatives). The first ones are called true positive set, and the second one – negative set (specificity). Obviously, the cut-off threshold value also influences the errors of type I and type II. Among other criteria are the following: True Positives Rate (TPR), False Positives Rate (FPR), sensitivity (Se), specificity (Sp), and Gini index. The last one is determined by the area under ROC-curve [Bidyuk, Matros, 2008]. Table 1 shows relation between area under curve (AUC) and Gini index.

Table 1. Relation between AUC and Gini indexes

AUC interval	Gini index	Model quality
0,9 – 1,0	0,8 – 1,0	Excellent
0,8 – 0,9	0,6 – 0,8	Very high
0,7 – 0,8	0,4 – 0,6	Acceptable
0,6 – 0,7	0,2 – 0,4	Medium
0,5 – 0,6	0 – 0,2	Unacceptable

The ROC-curve can also be used to find optimum cut-off value as a compromise between sensitivity and specificity of a model. The following criteria can be used for cut-off value selection: 1 – the requirement of minimum sensitivity, Se , (or specificity, Sp); 2 – the requirement of maximum total sensitivity and specificity of a model, $cutoff = \max_k (Se_k + Sp_k)$, where $k = 1,2,3,...$ is a number of a client; the requirement of a balance between sensitivity and specificity, i.e. when $Se \approx Sp$: $cutoff = \min_k |Se_k - Sp_k|$.

Example of DSS application

In this example we consider the problem of credit borrower's classification or estimation of their solvency. The following characteristics of clients were used: solvency, age, credit sum, type of currency (local or USD), term of crediting, term of residing in defined area, marital status, number of children, type of credit, and gender. The uncertainty met in this example was in the form of incomplete records. Thus, before using the data all the gaps were appropriately filled in. We used the database consisting of 5000 records that was divided into learning sample (4000 records), and test sample (1000 records). The probabilities of clients' default were computed and compared to the actual data. Also errors of the first and second type were computed using different cut-off values. It was established for Bayesian network that maximum model accuracy reached was 0.764 with the cut-off value 0.3. The Bayesian network is "inclined to over insurance", i.e. it rejects more often the clients who could return the credit. The model accuracy and the errors of type I and type II depend on the cut-off value. The cut-off value determines the lowest probability limit for client's solvency, i.e. below this limit a client is considered as such that he

will not return the credit. Or the cut-off value determines the lowest probability limit for client's default, i.e. below this limit a client is considered as such that he will return the credit. As far as the cut-off value 0.1 or 0.2 is considered as not important, in practice it is reasonable to set the cut-off value at the level of about 0.25 – 0.30. After a series of computational experiments, we stopped at the cut-off value of 0.3.

Statistical criteria characterizing quality of the models constructed are given in Table 2. Using this table, it is easy to compare the techniques used in this example.

It follows from the Table 1 that the best models for estimation of credit return probability turned out to be discriminant analysis, logistic regression and Bayesian network. The best common accuracy was shown by discriminant analysis (0.837), and logistic regression (0.798), though Bayesian network showed higher Gini index than logistic regression (0.689). The decision tree used is characterized by Gini index of about 0.583, and CA = 0.763. It should be stressed that acceptable values of Gini index for developing countries like Ukraine are in the range 0.4 – 0.6. The Bayesian network and nonlinear regression constructed showed rather high values for Gini index that are acceptable for the Ukrainian economy in transition.

Table 2. Quality of the models constructed

Model type	Gini index	AUC	Common accuracy	Model quality
Discriminant analysis	0.723	0.891	0.837	Very high
Bayesian network	0.689	0.845	0.764	Very high
Logistic regression	0.678	0.847	0.798	Very high
Decision tree (Chaid)	0.583	0.791	0.763	Acceptable
Linear regression	0.386	0.647	0.616	Unacceptable

The results of computing experiments lead to the conclusion that today scoring models and Bayesian networks are the best instruments for banking system due to the fact that BN provide a possibility for detecting "bad" clients and to reduce substantially the financial risks caused by the clients. It also should be stressed that DSS constructed is very useful instrument for a decision maker that helps to perform quality processing of statistical data using different techniques, generate alternatives and to select the

best one with a set of appropriate criteria. The system performs tracking of the whole computational process using separate sets of statistical quality criteria at each stage of decision making: quality of data, models and forecasts (or risk estimates).

Conclusions

The general methodology was proposed for constructing DSS for mathematical modeling and forecasting of economic and financial processes, and financial risk estimation that is based on the following system analysis principles: hierarchical system structure, taking into consideration of probabilistic and statistical uncertainties, availability of adaptation features, generating of multiple decision alternatives, and tracking of computational processes at all the stages of data processing and model constructing with appropriate sets of statistical quality criteria. As instrumentation for fighting possible uncertainties the following techniques are used: Kalman filter, missing data imputation techniques, multiple methods for model parameter estimation, and Bayesian programming approach.

The system proposed has a modular architecture that provides a possibility for easy extension of its functional possibilities with new parameter estimation techniques, forecasting methods, financial risk estimation procedures, and alternatives generation. High quality of the final result is achieved thanks to appropriate tracking of the computational processes at all data processing stages: preliminary data processing, model structure and parameter estimation, computing of short- and middle-term forecasts, and estimation of risk variables (parameters) as well as thanks to convenient for a user intermediate and final results representation. The system is based on the ideologically different techniques of modeling and risk forecasting what creates a convenient basis for combination of various approaches to achieve the best results. The examples of the system application show that it could be used successfully for solving practical problems of financial risk estimation. The results of computing experiments lead to the conclusion that today scoring models such as nonlinear regression, Bayesian networks and the models resulted from application of discriminant analysis are the best instruments for banking system due to the fact that they provide a possibility for detecting “bad” clients and to reduce financial risks caused by the clients. It also should be stressed that DSS constructed turned out to be very useful instrument for a decision maker that helps to perform quality processing of statistical data using different techniques, generate alternatives and to select the best one with a set of appropriate quality criteria. The system performs tracking of the whole computational process using separate sets of statistical quality criteria at each stage of decision making: quality of data, models and forecasts or risk estimates.

The DSS proposed could be used for support of decision making in various areas of human activities including strategy development for banking system and industrial enterprises, investment companies etc. Further extension of the system functions is planned with new forecasting techniques based on probabilistic techniques and fuzzy sets.

Acknowledgment

The research has been conducted in the frames of the grant of Ukrainian Ministry for Education № 2813p.: "Development of systemic methodology for modeling and estimating of financial risks (2015 – 2016)".

Bibliography

[MCNeil, 2005] A. McNeil, R. Frey and P. Embrechts. Quantitative risk management, Princeton: Princeton University Press, 2005.

[Basel Committee on Banking Supervision, 2006] International Convergence of Capital Measurement and Capital Standarts. A Revised Framework, Comprehensive Version, Basel Committee on Banking Supervision, Bank for International Settlements, http://www.bis.org/publ/bcbs128b.pdf (Accessed June 2006).

