ISSN 1992-6502 (Print)
Vol. 18, no. 5 (66), pp. 20-25, 2014

Вестник УГАТУ
Vestnik UGATU

ISSN 2225-2789 (Online)
http://journal.ugatu.ac.ru

UDC 681.142.2

THE PROCESS OF DETECTING FACTS IN A NATURAL TEXT

A. VOKHMINTCEV[1], A. MELNIKOV[2]

[1] vav@csu.ru, [2] mav@csu.ru

[1]Chelyabinsk State University (CSU), Russia

Submitted 2014, June 15

Abstract. In this article we present a method of extracting factual data (objects, their attributes and relationships) from natural language texts using ontological knowledge base model. Subject area is represented through ontological model extended by fuzzy links between objects

Key words: annotated image database; fact analysis; Image Net; Word Net; ontology; text mining; applied linguistics; extracting facts.

Considerable part of scientific and technical information in the modern world is open, especially during the stage of initial idea formulation, discussion, and approbation. Governmental counterespionage services and private enterprises both avail of this information in industrial design or marketing of science intensive products. Intelligence experts mostly use electronic resources: Internet (scientific articles, conference proceedings, industry and business news, special-purpose informational resources), design documentation, offline article storages in scientific libraries, social network sites (online conferences, forums, and blogs).

Existing natural language processing technologies have limited applicability due to the fact that search engines, document management systems or text mining systems that use these technologies, have very wide scope and cannot handle text semantics. Moreover, morphological search, topic search or syntactical-semantic tree-based search don't allow for search by meaning. In applied linguistics, semantics is understood as information linked with the word by means of thesaurus or explanatory dictionary. Research in the field of semantic analysis of full-text documents has been carried on since long ago: Agarwal S., Awan A., Fellbaum C, Chomsky's transformational grammar, Fillmore's predicate-argument structures, Schenk's concept model and meaning-text theory by Mel'čuk. Chomsky's approach could be explained as study of deep syntactical structure of a sentence, building a dependency tree, and detection of semantic anomalies. Fillmore and Schenk were the first to introduce ideas of concept and frame, that is, predicate-argument structures with roles, such as agent, object, addressee, source, and medium, assigned to components. Works of Mel'čuk, the author of language theory dealing with multi-level transformations between meaning and text, and vice versa, are also of great interest. One of characteristic features of this theory is its usage of dependency syntax and explanatory combinatorial dictionary, ancestor of modern thesauruses and ontologies. Meaning-text theory lies in the ground of modern semantics.

The task of semantic analysis of natural language documents is rather complex and is generally solved through building artificial intelligence systems, performing deep semantic analysis of the text using subject field knowledge base. Creating such a knowledge base containing the knowledge of mankind is one of the fundamental problems of applied linguistics and knowledge engineering. Complexity of the task is not the only problem: there are issues associated with changing of study object over time and quality of models built by experts. Nowadays all the projects in knowledge base construction are either limited by subject or targeting some specific search needs [1]. In the framework of already specified problem (analysis of relationship between the person and organization) it is possible to describe specific subject field and to build a knowledge base. To integrate professional scientific and technical knowledge it is necessary to develop:

- a metatext model of natural language that can represent textual data in a formalized way;
- methods to convert metatext into a knowledge base;
- ontology as a knowledge describing the subject field;
- methods to extract facts from the knowledge base.

NATURAL LANGUAGE METATEXT MODEL

Metatext model is based on the principles of communicative grammar of Russian and English languages. Its main principle is interconnection between syntax and semantics: syntax studies sensible speech and words' meaning should be used during syntactical analysis of the text. Metatext model is based on the concept of fact, which is the smallest unit of meaning in the utterance [2]. Let's define a fact) as an elementary syntactic and semantic language unit that corresponds to some elementary meaning in a knowledge model and has a set of morphological, syntactical, semantic, and functional features.

Facts could be connected by different types of relationships: hierarchy (HIR), transitivity (TRA), inclusion (TAR), union (UN), intersection (INSEL), subject (SUB).

Relationships in a set of facts represent their semantic connections, thus sentence semantics depends on the set of facts in that sentence.

Let's describe the process of natural language text document analysis. To solve this applied problem we need to develop a formalized natural language text model which i will be used as a basis for transforming the text into dependency tree and detecting facts (factual analysis of text data). Basic elements of linguistic descriptions are as follows:

- key objects of the fact;
- additional objects of the fact;
- fact scenario.

