The Information and Analytical Using of Non-Structured Information Resources

Serhii Lienkov[1], Viacheslav Podlipaiev[2], Igor Tolok[3], Igor Lisitsky[4], Oleksii Fedchenko[5], Nataliia Lytvynenko[6] and Svitlana Kuznichenko[7]

[1,3,5,6] *Military Institute of Taras Shevchenko National University of Kyiv, Lomonosova Str., 81, Kyiv, 03189, Ukraine*
[2] *Research Institute of Geodesy and Cartography, Velyka Vasyl'kivs'ka Str., Kyiv, 03150, Ukraine*
[4] *Central Research Institute of Armament and Military Equipment of the Armed Forces of Ukraine, Povitroflots'kyy Avenue, Kyiv, 03049, Ukraine*
[7] *Dept.of Information Technologies Odessa State Environmental University, Odessa, Ukraine*

Abstract

Following research article describes the conditions for the formation of interactive knowledge bases, that are based on the formation of growing pyramidal networks in the analysis of textual narratives. The stability conditions of knowledge systems on the basis of their representation in the format of logical-linguistic models are determined. The authors also determined the conditions of atypical representation of linguistic constructs knowledge in the process of their transformation into a system. The use of lambda-calculus notation for the formation of stable logical-linguistic models of narrative descriptions is proposed.

Keywords

logical-linguistic model, growing pyramidal networks, concepts, linguistic constructs, term, knowledge, narrative.

1. Introduction

The use of modern information in the activities of various specialists today is quite deep interdisciplinary. Moreover, the use of various information resources in solving applied problems requires the availability of service-developed interactive knowledge bases. And the effectiveness of their use depends on the truth of the content, which is determined by the information component.

The practical main part of productive knowledge today is concentrated in the form of text descriptions. At best, these narratives have their digital image in the form of their presentation in the formats of various editors and means of displaying texts in computer systems. However, these digital images don't have interactive services. Therefore, it's quite important to create intelligent services that can turn these texts into structurally organized knowledge bases.

There is already the problem of using a large number of narratives, which should sufficiently expand intertextual connections. It allows to create a digital image of knowledge systems used in a single display format.

The first stage of the process of transforming narrative descriptions into the format of interactive knowledge bases that are able to interact with each other is the formation of logical-linguistic models of text descriptions.

ISIT 2021: II International Scientific and Practical Conference «Intellectual Systems and Information Technologies», September 13–19, 2021, Odesa, Ukraine EMAIL: lenkov_s@ukr.net (1); pva_hvu@ukr.net (2); igortolok@72gmail.com (3), igor.lisitsky@gmail.com (4), a_fedchenko@ecomm.kiev.ua (5), n123n@ukr.net (6), skuznichenko@gmail.com (7)
ORCID: 0000-0001-7689-239X (1); 0000-0002-7264-0520 (2); 0000-0001-6309-9608 (3); 0000-0002-1505-199X (4); 0000-0003-1343-3828 (5); 0000-0002-2203-2746 (6); 0000-0001-7982-1298 (7)

© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3126/paper12.pdf

2. Research results and analysis

2.1. The constructive of logical-linguistic models formation

The information base of any interactive knowledge system consists of different data types [1,2]. These data have certain functional properties and form a rather complex structure of interdependent relations. Moreover, the very information base of systems of this class is dual in nature - the data that make it up have certain logical relationships on the one hand, and also some of them are certain concepts and linguistic constructs (hereinafter concepts) on the other hand, so data have linguistic attributes [3]. The functionality of these data is displayed in the form of symbolic and numerical formulas, and we present certain sequences of computational operations [1-3]. The linguistic structures of these data are presented in the form of a sequence of certain words in the form of sentences, statements, etc. [2].

However, it should be noted that everything related to the data will be presented through the concept of the term [3]. It follows that each sequence of symbols of finite length (SSFL), including numbers, as well as their representation in the form of formulas, can be considered as a rule and can also be represented as a term. From these formulas-rules it's possible to form in the future certain linguistic structures of the formal kind that are displayed according to the syntax defined for them.

