An Ontology-Based Model of Technical Documentation Fuzzy Structuring

Alexey M. Namestnikov[1], Alexey A. Filippov[1], and Valeria S. Avvakumova[2]

[1] Ulyanovsk State Technical University
nam@ulstu.ru
al.filippov@ulstu.ru
http://www.ulstu.ru
[2] FRPC JSC 'RPA 'Mars'
valeria.avvakumova73@gmail.com
http://www.npomars.com

Abstract. The article is concerned with the method for structuring the electronic archive of technical documentation on the basis of the domain-specific ontology. The ontology formal model, the technical document model, and the algorithm for clustering electronic archive content that has its origins in the modified fcm-method are presented. The authors are pioneered in offering the formalization of the measure of distance between ontological representations of the archive technical documents on the basis of hierarchy transformation complexities comparison. Different types of semantic relations between ontology concepts should be taken into account. Thus, the article considers the experimental results of the subset of the electronic archive technical documentation of the large project organization.

Keywords: ontology, clustering, technical document, fuzzy model, graph

Introduction

A modern large project organization possesses a sizable electronic archive of design and engineering documentation and engineering documentation. Its greater part is represented in unstructured text files. In actual truth, such an electronic archive contains the totally experience and knowledges of a great number of highly trained specialists that have been developing and designing complex systems over many years. In case of expanding the electronic archive, difficulties related to document analysis on the basis of predetermined properties ensure. Also skills of semantic processing of a great number of technical documentation and intimate knowledges of the subject area are required for persons who involved in complex technical systems designing. As a result, the important experience of previous developments fixed in electronic archives often becomes non-demanded. Thus, R&D cycle runtime increases.

The solution of the specified problem can be based on the use of intelligent methods and algorithms of text documents analysis in order to create the navigation structure of the technical documentation electronic archive. The paper [1] suggests using ontologies in intelligent document analysis.

Evaluating the specific character of project knowledges leads to the necessity of forming the project organization ontology with the special structure including features of a project process in the form of a subject area concept system, relations between these concepts, and interpretation functions. In such a manner the electronic archive should possesses properties of an intelligent system. At the moment mathematical methods and algorithms providing the means for structuring an electronic archive of technical documentation with consideration for its content and the specific character of a project organization subject area are not available.

Consequently, currently central problems include development of models, methods and algorithms for construction of the navigation structure of the technical documentation electronic archive on the basis of domain-specific clustering of partially formalized information resources.

In Section 1, the authors decribe the formal model of electronic archive ontology structure. Section 2 considers a technical document as an electronic archive resource and presents the ontological model. In its turn, Section 3 proposes the algorithm for ontology-oriented indexing of technical documents. The measure of distance in the context of ontology relating to the level of designing standards is formalized in Section 4. Section 5 offers some experimental results.

1 The structural model of an electronic archive ontology

A subject area of complex system designing places some constraints on the structure of an applied ontology. The rigid binding to standards and systems life cycle models applied at different stages of designing implies the necessity of forming the ontology that consists of a lot of levels, as indicated by 1.

Formally, the electronic archive ontology consists of two applied ontologies and may be written as the equation 1:

$$O = \langle O^D, O^{LC}, R_A \rangle, \tag{1}$$

where O^D is a subject area ontology component, O^{LC} is an ontology of designing systems life cycles, R_A is a unidirectional association relation between the ontology components. Let us consider the electronic archive ontology components in more detail (1).

In this way, let us write the domain-specific ontology as the following sequence:

$$O^D = \langle C, W, R^D, F^D \rangle,$$

where C is the set of electronic archive concepts that makes up a bulk of a conceptual apparatus of an automated system designing, $W = W^S \cup W^P$ is a set of subject area concepts, here W^S is a set of concepts on the level of standards, W^P is a set of concepts on the project level, R^D is a set of relations. Symbolically,

$$R^D = \{R_G^D, R_C^D, R_A^D\},$$

where R_G^D is anti-symmetric, transitive, irreflexive binary generalization relationship ('subclass_of'), R_C^D is a binary transitive composition relation ('part_of'), R_A^D is a binary relationship of unidirectional association.