Let's describe the process of detecting facts in a text and building a set of fact relationships. We will need the following definitions:

Lexeme – string of characters representing elementary text unit.

Object – sequence of lexemes or objects meeting certain restrictions and being analyzed as a single entity. Objects could include other objects.

Object attribute – property of an object that can be used in operations of comparison and assignment. Attributes could be linked with objects automatically during the text analysis, or defined by the user.

Object description – set of all object attributes carrying information on orphographical peculiarities of the object, its morphological, grammatical, and semantic characteristics.

Object isolation – joining the chain of lexemes corresponding to the identified object into a new object and assigning a new description to it.

Object isolation rule – "pattern-description" pair written in formalized language and used to produce object decription in case given object is detected using the pattern.

Target object – finite object, isolated by the component for user specified tasks according to the isolation rule. Target objects (TO) can be subdivided into two classes: significant TO and insignificant TO. Insignificant TO include all auxiliary sentence elements not having any meaning by themselves (service parts of speech — unions, prepositions and punctuation). Significant target objects, in their turn, fall into three categories:

- named objects – this class includes following semantic types: persons, organizations, geographic objects, technology and product names, and other proper nouns;
- unnamed target objects - this class includes full words of following parts of speech: common names, adjectives, auxiliary verbs, animated and unanimated objects, object attributes, events;
- special target objects – entities encountered in some special constructions in text, consisting of alpha-numeric characters: dates, adverbial modifiers of time, money amounts, identification data of people and organizations, etc.

Stage 1. Pre-syntactical analysis. Words, separators, terminators and stop-words in the text are detected at this stage. Then all the possible grammatical forms are determined using the detected morphology. Word forms corresponding to one triple (normal form, part of speech, grammatical number) are joined into lexemes.

Pre-syntactical analysis:

- morphological text analysis;
- processing of word forms not present in the dictionary;
- identification of stop words;
- preliminary analysis of typical structures in text document.

The result of 1st stage of processing is a set of sentences, where each sentence contains ordered list of words with variants of homonymous lexemes.

Stage 2. Syntactical analysis. In our case, the main task of syntactical analysis is to find dependencies between lexemes and isolate objects. Syntax analyzer takes a sentence with corresponding lexemes, derived during morphological analysis. Syntactical analysis is performed by building a list of syntax subordination trees, corresponding to different variants of a sentence, and then euristically selecting one variant from this list. Result of this action is a syntax dependency tree, representing a sentence.

Syntactical analysis:

- extraction of standard sentence constructions using the morphology data, building of word-combinations;
- identification of syntactic and semantic constructions in the text;
- construction of syntactic and semantic dependency tree (Fig. 1).

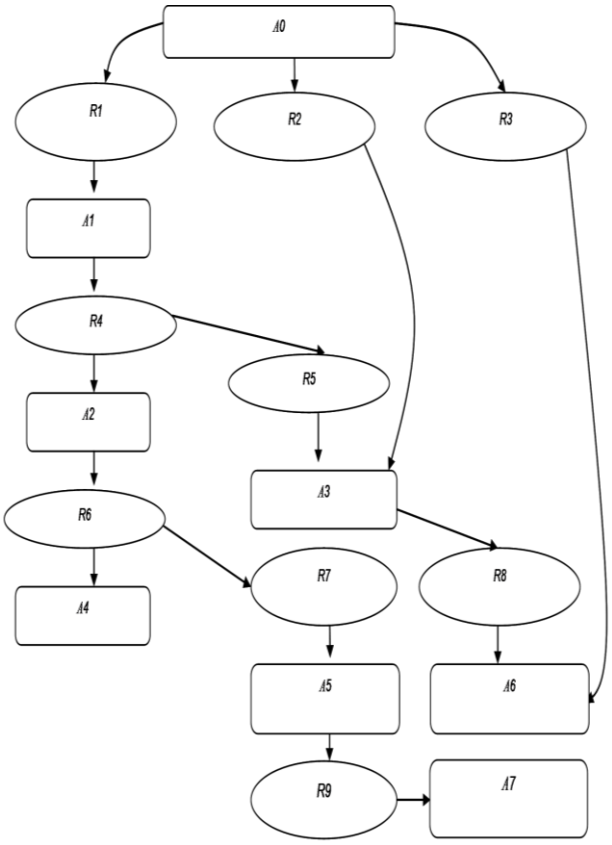

Fig. 1. Syntactic and semantic dependency tree

A0: Name =”make” and Semantic Type=”Verb’;