Further we will consider the final sequences of characters that are plural in nature, that is, they can be combined into plurals on certain grounds. Moreover, these sets can be represented as hierarchically related classes. Each such class includes sequences that have at least one common property [1, 3]. Such classes of SSFLs with properties form the certain topology, and therefore they can be represented as trees [2, 3]. One of such tree types is a growing pyramidal network (GPN) [4, 5]. Their attractiveness is the ability to automatically divide the SSFL into appropriate classes based on the specified properties of each SSFL.

The condition that SSFLs are divided into classes according to certain properties defines them as intentional [2], that is those that have signs-meanings, that we will define as the contexts of SSFLs. Then SSFLs that have a defined non-empty set of contexts will be defined as concepts and denoted by the variables $x, y, z, \ldots$ and the classes they form with letters $X, Y, Z, \ldots$ and so on. The presence of certain contexts in SSFL-concepts will be represented according to the notation of λ -calculus (lambda-calculus), namely - $X[\ ]$ [3]. The bracket $[\ ]$ is called "context holes". It's clear that the presence of the hole determines that the concepts aren't connected. Once we determine the term that can fill the hole, we get the connected SSFL terms.

Then all classes formed by SSFL concepts are extensional [3]. We'll define properties of SSFL-concepts by the letter r, and set of properties through R.

The hierarchical structures formed from SSFL in the form of GPN are marked trees. Their labels are SSFL concepts, that are class names, and SSFL-concepts, which aren't extensional, that is have only one semantic meaning. SSFL-concepts that have only one meaning can't be reduced, that is broken down into simpler concepts. Such SSFL concepts will be defined in the future as terminal [4, 5].

All SSFL-concepts form a certain set of names Σ, that are labels of all GPN nodes. Under such conditions, GPN is unique to the set of Bohm trees [1-3]. That is, the topology of the interaction of SSFL sets concepts can be represented as a set of Σ - labeled trees formed by GPN nodes.

$$\Sigma = \{X_1, X_2, \ldots, X_n, a_1, a_2, \ldots, a_m\}, \qquad (1)$$

where X_i – class of SSFL-concepts, a_j – terminal node (the non-extensional SSFL-concept).

Having determined the property classes $R1, R2, \ldots, Rm$, that implement the division of all GPN concepts into classes, and determine the relationship between the concepts, we obtain the corresponding GPN. According to [4-6], each GPN is a taxonomy.

Based on the condition formulated at the beginning, namely that an arbitrary type of SSFL is a term, it can be argued that all names of SSFL-concepts can form the set of terms Λ, that's represented in the notation of lambda calculus [3]. This allows us to consider all SSFL-concepts and their meanings nominally. This condition is met on the basis that all the SSFL-concepts presented in expression (1) aren't related by a strict ordering relationship. Moreover, when we move on to the GPN, it's always possible to distinguish many sets of SSFL-concepts, that also aren't related to the relationship of strict ordering.

We'll note also one more constructive property of GPN. Nodes that are hierarchically interconnected can form truth statements that can be calculated. Thus, based on the construction of the GPN from SSFL-concepts, a certain system of knowledge in terms of Λ-terms is formed. Its information base consists of certain linguistic structures formed from SSFL-concepts, that are terms. The values of these terms required for calculations are determined in the process of assigning them the appropriate contexts. This process is interactive. According to [3], each term representing the certain SSFL-concept will be represented in the form of the Bohm tree of the form (1). Then we can say the following - there is a meta-procedure that can turn the whole set of linguistic constructs into GPN, which is a composition of Bohm trees, that in turn is also a composition of many Λ-terms, formed by SSFL-concepts of the same GPN. Therefore, in fact, the set of Λ-terms can be represented as a certain interactive knowledge base (IKB).

It's clear that both functional data and linguistic structures that make up an interactive system of knowledge, that we present in the form of a set of Λ-terms, have certain relationships with each other, that is in a certain way logically and functionally characterize each other. Therefore, it's most effective for further consideration of the information base of arbitrary IKB to present in aggregate form, which is implemented in the form of the logical-linguistic models (LLM) class. This class of models is implemented on the basis of predicative representation of information structures of arbitrary type [7-15]. This allows us to consider them together in an arbitrary sequence without defining the relationship strictly and not strictly. Also, all LLM objects are atypical. This atypicality provides the definition of procedures that can jointly process the entire complex data structure that make up the information base of interactive knowledge systems. Then the whole set of such data will be defined as a separate class of atypical data, that allows to interpret as nominal [3, 4].

