The Fuzzy Ontology as a Core of the Knowledge Base for the Technical Electronic Archive

A. M. Namestnikov
Ulyanovsk State Technical University, Ulyanovsk, Russia
e-mail: am.namestnikov@gmail.com

***Abstract*—The article describes the structure of the ontology presented on the basis of the fuzzy extension of the description logic $\mathcal{SHOIN}(\mathcal{D})$. A set of metalevels of the ontology reflects the specificity of the domain – development of complex automated systems. The results of computational experiments are presented.**

***Keywords*—ontology, electronic archive, description logic, fuzzy sets, knowledge base.**

I. Introduction

Development of intelligent electronic archives of design organizations entails construction of knowledge bases that represent integral parts of any intelligent system. According to the up-to-date point of view on constructing intelligent information systems presented in [1],[2],[3], the ontology can be considered as a tool for expert knowledge representation. Nowadays, a wide range of languages for representing applied ontologies are known. Taking into account the W3C (The World Wide Web Consortium) support of the languages based on the OWL group on the level of standards, $\mathcal{SHOIN}(\mathcal{D})$ formalism will be used as a logic basis of the description language for the ontology of automated system (AS) designing information support [4]. The description logic $\mathcal{SHOIN}(\mathcal{D})$ has a lot of possibilities of representing the domain model. Nevertheless, in order to represent knowledge about semistructured information resources, it is not quite sufficient to use this formalism. The natural language features and incompleteness in description of classes, entities, and relationships between them in project diagrams require to use formalisms that are able to work with fuzzy and incomplete data. One of the extensions of $\mathcal{SHOIN}(\mathcal{D})$ is $fuzzy\mathcal{SHOIN}(\mathcal{D})$ formalism (see [5], [6]), combining linguistic possibilities of the basic description logics and advanced mathematical tools of the fuzzy set theory.

II. The Ontology Structure

The domain of complex ASs designing imposes requirements to the structure of an applied ontology [7] . The specificity of the structure and content of information resources of electronic archives and project activity taken as a whole brings about the necessity of constructing the ontology including the set of metalevels shown in Fig. 1.

Formally, the set of the electronic archive ontology components may be written as the following sequence:

$$O = \langle PL, DL, CL, AL, R, F \rangle,$$

where PL is a metalevel of projects including information about implementing projects i.e. the taxonomy of projects