The set of concepts C is defined by the following equation:

$$C = \left(C^{S_1} \cup C^{S_2} \cup \ldots C^{S_k}\right) \cup C^P,$$

where $C^{S_i}, i = \overline{1,k}$ is the set of subject area concepts for the standards of the i^{th} group, C^P is the set of subject area concepts extracting from the technical documentation of projects realized.

The set of interpreting functions is denoted as follows:

$$F^D = \{F_{WC^P}^D, F_{C^P C^S}^D\},$$

here $F_{WC^P}^D : \{W\} \to \{C^P\}$ is a function correlating a set of terms and a set of subject area concepts, $F_{C^P C^S}^D : \{C^P\} \to \{C^S\}$ is an interpretation function of the set of concepts allowing to go to the level of concepts defined in standards.

The ontology on a life cycle as a sequence component (eq. 1) consists of three sets and is denoted by the following equation:

$$O^{LC} = \langle M^{LC}, St^{LC}, R^{LC} \rangle,$$

here M^{LC} is a set of models of designing systems life cycles, St^{LC} is a set of life cycle stages.

Definition 1. *Terminological environment of concepts is the set of terms (layers) from the electronic archive technical documentation of projects realized.*

According to the paper [1], a semantic distance between the concept and terms in the technical document should be defined on the basis of the semantic relation idea. The idea encloses the use of 'distance' between words.

The semantic coefficient of the relation between the concept and the term (the semantic distance) is defined by the following equation:

$$S\left(c_i^{P(S)}, w_j\right) = \frac{\sum_{occur\left(c_i^{P(S)}, w_j\right)} \frac{1}{exp(sentence \cdot (paragraph+1))}}{num\left(occur\left(c_i^{P(S)}, w_j\right)\right)} \cdot \\ \cdot \frac{num\left(paragraph - cooccur\left(c_i^{P(S)}, w_j\right)\right)}{num\left(totalparagraph\right)},$$

here $c_i^{P(S)}, w_j$ is the i^{th} concept on the level of projects (standards) of the ontology and the j^{th} term, *sentence* is the distance expressed in the form of the

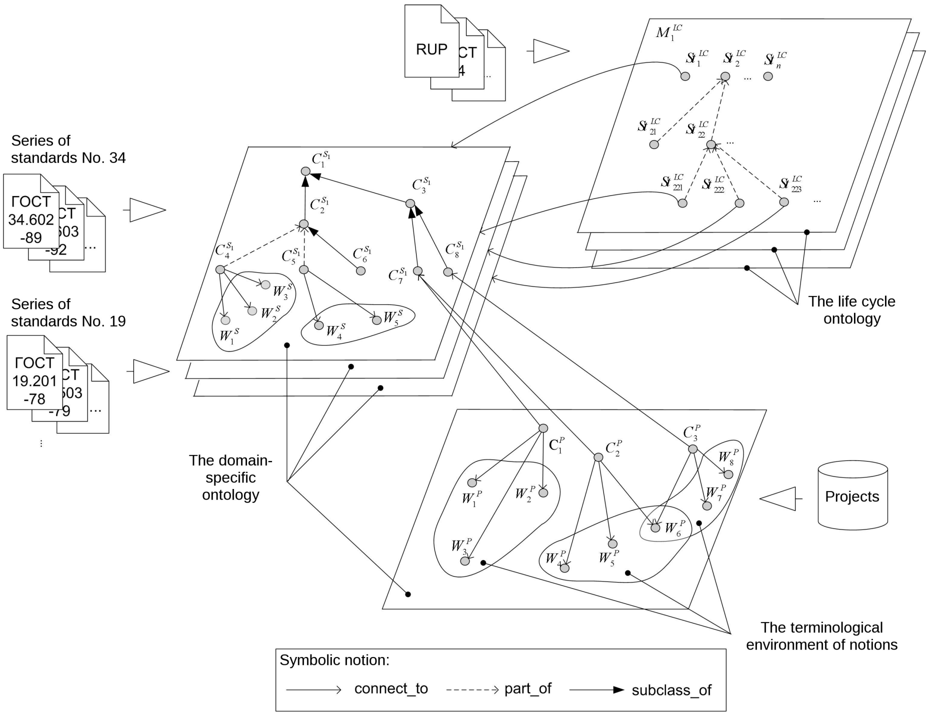

Fig. 1. The structure of the electronic archive applied ontology.