[Mays, 2001] E. Mays. Handbook of Credit Scoring, Chicago: Glenlake Publishing Company, Ltd., 2001.

[Neil, 2005] M. Neil, N.E.Fenton, M. Tailor. Using Bayesian networks to model expected and unexpected operational losses, Risk Analysis, 2005.

[Shakhov, 2002] V.V. Shakhov, V.G. Medvedev, A.S. Millerman. Theory and Management of Insurance Risks, Moscow: Finances and Statistics, 2002.

[Burstein, 2008] F. Burstein, C.W. Holsapple. Handbook of Decision Support Systems, Berlin: Springer-Verlag, 2008.

[Hollsapple, 1996] C.W. Hollsapple, A.B. Winston. Decision Support Systems, Saint Paul: West Publishing Company, 1996.

[Bidyuk, 2012] P.I. Bidyuk, O.P. Gozhyj, L.O. Korshevnyuk. Development of Decision Support Systems, Mykolaiv: Black Sea State University named by Petro Mogyla, 2012.

[Tsay, 2010] R.S. Tsay. Analysis of Financial Time Series, Hoboken: Wiley & Sons, Inc., 2010.

[Bidyuk, Menyailenko, 2008] P.I. Bidyuk, O.S. Menyailenko, O.V. Polovcev. Methods of Forecasting, Lugansk: Alma Mater, 2008.

[Bidyuk, Matros, 2008] P.I. Bidyuk, O.Y. Matros. Models for credit risk estimation, Cybernetics and Computations (Kyiv), vol. 153, 2008.

[Jong, 2008] P. De Jong, G.Z. Heller. Generalized Linear Models for Insurance Data, New York: Cambridge University Press, 2008.

[Gilks, 2000] W.R. Gilks, S. Richardson, D.J. Spiegelhalter. Markov Chain Monte Carlo in Practice, New York: Chapman & Hall/CRC, 2000.

[Jensen, 2007] F.V. Jensen, Th.D. Nielsen. Bayesian Networks and Decision Graphs, New York: Springer, 2007.

[Zgurovsky, 2015] M.Z. Zgurovsky, P.I. Bidyuk, O.M. Terentyev, T.I. Prosyankina-Zharova. Bayesian Networks in Decision Support Systems, Kyiv: Edelwais, 2015.

[Bernardo, 2000] J.M. Bernardo, A.F.M. Smith. Bayesian theory, New York: John Wiley & Sons, Ltd., 2000.

[Bolstad, 2010] W.M. Bolstad. Understanding computational Bayesian statistics, Hoboken (New Jersey): John Wiley & Sons, Ltd, 2010.

[Chui, 2009] C.K. Chui and G. Chen. Kalman filtering with real-time application, Berlin: Springer-Verlag, 2009.

Authors' Information

Petro Bidyuk *– Dr. of Eng Sci., professor at the Institute for Applied System Analysis, NTUU "KPI", Peremohy avenue, 37, Kyiv - 03056, Ukraine; e-mail: pbidyuke@gmail.com*

Major Fields of Scientific Research: Mathematical modeling of processes of various nature, Statistical data analysis, Adaptive forecasting, Automatic control of industrial processes, Technical systems.

Svitlana Trukhan *– PhD student; Institute for Applied System Analysis, NTUU "KPI", Peremohy avenue, 37, Kyiv - 03056, Ukraine; e-mail: svetlana.trukhan@gmail.com*

Major Fields of Scientific Research: Mathematical methods of system analysis, Applied statistics, Time series analysis, Software engineering.

ON THE OPEN TEXT SUMMARIZER

Filip Andonov, Velina Slavova, Georgi Petrov

Abstract*: This paper presents proposed improvements to the Open Text Summariser (OTS) based on a heuristic approach. It describes the ten steps of the implemented algorithm and outlines further ideas for its development. The authors discuss valuable feedback gained from four experts evaluating the summaries generated by the OTS And the improved OTS and propose paths for future improvements. Then the results of the experts' evaluation of the two summarizers are presented and analyzed. Problems affecting the precision of both summarizing algorithms are discussed.*

Keywords: *Summary, extraction-based method, heuristics*

ACM Classification Keywords*: 1.2.7. Natural language processing, 1G.3 Correlation and regression analysis,*

Introduction

The Open Text Summarizer (OTS) is an implementation of a grammar-agnostic method for creating a summary of a text. It is a simple yet powerful method. The idea behind it is good enough to make it compete in terms of quality of results with much more complicated methods using advanced techniques. [Open Text Summarizer, 2016] Still the fact that it is independent of the language of the text makes some space for improvements by adding heuristics without compromising its language independence (much). The main approaches to analyzing text are abstraction-based and extraction-based. The former analyzes the text and rephrases it by omitting details. This is what humans do when solving text summarization tasks. The latter identifies key sentences and selects them for inclusion in the summary, trying to keep the text coherent. A famous algorithm that uses this approach is Google's TextRank. Both OTS and our proposed algorithms are extraction-based.

Description of the proposed algorithm

We present a summary-generating algorithm based on OTS. Our aim is to improve the characteristics of the OTS, namely the quality of the generated summary and of the extracted keywords. At the same time we have tried to not complicate the algorithm with procedures that use "hard" NLP such as syntactic analysis, but by using some heuristics in order to include in the summary more semantically important

parts of the text. We propose a different way for identifying word-forms representative for the given text (and representative sentences based on that) by comparing the word frequency in the text to the word frequency in a large corpora.

The algorithm we propose makes use of the following additional to the text information:

stem of nouns, verbs, adjectives; abbreviations list for sentence segmentation; a stop-word list; a list of personal pronouns; a list of linking words and a list of substitutions for uniforming non-alphabetic characters.

Our algorithm consists of 10 main steps.

1. First, we make sentence segmentation. 2. Then we do string-replacements in two stages: a) We replace all non-traditional quotation marks with their most commonly used variant; b) we replace all short forms such as they've, doesn't, weren't, etc. with their full forms they have, does not, were not, etc. This is needed as the next step, 3., removes all the punctuation symbols and if the short forms are left unmodified, the part after the apostrophe will be left as a separate word and will not be recognized when compared against a language dictionary. 4. We stem all the word-forms. This is a common practice in many natural language processing tasks and is needed here because we are interested in the meaning, not the grammatical form of the word, as discussed in [Andonov et al. 2016]. 5. Stop-words are removed. These are short words such as conjunctions, pronouns, etc. that are frequent, but whose occurrence in the text does not give us information about its meaning. 6. After these preliminary steps, the score of every stemmed word in the text is calculated. The calculations use the frequency of the basic form of the word in an English corpora [Heuven et al. 2014] and the frequency in the text.

The formula for calculating the score is the following:

$$s_v = max(\frac{h_v}{max(h)} - \frac{g_v^{1.5}}{max(g)}, 0) \quad (1)$$

A default value is assigned to the words-forms in the text that do not exist in the corpora - the median of the corpora.

