HIERARCHICAL THREE-LEVEL ONTOLOGY FOR TEXT PROCESSING

Victor Gladun, Vitalii Velychko, Leonid Svyatogor

Abstract: The principal feature of ontology, which is developed for a text processing, is wider knowledge representation of an external world due to introduction of three-level hierarchy. It allows to improve semantic interpretation of natural language texts.

Keywords: ontology, text processing, thematic text analysis.

ACM Classification Keywords: I.2.7 Natural Language Processing - Text analysis

Introduction

One of the practical purposes of an artificial intelligence methods use is *Text Processing* - the semantic analysis of naturally-language texts both general thematic orientations and concerning to various domains [1, 2, 3, 4]. Language thesauruses and linguistic ontologies are developed for this purpose and set the certain conceptually-expressed system of knowledge. Distinction between thesauruses and linguistic ontologies consists in volume knowledge representation and methods of classification (structurization) the conceptual environment. Among significant achievements in the field of lexicographical representation of natural languages it is possible to specify: Roget P.M. and Dornzaif F. thesaurus [5], the Ideographic dictionary of the Russian language [6], the Thesaurus of the Russian language *RuThes* [1], resource *WordNet* [7]. However, ontology is more adequate to the conceptual texts analysis, because by means of ontology connections between objects, concepts and its properties can be most full presented by all of them a variety.

As is known, ontology in a general view represents the sets of the domain terms, relations between terms and the domain interpretation functions on the terms and the relations. It is possible to point out the next advantages of ontologies: a) deep interaction between both described objects and phenomena and the contextual environment; b) the economical storage of the information demanding storing domain terms and relations, instead of stages memorizing; c) universal character of ontology, which allows within its structure to solve both problem – knowledge synthesis and analysis; d) "flexibility" of the knowledge structure which is adjusted on diverse domains.

Ontologies are developing, more often, for concrete domains ("narrow ontologies"), at that process begins with the analysis of a text collection [1, 2]. However, in a context of the given work the principal interest have linguistic ontologies ("wide ontologies"), which cover the knowledge continuum in different spheres of human activity, so as *Mikrokosmos* [8], *SUMO* [9] *and Knowledge Representation* by J. Sowa [10].

Two approaches to creation of ontologies. Topicality of an compromise

Now two opposite approaches to construction of the linguistic ontologies were formed. First of them – *"search ontologies"* – is purposeful on problems of automatic text processing in various domains. For example, Sociopolitical thesaurus is focused on the social and political life and used in such applications of automatic text processing as conceptual indexing, automatic text categorization [1]. The ontology synthesis technology is based on the analysis of the representative text collection (sometimes – tens thousands sources). Sizeable difficulties are in the fact that is impossible "to fish out" the *context knowledge* out of professional text collections, what is quite necessary to texts understanding. Therefore, it requires substantial experts' efforts for completion of the ontology by complementary concepts and to organisation semantic relationships between domain and external world. Moreover, in many cases the semantic communications do not lend its to formalization. As a result, it is impossible to avoid of essential intervention of expert. The final structure becomes too bulky in use and difficult for tuning.

On the other hand, there is an opposite approach. As a starting point of construction "abstract ontologies" consider *Universe* [10], *Essence* [9], *All* [8]. The knowledge representation is set by the branched out hierarchical structure. Concepts ontology reaches the maximal generalization and abstraction, and on this height, more often is remaining. By virtue of it, application of universal ontology to concrete texts analysis seems to be very problematic. Besides that, the formal relations used in abstract ontologies far not always describe the properties of real world in its reflection by man. Therefore, some authors introduce into consideration "flexible" relations: "*conceptual dependence relations*" [11], "*role relations*" [1], "*symmetric and asymmetrical associations*" [2] or "*associative relations*" [12]. Using of similar relations to thematic analysis NL text seems very constructive.

Thus, in a context of the lead analysis, the problem consists in separation between professional-focused and abstract ontologies, i.e. – between a particular description and abstract presentation of knowledge. Resources, which are available in a free access, do not allow to realize thematic text analysis in Russian. The way out from this opposition consist, as it seems, in synthesis "wide" and "narrow" ontologies. It is necessary to create an integrated structure, where it can be distributed and balanced described both the general, meaningful knowledge about nature and society, and concrete domain resources. Such structure can reflect a hierarchical picture of the whole world.