number of sentences between the concept and the term, $paragraph$ is a distance expressed in the form of the number of paragraphs between the concept and the term, $num\left(paragraph - cooccur\left(c_i^{P(S)}, w_j\right)\right)$ is the number of paragraphs where coocurrence $c_i^{P(S)}$ and w_j exist, $num\left(occur\left(c_i^{P(S)}, w_j\right)\right)$ is the number of rencontres between $c_i^{P(S)}$ and w_j, $num\,(totalparagraph)$ is the number of paragraphs in the document.

After defining semantic distances between the concept and the document terms, its necessary to define the subset of terms that are appreciably semantically close to the concept. In case of defining the terminological environment, according to the paper [2], the hypothesis of λ-compactness that leans up the λ-distance, taking into account a normalized distance d between terms and the characteristics of a local density of terms τ about these elements.

If the semantic distances between all the pairs of terms with the terminological environment are defined, the graph connecting all terms can be plotted. After that, the most long edge (the graph diameter D) should be defined. Consider two terms w_i and w_j and denote the length of the edge connecting them (the semantic distance) as $\alpha\,(w_i, w_j)$. We obtain the normalized distance between terms $d = \frac{\alpha}{D}$.

Further, let us find the shortest edge between the ones adjusted to the edge (w_i, w_j). Its length is denoted by β_{min}. The ration between the lengths of adjusted intervals is denoted by $\tau^* = \frac{\alpha}{\beta_{min}}$. In order to normalize this value, let us find the largest value τ_{max} in the entire graph. The value $\tau = \frac{\tau^*}{\tau_{max}}$ is a normalized characteristic of a set local density nonhomogeneity about the ontology terms w_i and w_j. $\lambda = f(\tau, d)$ is a λ-distance between the terms w_i and w_j. According to the paper [2], the use of $\lambda = \tau^2 \cdot d$ as such a distance measure is suggested.

In order to define the terminological environment of the ontology concept on the level of realized projects, it is necessary to mark such an edge (w_i, w_j) that can be a boundary between terms related to the ontology concept and terms that are not included in the terminological environment of the concept. With the use of λ-KRAB algorithm, the final criteria characterizing the quality of such a disjunction of terms is denoted by the following equation:

$$F = h^4 \tau^2 d \to max,$$

where $h = 2 \cdot \frac{m^+}{m} \cdot \frac{m^-}{m}$, is the equinumerosity criteria of the specified classes of terms. Here m^+ is the number of terms included on the terminological environment of the concept, m^- is the number of other ones.

Thus, with the use of the λ-compactness hypophysis, the subset of terms that is included in the terminological environment of the concerned concept is defined.

Every terminological environment W_k of the concept $C_k^{P(S)}$ can be denoted by the following equation

$$\{(w_{1k}, f_{1k}), (w_{2k}, f_{2k}), \ldots, (w_{ik}, f_{ik}), \ldots, (w_{lk}, f_{lk})\},$$

here w_{ik} is i^{th} term k^{th} ontology concept, l_k is the total amount of term associated with the the k^{th} concept, f_{ik} is a normalized semantic weight of the i^{th} term in the terminological environment of the k^{th} concept (normalized semantic distance between the term and the concept in the context of the one ontology environment).

2 The ontology model of the technical document as an electronic archive resource

A technical document in the context of an electronic archive is considered is an information resource. Any one of technical documents can be considered as a container of partially structured information. On the one hand, we deal with a natural language text, but on the other hand, a technical document is proper structured. The structure is defined in different standards.