Thus, we obtain a higher score for those words that are more frequent in the text than in the corpora. Negative values for the score cannot be used so we replace them with zero, as given in (1). In the previous paper [Andonov et al. 2016], the heuristic formula we proposed was based on division of the two members given in (1). Our more extensive experiments after the first publication showed that it does not give satisfying results for words that have low frequencies in the corpora. As a result, now we apply subtraction, which has given much better results.

7. Once we have obtained the scores of individual words, we calculate the score of the sentences. In order to track the rule of the usual application of the semantic focal point in writing, here we use a heuristic that sentences in the beginning of the paragraph are slightly more important than those in the middle. Thus we give bigger weight to the first sentence of the text and to all first sentences in paragraphs. After that a link score is calculated, based on a list of expressions that we have created. The idea is that linking words and phrases such as 'for example', 'in addition', 'for instance', 'in particular', 'in fact', etc., especially when placed at the beginning of a sentence, mean that the sentence containing them is referencing to something said in the previous one. Thus if a sentence starts or has near its beginning such an expression then this score is higher. The reason for being interested in such sentences is that they should not be separated from the ones they are linking to, otherwise the coherence of the summary will deteriorate.

Our algorithm assigns different weights to different parts of speech.

From a linguistic point of view the subject and predicate are the most important. Without sentence structure analysis however the best we can do is to identify the objects and characters the sentence talks about. For that reason we consider nouns more important and verbs and adjectives less so.

$$S_s = pb_s * \sum_{i=1}^{|s|}(h_i * w_i) \qquad (2)$$

where S_s is the score of sentence s, pb_s is the paragraph beginning score of s, $|s|$ is the number of words in the sentence s, h_i is the score of every word i in the sentence s and w_i is the part of speech weight of every word i in sentence s.

8. At this step we mark the sentences for the summary. All sentences are sorted by their score and then the first one third with the highest scores are marked for inclusion. We use 1/3 of the original text as the selected length of the summary as this is a common practice when humans summarize text. However, this value can be changed. At this step, are also marked the sentences with high link score, as this indicates a high probability that a consequent sentence is referencing to content of the current one. We take this into account in order to improve the coherence of the generated summary.

9. After having extracted and assigned all this additional information to the text, the summary is created by adding all sentences marked for inclusion. Also the quotation marks of partially included in the summary quotes are fixed.

10. The final step is to list the words with the highest score as keywords.

Ideas for future development

In order to verify the hypothesis that our changes improve the quality of the summary we made a blind evaluation test by human experts of the summaries generated by both algorithms. One benefit of this is that the experts gave us not only their ratings to the summaries, but also valuable feedback. The classical understanding of a summary is 1/3 of the text, however when the text is large – for example for news and opinion articles, a summary of 5-7 pages is still pretty large and, for practical reasons, one expert pointed out that it would be better to make this fraction smaller. A similar remark was made about the number of keywords, as the fixed number of 5 keywords is inadequate for very small or very large articles.

After the analysis of the results we also came to some general conclusions concerning the strategy to be used in the future. From a data mining perspective, both algorithms do not use a potentially valuable bit of information – the topic of the article. The keywords and the summaries will probably be of better quality if a reliable entities detection algorithm is used, which is not the case at the moment. When the text is fairly long and/or consists of self-contained sections, our tests with manual splitting of the text and generating separate mini-summaries to be combined in a single one made us believe that summarizing the paragraphs one by one with local centers or utilizing a sliding window for the algorithm will be a better strategy.

Comparative analysis and some statistical parameters of the summaries

In order to evaluate the effect of the changes performed on the original algorithm, we did a comparative study using the following procedure:

The authors submitted the summaries followed by the full text of the articles/news items and the keywords to 4 evaluators. Each one had to read at least 7 summaries and keywords and evaluate their quality by reading beforehand the full texts. The experts had to use a scale from 1 to 5 where 1 is the worst and 5 is the best mark. The texts and what algorithm was used to generate the summary were chosen at random and the experts had no information about which summary is created by means of which algorithm.

After the evaluation, we performed a trivial procedure of scaling the assigned marks to the global mean mark as the evaluators had different requirements and different mean marks.

In order to study the influence of some parameters of the original text on the quality of the summaries, we measured the following characteristics of the original text:

1. NumberWords is the number of words in the text after segmentation

2. NumberParagraphs is the number of paragraphs in the text as detected by our segmentation procedure
3. Complexity – we are using the Automated Readability Index (ARI). This is a readability test designed to assess the understandability of a text. At the output ARI gives a number which approximates the grade level needed to comprehend the text. The formula it uses is shown below:

$$4.71 * (\frac{characters}{words}) + 0.5(\frac{words}{sentences}) - 21.43 \quad (3)$$

Richness – we measure the richness by using the Yule's *I* index

$$I = \frac{M_1^2}{M_2 - M_1} \quad (4)$$

where M_1 is the number of all word forms a text consists of and M_2 is the sum of the products of each observed frequency to the power of two and the number of word types observed with that frequency The larger Yule's *I*, the larger the diversity of the vocabulary (and thus, arguably, the more difficult the text) [Teller, 2011].

Table 1 gives the description of the measures characteristics of the sample of 50 papers and the marks, given by the evaluators (after correction).

Table 1. Descriptive statistics of the sample

	N	Minimum	Maximum	Mean	Std. Deviation
Number of Words	50	263	6706	978,94	1112,825
Number of Paragraphs	50	1	67	16,18	12,462
Complexity	50	7	20	13,03	3,033
Richness	50	6,20	42,72	20,7705	7,99629
Text Summary Mark	50	2,55	5,55	4,1224	0,73468
KeyWords Mark	35	1,86	5,29	3,8603	0,92854

Comparison of the results for the two summarizers.

Our relatively small sample has shown some promising tendencies. The comparative statistics are given in table 2.

Table 2. Comparative parameters of the evaluation of the two summarizers – OTS and the developed Improved OTS (IOTS)

Sumarizer		TextSunnaryMarkr	KeyWordsMark
IOTS	Mean	4,155	3,911428571
	N	22	14
	Std. Deviation	0,751	1,068520596
OTS	Mean	4,096	3,826190476
	N	28	21
	Std. Deviation	0,734	0,84884319
Total	Mean	4,122	3,860285714
	N	50	35
	Std. Deviation	0,734	0,928539288

As it is seen from the results in table 2, the mean mark for the quality of the text of the summary and the mean mark for the keywords extraction are better for the reported here Improved OTS (IOTS) summarizer reported here.

Analysis of the influence of the parameters of the text on the quality of the summary

In order to check whether the diffidence is statistically significant, we performed an independent sample T-test and ANOVA. Unfortunately, with such a small sample, having a small difference in the means and big variances of the evaluations, both tests cannot reject the hypothesis that the means of marks for the two summarizers are equal.

