Ontology-based classification model of text resources of an electronic archive

Vadim Moshkin
Information Systems department,
Ulyanovsk State Technical University,
Ulyanovsk, Russia
v.moshkn@ulstu.ru

Anton Zarubin
The Bonch-Bruevich Saint - Petersburg
State University of Telecommunication
St. Petersburg, Russia
azarubin@sut.ru

Albina Koval
The Bonch-Bruevich Saint - Petersburg
State University of Telecommunication
St. Petersburg, Russia
akoval@sut.ru

***Abstract* —This paper presents an ontological model of a text document as an electronic archive resource. The article also presents an ontology-based algorithm for the classification of technical documents. In conclusion, the results of experiments confirming the effectiveness of models and algorithms in solving the problem of classification of documents of an electronic archive are presented. An assessment was also made of the use of linguistic and statistical algorithms for determining the terms of poorly structured information resources.**

Keywords—ontology, classificationc, ontological representation, machine learning, electronic archive

I. INTRODUCTION

The task of categorizing text documents to simplify the search for information in an electronic archive of an organization is more relevant than ever. In most cases, archiving is structured manually by archive specialists. Specialists should have knowledge in the subject area and take into account the specifics of the stored documentation.

Automation of categorization of the archive of electronic text documents should be carried out taking into account the semantics of information in the documents. Otherwise, the experience of highly qualified specialists developing this documentation will be difficult to extract from unstructured resources for further use.

Currently, researchers offer various ways to solve this problem. In [1], the ant colony classification algorithm is used to classify data and is used to quickly search for large amounts of data from intelligent archives.

Characteristics of a text document are taken into account during its analysis and processing and are included in the document model.

The extended Boolean model of the document does not represent terms with values of 0, 1, but with weighting coefficients using the theory of fuzzy sets [2-5]. In this case, the value of the weight coefficient is determined from the interval [0, 1], thus we obtain that $D \in [0, 1]^n$.

The vector model formally presents text documents as a matrix of terms and documents [6]:

$$M = |F| \times |D|,$$

where $F = \{f_1, \ldots, f_k, \ldots, f_z\}$; $D = \{d_1, \ldots, d_i, \ldots, d_n\}$, d_i is a vector in the z-dimensional space R_z.

In [7] [8], the authors present an ontology designed to model archival descriptions of collections of historical documents. In [9], the authors present aspects of the current activities of the digital library related to the Semantic Web and present their functionality. They show examples ranging from general architectural descriptions to the detailed use of specific ontologies. In [10], a semantic search portal is proposed for intercultural archives, including documents, images, audio and video.

One of the solutions to this problem is the use of intelligent algorithms for the analysis of text documents with the division of the archive into classes in accordance with the semantics of the subject area. The semantics of the subject area will be concluded in the subject ontology, formed through the analysis of textual documentation.

Domestic and foreign researchers (Gavrilova T.A. [11], Zagorulko Yu.A. [12], Khoroshevsky V.F., Soloviev V.D., Lukashevich N.V., Dobrov B.V., Smirnov S. V., Guarino [13], Uschold M. et al.) note the relevance of applying the ontological approach to the automatic structuring of large text archives using the ontological approach and extracting the semantic basis of project documentation.

II. MODELS OF APPLIED ONTOLOGY OF TEXT DOCUMENTS

The construction of an ontology in the classification of documents in electronic archives is necessary to take into account the characteristics of the subject area and to increase the speed of document search. Ontology defines a semantic scale that defines whether a document belongs to one class [14] [15].

Thus, the formal model of applied ontology of the electronic archive of project documentation is:

$$O_{ARC} = \langle T, T_{ORG}, Rel, F \rangle,$$

where T is the set of terms of design documentation for an electronic archive; T_{ORG} is a set of terms of a problem area; Rel is a set of ontology relationships. Many relationships include the following:

$$Rel = \{R_H, R_{PartOF}, R_{ASS}\},$$

where R_H is the hierarchy relation; R_{PartOF} is a part-to-whole relationship; R_{ASS} is an association relation.