We compare a frequency of occurrence of terms in one technical document with a frequency of occurrence of the same terms in the whole set of documents. It is originally conceived that the terms are not valuable if the frequency of terms in the document analyzed is far in excess of the frequency in the whole set of documents. Symbolically, such a dependence can be denoted as follows:

$$f_i = tfidf_i = tf_i \cdot log\left(\frac{N}{df\,(w_i)}\right),$$

here $tfidf_i$ is a relative importance of the term w_i in a document, tf_i is a normalized frequency of term w_i occurrence, N is a number of documents, $df\,(w_i)$ is a number of documents containing a term w_i.

An ontological model of a technical document is such a document representation that corresponds to the applied ontology state of an electronic archive. By [3], it follows that the notion of electronic document passport including a semantic index can be an analog of such a model.

A section of a technical document can be shown as follows:

$$s_i^d = \langle ch_{s_i^d}, C^P_{s_i^d}, C^S_{s_i^d} \rangle,$$

where s_i^d is the i^{th} section of a technical document d, $ch_{s_i^d}$ is a unique name of the i^{th} section of a technical document d, $C^P_{s_i^d}, C^S_{s_i^d}$ is a subset of subject area concepts, defined in the context of the i^{th} section of a technical document d.

Let us denote the j^{th} term of the i^{th} section of a technical document d by $w_j^{s_i^d}$, than a set of terms of the i^{th} section of a technical document d can be defined as:

$$W_{s_i^d} = \{w_1^{s_i^d}, w_2^{s_i^d}, \ldots, w_{l_{s_i^d}}^{s_i^d}\},$$

where $l_{s_i^d}$ is a number of terms of the i^{th} section of a technical document d.

With the use of an interpretation function of the ontology $F^D_{WC^P} : \{W\} \rightarrow \{C^P\}$ on the stage of technical document indexing, we obtain the ontological representation of the document section:

$$oV^d_{s_i^d} = \langle ch_{s_i^d}, C^P_{s_i^d}, C^S_{s_i^d} \rangle, C^P_{s_i^d} \subseteq C^P, C^S_{s_i^d} \subseteq C^S \mid_{St_k^{LC}} .$$

$C^S_{s_i^d} \subseteq C^S \mid_{St_k^{LC}}$ means that the ontological representation of the document includes only ontology concepts of a subset C^S (on the level of standards using in automated systems designing) that correspond to the k^{th} stage of designing St_k^{LC}.

With the use of function $F^D_{C^P C^S} : \{C^P\} \rightarrow \{C^S\}$, we can get the final representation of a technical document section that considers the state on an electronic archive applied ontology:

$$\overline{oV^d_{s_i^d}} = \langle ch_{s_i^d}, \{C^P_{s_i^d} \cup C^S_{s_i^d}\} \rangle, C^P_{s_i^d} \subseteq C^P, C^S_{s_i^d} \subseteq C^S \mid_{St_k^{LC}} .$$

A formal ontology model of a technical document can be defined as follows:

$$oV^d = \langle S^d, \{C^P_d \cup C^S_d\} \rangle,$$

The two main parts can be marked in the above equation: a structural one (S^d) and a conceptual one ($\{C_d^P \cup C_d^S\}$) in the context of realized projects of the archive and standards applied in the process of automated system designing with regard to the stage of a life cycle.

3 Ontology-oriented indexing of technical documents

The ontology indexing of a technical document has in its basis the following function:

$$F_{oV^d} : s_i^d \to oV_{s_i^d}^d,$$

here s_i^d is the i^{th} section of a technical document d, $oV_{s_i^d}^d$ is an ontological representation of the i^{th} section of a technical document d.

Notice that the method of computing a normalized weight of a term $w_j^{s_i^d}$ in the i^{th} section of a technical document d has in its basis the following equation:

$$f_j^{s_i^d} = 1 + log\left(tf_{w_j^{s_i^d}}\right) \cdot log\left(\frac{N}{dt}\right) \cdot \frac{1}{\sqrt{tf^2_{w_1^{s_i^d}} + tf^2_{w_2^{s_i^d}} + \ldots + tf^2_{w_n^{s_i^d}}}}, 1 \leq j \leq n,$$

here $f_j^{s_i^d}$ is a normalized weight of a term $w_j^{s_i^d}$ in the i^{th} section of a technical document d, $tf_{w_j^{s_i^d}}$ is a term $w_j^{s_i^d}$ frequency of occurrence, N is the total amount of documents, dt is a number of documents including a term $w_j^{s_i^d}$, n is a number of terms in the j^{th} section of a technical document d.