In order to investigate the possible means for further improvement, we examined the influence of the parameters of the original text on the quality of the generated summaries and the extraction of keywords (assuming that the experts' marks show this quality correctly).

One expects both algorithms explained in the previous section to be sensitive to the length of the text to be summarized. However, as shown in table 3, there is no effect of the number of words in the original text on the quality of the summary and the keywords.

Table 3. Correlation matrix (Pearson) – effect of the parameters of the text on the quality if the summary and on the keywords extraction

		TextSummary Mark	KeyWords Mark	Complexity	NumberWords	NumberParagraphs	Richness
TextSummary Mark	Correlation	*1*	-,196	-,297	,185	,168	-,256
	Sig. (2-tailed)		,260	,036	,199	,243	,073
	N	50	35	50	50	50	50
KeyWordsMark	Correlation	-,196	1	-,177	,170	,208	-,190
	Sig. (2-tailed)	,260		,308	,328	,231	,274
	N	35	35	35	35	35	35
Complexity	Pearson Correlation	-,297	-,177	1	,024	,037	,043
	Sig. (2-tailed)	,036	,308		,866	,798	,766
	N	50	35	50	50	50	50
NumberWords	Correlation	,185	,170	,024	1	,866	-,489
	Sig. (2-tailed)	,199	,328	,866		,000	,000
	N	50	35	50	50	50	50
NumberParagraphs	Correlation	,168	,208	,037	,866	1	-,527
	Sig. (2-tailed)	,243	,231	,798	,000		,000
	N	50	35	50	50	50	50
Richness	Correlation	-,256	-,190	,043	-,489	-,527	1
	Sig. (2-tailed)	,073	,274	,766	,000	,000	
	N	50	35	50	50	50	50

We observe another dependency – the quality of the summary is negatively correlated with the complexity of the original text (Parson Correlation -0.297, p-value 0.03). It seems the more the text is evaluated as complex (following equation (3), as explained in this section), the worse is the quality of the summary.

This dependency can be interpreted as follows: The term $\frac{words}{sentences}$ in ARI readability index (see equation (3)) expresses a measure of the length of an ‘’average” sentence in the text. As explained, the summarizers extract (leave in the summary) entire sentences, having calculated first scores concerning the words inside of each sentence, sentence by sentence. From where - the dependency we observe.

Conceder we have to express T simple thoughts in a text with N words. The ideal number of sentences would be T – one simple thought in one sentence. This is an ideal nonrealistic case in which the average length of a sentence would be N/T and the complexity (difficult readability) measure ARI would be small. Unfortunately for the summarizers of the considered type, the writing style contains complex subordinated, fused, concatenated etc. sentences, expressing more than one simple thought within a sentence. In result the text is judged as “complex” using expression (3), and the summarizers extract 1/3 not of T sentences, but of T-K longer complex sentences. That means: 1. each of the longer sentences is rated (see step 7 of the algorithm) using words which express more than one simple thought; 2. The summary contains less than T/3 longer sentences which risk being semantically unrelated, so the coherence of the summary is damaged.

Unfortunately, to solve this problem, one needs syntactic analysis to separate the sub-sentences, something that we want to avoid.

Analogical reasoning can be applied for the number of paragraphs.

The other path of reasoning in our analysis is related to the result concerning the dependency of the quality of the summary with the richness of the original text.

The formula (1) applied to calculate the scores does not take into consideration the use of synonyms. However, the writings of higher quality avoid the use of one and the same word several times, writers look for synonyms to make the text less repetitive and following expression (4) it becomes “richer”. Obviously that influences the frequencies used to calculate the scores and the quality of the summary. Once more time the problem touches semantic questions. However, this behavior of the algorithm can be adjusted using synonym dictionaries in a convenient way.

Conclusion

We have introduced several modifications in OTS that are showing promising results for improved quality of the summaries and keyword extraction. We found that the results are still not statistically significant and we did an analysis of the parameters of the original text and the quality of the summaries of *extraction-based* type. Our results show paths for further improvement of this type of summary generation.

Acknowledgment

This paper is published with partial support by the ITHEA ISS (www.ithea.org) and the Central Fund for Strategic development, New Bulgarian University.

We are grateful to the colleagues and friends who helped us with the summary evaluation. They are Galina Velichkova, Dessislava Petkova and Krassimir Todorov. We are grateful to our colleague Dimitar Atanasov for consulting our statistical approaches.

Bibliography

[Andonov et al. 2016] Filip Andonov, Velina Slavova, Marouane Soula. Heuristics-based classifier in a framework for sentiment analysis of news. International Journal "Information Content and Processing" Volume 3, Number 3, ITHEA, 2016. ISSN 2367-5128 (printed), ISSN 2367-5152 (online). pp. 224 - 234

[Heuven et al. 2014] Walter J. B. van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert, SUBTLEX-UK: *A new and improved word frequency database for British English*, The Quarterly Journal of Experimental Psychology Vol. 67, Iss. 6,2014

[Open Text Summarizer, 2016] Open Source Text Processing Project: Open Text Summarizer. http://textprocessing.org/2016/01, visited 08.2016

[Teller, 2011] Swizec Teller. Measuring vocabulary richness with python. September 28, 2011. https://swizec.com/blog/measuring-vocabulary-richness-with-python/swizec/2528, visited 08.2016

Authors' Information

Filip Andonov **-** *New Bulgarian University, department of Computer Science, fandonov@nbu.bg.*

Major Fields of Scientific *Research: multicriteria optimization, data mining, text processing, Python language*

Velina Slavova **-** *New Bulgarian University, department of Computer Science, vslavova@nbu.bg*

Major Fields of Scientific Research*: AI, Cognitive Science*

Georgi Petrov *- New Bulgarian University, department of Telecommunications, gpetrov@nbu.bg*

Major Fields of Scientific Research*: automation of processes*

АВТОМАТИЗИРОВАННЫЙ АНАЛИЗ ТЕМАТИКИ ТЕКСТОВ НОВОСТЕЙ

Алла Заболеева-Зотова, Алексей Петровский, Юлия Орлова, Татьяна Шитова

***Аннотация**: В работе рассматриваются возможные подходы к автоматизированному анализу текста новостей с помощью алгоритмов кластеризации текстов, извлечения ключевых слов и формирования семантической связности блоков текста. Особое внимание уделено выявлению тематики и сюжетов новостей.*

***Ключевые слова**: поток новостей, тематическая кластеризация текстов, извлечение ключевых слов, семантическая связность блоков текста*

***ACM Classification Keywords**: A.0 General Literature - Conference proceedings*

Введение

Проблема снижения информационной перегрузки людей, и в особенности пользователей Интернета и социальных сетей, становится всё более актуальной. Как сообщает американская исследовательская служба Cyveillance, в начале XXI века количество страниц в Интернете превысило 4 млрд, и с каждым днем увеличивается на 7 млн. Темпы роста аудитории онлайновых новостных ресурсов практически вдвое превышают темпы роста общей численности пользователей Интернета. Большую часть информации, с которой имеют дело пользователи, составляют «сырые» неструктурированные данные. Поэтому велика потребность в эффективных технологиях автоматизированного анализа информации, представленной на естественном языке, выявления групп семантически похожих текстов.