A1: Name=”transaction” and SemanticType =”Event”;

A2: (Name=”Purchase” or Name=”Purchase of the action” or Name=”purchase of actions of Lukoil”) and SemanticType=”Event”;

A3: Name=”Ivanov” and SemanticType=”Person Name;

A4: Name=”Petrov” and SemanticType=”Person Name”;

A5: (Name=”action” or Name=”actions of Lukoil”) and SemanticType=”Event”;

A6: Name = “In November 2003” and SemanticType=”Time”;

A7: Name=”Lukoil” and SemanticType=”Organization;

R1: RelationName =”argument” and RelationCase=”V”;

R2: RelationName=”argument” and RelationCase= “and” and RelationRole=”subject”;

R3: RelationName=”circumstance”;

R4: RelationName=”argument” and RelationCase=”D” and RelationConnector=”on”;

R5: RelationName=”argument” and RelationRole=”subject”;

R6: RelationName = “argument” and RelationCase= “R” and RelationConnector=”at;

R7: RelationName =”argument” and RelationCase =”R” and RelationRole=”object”;

R8: RelationName=”Accessory” and RelationCase =”R”.

After the 2nd stage, the set of syntactic and semantic dependency trees is formed. The most probable parsing variant is then chosen using heuristic algorithms.

Stage 3. Semantic analysis. Main task of semantic analysis is to extract facts from text and determine semantic connections between those facts. Basic structure of a fact is an action, usually represented in a sentence by a verb or participle, but in some cases by complex clauses — anaphoric verbal connections, noun clauses etc. Semantic analysis comprises three main substages. During the first substage objects get isolated and their properties get defined, fact objects are divided into key objects and additional objects. During the second substage the syntactical compatibility of each verb with isolated objects is determined and probable connections between objects corresponding to noun lexemes are derived. During the third substage objects get marked by roles and fact scenario is selected.

Semantic analysis:

- extraction of target objects from the text;
- building the logical scheme of the situation;
- classification of semantic networks;
- context analysis;
- chronological analysis, elimination of collisions.

The result of 3th stage of processing is a set of target objects with semantic connections. To extract target objects, it is necessary to build the logical scheme of the situation. Resulting set is then classified into categories.

Factual analysis of the text can find descriptions of situations corresponding to certain templates (rule of extraction of target objects), for example invention of address bus or stock purchase. Fact search is performed within semantic network of text document. Logical scheme of the situation identifies class of possible situations and contains slots that unambiguously separate it from other sit-

uations. Semantic analysis module fills the slots with values; some slots can be left empty. Further, semantic module relates the set of sentence semantic networks of text document with corresponding template than defines a fact, using modified decision tree algorithm C4.5 according to the situation templates in the knowledge base. Factual analysis also employs statistic methods to calculate frequency weight of target object or its connections in the document. Statistical methods are based on the calculation of TF*IDF.

ONTOLOGY AS A KNOWLEADGE MODEL

Knowledge bases can be built by subject area experts, but also automatically using Text Mining algorithms that extract data on objects, their attributes and connections using formal rules. Among tools for semantic analysis we could note Convera RetrievalWare and Russian Context Optimizer (for documents in Russian language). In this work we will discuss the factual analysis method of natural language texts, which allows to extract information on monitored objects using predefined semantic inference rules.

Let $X = \{x_i\}, i = \overline{1,n}$ is a finite set of semantic objects, where n is a number of objects in a knowledge model, and $\tilde{E} = \left\{\tilde{e}_k\right\}, k = \overline{1,m}$ is a set of semantic relationships, $A = \{a_i\}, i = \overline{1,p}$ is a set of possible attributes, and $\Psi = \{\psi_S\}, s = \overline{1,t}$ – set of inference rules. We will call tuple $\tilde{H} = \left\langle X, \tilde{E}, A, \Psi \right\rangle$ an ontology.

Semantic network of a sentence is an example of fact in its simple form. More complex facts could be defined through anaphoric or other references in the paragraph, text block, text document or collection of text documents.