The purpose of the given work consists in offering the concept of integrated hierarchical (three-level) representation model of environment, which: a) in the compressed kind and with a different degree of generalization (or detailed elaboration) reflects actual knowledge about structure of an external world; b) is focused on text processing both the general subjects and separate areas of knowledge and c) allows to integrate a professional knowledge into a conceptual network without reorganization of the upper and middle ontology levels.

This conception develops the *semantic (thematic) analysis method* of NL text processing by creation of the document synopsis. Procedure of making a synopsis is based on disclosing of the given theme by means of sequence of keys words, which are automatically generated with the program "KONSPEKT" [13]. However, defect of this method was in strongly simplified one-level model of the real world representation, which was not taken into account hierarchy and depth of external world.

Substantiation of the approach to construction of hierarchical ontology

For construction of multilevel ontology, the methods and mathematical models that contain models of ontology, knowledge and domain are used [4]. Offered here three-level associations ontology is intended for the decision of more specific problem - the thematic texts analysis. It defines a number of preconditions and features.

1. *Gnosiological conceptions in ontologies. Paradigm acceptance*

There are many approaching methods to problem of universe representation in philosophy, natural sciences and linguistics. The authors of different universal ontologies describe a world with such general categories as: *Essence = Material, Abstract* [3, 9]; *All = Object, Event, Property* [3, 8]; *Universum* (is divided on seven components) [3, 10]. That is not exclude and others, exotic variants: for example, classification of *All* into *Goodness* and *Evilness* may be successful... The question of substantiation usually is not considered.

However, more practical and pragmatic methods recommend the scientific methodology of the system analysis. It operates the following types of resources: *Substance; Energy; Information; Man; Organisation; Space; Time* [14].

These categories, in our opinion, possess a necessary diversity and they are objects of researches in physical and social sciences. This world outlook may be taken as a base of description of the external world.

However, the most modern is the materialistic idea proposed by academician V.I. Vernadsky. In accordance with it, all, what is just known, may be divided on two fundamental categories: *Inert substance* and *Alive substance*. The *Alive substance* is realized in *Biosphere* and *Noosphere* - sphere of Human activity. Both categories characterize the *Matter* fundamentally. This materialistic paradigm is put in a basis of offered ontology. It works out in detail in partition "Choice of categories for upper level of ontology".

2. *A choice of the general structure*

In correspondence with hierarchical picture of the knowledge about the world, we are distinguishing three levels in ontology construction. On the upper level are the general categories of universe; here the strict taxonomy is possible. Upper level summarizes the concepts of general knowledge and reflects the ontological basis.

The middle level is disclosing the base terms in more detail - by concepts having lesser level of commonality. Concepts, used on middle level, reflect the universal, popular and well-known terms existing in nature, society and environment; there are presented relations between them also.

At last, on the lower level is concentrated knowledge, which characterizes concrete situations and describe some environment. At this level is presented knowledge of the problem-focused area. Therefore, on this level two intersected blocks exist: one of them contains the concepts and words of common usage, which exist in general, and interdisciplinary texts; other block serves the domains. Due to this two-blocks structure the ontology may grow at the expense of new domains. In principle, the domain block may be empty - the working capacity of ontology is remained fully.

Corresponding to hierarchical structure of ontology we will use the term **HiO**.

3. *A choice of connections*

At upper level **HiO** the formal connection of type «the whole - a part» is applied. At middle and upper levels, except the formal connections, following specific types of connections are entered into consideration:

a) **A** reveals through **B**; **B** explains **A**;

b) **A** is characterized by property **E**;

c) **A** is *associatively* connected with **C** .

Widely used in **HiO** the term "associative connection" is not formal. It is necessary to reflect individual semantic correlation of two (any) concepts if it takes place. Associations have a situational and dynamical character. At an *information level* they can be able to open unexpected properties (or laws) of some object (or phenomena). At a *functional level,* they fix some dependence. At a *cognitive level* associative connection of two concepts means that one image has excited another. At a *logic level* association are predicative and implicative, but in most cases – not transitive.

Constructive properties of associations ontology

Based on a practical orientation of the text processing, we shall specify some functional **HiO** features.