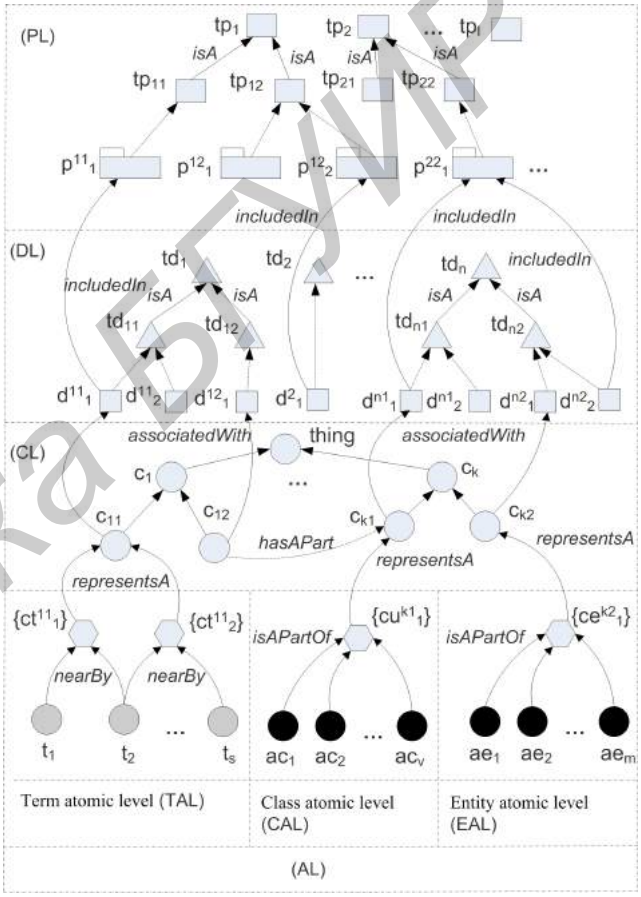

Figure 1. The structure of ontology for the electronic archive

classes and instances including technical documents; DL is a metalevel of documents including the taxonomy of documents classes and instances; CL is a metalevel of concepts based on the taxonomy of concepts related to the domain of the design organization and implementing projects; such relationships as «hasAPart», «associatedWith» an others are additionally used; AL is an atomic metalevel including term atomic level (TAL), class atomic level (CAL), entity atomic level (EAL); R is a set of relationships between concepts and/or instances related to different metalevels of the ontology.

Separation of the atomic metalevel into three ones $\{TAL, CAL, EAL\}$ (Fig. 1) means that in the procedures of conceptual design of AS, the intelligent analysis is carried out for the following electronic archive information resources: textual technical documents, class diagrams, fragments of

program subsystems models and data models.

III. The Knowledge Base Description

In the context of the description logic $\mathcal{SHOIN}(\mathcal{D})$, an ontology represents the knowledge base defined by the following equation $\mathcal{KB} = \{TBox, ABox\}$,

where $TBox$ is a set of terminological axioms representing the common knowledge about the concepts of the design organization electronic archive and their relationships; $ABox$ is a set of statements (facts) about the individuals.

Taking into account the ontology structure, let us denote $TBox^{arch}$ as the terminology of the project archive, $TBox^{dom}$ as the terminogy of the design organization domain; $ABox^{arch}$, $ABox^{dom}$ as the corresponding sets of facts: $TBox = TBox^{arch} \cup TBox^{dom}$, $ABox = ABox^{arch} \cup ABox^{dom}$.

$TBox^{arch}$ (accordingly, $ABox^{arch}$) includes terminology (facts) of the metalevels of projects and the ontology documents (Fig. 1). The metalevel of concepts and atomic metalevel defines as $TBox^{dom}$ and $ABox^{dom}$.

Let us write the $TBox^{arch}$ terminology content based on the ontology structure (Fig. 1).

$\mathbf{TBox^{arch}}$ terminology:

$$\begin{array}{ll} tp_{11} \sqsubseteq tp_1 & tp_1 \sqsubseteq tp \\ tp_{12} \sqsubseteq tp_1 & tp_2 \sqsubseteq tp \\ tp_{21} \sqsubseteq tp_2 & \vdots \\ tp_{22} \sqsubseteq tp_2 & tp_l \sqsubseteq tp \end{array}$$

$$tp \equiv \top \sqcap \leq 1 hasATypePrjName.String,$$

here $hasATypePrjName$ is the name for the functional role «has a name of the project type», $String$ is a concrete domain of the string type.

The concept «Project» can be defined as

$$P \equiv \top \sqcap \leq 1 hasAPrjName.String \sqcap$$
$$\sqcap \leq 1 hasADeveloperName.String \sqcap$$
$$\sqcap \exists hasAInitialDate.Date \sqcap \exists hasAType.tp,$$

here $hasAPrjName$, $hasADeveloperName$, $hasAInitialDate$, $hasAType$ are the names for the corresponding roles «has a project name», «has a developer name», «has the initial date of the project», «has a type». $Date$ is the concrete domain of data type.

$$\begin{array}{ll} td_{11} \sqsubseteq td_1 & td_1 \sqsubseteq td \\ td_{12} \sqsubseteq td_1 & td_2 \sqsubseteq td \\ \vdots & \vdots \\ tp_{n1} \sqsubseteq td_n & td_n \sqsubseteq td \\ tp_{n2} \sqsubseteq td_n & \end{array}$$

$$td \equiv \top \sqcap \leq 1 hasADocTypeName.String,$$

here $hasADocTypeName$ is the name for the functional role «has the name of the document type».