We introduce here our own way of representing object relationships as relationship matrix. For ontology object x_α let's call a matrix:

$$R(\tilde{H}) = \| r_{i\,k} \|_{n \times m}$$

– the matrix of semantic relationships (Fig. 2), where n row vectors represent objects in ontology semantic network and m column vectors represent semantic relationships between ontology objects. Here $k = q'$ for integral relationships $R(\tilde{H})^{\Sigma}$ and $k = q' + j$ for differential relationships $R(\tilde{H})_j^{\Delta}$, $j = \overline{1,p}$, where p is a number of typical relationships, q is a basis of x_α object. Matrix elements are degrees of adjacency of ontology object x_α and other ontology objects x_i: $r_{i,k} = \mu_{R(\tilde{H})^{\Sigma-\Delta}}(x_\alpha, x_i)$. For each relationship $R(\tilde{H})^{\Sigma-\Delta}$, as defined by linguistic variable A, $B^1, B^2, \ldots, B^p$, relationship matrix contains a term of linguistic variable with the maximum value of affiliation function μ_F.

	$\mu_{R(\tilde{H})^{\Sigma}}(x_\alpha,x_i)/(x_\alpha,x_i)$ $\tilde{e}_{q'}$	$\mu_{R(\tilde{H})_1^{\Delta}}(x_\alpha,x_i)/(x_\alpha,x_i)$ $\tilde{e}_{q'+1}$	…	$\mu_{R(\tilde{H})_p^{\Delta}}(x_\alpha,x_i)/(x_\alpha,x_i)$ $\tilde{e}_{q'+p}$
x_α	0	0	…	0
x_β	$\vee_{q=1}^{7} A_q : \mu_{R(\tilde{H})^{\Sigma}}(x_\alpha,x_\beta)$	$\vee_{q=1}^{7} B_q^1 : \mu_{R(\tilde{H})_1^{\Delta}}(x_\alpha,x_\beta)$	…	$\vee_{q=1}^{7} B_q^p : \mu_{R(\tilde{H})_p^{\Delta}}(x_\alpha,x_\beta)$
x_χ	$\vee_{q=1}^{7} A_q : \mu_{R(\tilde{H})^{\Sigma}}(x_\alpha,x_\lambda)$	$\vee_{q=1}^{7} B_q^1 : \mu_{R(\tilde{H})_1^{\Delta}}(x_\alpha,x_\lambda)$	…	$\vee_{q=1}^{7} B_q^p : \mu_{R(\tilde{H})_p^{\Delta}}(x_\alpha,x_\lambda)$
…	…	…	…	…
x_δ	$\vee_{q=1}^{7} A_q : \mu_{R(\tilde{H})^{\Sigma}}(x_\alpha,x_\delta)$	$\vee_{q=1}^{7} B_q^1 : \mu_{R(\tilde{H})_1^{\Delta}}(x_\alpha,x_\delta)$	…	$\vee_{q=1}^{7} B_q^p : \mu_{R(\tilde{H})_p^{\Delta}}(x_\alpha,x_\delta)$

Fig. 2. Matrix of semantic relationships

Methods to extract facts from the knowledge base. In this work we will describe methods to extract facts from the knowledge base built on top of ontology. Let's consider following methods [3]:

- method based on the transitive relationships;
- method based on the object membership in a class;
- method based on the strength of the relationship;
- method based on the type of the relationship.

Method based on the transitive relationships defines the set of objects accessible from object x_i (analyzed object) with the help of fuzzy chains $\tilde{C}(x_i, x_{q+1})$ with maximal length q. Length of the chain should be limited because majority of semantic objects are transitively linked with each other in ontology through 5 to 7 relationships.

Method based on the object membership in a class determines membership of the object classes in a description of a semantic object $a_{class\ object}^{x_i}$ in classes of the knowledge model $c = \{c_1, c_2, \ldots, c_M\}, c_{i^2}, i^2 = \overline{1,M}$ for each depth level of transitive relationships, where M is a number of classes in a knowledge model.

Method based on the strength of the relationship in the general case if applied separately for each relationship $R(\tilde{H})^{\Sigma-\Delta}$ between semantic objects x_α, x_β. This method determines membership of relationship strength between objects

$a^{\tilde{e}_k}_{weight\ rel} = \mu_{R(\tilde{H})^{\Sigma-\Delta}}(x_\alpha, x_\beta)$ in a term of linguistic variable $(A, B^1, B^2, \ldots, B^p)$ corresponding to this relationship and having a connection measure between objects as a value. Additionally, for each term of a linguistic variable, a degree of affiliation could be determined, transforming fuzzy network of ontology relationships $\tilde{H}$ for given relationship $R(\tilde{H})^{\Sigma-\Delta}$ into fuzzy network of semantic relationships of level Ω $\tilde{H}^{R(\tilde{H})^{\Sigma-\Delta}}_{\Omega} = \left\langle X, \tilde{E}^{R(\tilde{H})^{\Sigma-\Delta}}_{\Omega} \right\rangle$.