Существует достаточно много программных продуктов, предоставляющих функции анализа текстовых документов. Среди отечественных систем отметим TextAnalyst, Galaktika-ZOOM, из зарубежных – мощный инструмент анализа текстов IBM Text Miner. В TextAnalyst имеются опции создания семантической сети большого текста, подготовки аннотации, автоматической классификации и кластеризации текстов. IBM Text Miner содержит утилиты классификации, кластеризации, поиска ключевых слов и составления аннотации текстов. Однако эти программы не направлены на обработку новостных статей.

Российская система Яндекс Новости позволяет автоматически группировать данные в новостные сюжеты и составлять аннотации статей на основе кластеризации документов. Сервис InfoStream,

обеспечивает доступ к оперативной информации в поисковом режиме с учетом семантической близости документов. Мобильный агрегатор новостей Summly, используемый компанией Yahoo!, также осуществляет группировку новостей по темам. Однако приложение абсолютно неприменимо для обработки текстов на русском языке.

Таким образом, существующие программные системы не решают поставленную проблему полностью. В работе предложена методика комплексного анализа текста новостей, основанная на комбинации алгоритмов тематической кластеризации текстов, статистических алгоритмов извлечения ключевых слов и алгоритмов формирования семантической связности блоков текста. Проведена апробация методики при он-лайн обработке потоков новостных интернет-статей.

Обработка и анализ текста новости

Анализ новостных текстов включает в себя тематическую кластеризацию текстов и последующую комплексную обработку статей. В основу предлагаемой обобщенной структуры текста новости (рисунок 1) положен принцип «перевернутой пирамиды», который требует размещение основной информации в самом начале материала и последующее ее раскрытие в деталях далее по тексту.

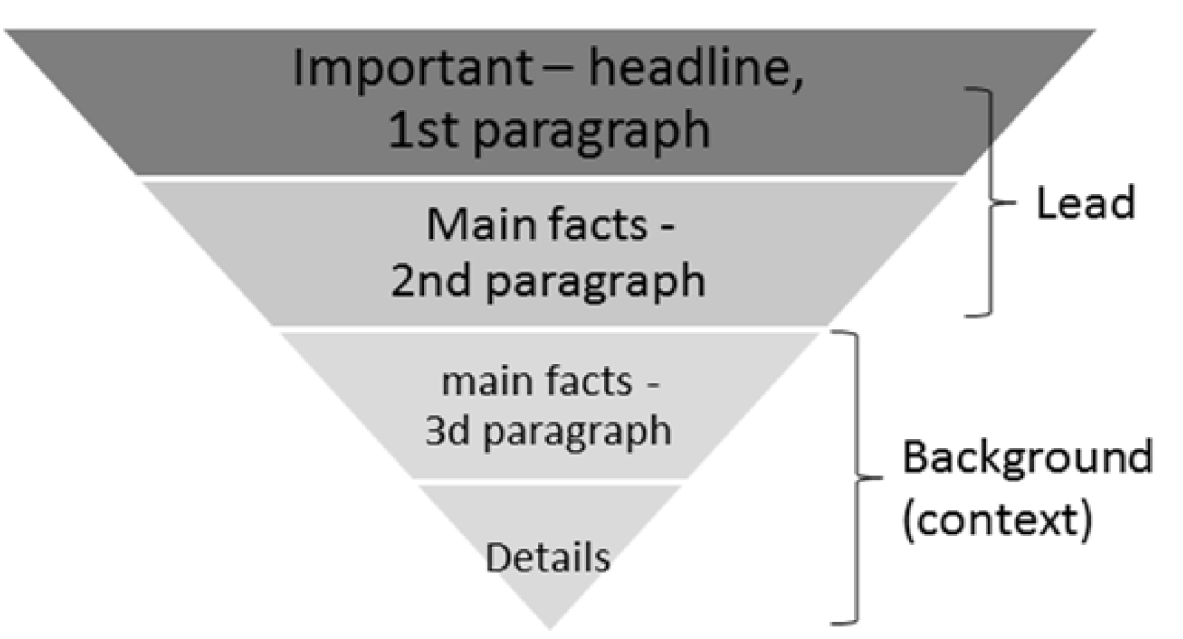

Рисунок 1 – Структура новостного текста.

Заголовок новости отражает ее тему и содержит не более 10 слов (около 80 символов). Так, для примера, в Яндексе отображается в заголовке не более 15-ти слов, Google показывает до 70 слов.

Основные факты, касающиеся события, представляются в 1-2 абзацах, и составляют так называемую вводную часть (lead), которая освещает главную тему новости.

3-й и последующие абзацы составляют контекст (background) новости. Как правило, здесь раскрываются детали происходящего, дается информация, напрямую касающаяся новости.

Таким образом, для содержания новости справедлива формула: (Who? + What? + Where? + Why? + When? + How?) [Добров, 2010]. Это так называемый закон «пяти W и одного H», приписываемый Р.Киплингу. Если бы все новостные сообщения строились по единой структуре, то решение задачи тематического анализа текстов новостей могло бы значительно упроститься.

Процесс обработки и анализа текста новости можно разбить на несколько этапов.

Графематический анализ представляет собой начальный этап обработки текста, в ходе которого вырабатывается информация, необходимая для дальнейшей обработки морфологическим и синтаксическим анализаторами. В задачу графематического анализа входит внутреннее представление структуры новости: T = <P, S, W>, где P – абзацы, S – предложения, W – слова. При этом необходимо корректно выделить заголовок и первое предложение абзаца, содержащее основные факты статьи.

Следующим этапом является морфологический анализ, цель которого – построение морфологической интерпретации слов входного текста. Все методы можно разделить на словарные и вероятностно-статистические (без использования словаря). Недостатками вторых являются большой объем лексиконов, плохая работа на малой выборке, отсутствие точных лингвистических методов. Словарный метод основан на подключении словаря (тезауруса), дает максимально полный анализ словоформы. Для данного блока целесообразно использовать морфологические библиотеки, например, Lemmatizer, FreeLing, NLTK, MCR, tokenizer [Михайлов и др., 2009].

Синтаксический анализ рассматривается как задача построения дерева зависимостей предложения, в ходе которого происходит выделение синтаксических конструкций, определение связности и подчинения фрагментов [Grune, 2012]. Для поиска ключевых слов разработан алгоритм [Soloshenko et al, 2014], сочетающий выделение именованных сущностей из текста

новости (на основе результатов морфоанализа и подключаемого модуля PullEnti), подсчет веса слова с учетом частоты его встречаемости (рисунок 2). Пороговое значение для признания слова ключевым задается значением относительной частоты встречаемости слова-кандидата в ключевые слова, с индексом, равным 0,2×количество сущностей. Такое значение определено экспериментально на выборке из 100 текстов.