The predicativeness of the linguistic constructs of IKB, as the composition of Bohm trees, determines the nature of the formation of statements from the nodes of these trees. Moreover, the formation of GPN as the composition of Bohm trees is also predictive.

However, the process of LLM formation is realized on the basis of determining the order relation over certain sets of Λ-terms, that leads to the loss of the nominal value of their terms. It gives the calculation of the contextual meanings of the terms semantic nature and thus implements an interactive act of interaction with the information base.

$$\{X_1, X_2, ..., X_n, a_1, a_2, ..., a_m\} \to \Lambda \to$$
$$\to \psi \to \bar{\Sigma} \to \Sigma = \quad (2)$$
$$= \{X_1, X_2, ..., X_n, a_1, a_2, ..., a_m\};$$
$$\{X_1[\ ], X_2[\ ], ..., X_n[\ ]\} \to$$
$$\to \{X_1[B], X_2[D], ..., X_n[V, P]\} \to \Psi; \quad (3)$$
$$\Sigma = \{\perp\} U \{\lambda x_1, \lambda x_2, ..., \lambda x_n, \lambda a_1, \lambda a_2, ..., \lambda a_m\}, \quad (4)$$

where $\perp$ - the smallest element of all SSFL-context values; B, D, V, P - context values.

Expressions (2) - (4) reflect the generalized metaprocedure of IKB formation on the basis of definition of context values of SSFL-concepts and their transformation.

The introduction of the smallest value of the context and the definition of the contexts themselves passively determines the order relation over the set of λ-terms, and thus creates the conditions for the formation of the GPN Ψ. That is, expressions (2) - (4) are recursive.

It can then be argued that an arbitrary LLM has a nonempty structure of relationships between SSFL-concepts, which has a hierarchical form and can be represented as a tree. LLM is also an open structure. This means that the information base, the logical and linguistic characteristics of which it represents, can be supplemented at any time with the latest concepts and their relationships. The open nature of LLM determines that this class of models has the property of inductance. That is, their graph model in the form of a tree can grow due to the latest concepts and their relationships. One of the effective types of graph models of LLM is a growing pyramidal network (GPN) [4, 5]. Their positive distinguishing feature is the fact that an arbitrary GPN is equivalent to an arbitrary taxonomy of narrative description [1, 2, 6].

The attributes of the concepts that make up the GPN nodes can be contexts that describe their semantics; belonging to a certain thematic class, that is determined by their semantics; relations between concepts, etc. That is, the inductive process of forming the new nodes of the GPN can be represented as a sequence of statements that are formed on the basis of the contexts of each inductively active concept. Thus, in the process of forming GPN, as a structural reflection of LLM, the formation of logical expressions of a certain

set of statements is realized. Using the attributes of each concept of these statements, it's possible to form a formal expression in the form of a record of the algebra of statements calculus [3]. And the names in this expression will be the names of concepts. This determines that the GPN is structurally unique in the formula of the algebra of expressions, which is formed in terms of the concepts of the GPN, that are propositional variables, using logical operations: conjunction "Λ", disjunction "V", negation "$\neg$" and following "$\rightarrow$".

2.2. The operational components of text transformation processes

All constructs of LLM, namely: statements, chains of knots of GPN, logical formulas are certain terms. Linguistic constructs from terms have an atypical representation and can also have a propositional character, that determines the nominal value of SSFL-concepts, which are interpreted by formulas in the notation of statements algebra. Moreover, contexts that semantically define concepts that are propositional variables also characterize these concepts as dichotomous. This means that each statement that is formed on the basis of the concepts of the GPN is characterized by one of two meanings, that is to answer arbitrary questions in the format of "YES" or "NO".

For expressions (1) - (4), this means that they are significant in the case of "YES", and may not be taken into account in the case of "NO". That is, provided that the contexts of the GPN form a true expression formula (2) - (4), an interactive knowledge base is formed. If there is a case of "NO", which means that the true statements haven't been formed, IKB or a fragment of these GPN isn't implemented.