The concept « Document » can be defined as

$$D \equiv \top \sqcap \leq 1 hasADocDecimal.String \sqcap$$
$$\sqcap \exists hasAAuthor.String \sqcap$$
$$\sqcap \exists hasADate.Date \sqcap$$
$$\sqcap \exists hasAType.td \sqcap \forall includedIn.P,$$

here $hasADocDecimal$, $hasAAuthor$, $hasADate$, $hasAType$, $includedIn$ are the names for the corresponding roles «has a decimal number», «has an author», «has a date», «has a type» and «included in».

The set of **$\mathbf{ABox^{arch}}$ facts:**

$$\begin{array}{ll} p_1^{11} : P & \langle p_1^{11}, tp_{11} \rangle : hasAType \\ p_1^{12} : P & \langle p_1^{12}, tp_{12} \rangle : hasAType \\ p_2^{12} : P & \langle p_2^{12}, tp_{12} \rangle : hasAType \\ p_1^{22} : P & \langle p_1^{22}, tp_{22} \rangle : hasAType \\ & \\ d_1^{11} : D & \langle d_1^{11}, td_{11} \rangle : hasAType \\ & \langle d_1^{11}, p_1^{11} \rangle : includedIn \\ \vdots & \vdots \end{array}$$

$\mathbf{TBox^{dom}}$ terminology:

In case of defining $TBox^{dom}$ terminology, the use of the concrete domain is not quite sufficient. The question at issue is defining the degree of expression of the ontology concepts (in the context of the metalevel of terms) in the documents of the design organization electronic archive. Each concept c_i can relate to any document fragment d_j with different membership degrees. For this purpose, fuzzy predicates with predefined membership functions are offered to use.

The trapezoidal and triangular functions, the L-functions and R-functions are not only computationally simple but most frequently used to specify membership functions of fuzzy variables. In this paper, the functions are defined on $[0, 1]$. The trapezoidal function $trz(x; a, b, c, d)$ is defined as follows: suppose $a < b \leq c < d$ d from the set of $[0, 1]$, then:

$$trz(x; a, b, c, d) = \begin{cases} 0, & \text{if } x \leq a; \\ (x-a)/(b-a), & \text{if } x \in [a, b]; \\ 1, & \text{if } x \in [b, c]; \\ (d-x)/(d-c), & \text{if } x \in [c, d]; \\ 0, & \text{if } x \geq d. \end{cases}$$

The triangular function $tri(x; a, b, c)$ is defined as

$$tri(x; a, b, c) = \begin{cases} 0, & \text{if } x \leq a; \\ (x-a)/(b-a), & \text{if } x \in [a, b]; \\ (c-x)/(c-b), & \text{if } x \in [b, c]; \\ 0, & \text{if } x \geq c. \end{cases}$$

The *L-function* $L(x; a, b)$ is defined as

$$L(x; a, b) = \begin{cases} 1, & \text{if } x \leq a; \\ (b-x)/(b-a), & \text{if } x \in [a, b]; \\ 0, & \text{if } x \geq b. \end{cases}$$

Finally, the *R-function* $R(x; a, b)$ is defined as

$$R(x; a, b) = \begin{cases} 0, & \text{if } x \leq a; \\ (x-a)/(b-a), & \text{if } x \in [a, b]; \\ 1, & \text{if } x \geq b. \end{cases}$$

Let us write the terminology of the concept metalevel (the structure is shown in Fig. 1).