1. *Completeness and taxonomy* on upper level of the ontology signify that the chosen categories in aggregate are representing the Matter in the exhaustive way. Outside these categories, there should be no manifestation of the reality. Categories are subordinated to strict hierarchy and classification that excludes logic ambiguity of concepts.

2. *Natural-scientific lexicon.* The ontology categories and concepts should be common, simple and clear. They lexically expressed by those concepts and terms which were established in sphere of the general

knowledge, in natural and social sciences, in the socio-cultural environment. The upper level **HiO** can be supplemented with a special terms.

3. *Connectivity on association*. Connections between concepts inside of levels and between ones include both formal and informal (associative) connections. Associative connections are actively used at middle and upper levels of **HiO** to fuller description of theme.

4. *Antagonism reflection*. Concepts, reflecting properties, which have (inside of the given measure) its contrast, can be designated by pair's words (antonyms).

Synthesis of hierarchical associations ontology

On the base of the formulated preconditions there is clear the following prospect of actions. It is necessary to choose categories, concepts and connections between them and distribute on the upper, middle and lower levels of hierarchical associations ontology. Thus on upper level the simplified strict model of the world is presented. On the middle level, it is necessary to disclose the categories of upper level, using wide concepts of interdisciplinary dialogue. On the lower level, the conceptual basis of middle level is to be described in detail. In addition, here professional domains knowledge is localized. Received three-levels network ontology should be later connected to linguistic database and integrated in the working system of NL texts analysis. On the Fig. 1 general block diagram of three-level hierarchical associations ontology is presented.

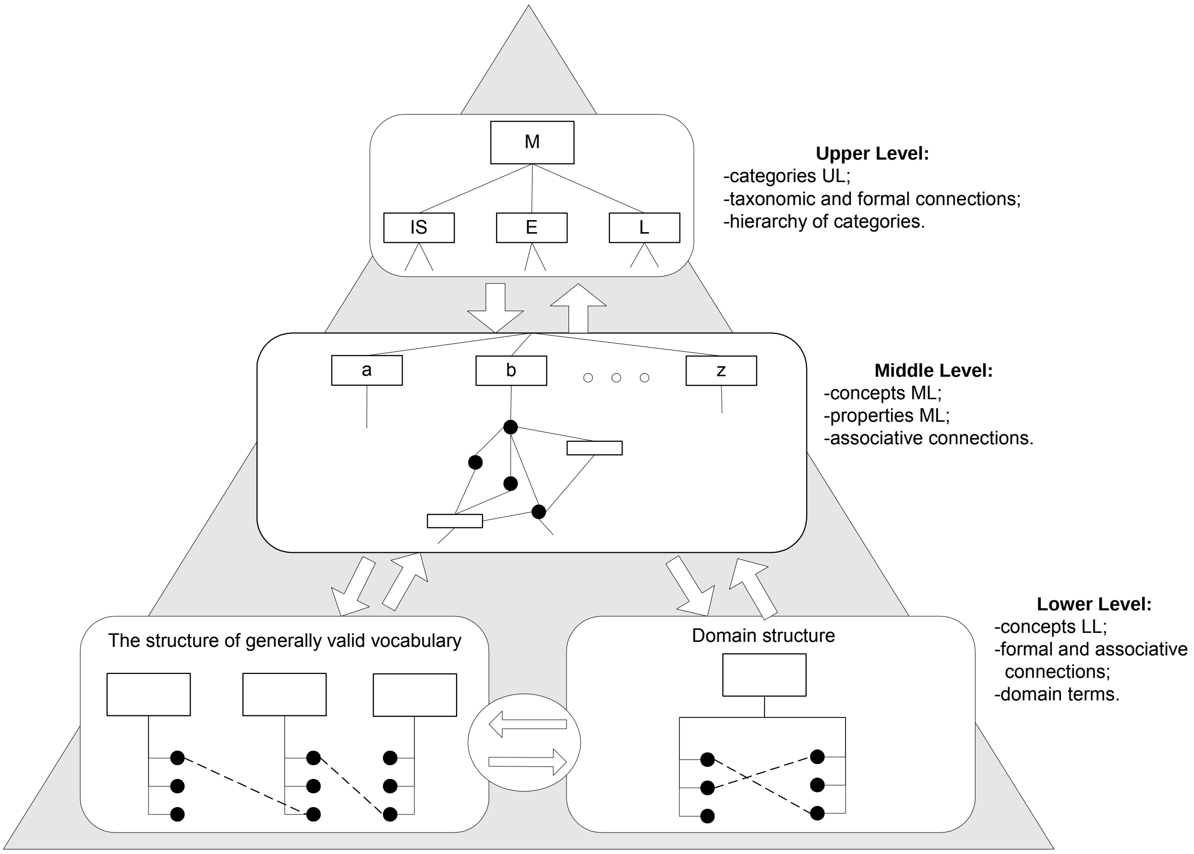
M
IS
E
L
Upper Level:
-categories UL;
-taxonomic and formal connections;
-hierarchy of categories.
a
b
z
Middle Level:
-concepts ML;
-properties ML;
-associative connections.
The structure of generally valid vocabulary
Domain structure
Lower Level:
-concepts LL;
-formal and associative connections;
-domain terms.