Formally the set of terms of design documentation for an electronic archive is:

$$T = (T^{D_1} \cup T^{D_2} \cup \ldots \cup T^{D_k}) \cup T^{ARC},$$

where $T^{D_i}, i = \overline{1,m}$ is the set of terms of the i-th problem area; T^{ARC} is a set of terms of the problem area extracted from the documents of the electronic archive.

Formally, the functions of interpretation of the subject ontology are:

$$F = \{F_{T_{ORG}T}, F_{T^{ARC}T^D}\},$$

where $F_{T_{ORG}T}: \{T_{ORG}\} \to \{T\}$ is an interpretation function that defines the correspondence between the terms of the problem area and the terms of the design documentation of the electronic archive; $F_{T^{ARC}T^D}: \{T^{ARC}\} \to \{T^D\}$ is an interpretation function that defines the correspondence between the terms of the problem domain extracted from electronic archive documents and the terms of the problem domain.

The main in the ontology of the electronic archive is the relation "associate_with". This relation determines the subject area to which the project document of the electronic archive belongs and determines the subject of the document.

The characteristic of the weight of the term f_i in a text document is the frequency of the i-th term in the document. Hence, the following patterns are relevant:

- high-frequency terms in a document are system-wide;
- terms with a low frequency in a particular document do not provide an improvement in the quality of search for documents in the archive.

The most indicative are terms that have an average frequency of occurrence in a document, but most fully characterize a document in a problem area [16, 17].

If the frequency of occurrence of one term is significantly higher in the document than the frequency of its occurrence in all analyzed documents of the electronic archive, then this term is semantically significant. Formally, this rule is

$$s_i = tf_i \cdot log\left(\frac{M}{df(t_i)}\right),$$

where s_i is an indicator of the semantic significance of the term t_i in this document; M is the total number of all documents in the electronic archive; tf_i is the value of the index of the normalized frequency of the term t_i; $df(t_i)$ is the total number of documents containing the term t_i.

Thus, the ontological model of an electronic archive document is:

$$V_j^{doc} = \langle T^{ARC}, T^D \rangle,$$

where T^{ARC}, T^D is the set of terms of the problem area of the j-th document of the electronic archive.
Hence

$$associate\_with(d, T_K) = 1.$$

This equality assumes that the document d is mapped into the space of terms $T, T_2, \dots, T_k$.. If t_i^d is the *i*-th term of the document d, then the set of terms of the document d can be represented as follows:

$$T^d = \{t_1^d, t_2^d, \dots, t_n^d\},$$

where *n* is the total number of terms in the document *d*.

III. The Ontological Index Model

The ontological indexing algorithm for text documents of the electronic archive is shown in Figure 1.

The degree of semantic significance of an electronic ontology term is the value of coincidence of the term context environment with the set of terms of the electronic archive document. Contextual environment is composed of terms that are semantically close to the analyzed concept of a problem area [19].

Hence the formally semantic index of the i-th document is:

$$\{(t_1^d, s_1), (t_2^d, s_2), \dots, (t_i^d, s_i), \dots, (t_n^d, s_n)\},$$

where *l* is the total number of terms in the *i*-th document of the electronic archive after text preprocessing.

The degree of expression of the concept l_k in the *i*-th document *d* will be calculated as follows:

$$\mu(l_k) = 1 - \frac{1}{l}\sum|s_k - s_i|,$$

where s_k, s_i are indicators of the frequency of the term t_i in the description of the *k*-th term of the ontology in the document *d*; *n* is a measure of the power of the text input of l_k.

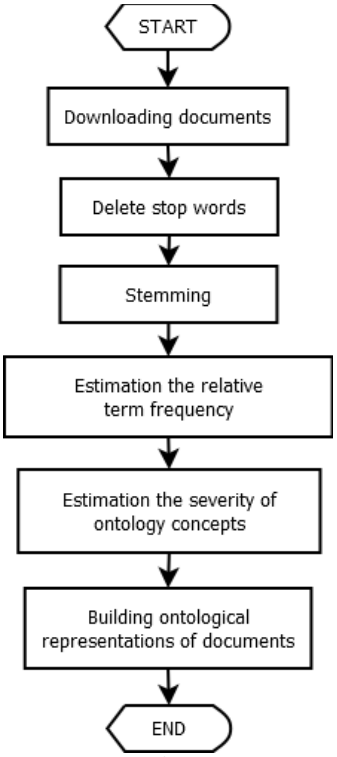

Fig. 1. Ontological indexing algorithm for text documents.