$$\begin{array}{ll} c_{11} \sqsubseteq c_1 & c_1 \sqsubseteq c \\ c_{12} \sqsubseteq c_1 \sqcap \exists hasAPart.c_{k1} & \vdots \\ c_{k1} \sqsubseteq c_k & c_k \sqsubseteq c \\ c_{k2} \sqsubseteq c_k & \end{array}$$

$$c \sqsubseteq \top \sqcap \forall associationWith.D \sqcap$$
$$\sqcap(\exists hasAExpValue.High \sqcup \exists hasAExpValue.Middle \sqcup$$
$$\sqcup \exists hasAExpValue.Low)$$

$$c^{exp} \sqsubseteq c \sqcap \exists hasAExpValue.High,$$

here $hasAPart$ is a name for the role «has a part», $hasAExpValue$ is a name for the role «has a value of a degree of expression». $High$, $Middle$ and Low are the fuzzy concrete predicates defined as

$$High, Middle, Low : [0,1] \to [0,1].$$

The individual concept c^{exp} represents the ontology concept with high degree of expression in any document.

The parametrically fuzzy predicates are defined as follows:

$$\begin{array}{rcl} Low(x) & = & L(x; 0.2, 0.4); \\ Middle(x) & = & trz(x; 0.2, 0.4, 0.6, 0.8); \\ High(x) & = & R(x; 0.6, 0.8). \end{array}$$

Let us define the $TBox^{dom}$ terminology related to the atomic metalevel and associated with the terminology of concept metalevel terminology as follows:

$$\begin{array}{c} \{ct_1^{11}\} \equiv \top \sqcap \leq 1 representsA.c_{11} \\ \{ct_2^{11}\} \equiv \top \sqcap \leq 1 representsA.c_{11} \\ \{cu_1^{k1}\} \equiv \top \sqcap \leq 1 representsA.c_{k1} \\ \{ce_1^{k2}\} \equiv \top \sqcap \leq 1 representsA.c_{k2} \\ T \equiv \top \sqcap (\exists nearBy.\{ct_1^{11}\} \sqcup \exists nearBy.\{ct_2^{11}\}) \\ \{cu_1^{k1}\} \equiv \top \sqcap \leq 1 representsA.c_{k1} \\ AC \equiv \top \sqcap \exists isAPartOf.\{cu_1^{k1}\} \\ \{ce_1^{k2}\} \equiv \top \sqcap \leq 1 representsA.c_{k2} \\ AE \equiv \top \sqcap \exists isAPartOf.\{ce_1^{k2}\} \end{array}$$

The set of $\mathbf{ABox^{dom}}$ **facts:**

$$\begin{array}{ll} ct_1^{11} : \{ct_1^{11}\} & \langle t_1, ct_1^{11} \rangle : nearBy \\ ct_2^{11} : \{ct_2^{11}\} & \langle t_2, ct_1^{11} \rangle : nearBy \\ cu_1^{k1} : \{cu_1^{k1}\} & \langle t_2, ct_2^{11} \rangle : nearBy \\ ce_1^{k2} : \{ce_1^{k2}\} & \langle t_s, ct_2^{11} \rangle : nearBy \\ t_1 : T & \langle ac_1, cu_1^{k1} \rangle : isAPartOf \\ t_2 : T & \vdots \\ \vdots & ct_1^{11} : c^{exp} \geq 0.75 \end{array}$$

$$\begin{array}{ll} ae_m : AE & ct_2^{11} : c^{exp} \geq 0.6 \\ & cu_1^{k1} : c^{exp} \geq 0.8 \\ & ce_1^{k2} : c^{exp} \geq 0.7 \end{array}$$

The facts as $a : C \geq \eta$ mean that the instance a pertains to the concept C with the membership degree not lower than threshold η.

IV. The Conceptual Index of the Electronic Archive

Suppose $C = \{c_i\}$, $i \in I = \{1, 2, 3, ..., n\}$ is a finite set of the domain concepts fixed in the ontology; $D = \{\tilde{d_j}\}$, $j \in J = \{1, 2, 3, ..., m\}$ is a family of fuzzy subsets in C. The pair $\widetilde{CI} = (C, D)$ is called a fuzzy nonoriented hypergraph if $\tilde{d_j} \neq \emptyset, j \in J$ and $\bigcup\limits_{j \in J} \tilde{d_j} = C$; herewith, $c_1, c_2, ..., c_n \in C$ are the graph vertices and a set D containing $\tilde{d_1}, \tilde{d_2}, ..., \tilde{d_m}$, is a set of fuzzy edges of the hypergraph.