Fig1. The block diagram of three-level associations ontology.

Choice of categories for upper level of ontology

As it was specified above, within the limits of offered ontology the general picture of a Universe is subordinated to materialistic idea by V.I. Vernadsky. Apex of hierarchy is the philosophical category the *Matter*. It shows itself in exhaustive manner as *Inert* and *Alive substance*. On the other hand, the Inert substance may become the forms either *Substance* or *Energy* (which are passing each another under the Preservation Law). Therefore, we realize trichoyomy of the *Matter* on: *Substance* (inert), *Energy* and *Life* (the substance of *Alive*). In this case dividing is made on base "The Form existence of Matter". Each of three categories is presented by a number of subcategories, as is shown on Fig.2. It is necessary to add some explanations to the resulted ontological structure of upper level.

The first. At a pyramid of upper level, there are no such general categories of knowledge, as, for example: the Being, the Consciousness, the Measure, the State, the Property, the Quantity, the Quality and others. In **HiO** some of them are transferred on the middle level, owing to what they are released from philosophical sense and "work" as a terms of natural sciences.

The second. It is possible to show, that **HiO** upper level possesses property of completeness of conceptual volume. Really: a) *Substance, Energy* and *Life* are *the forms of the Matter realization*; b) the *Space* and *Time* serve as *the forms of the Matter distribution* and c) the *Reason* is *the way of the Matter reflection*. Summary these three metacategories are exhausting a metacategory "Being of the Matter". If to accept the given statement for an axiom, **HiO** covers all known (or real) properties of the Matter. As a result, the given categorical system on upper level is complete and closed.

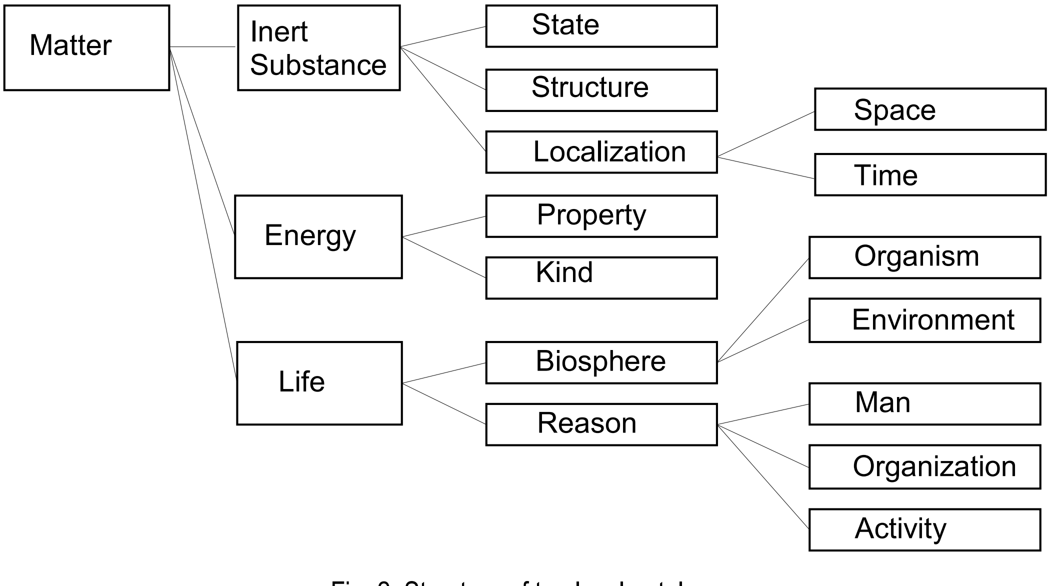

Fig. 2. Structure of top level ontology

In view of proposed above universal categories, concepts and comments the continuation of the ontological scheme may be formed.