Рисунок 2 – Алгоритм поиска ключевых фраз в тексте.

Структуру совокупности знаний S текста новости можно определить следующим образом : S = {M, F}, где M – множество всех понятий данной совокупности знаний, F – отношение «смысловая связь». В качестве формальной модели структуры знаний можно использовать семантическую сеть, определяемую как ориентированный граф G = (E,V), где E – множество вершин, поставленное во взаимно однозначное соответствие с множеством понятий; V – множество ориентированных дуг. Дуга выходит из вершины, соответствующей основному понятию A, и входит в вершину, соответствующую понятию, которое сочетается по смыслу с понятием A [Машечкин и др., 2011; Dmitriev et al, 2013]. Таким образом, содержание новости можно наглядно представить в виде ключевых понятий и связей между ними, либо в виде так называемой интеллектуальной карты (mind map).

Подсчет веса предложения при построении аннотации осуществляется в зависимости от его нахождения в тексте новости и рассчитывается по формуле:

$$Ws = N(kw)\cdot RF(kw)\cdot WP\ C. \quad (1)$$

Здесь Ws – вес предложения; N(kw) – количество вхождений ключевого слова в предложение; RF(kw) – относительная частота ключевого слова; WP – относительный вес параграфа в тексте, равный 0.35 для первого параграфа, 0.2 для второго параграфа, 0.1 для остальных параграфов (контекст); C – коэффициент значимости предложения внутри параграфа, равный 1.0 для первого предложения в абзаце, 0.8 для остальных предложений [Soloshenko et al, 2014]. В итоговую аннотацию в зависимости от заданного коэффициента сжатия включаются предложения с наибольшим весом.

Методы тематической кластеризации текстов

Напомним основные понятия. Кластеризация – разбиение множества объектов на подмножества (кластеры), число которых может быть произвольным или фиксированным. Основные группы алгоритмов кластеризации: иерархические и неиерархические, четкие и нечеткие. Иерархические алгоритмы строят несколько разбиений исходного множества объектов на непересекающиеся кластеры. Результатом является дерево кластеров, корень которого – все исходные объекты, а листья – итоговые кластеры. Неиерархические алгоритмы строят одно разбиение объектов на заданное число кластеров. Четкие алгоритмы определяют принадлежность каждого исходного

объекта только одному кластеру. Нечеткие алгоритмы ставят в соответствие каждому объекту степень его принадлежности к нескольким кластерам [Bandyopadhyay et al., 2013].

Иерархические алгоритмы разделяются на агломеративные (восходящие) и дивизимные (нисходящие). Первые строят кластеры снизу вверх, начиная с множества кластеров, содержащих по одному одиночному объекту, затем последовательно объединяют пары кластеров, пока не получат один кластер, содержащий все исходные объекты. Вторые разбивают кластеры сверху вниз, начиная с одного кластера, которому принадлежат все исходные объекты, затем этот кластер делится на два и так рекурсивно до тех пор, пока каждый объект не окажется в своём отдельном кластере. Основное их различие заключается в выборе критерия, используемого для принятия решения о том, какие кластеры следует объединить на текущем шаге алгоритма. Большое распространение получили следующие критерии:

- одиночная связь (минимальное расстояние или максимально сходство) – сходство между двумя наиболее похожими объектами/кластерами;

- полная связь (максимальное расстояние или минимальное сходство) – сходство между двумя наиболее непохожими объектами/кластерами;

- групповое усреднение всех показателей сходства – сходство двух кластеров есть среднее сходство всех пар объектов, включая пары объектов из одного кластера, исключая близость объекта самому себе;

- центроидный метод;

- метод Уорда.

Алгоритм k-средних начинает с некоторого начального разбиения объектов на заранее заданное число кластеров и уточняет его, оптимизируя целевую функцию – среднеквадратичную ошибку кластеризации как среднеквадратичное расстояние между объектами и центрами их кластеров:

$$e(D,C) = \sum_{j=1}^{k} \sum_{i:d_i \in c_j} || \vec{d}_i - \vec{\mu}_j ||^2 \qquad (2)$$

где μ_j – центроид кластера C_j. Обычно исходные центры кластеров выбираются случайным образом. Затем каждый объект присваивается тому кластеру, чей центр является наиболее близким документу, и выполняется повторное вычисление центра каждого кластера как центроида, или среднего своих членов. Такое перемещение объектов и повторное вычисление центроидов кластеров продолжается до тех пор, пока не будет достигнуто условие остановки, например: (а) достигнуто пороговое число итераций, (б) центроиды кластеров больше не изменяются, (в) достигнуто пороговое значение ошибки кластеризации.

Нечеткий алгоритм классификации k-средних FCM относит каждый объект к более чем одному кластеру. Как и его чёткий вариант, данный алгоритм, начиная с некоторого начального разбиения данных, итеративно минимизирует целевую функцию, которой является следующее выражение:

$$e_m(D,C) = \sum_{i=1}^{|D|}\sum_{j=1}^{|C|} u_{ij}^m \,||\, \vec{d}_i - \vec{\mu}_j \,||^2 \qquad (3)$$

где m – степень нечеткости, $1<m<\infty$, u_{ij} - степень принадлежности i-го объекта j-му кластеру.

Алгоритм минимального покрывающего дерева MST сначала строит на графе минимальное покрывающее дерево, а затем последовательно удаляет ребра с наибольшим весом. На рисунке 3 изображено минимальное покрывающее дерево, полученное для девяти объектов. Путём удаления связи, помеченной CD, с длиной равной 6 единицам (ребро с максимальным расстоянием), получаем два кластера: {A, B, C} и {D, E, F, G, H, I}. Второй кластер в дальнейшем может быть разделён ещё на два кластера путём удаления ребра EF, которое имеет длину 4,5 единицы [Pera et al., 2012].