Taking into account that an individual document has its ontological mapping as a result of conceptual indexing, a set $D = \{\tilde{d_j}\}$ can be defined as a set of documents in the conceptual index, $\tilde{d_j}$ is an individual ontological representation of the j-th document. We get that the fuzzy ontology hypergraph

$$\widetilde{CI} = (C, D) \tag{1}$$

formally defines the conceptual index of the document base.

The two concepts c_α and c_β (hypergraph vertices) of the conceptual index are fuzzy adjacent if a document (fuzzy hypergraph edge) exists and the document includes both notions where a degree of adjacency for the c_α and c_β concepts can be defined as follows:

$$\mu(c_\alpha, c_\beta) = \bigvee\limits_{d_i \in D} \mu_j(c_\alpha, c_\beta), \text{ where} \tag{2}$$

$$\mu_j(c_\alpha, c_\beta) = \mu_{d_j}(c_\alpha) \& \mu_{d_j}(c_\beta).$$

The value $1 - \mu(c_\alpha, c_\beta)$ represents the distance between the c_α and c_β concepts on the basis of the document content.

The index can be used for specifying the user's project query to the archive of documents in case when the concept of user's interest exists in the query but the result are leaved to be desired. In order to specify the query, the text input of the concept that has the shortest distance to the initial one is used.

Two documents $\tilde{d_\gamma}$ and $\tilde{d_\delta}$ are fuzzy adjacent if $\tilde{d_\gamma} \cap \tilde{d_\delta} \neq \emptyset$, moreover,

$$\mu(\tilde{d_\gamma}, \tilde{d_\delta}) = \bigvee\limits_{c \in (d_\gamma \cap d_\delta)} \mu_{d_\gamma \cap d_\delta}(c) \tag{3}$$

is a degree of adjacency between $\tilde{d_\gamma}$ and $\tilde{d_\delta}$. The value $1 - \mu(\tilde{d_\gamma}, \tilde{d_\delta})$ describes the distance between documents in the information base on the basis of documents content and the electronic archive ontology. The value can be used in fuzzy clustering in the information base content, i.e. in tasks where the distance between the cluster centre (a hypothetical document, for example) and analysed documents is of paramount importance for the target function.

V. Computational experiments

In case of analysis of the computational experiments results on the basis of the documentation of FRPC JSC 'RPA 'Mars' electronic archive, the domain-specific ontology was used.

The domain-specific ontology consists of 300 concepts. They include 219 concepts from standards used at the enterprise and 81 concepts and 10078 unique terms from realized projects.

The expert of FRPC JSC 'RPA 'Mars' prepared the selection involving 5017 technical documents. The selection is grouped into three main sections:

- the section based on the documentation class that consists of 34 groups;
- the section based on work sectors that consists of 28 groups (products discussed in documents);
- the section based on the documentation type that consists of 52 groups (GOST 2.601, 2.602, 2.102, 2.701, 3.1201).

In order to perform the experiment of quality evaluation of structuring FRPC JSC 'RPA 'Mars' electronic archive documentation, the index containing both ontological and traditional representations of technical documents (set of «term-frequency» pairs) was used. Further, the indices were structured with the use of different variants and subsequent quality evaluation according to the following list:

- structuring the traditional representations of technical documents with the use of Oracle Text tools;
- structuring the traditional representations of technical documents with the use of the modified FCM-algorithm of clustering;
- structuring the ontological representations of technical documents with the use of the modified FCM-algorithm of clustering;