This greatly simplifies the formation of a training sample for an interactive knowledge system. It can be based on concepts whose significance in relation to the question of belonging to certain classes is true. That's, to the question of the existence of the certain certainty that the concept of GPN belongs to certain class or group of classes, we will always get the answer "YES". But it is clear that when the latest concepts are included in the GPN, we will receive answers not only "YES" but also "NO". And this determines the conditions for expanding the training sample of the intelligent system.

According to the homotopy type theory [1, 2], GPN is unilateral to the decision tree. Therefore, the representation of the GPN in the form of formulas with propositional variables, that are the concepts of the GPN, can be represented in the form of the certain decision tree. Each formula of propositional variables and logical operations that is formed when interacting with the LLM of the interactive knowledge base is determined by the hierarchy of the classification structure of the subject of interaction. Depending on the attributes of the concepts of active LLM, we obtain the value of belonging of the propositional variable to certain classes of concepts, and thus form a formal notation in the notation of the statements algebra and further in the form of GPN.

The atypical nature of expressions (1) - (4), including the case of defining the contexts of SSFL using propositional variables, means that the type of meaning of these contexts isn't important for calculations. They can be both numerical and non-numerical. Moreover, the logical expressions from propositional variables are quite stable to the order of their positioning in the formal expression, so they can occupy an arbitrary position in the record. Also, the values that they receive in the calculation don't require determining the relationship of strict or non-strict order. That's, transformations (2) - (4) are always able to determine the truth and objectivity of LLM values [12].

Thus, the GPN is the primary LLM taxonomy of the narrative of the document being processed. The training sample, which is the primary basis of the process of machine learning of the interactive knowledge base, is formed from the concepts of this narrative. Then formed on this basis, the GPN provides a systematic reflection of all the narratives that make up the primary information base of the interactive knowledge system. The systemology of the interactive knowledge base follows from the systemology of LLM and GPN. This provides a complete and correct interpretation of the properties of all the concepts that make it up. And as a consequence, it implements the solution of problems of classification of concepts that determine the latest nodes of GPN, diagnosing the states of all concepts on the basis of the formation of logical formulas in the notation of the statements algebra. Also, the systemology and dichotomy of propositional expressions from the concepts of GPN create conditions for predicting the presence of certain properties in the newly formed nodes of GPN.

Prediction in our view of LLM can have a truncated form of expression (2), which is supplemented by a representation of the form (6), namely:

$$\{X_1, X_2,..., X_n, a_1, a_2,..., a_m\} \to \Lambda \to$$
$$\to \psi \to \breve{\mathrm{T}} \to \Sigma = \quad (5)$$
$$= \{X_1, X_2,..., X_n, a_1, a_2,..., a_m\},$$
$$(\Sigma) \to \Sigma, \quad (6)$$

where the contexts for all SSFL-concepts are defined. In this case, the set of λ -terms includes certain functional expressions that implement predictive calculations [12, 16].

The decision tree, that is based on the relationship between the concepts of the GPN, is a composition of Bohm trees, and can be converted into a propositional expression. Its elementary expressions, within the conditions of the specific problem, take the meaning of "true" or "denial". The calculation of these values is realized on the basis of determining the degree of belonging of the attributes of the new concepts to the characteristic descriptions that make up the contexts of the educational sample.

Expressions (5) - (6) define not only different functionalities, but also the systemic stability of the latest concepts of GPN. To do this, the procedure of discretization of λ -terms set is determined, which implements the definition of the corresponding numerical scales, that consist of intervals characteristic of the contexts values of SSFL-concepts in a particular state. These procedures also take into account the frequency distribution of concepts in different classes, thereby increasing their classification features in the GPN, and as a consequence, systemic accuracy. Another consequence is the formation of more effective propositional expressions with the use of the latest concepts of the GPN, which are unique to the decision tree, and as a result define more stable systemic rules.

$$(\Sigma) \to BT(M) = \{\perp\} U \{\lambda x_1, \lambda x_2,..., \lambda x_n, \quad (7)$$
$$\lambda a_1, \lambda a_2,..., \lambda a_m \}$$

where $BT(M)$ according to [4] - the marked tree, M - the term which has solvability, that is all statements formed from its SSFL -concepts are true.