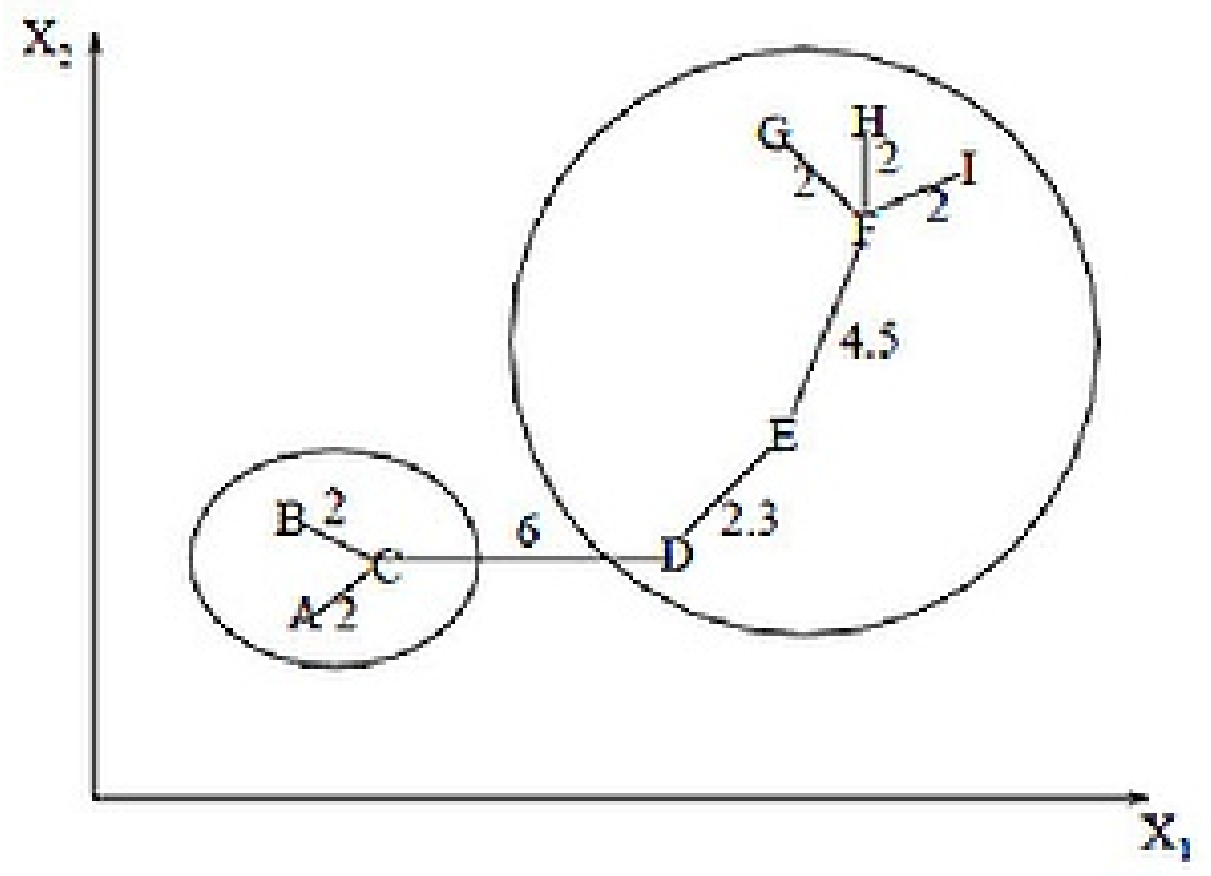

Рисунок 3 – Иллюстрация алгоритма MST.

Алгоритм самоорганизующихся карт (Self Organizing Maps – SOM) был предложены для визуализации и кластеризации данных. Визуализация данных осуществляется путём проецирования многомерного пространства данных в двумерное пространство – карту данных. Такая карта, построенная для массива полнотекстовых документов, может служить как поисковый механизм, альтернативный поиску по запросу, предлагающий обзор/навигацию по коллекции документов [Kiryakov, 2004]. Идея алгоритма заключается в том, чтобы обучить нейронную сеть без учителя. Сеть состоит из некоторого числа нейронов, упорядоченных по узлам двумерной сетки. Каждый нейрон имеет координаты в исходном $|\tau|$-мерном пространстве документов и двумерном пространстве карты. В процессе обучения нейроны упорядочиваются в пространстве документов так, чтобы наилучшим образом описать входной массив документов. Этот процесс является итерационным. На каждой итерации t случайным образом выбирают документ d_i из входного массива D; находят нейрон-победитель m_c, ближайший к документу d_i; корректируют веса соседей нейрона-победителя: $m_i(t+1) = m_i(t) + h_{ci}(t)[d_i - m_i(t)]$.

Алгоритмы групповой иерархической и неиерархической кластеризации объектов разработаны для случаев, когда объекты могут присутствовать в нескольких различающихся версиях [Petrovsky, 2003]. Объекты, описываемые многими количественными и/или качественными признаками, рассматриваются как точки метрического пространства мультимножеств [Петровский, 2003]. Примером может служить одна и та же новость, содержащаяся в разных текстах или опубликованная несколько раз в разное время.

При выборе оптимального алгоритма кластеризации потока текстов новостей необходимо учитывать его следующие особенности: постоянно растущая коллекция документов, одна и та же статья может отражать несколько сюжетов, новости имеют определенную структуру текста, разные части документа должны иметь различный вес при нахождении близости, сюжеты и документы могут иметь перекрестные ссылки друг на друга. Для работы с новостными текстами желательно, чтобы алгоритм был неиерархическим, нечетким, инкрементальным. Поэтому наиболее перспективным для данной задачи видится применение алгоритма FCM или нейронных сетей.

Апробация методического подхода

Для проверки предложенного подхода к тематической кластеризации текстов новостей была создан программный комплекс, основанный на принципе многокомпонентности программного обеспечения [Zaboleeva-Zotova et al, 2013], а также проведен эксперимент [Солошенко и др., 2014], цель которого – доказать, что за счет автоматизации обработки статей новостного потока снизилось время на обработку и повысилось качество обработки новостных интернет-статей. Были получены следующие результаты.

Время обработки текстов новостей уменьшилось как минимум в два раза. При этом время с помощью программы учитывается не только непосредственно время составления аннотации, но и время, необходимое для окончательной корректировки текстов (рисунок 4). Качество аннотации оценивалось экспертами по следующим критериям: сохранение ключевых фактов, связность новостной статьи, сохранение синтаксической структуры текста после удаления незначащих частей. Каждый из критериев имел шкалу оценок от 0 до 10 баллов. Качество полученной аннотации оценивалось по среднему арифметическому трех показателей для каждого текста. Качество автоматически обработанных текстов новостей осталось на том же уровне, что и при анализе текста человеком.