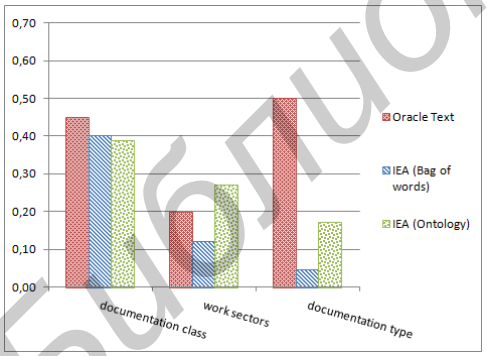

Figure 2. Quality evaluation of FRPC JSC 'RPA 'Mars' electronic archive documentation structuring

As indicated by Fig. 2, the most appropriate values of the evaluation function for ontological results were obtained in case of structuring the technical documentation selection in work sectors as it performs structuring in individual documents content. In case of structuring according to the document type, Oracle Text outperforms the others. The algorithm works well in case of structuring in accordance with the document type when Oracle Text gives the best results. The modified FCM-algorithm of clustering ontological representations of technical documents provides structuring of highest quality in accordance with work sectors regarding to the content.

VI. Conclusion

The computational experiments show that the results of structuring the ontological representations of technical documents is 40% better than results structuring with the use of Oracle Text. The time spending on indexing and structuring processes of technical documentation ontological representations is, on the average, 7% less than the total time spending on indexing and structuring processes of technical documentation traditional representations. The ontological approach to indexing and structuring technical documentation makes possible structuring the electronic archive for less time.

This work is supported by the Russian Foundation for Basic Research (Grant No. 16-47-732033 «Development of models and tools for ontological analysis of project diagrams based on the machine learning methods»).

References

[1] Serrano-Guerrero J., Olivas J. A., de la Mata J., Garces P. Physical and Semantic Relations to Build Ontologies for Representing Documents. Fuzzy logic, Soft Computing and Computational Intelligence (Eleventh International Fuzzy Systems Association World Congress IFSA), Beijing, China. Tsinghua University Press, 2005, vol. 1, pp. 503-508.

[2] Zagoruyko N.G. Prikladnye metody analiza dannykh i znanii [Applied Approaches to Data and Knowledge Analysis]. Novosibirsk, IM SO RAN Publ., 1999. 270 p.

[3] Zagorulko Yu.A., Kononenko I.S., Sidorova E.A.: Semanticheskii podkhod k anal- izu dokumentov na osnove ontologii predmetnoi oblasti [A Semantic Approach to the Document Analysis on the basis of Domain Ontology]. Available at: www.dialog21.ru/digests/dialog2006/materials/html/SidorovaE.html.

[4] Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, 2003, 624 p.

[5] Umberto Straccia. Transforming fuzzy description logics into classical description logics. In Proceedings of the 9th European Conference on Logics in Artificial Intelligence (JELIA-04), number 3229 in Lecture Notes in Computer Science, pages 385–399, Lisbon, Portugal, 2004. Springer Verlag.

[6] Umberto Straccia. Fuzzy description logics with concrete domains. Technical Report 2005-TR-03, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy, 2005.

[7] Namestnikov A.M., Filippov A.A., Avvakumova V.S. An ontology-based model of technical documentation fuzzy structuring. Proceedings of the 2nd International Workshop on Soft Computing Applications and Knowledge Discovery (SCAKD 2016) colocated with the 13th International Conference on Concept Lattices and Their Applications (CLA 2016), Moscow, Russia, 2016, pp. 63-74.

НЕЧЕТКАЯ ОНТОЛОГИЯ КАК ЯДРО БАЗЫ ЗНАНИЙ ДЛЯ ТЕХНИЧЕСКОГО ЭЛЕКТРОННОГО АРХИВА

Наместников А.М.

В статье приводится структура онтологии на основе нечеткого расширения дескрипционной логики SHOIN(D). Множество метауровней онтологии отражает специфику предметной области разработки сложных автоматизированных систем. В работе определены основные логические аксиомы, на основе которых выполняется логический вывод.