Definition 2. *A degree of manifestation of an electronic archive ontology concept is a degree of conjunction between a terminological environment and a set of concepts of a technical document fragment subject to the condition that a terminological environment includes terms that are semantically close to the concept.*

Computing the degrees of manifestation of ontology concepts for every section of a technical document is performed with the use of the apparatus of fuzzy irrelevance [4]. Fuzzy irrelevance between a set W (a set of ontology terms on the level of projects (standards) included in the terminological environment of concept) and a set $C^{P(S)}$ (a set of concepts of an applied ontology on the level of projects (standards)) denoted by $\tilde{\Gamma} = \left(W, C^{P(S)}, \tilde{O}\right)$ where W and $C^{P(S)}$ are crisp sets, $\tilde{O}$ is a fuzzy set in $W \times C^{P(S)}$. A set W is a domain of a function, a set $C^{P(S)}$ is a range of a function, and $\tilde{O}$ is a fuzzy graph of a fuzzy relevance.

The crisp relevance $\Gamma = \left(W, C^{P(S)}, O\right)$ with a chart O as a carrier of a fuzzy chart $\tilde{O}$ is called the carrier of fuzzy relevance $\tilde{\Gamma} = \left(W, C^{P(S)}, \tilde{O}\right)$. In the context of an ontology, a chart O defines parts of unidirectional associations R_A^D between a project concepts and terms in an ontology.

In order to find the meaning of concept domination, the method comparing the terminological environment of every concept in the ontology of a subject area ontology on the project level with the text analyzed. Let us remark that the minimal fragment of a text analyzed is a sentence and a maximal one is the whole document, as in different fragments of the text different concepts of the subject area are layed an emphasis on [5].

The algorithm of computing a degree of dominance of a concept in the text fragment consists of the following steps:

Step 1. Defining the maximal degree of manifestation of ontology concepts in the text fragment of a technical document d:

$$\mu_{\hat{fr_p^d}}\left(c^{P(S)}\right) = max_c\left(\mu_{fr_p^d}\left(c^{P(S)}\right)\right).$$

Step 2. Defining the mean of a degree of manifestation of ontology concepts without the concept with the maximum degree of manifestation (defined at the previous step):

$$\mu_{\tilde{fr_p^d}}\left(c^{P(S)}\right) = \frac{1}{n-1}\sum_{i=1}^{n-1}\mu_{fr_p^d}\left(\hat{c}_i^{P(S)}\right),$$

where $\hat{c}_i^{P(S)} \in c^{P(S)} - c_{max}^{P(S)}$, $c_{max}^{P(S)} = argmax_{c^{P(S)}}\left(\mu_{fr_p^d}\left(c^{P(S)}\right)\right)$, n is a number of concepts with a non-zero degree of manifestation for a text fragment fr_p^d.

Step 3. Defining a degree of manifestation of a concept in a text fragment fr_p^d:

$$\Delta_{fr_p^d}\left(c^{P(S)}\right) = \mu_{\hat{fr_p^d}}\left(c^{P(S)}\right) - \mu_{\tilde{fr_p^d}}\left(c^{P(S)}\right). \quad (2)$$

The equation 2 defines a quality of selection of a text fragment in a technical document in order to constrain the subject area concept that is fixed in an electronic archive ontology.

Having applied the ontology interpretation function $F_{WC^P}^D : \{W\} \to \{C^P\}$, we obtain an initial ontological representation of each segment. The representation consists of initial sets of concepts on the levels of projects and standards that require correction.

The results of the experiments with extracting text fragments on the basis of the genetic optimization show that averages 30% of concepts add up to 70% of the total degree of manifestation of all the concept of the text fragment.