Thus, the interactive system of knowledge, that is implemented on the basis of the formation of GPN in the process of processing documents and narratives, is determined by the high stability of the systemic features of the GPN concepts and their relations. This is ensured by the following procedural interpretation of the properties of the GPNs themselves, as certain objects of a complex hierarchical structure.

1) Formation of propositional expressions in the notation of the algebra expressions that determine the classes of GPN concepts based on the optimal definition and selection of attributes combinations that are significant in the interval of a certain scale. At the same time, due to the application of the operation "negation", the procedure of minimizing the descriptions length of each class defined in the GPN is also implemented.

2) Reliable classification of all concepts included in the training sample for GPN, and as a consequence of the formation of propositional expressions that dynamically reveal the patterns of both relevant classes of concepts and the relationship between them, while regulating the compactness of the training sample, excluding quality assessment of patterns, that were discovered.

3) Defining the membership function, which implements the mechanisms of fuzzy logic in calculating the characteristic characteristics of GPN concepts and their classes, and obtaining clear and fuzzy levels of reliability and their ranks, the validity of attribute features of concepts and their properties and relationships, including zero value type "I don't know".

All these procedural actions ensure the formation of GPN and on its basis LLM, that determines the functional structure of the interactive knowledge system. Based on them, the linguistic-semantic and conceptual analysis and processing of multilingual natural-language narrative descriptions are realized in the environment of the specified system. The selection of linguistic constructs of different length and complexity, identification and selection of intercontextual relations for all concepts that determine the semantic features of GPN and LLM, including the educational sample, is provided.

GPN and as a consequence of LLM, that are built on the basis of the above-described machine learning procedures, are characterized by the property of inductance. The further development of GPN, based on the encapsulation of new concepts, also expands the set of propositional expressions, that are in fact certain linguistic constructs, built on the application of logical operations to disordered elementary records -

statements that don't have logical operators inside.

This is functionally represented by expressions (2) - (7). When forming Bohm trees of the form (4) under conditions that the contexts of their nodes determine only the true values, we implement recursion from expressions (2) - (4).

The identification of intercontextual relations in the process of the latest concepts encapsulation and further inductive growth of the pyramidal network, realizes the discovery of new statements as systems of knowledge. The intercontexts of the relationship are revealed through the logical operation "conjunction", and the direct growth of GPN is realized by the use of logical operations "disjunction", "negation" and "following" both direct and reverse.

If we apply the rule of Godel's theorem on incompleteness [4], we can determine that no matter how many concepts aren't encapsulated in GPN and LLM, and no matter how many of their contexts in GPN aren't related, GPN, LLM and interactive knowledge system are never will be complete. The result is the formation of indeterminate nodes, which are the result of applying the "conjunction" operation to selected sets of the training sample.

All undefined nodes are concepts of complex structure. Their concepts, like linguistic constructs, have logical operations inside them and can therefore take the form of complex statements. Then such concepts can also be presented in the form of propositional expressions, that are able to define and classify the latest concepts with a complex structure.

Also, uncertainty concepts based on the use of inductance properties implement the clustering procedure, that provides identification of semantically equivalent concepts and their classes. The degree of this equivalence is determined based on the application of the membership function. Depending on the significance of the degree of equivalence, the concepts of uncertainty either form the newest class or are included in an existing thematic class.

After all, the measure of equivalence allows us to apply the rule of logical inference "following" by analogy. With a predetermined degree of equivalence, it's possible to draw conclusions about the belonging of new concepts and their classes to those already defined, and also to determine the degree of certain statements validity that are formed on the basis of concepts whose contexts are relevant.

3. Conclusions

The methodology and formation of growing pyramidal networks constructively ensures the transformation of narrative texts into the format of interactive knowledge bases. GPNs are able to determine the conditions for the stability of information databases of interactive knowledge systems, to implement the transformation into their formats of unstructured narrative descriptions of various types, from scientific articles to catalogs of scientific and technical products, monographs and more.

The conceptual basis of such transformations in the form of atypical expressions provides the implementation of intellectual services for processing narratives by means of linguistic-semantic and conceptual analysis with their subsequent transformation into the format of logical-linguistic models and interactive knowledge bases.