Method based on the type of the relationship determines membership of the relationship type $a^{\tilde{e}_k}_{type\ rel}$ between semantic objects x_α, x_β in one of the standard relationship types in the knowledge model.

If semantic object x^l_i is a member of at least one class in a set, this object is selected for further analysis, otherwise the object gets deleted from the knowledge base. Classes in a c^l set are marked in a property $a^{x_i}_{class\ object}$ of semantic object.

Method "strength of the relationship" is applied separately for each relationship $R(\tilde{H})^{\Sigma-\Delta}$. Analyst determines the value of a linguistic variable for the given relationship:

for integral relationship $R(\tilde{H})^{\Sigma}: A_q$;

for differential relationship $R(\tilde{H})^{\Delta}_1: B^1_q$;

. . . . . ;

for differential relationship $R(\tilde{H})^{\Delta}_p: B^p_q$.

To effectively extract facts from the knowledge base, method based on the transitive relationship should be applied first; the order of application of other methods is not relevant. Fact extraction methods are controlled by the corresponding factor extraction parameters, and the query to the ontology knowledge base could be represented as a tuple:

$$I = (x^*_\alpha, q^*, \Gamma_1(x_\alpha)^*, \ldots, \Gamma^q_l(x_\alpha)^*, A^*, B^{1^*}, B^{2^*}, \ldots, B^{P^*}, R^{/^*}, X^*)$$

where x^*_α is parameter "semantic object for analysis";

q^* is parameter "length of fuzzy chain";

$\Gamma^q_l(x_\alpha)^* = < c^{l^*}_1, c^{l^*}_2, \ldots, c^{l^*}_s, x^*_\rho >$ is tuple of parameters, where $c^{l^*}_i$ is class of level l, x^*_ρ is semantic object in question;

$A^* = \left\langle A^*_q, \Omega^*, \Psi^* \right\rangle$ is tuple of parameters, where A^*_q – value of linguistic variable A, Ω^* is grade of membership, $\Psi^* = \{=, >, <, \geq, \leq\}$ – selection strategy for "strength of relationship" method;

$B^{P^*} = \left\langle B^{p^*}_q, \Omega^*, \Psi^* \right\rangle$ is tuple of parameters, where $B^{p^*}_q$ is value of linguistic variable B^p_q, Ω^* is grade of membership, $\Psi^* = \{=, >, <, \geq, \leq\}$ is selection strategy for "strength of relationship" method;

$R^{/^*} = \{R^{/^*}_1, R^{/^*}_2, \ldots, R^{/^*}_z\}$ $R^{/^*}_j$ is standard relationship;

$X^* = \{x^*_1, x^*_2, \ldots, x^*_y\}$, $x^*_j, j = \overline{1, y}$ is set of objects to exclude from analysis.

Method based on the transitive relationships allows getting detailed information on queried object (organization, person, product or technology) and its connections with other objects.

Problem definition:

Determine connections between legal and natural persons in the neighborhood of Viag with Rurghas.

Retrieve all existing connection paths between companies OGK-4 and E. ON in years from 1999 to 2011. Maximum path length greater or equal 2, path should consist of objects of type person or its derivaties. Return following fact types: stock purchase, energy production, investment, energy sell.

Example:

Get information about OGK-4 and E. ON and his relations with Russian energy companies in years from 2010 to 2012.

Resulting scheme (Fig. 3) is used to solve this analytic problem. After studying the structure of relationships between semantic objects analytic turns to natural-language texts containing the relevant information.

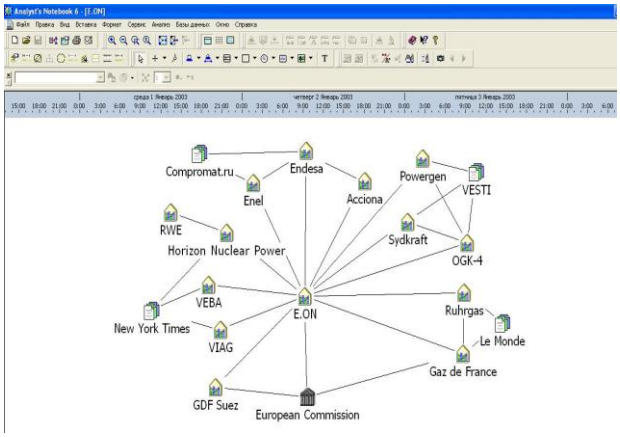

Fig. 3. Method "Transitive relationship"

CONCLUSION

In the context of research jobs on the given subjects taking place in laboratory of applied programming, the method of factual analysis have been received.