The final step of forming the ontological representation of a technical document is the use of interpreting function $F_{C^P C^S}^D : \{C^P\} \to \{C^S\}$ that allows to specify a set of concepts on the level of standards resting on the subset of ontology concepts found in a technical document. The concepts correspond to the realized projects.

In case of realizing the above procedures, we get the final ontological representation for every i^{th} section of a technical document.

4 The ontological measure of distance between documents

Let us consider the formal measure of distance between documents in the context of ontology concepts relating to the level of designing standards. Every ontological representation can be illustrated in a form of a tree (a hierarchy) of subject area concepts. Such an hierarchy can be defined by finding a minimal tree including all concepts from the ontological representation [2].

The Levenshtein distance between hierarchies can be defined on the basis of computing an edit operation cost that should be found for each type of a semantic relation. Thus, an edit operation for a generalization relation is denoted by $\phi_{S_i}\left(R_G^D\right)$ and a 'part_of' one is denoted by $\phi_{S_i}\left(R_C^D\right)$. S_i shows belonging the value of an edit operation to the the i^{th} group of standards. Actually, in case of clustering, an edit operation is defined as a weight of a certain relation. The weight value lies in the range between 0 and 1 and have different values within the framework of every group of standards.

The total edit distance between the hierarchies is defined as the following equation:

$$\tau_{oV}^{*} = max_i\left(\sum_{s=1}^{m}\phi_{S_i}\left(R_G^D\right)_s + \sum_{l=1}^{n}\phi_{S_i}\left(R_C^D\right)_l\right),$$

where i is a group of standards number, s is an adding generalization relation number, l is an adding 'part_of' relation number. The total edit distance can be computed as a maximum one from all edit distance defined for every group of standards.

A normalization coefficient T_{oV} is defined on the basis of all semantic relation of a generalized hierarchy. Thus, a measure of distance between ontological representations of technical documents can be defined as follows:

$$\parallel oV^{d_1} - oV^{d_2} \parallel = \frac{\tau_{oV}^{*}}{T_{oV}}.$$

In order to create the navigation structure in the form of a nested set of clusters of technical documents, it is necessary to solve the problem of setting the weights of semantic relations between ontology concepts on the level of standards. As noted above, weight coefficients are defined as $\phi_{S_i}\left(R_G^D\right)$ and $\phi_{S_i}\left(R_C^D\right)$ for a generalization relation and 'part_of' relations respectively.

In view of the fact that the specified relations are used in the ontology concepts for different groups of standards, let us suppose that their optimal values for each group (in the context of their concept hierarchies) are generally different. Let us formulate the principle of the best value for weight coefficients of ontology semantic relations.

Let $\{oV^d\}^*$ be a set of ontological relations of documents included in the model sampling (the expert division of documents between classes). The following equation is true:

$$\{oV^d\}^* \subset \{oV^d\},$$

where $\{oV^d\}$ is a full set of ontological representation of electronic archive technical documents. The ontology is defined by the equation (1). On the level of standards, the generalization and 'part_of' relations are defined on the basis of concepts with corresponding weight coefficients $\phi_{S_i}\left(R_G^D\right)$ and $\phi_{S_i}\left(R_C^D\right)$, where S_i is the i^{th} group of designing standards used in ontology creation.

A set $\{oV^d\}^*$ consists of two subsets $\{oV^d\}_+^* \cup \{oV^d\}_-^*$ that correspond to the expert division of documents between two predetermined classes. The optimization problem of weight coefficients of semantic relations consists of finding such a set of coefficients as follows:

$$\{\langle\phi_{S_1}^*\left(R_G^D\right),\phi_{S_1}^*\left(R_C^D\right)\rangle,\langle\phi_{S_2}^*\left(R_G^D\right),\phi_{S_2}^*\left(R_C^D\right)\rangle,\ldots,\langle\phi_{S_n}^*\left(R_G^D\right),\phi_{S_n}^*\left(R_C^D\right)\rangle\}.$$

The clustering coefficient defined by the equation 3 should be as low as possible.

$$F^* = \frac{max\left(\bar{K_+} + \bar{K_-}, \hat{K_+} + \hat{K_-}\right)}{N} \to min \tag{3}$$

where $\bar{K_-}$ and $\hat{K_-}$ are sets of absent documents respectively in the first and the second clusters, $\bar{K_+}$ $\hat{K_+}$ are sets of redundant documents respectively in the first and the second clusters, N is the number of documents.