The principles building the middle level of ontology

Purpose of the middle level of ontology (MLO) is, on the one hand, to disclose all categories of upper level - to give them semantic filling, and with another – to form a semantic environment for the concordance with the lower level's concepts.

MLO represents such level of knowledge, which is common to a various areas; that is interdisciplinary knowledge. Per se, the middle level of hierarchy fixes itself a layer of valid human knowledge, which is generalized by collective experience in science, culture, practice – out of professional sphere. It operates by generally accepted words. A material to this level is formed by the knowledge engineer. The middle level is "conservative": it is a "constant" **HiO**s component. At the given level informal (associative) connections of type "object – property", which (in opinion of the expert) bear the helpful information for disclosing internal structure of ontology, are actively used. It is necessary to emphasize, that occurrence doubtful connections is not lack of associations ontology, quite the contrary – they open an opportunity to additional adjustment.

Brief description MLO. Middle level of ontology represents set of network structures: as a name (and initial node) of each structure serves a category of upper level; internal nodes are the concepts of the middle level; internal connections between concepts disclose the important characteristical properties of category to be done.

Here, with a view of place economy, concepts and the full structures making middle level of **HiO** are not presented. However, as an example, the structure of the *Organization* category from the *Reason* cluster (see Fig. 3) is shown.

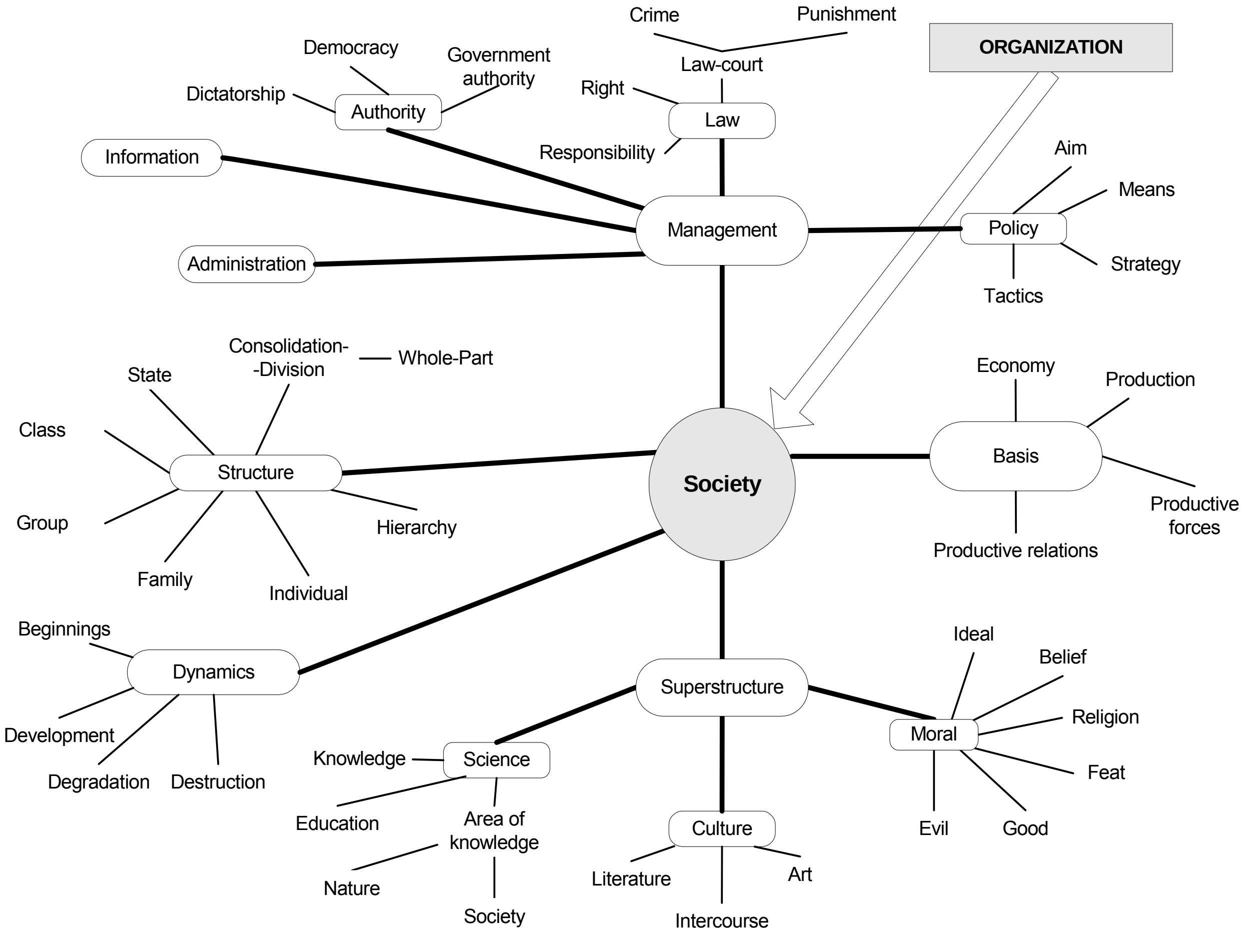

Fig. 3. A fragment of associations ontology structure of a middle level for a category "Organization".

The note. The interpretating opportunities of semantic analysis system, as a whole, essentially depend on successful construction of a middle part of the ontological structure. On the other hand, due to interaction of the middle level with upper one, text interpretation becomes more deeply and full: the explanatory resource **HiO** is used. If a new domain will be added, the middle level can be corrected.

The lower level of ontology

The **HiO** lower level is produced to bind together text keywords to concepts of the middle level of ontology: here the concepts of the middle level can be determined by means of text words. On this level, two blocks of the interconnected concepts are stipulated.