4. References

[1] V. Voevodsky, Univalent Foundations of Mathematics: Proceedings of Logic, Language, Information and Computation, WoLLIC 2011, in: Lev D. Beklemishev, Ruy de Queiroz (Eds.), Lecture Notes in Computer Science, volume 6642, Berlin, Heidelber, Springer, 2011, p. 311. doi:10.1007/978-3-642-20920-8.

[2] Homotopy Type Theory: Univalent Foundations of Mathematics, Princeton: Institute for Advanced Study, 2013, 603 p.

[3] van Dalen, Dirk (2013). Logic and Structure. Universitext. Berlin: Springer. doi:10.1007/978-1-4471-4558-5.

[4] Dybjer, Peter & Kuperberg, Denis. (2012). Formal neighbourhoods, combinatory Böhm trees, and untyped normalization by evaluation. Ann. Pure Appl. Logic. 163. 122-131. 10.1016/j.apal.2011.06.021.

[5] V. Gladun, Processes of formation of new knowledge, Sophia, SD "Teacher 6", 1994.

[6] O. Strizhak, Transdisciplinary integration of information resources (Information Technology), PhD thesis, The National Academy of Sciences of Ukraine, Inst. of Telecommunications and Global. inform. Space, Kyiv, 2014.

[7] M. Dymarsky Ways to embody the predicative relationship: Acta Linguistica Petropolitana, in: Proceedings of the Institute

of Linguistic Research of the Russian Academy of Sciences, in: N. N. Kazansky (Eds.), T. XI. Part 1, Categories of nouns and verbs in the system of functional grammar, in: M. D. Voeikova, E. G. Sosnovtseva (Eds.), Nauka, 2015, pp. 41-62.

[8] Understanding Predication. Series: Studies in Philosophy of Language and Linguistics / Edited By Piotr Stalmaszczyk // Peter Lang: Frankfurt am Main, Bern, Bruxelles, New York, Oxford, Warszawa, Wien, 2017. – 292 pp. DOI: https://doi.org/10.3726/b11243.

[9] O. Kulbabska, Modern interpretations of the category of predication in linguistics, Ukrainian language, № 1, 2009, p. 61-73. ISSN 1682-3540.

[10] Ivanova K.B., Vanhoof K., Markov K., Velychko V. Introduction to the Natural Language Addressing International Journal "Information Technologies & Knowledge". 2013. Volume 7, Number 2. p. 139-146.

[11] N. B. Cocchiarella, Philosophical Perspectives on Formal Theories of Predication. In: Handbook of Philosophical Logic. Synthese Library (Studies in Epistemology, Logic, Methodology, and Philosophy of Science), volume 167, Springer, Dordrecht, 1989, pp. 253-326. doi.org/10.1007/978-94-009-1171-0_3.

[12] V. Velichko, Logical-linguistic models as a technological basis of interactive knowledge bases, International Journal "Information Models and Analyses", volume 8, number 4, 2019, pp. 3 25-340.

[13] O. Stryzhak, S. Dovgyi, M. Popova, R. Chepkov, Transdisciplinary Principles of Narrative Discourse as a Basis for the Use of Big Data Communicative Properties, in: Arai K. (Eds) Advances in Information and Communication, FICC 2021, Advances in Intelligent Systems and Computing, volume 1364, Springer, Cham, 2021. doi.org/10.1007/978-3-030-73103-8_17.

[14] O. Stryzhak et al. Decision-making System Based on The Ontology of The Choice Problem, J. Phys.: Conf. Ser. 1828 012007, 2021. doi:10.1088/1742-6596/1828/1/012007.

[15] S. Dovgyi, O. Stryzhak, Transdisciplinary Fundamentals of Information-Analytical Activity, in: M. Ilchenko, L. Uryvsky, L. Globa (Ed.), Advances in Information and Communication Technology and Systems. MCT 2019, Lecture Notes in Networks and Systems, volume 152, Springer, Cham, 2021. doi.org/10.1007/978-3-030-58359-0_7.

[16] A. Gonchar, O. Strizhak, L. Berkman. Transdisciplinary consolidation of information environments, Communication, № 1 (149), 2021, pp. 3–9.