5 The analysis of computational experiments result on the basis of FRPC JSC 'RPA 'Mars' electronic archive documentation

In case of analysis of computational experiments result on the basis of the documentation of FRPC JSC 'RPA 'Mars' electronic archive, the domain-specific ontology was used. The ontology consists of two series of standards used at the enterprise:

1. GOST 34. Information technologies. Open systems interconnections. (It consists of 108 ontology concepts at the level of standards).
2. GOST 19. Unified system for design documentation. (It consists of 111 ontology concepts at the level of standards).

The ontology level appropriate to the realized projects is based on the selection of FRPC JSC 'RPA 'Mars' electronic archive documentation that includes 5017 technical documents. The level consists of 81 concepts and 10078 unique terms comprising the terminological environment of concepts.

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

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

– the section based on the documentation type that consists of 52 groups (GOST 2.601, 2.602, 2.102, 2.701 3.1201);

– the section based on work sectors that consists of 28 groups (products discussed in documents).

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 'termin-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;
– structuring the ontological representations of technical documents with the use of the modified FCM-algorithm of clustering with regard to the life cycle models of the designing system.

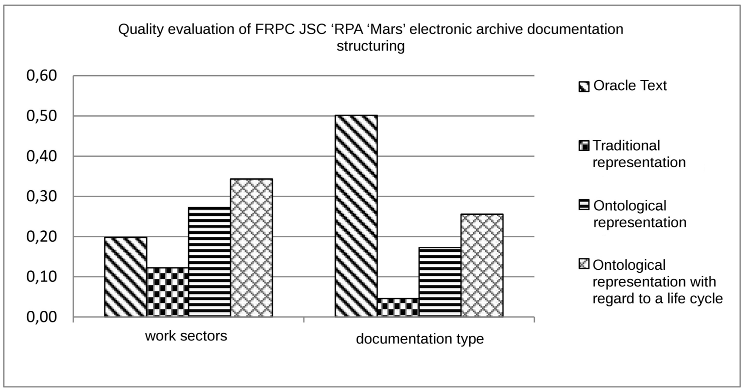

Fig. 2. Quality evaluation of FRPC JSC 'RPA 'Mars' electronic archive documentation structuring.

As indicated by Fig. 2, the most appropriate values of the evaluation function for ontological results with regard to the life cycle models of the designing system 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 function of documentation structuring with the use of Oracle Text is based on the clustering algorithm considering a frequency of term occurrence in documents. 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 with regard to the life cycle models of the designing system provides structuring of highest quality in accordance with work sectors with regard to the content.

Conclusion

The computational experiments show that the results of structuring the ontological representations of technical documents with regard to the life cycle models of the designing system 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. As this takes place, the most time spending is related to the process of documentation indexing.

Acknowledgments The research was carried out within the state assignment No. 2014/232 for the accomplishment of state works in the sphere of scientific affairs of the Ministry of Education and Science of the Russian Federation (theme 'Developing a new approach to the analysis of partially structured information resources').

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. 503508.
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 analizu 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. Bershtein L.S., Bozhenuk A.V.: Nechetkie grafy i gipergrafy [Fuzzy Graphs and Hypergraphs]. Moscow, Nauchnyi Mir Publ., 2005. 256 p.
5. Namestnikov A.M., Filippov A.A.: Metod geneticheskoi optimizatsii ontologicheskikh predstavlenii proektnykh dokumentov v zadache indeksirovaniia [The Method of genetic optimization of project documentation ontological representations in case of indexing]. Trinadtsataia nats. konf. po iskusstvennomu intellektu s mezhd. uchastiem KII-2012 [The XIII National Artificial Intelligence Conference with international participation KII-2012]. Belgorod,BSTU Publ., 2012, pp. 84-91.