REFERENCES

1. **Fellbaum C.**, *WordNet an electronic lexical database.* Cambridge: MIT Press, 1998. Vol. 1.

2. **Vokhmintsev A., Melnikov A.** "The knowledge on the basis of fact analysis in business intelligence," in *Proc. Digital Product and Process Development Systems: IFIP TC 5 Int. Conf. NEW PROLAMAT 2013*. Dresden: Springer book, 2013, pp. 354-363.

3. **Felzenszwalb P., Girshick R., McAllester D., Ramanan D.** "Object detection with discriminatively trained part-based models," In *IEEE Pattern Analysis and Machine Intelligenc*e (*PAMI*). Los Alamitos, 2009.

METADATA

Title: The process of detecting facts in a natural text.

Authors: A. V. Vohkmintcev [1], A. V. Melnikov[2]

Affiliation:
Chelyabinsk State University (CSU), Russia.

Email: [2] vav@csu.ru.

Language: English.

Source: Vestnik UGATU (scientific journal of Ufa State Aviation Technical University), vol. 18, no. 5 (66), pp. 15-20, 2014. ISSN 2225-2789 (Online), ISSN 1992-6502 (Print).

Abstract: In this article we present a method of extracting factual data (objects, their attributes and relationships) from natural language texts using ontological knowledge base model. Subject area is represented through ontological model extended by fuzzy links between objects.

Key words: annotated image database; fact analysis; Image Net; Word Net; ontology; text mining; applied linguistics; extracting facts.

About authors:

VOKHMINTCEV, Aleksandr Vladislavovich, PhD. In 2002 defended candidate thesis "Fuzzy semantic hypernetwork method for extraction of structured knowledge from natural language texts." Since 2013 is a head of Research Laboratory "Intellectual information technology and systems" of Russian Academy of Sciences and Chelyabinsk State University. Scientific interests are related to development of methods for linguistic analysis of full-text databases and visualization of big data.

MELNIKOV, Andrey Vitalyevich, Professor. In 1985 defended his Candidate of Science thesis. In 1985 elected as head of department of electronic computers in Chelyabinsk Polytechnic Institute. In 1995 defended doctoral thesis "Design automation for distributed information processing systems based on further development of formal attribute grammar theory**"**, and in 1997 was awarded professorship. His scientific interests lie in the area of developing new technologies for processing and visualization of big data using artificial intelligence methods. Since 2004 he is pro-rector for research in Chelyabinsk State University.

МЕТДАННЫЕ

Заголовок: Обнаружение фактографических данных в текстах на естественном языке.

Авторы: А. В. Вохминцев [1], А. В. Мельников[2]

Организация: Челябинский государственный университет (ЧелГУ), Россия.

Email: [2] vav@csu.ru.

Язык: Английский.

Источник: Вестник УГАТУ. Т. 18, № 5 (66), pp.15-20, 2014. ISSN 2225-2789 (Online), ISSN 1992-6502 (Print).

Аннотация: Представлен метод извлечения фактографических данных (объекты и их атрибуты, связи) из текстовых документов на русском языке, основанный на использовании онтологических моделей знаний. Предметная область представлена в виде онтологии, расширенной за счет введения нечетких связей между объектами.

Ключевые слова: аннотированные базы данных изображений; фактографический анализ; Image Net; Word Net; онтологии; text mining; прикладная лингвистика; извлечение фактов.

Об авторах:

ВОХМИНЦЕВ Александр Владиславович, зав. науч.-иссл. лаб. интеллектуальных информационных технологий и систем РАН и ЧГУ. Канд. техн. наук (2002). Иссл. в обл. лингв. анализа полнотекстовых баз данных и методов визуализации больших данных.

МЕЛЬНИКОВ Андрей Витальевич, проректор по науч. работе, проф. Д-р техн. наук (1995). Иссл. в обл. обработки и визуализации больших данных на основе методов иск. интеллекта.