One of them – *the interdisciplinary block* – is intended for text processing of the general-thematic discourses. Other – *the domain block* – is synthesized under concrete domain. These two blocks are connected mutually and correlate to concepts of middle level too. When on an input of ontology appears some text interdisciplinary block always works; the domain block becomes more active on the professional text.

The choice of concepts field and connections between them for the lower level makes by engineer on knowledge and domain expert together. There are technologies to process linguistic resources, for example, the software complex "RuThes" and other [1, 2, 4] which allow to synthesize domain ontology. However, if to be oriented on more simple problem – the localization of the document theme – much more simple procedures can be offered.

One of them consists in indexing: the words and terms of common and professional glossaries (which are represented in database) can be connected with concepts of lower level of ontology. The other way is proposed by T.Taran: some situation or scene is determined by concept lattices [15].

After construction the lower level of ontology, **HiO** synthesis comes to an end. However, received ontology represents only the theoretic-descriptive scheme, instead of the analysis system. For a text processing, it is necessary to connect ontology with a special dictionary of natural language, which is contained in ones memory – database.

Text-processing procedure

The binding together of the text and the ontology is made through the dictionary of Russian. The dictionary reflects lexicon of a natural language. When the real NL text is analysed, first of all the set of keywords is discover. These keywords should make active some elements of the lower level of ontology. Which elements will be made active, will specify the list of the indexes, appropriated to dictionary elements. This list in an obvious kind sets associative connections between the given elements and concepts of the lower level. Hence, the concrete text word through the dictionary stimulates a subset of concepts and connections of the lower level, and through these concepts the signal transfers to middle and upper levels of **HiO**. As a result, in all ontological network automatically some semantic trajectory of an entrance word is localised.

This trajectory is possible to use in the text-processing system, namely: a) for deeper interpretation of the text, or b) – as an initial material for repeated, purposeful disclosing a theme in enriched context.