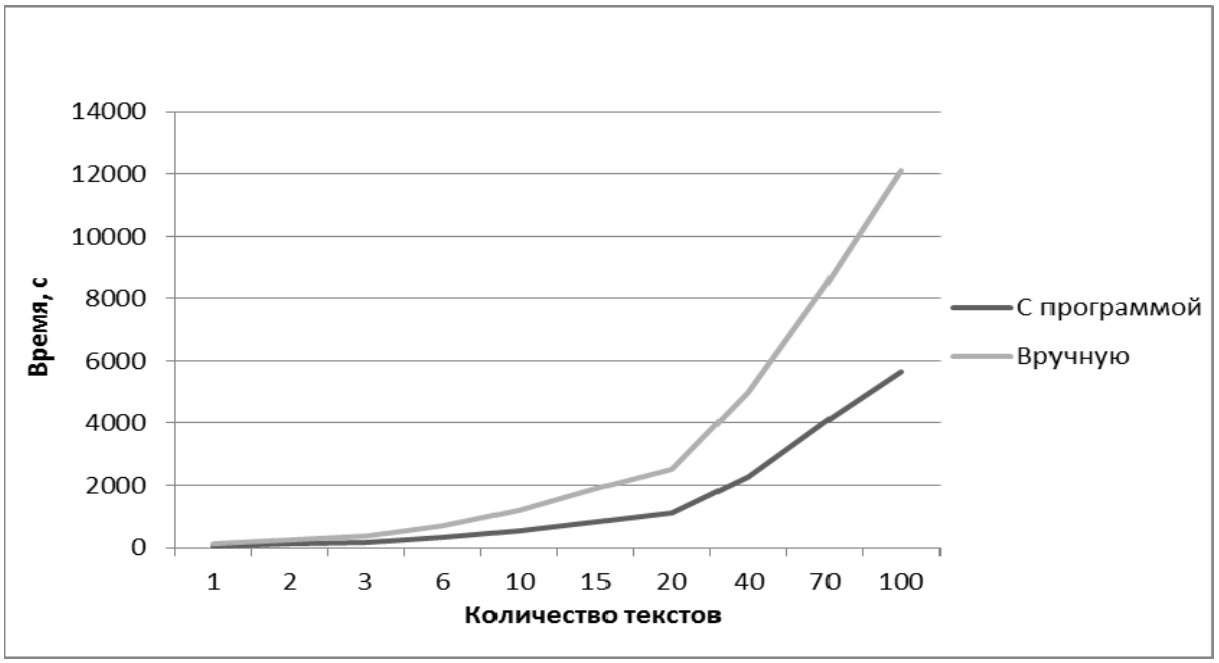

Рисунок 4 – Зависимость времени обработки от количества текстов.

Заключение

Анализ новостных текстов включает в себя задачу кластеризации и последующую комплексную обработку статей. Проведен анализ потока новостей, выделены особенности документов. Проанализированы методы кластеризации объектов, предложены алгоритмы, наиболее подходящие для анализа текстов новостей. Реализована часть программной системы для он-лайн агрегации новостей из интернет-источников,и проведены исследования эффективности ее работы.

Благодарности

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Работа опубликована при частичной поддержке проекта ITHEA XXI общества ITHEA ISS (www.ithea.org) и ADUIS (www.aduis.com.ua).

Библиография

[Добров, 2010] Добров Б.В. Исследование качества базовых методов кластеризации новостного потока в суточном временном окне // Труды 12-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – Казань, 2010. – 287–295.

[Машечкин и др., 2011] Машечкин И.В., Петровский М.И. Латентно-семантический анализ в задаче автоматического аннотирования // Программирование. – 2011. – Т. 37, № 6. – 67-77

[Михайлов и др., 2009] Михайлов Д.В., Емельянов Г.М. Морфология и синтаксис в задаче семантической кластеризации // Математические методы распознавания образов. – Суздаль, 2009. – 1-4.

[Петровский, 2003] Петровский А.Б. Пространства множеств и мультимножеств. – М., Едиториал УРСС, 2003.

[Солошенко и др., 2014] Солошенко А.Н., Розалиев В.Л., Орлова, Ю.А. Автоматизация семантического анализа новостных Интернет-текстов. // Открытые семантические технологии проектирования интеллектуальных систем: матер. IV междунар. науч.-техн. конф. – Минск, БГУИР, 2014. – 435-438.

[Bandyopadhyay et al., 2013] Bandyopadhyay S., Saha S. Unsupervised Classification. – Berlin, Springer, 2013.

[Grune, 2012] Grune D. Tokens to Syntax Tree – Syntax Analysis. – New York, Springer, 2012.

[Dmitiev et al, 2013] Dmitiev A.S., Zaboleeva-Zotova A.V., Orlova Yu.A., Rozaliev V.L. Automatic identification of time and space categories in the natural language text // Applied Computing 2013: Proceedings of the IADIS International Conference. – Fort Worth, 2013. – 187-190.

[Kiryakov, 2004] Kiryakov A. Semantic annotation, indexing, and retrieval // Web Semantics: Science, Services and Agents on the World Wide Web. – 2004. V.2. № 1. – 49-79.

[Petrovsky, 2003] Petrovsky A.B. Cluster analysis in multiset spaces. // Information Systems Technology and its Applications. – Bonn, Gesellschaft für Informatik, 2003. – 109-119.

[Pera et al., 2012] Pera, M.S., Ng, Y.-K.D. Using maximal spanning trees and word similarity to generate hierarchical clusters of non-redundant RSS news articles. // J. Intell. Inf. Syst. – 2012. V.39. – 513-534.

[Soloshenko et al, 2014] Soloshenko A.N., Orlova Yu.A., Rozaliev V.L., Zaboleeva-Zotova A.V. Thematic clustering methods applied to news texts analysis // Knowledge-Based Software Engineering: Proceedings of 11th Joint Conference. – Springer, 2014. – 294-310.

[Zaboleeva-Zotova et al, 2013] Zaboleeva-Zotova A.V., Orlova Yu.A., Rozaliev V.L., Fomenkov S.A., Petrovsky A.B. Formalization of initial stage of designing multi-component software // Multi Conference on Computer Science and Information Systems: Proceedings of the IADIS International Conference – Prague, IADIS, 2013. – 107-111.

Сведения об авторах

Заболеева-Зотова Алла Викторовна *– д.т.н., профессор, начальник управления Российского фонда фундаментальных исследований, старший научный сотрудник Института системного анализа ФИЦ «Информатика и управление» РАН, Россия, Москва 119991, Ленинский пр-т, 32А, e-mail::* zabzot@rfbr.ru

Петровский Алексей Борисович *– д.т.н., профессор, заведующий лабораторией Института системного анализа ФИЦ «Информатика и управление» РАН, Россия, Москва 117312, пр-т 60-летия Октября, 9, e-mail::* pab@isa.ru

Орлова Юлия Александровна *– к.т.н., к.п.н., доцент, доцент Волгоградского государственного технического университета, Россия, Волгоград 400005, пр-т Ленина, 28, e-mail:: yulia.orlova@gmail.com*

Шитова Татьяна Алексеевна *– экономист Института системного анализа ФИЦ «Информатика и управление» РАН, Россия, Москва 117312, пр-т 60-летия Октября, 9, e-mail:: tanya-petrovskay@yandex.ru*

Automated Analysis of Thematic of News Texts

Alla Zaboleeva-Zotova, Alexey Petrovsky, Yulia Orlova, Tatiana Shitova

Abstract*: The paper deals with possible approaches to the automated analysis of news texts using algorithms for clustering texts, keywords extracting and forming the semantic coherence of text blocks. Particular attention is paid to the discovery of themes and subjects of news.*

Keywords: *stream of news, thematic clustering texts, keywords extracting, semantic coherence of text blocks.*

TABLE OF CONTENTS