Conclusion

As a result of researches is developed hierarchical structure of three-level associations ontology, which differs by the following:

– unites in uniform structure the general categories of the description of the world (on upper level) with the conceptual environment of interdisciplinary knowledge (on a middle level) and with the topical concepts at the lower level. Ontology supposes inclusion of the new blocks – models of domains – without alteration of upper level and with expansion of a middle level of a network;

– the network model of associations ontology is the simple and constructive scheme, which allows to trace in the text the theme that was given. Synthesis HiO practically excludes greater expenditures of labour on viewing of the texts collection, because a priori is based on known natural-scientific knowledge;

– the HiO serves as a construct with well-founded basis of scientific general categories. At a level metaontology the bases of categories dichotomy are precisely well founded and their conceptual completeness is proved. For a semantic description of environment, informal (associative) connections are widely used. The semantic trajectories of the conceptual analysis, received as a result the text-processing, help to interpret any theme in a context of universal human knowledge.

Based on hierarchical three-level associations ontology the new version of "Konspekt" system is developing.

The literature

1. Dobrov B.V., Loukachevitch N.V. Linguistic ontology of natural sciences and technologies for applications in sphere of information search. Scientific notes of the Kazan state university. Series Physical and mathematical sciences. – 2007. Volume 149, book 2. pp. 49-73. In Russian.
2. Loukachevitch N.V., Nevzorova O.A. Aviaontology-2004: the analysis of a modern state of a resource // Computer linguistics and intellectual technologies: Works of the International seminar Dialogue'2004 (June, 2-7nd 2004г.) // Edited by I.M.Kobozeva, A.S.Narinjani, V.P.Selegej. – M.: Nauka – 2004. – Vol.2 - pp. 424-430. In Russian.
3. Palagin O.V., Petrenko M.G. Construction of abstract model of language- ontological information system. // Mathematical machines and systems. - 2007. - №1. - pp.42-50. In Ukrainian.
4. Artemieva I.L. Multilevel ontologies for domains with complicated structures. // Proceedings of the XIII-th International Conference “Knowledge-Dialogue-Solution” – Varna, 2007 Volume 2 Sofia, Institute of Information Theories and Applications FOI ITHEA, Bulgaria– 2007 pp. 403-410
5. Morkovkin V.V. Ideographic dictionaries. M.: Mosc. st. uni-ty, 1970. – 71 p. In Russian.
6. Baranov O.S. The Ideographic dictionary of Russian, M. – 2002, – 1200 p. In Russian.
7. Miller G., Beckwith R., Fellbaum C., Gross D. and Miller K. Five papers on WordNet. - CSL Report 43. Cognitive Science Laboratory, Princeton University, 1990.
8. Mahesh, Kavi. Ontology development for machine translation: ideology and methodology. Memoranda in Computer and Cognitive Science MCCS–96–292, New Mexico State University. 1996.
9. SUO, (2001), The IEEE Standard Upper Ontology web site, http://suo.ieee.org.
10. Sowa, John F. Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, © 2000. - 594pp.
11. Guarino N. Formal Ontology and Information Systems. In N. Guarino (ed.) Formal Ontology and Information Systems. //Proceedings of FOIS'98. - Trento, Italy. - 1998. - 6-8 June. – IOS Press, Amsterdam. – pp.3-15.
12. Grishina O.V. Information search in Internet // Intellectual technologies and systems. Collected articles of post-graduate and students. Issue 2 // Edited by Yu.N.Filippovich. – M.: Publishing house MGUP, – 1999. – pp. 18-24. In Russian.
13. Gladun V.P., Velychko V. Yu., Svyatogor L.A. Thematic analysis of natural language texts. // Computer linguistics and intellectual technologies: Works of International Conference “Dialogue 2006” (Bekasovo, 31 May – 4 June, 2006) // Edited by A.S.Narinjani, V.P.Selegej. – M.: RSUH Publishing Centre. – 2006. pp. 115-118. In Russian.
14. Kaziev V.M. Introduction in the analysis, synthesis and modelling of systems. – INTUIT.RU, Binom. Laboratory of knowledge. – 2006. – 248p. In Russian.
15. Taran T.A., Zubov O.A. Artificial intelligence: the theory and applications. Lugansk. – 2006. – 240 p. In Russian.

Authors' Information

Victor Gladun, Vitalii Velychko, Leonid Svyatogor – V.M.Glushkov Institute of cybernetics of NAS of Ukraine, Prospekt akad. Glushkova 40, 03680 Kiev, Ukraine; e-mail: aduis@rambler.ru