Thus, after indexing document d is an ontological representation − a fragment of the domain ontology, in which the degree of expression from 0 to 1 is defined for each ontology concept.

IV. Classification of Ontological Representations of Electronic Archive Documents

First you need to determine the classes by which the electronic archive documents will be split. For this, it is necessary to define many concepts of ontology and a linguistic label to determine the degree of expression of a concept in a class.

The linguistic label defines the meaning for the interval of the expression degree of the ontology concept. For example, the linguistic label "High" may correspond to the value of the expression degree of the concept from the interval from 0.7 to 1.0.

Thus, at the first step of the classification algorithm for the contents of the electronic archive, it is necessary to specify a set of classes *G* and determine their properties:

$$G = \{g_1, g_2, \dots, g_i, \dots, g_n\},$$
$$g_i = \{\langle c_1, m\rangle, \dots, \langle c_k, m\rangle\},$$
$$m \in [High, Middle, Low],$$
$$High = [0.7 \dots 1.0],$$
$$Middle = [0.5 \dots 0.7),$$
$$Low = [0 \dots 0.5),$$

where g_n is the *n*-th class of documents (classification basis);
c_k is the *k*-th concept of ontology; *m* is the linguistic label.

At the second step, the degree of belonging of the document d to each class g_i is calculated using the following expression:

$$s(g_i) = k - \sum_{i=1}^{k}(1 - \theta_k),$$

where k is the number of parameters of the class g_i; θ_k is a sign of a document matching the d -th property of the class g_i, which is calculated using the following expression:

$$\theta_k = \begin{cases} 1, c_k \in d, \mu(c_k) \in m \\ 0 \end{cases}.$$

Thus, the document d corresponds to the characteristic θ_k if it contains concepts characterizing the given attribute and its degree of expression is included in the interval of the linguistic label.

V. Experiment Results

As part of this study, a series of experiments were conducted to assess the quality of the classification of documents in the electronic archive of the Federal Scientific Production Center JSC Mars. JSC Mars is an organization engaged in the design, development and maintenance of automated systems, software and hardware for the Russian Navy.

For the experiments the following sets of documents were selected:

- technical specifications;
- patent research reports;
- specifications;
- testing programs and techniques;
- programmer, user, system administrator, etc. manuals.

1037 design documents were selected for the experiments. Figure 2 shows the signs of expert dividing documents into classes according to certain criteria.

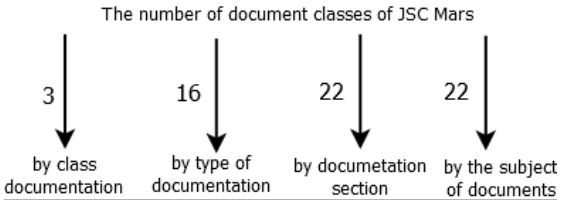

Fig. 2. Expert classification of documents of JSC Mars.

A software system was developed and experiments were conducted to select the preferred method for defining the term.

The test set consisted of a public data set containing about 10,000 tweets. Each tweet contains a text describing either a disaster or some other information. Each tweet has a label that determines whether the tweet belongs to the “disaster” or “other” class.

The first step is data preprocessing. Pretreatment consists of:

1) Delete stop words.
2) Tokenization according to words - splitting the analyzed text into separate words.
3) Bringing the register.
4) Lemmatization.

At the second step, many indices of the analyzed poorly structured information resources were erased, which were based on the following models of information resource representation: a word bag (Fig. 3), a statistical model based on TF-IDF (Fig. 4), and a linguistic model based on Word2Vec (Fig. 5).

The following assessments of the classification quality were obtained: “Bag of words” - 62.23%, “TF-IDF” - 74.78%, “Word2Vec” - 81.7%.

Thus, the linguistic model will be the best method for defining the terms of a semi-structured information resource. However, it is recommended to use a statistical model if low computational complexity of the algorithm and high speed of operation are required.

An ontological set of document indixes and classic indexes, which include the term-frequency values, were built. The classification quality assessment model from [20] was used as an evaluation function. The results of the experiments are presented in figures 6 and 7.

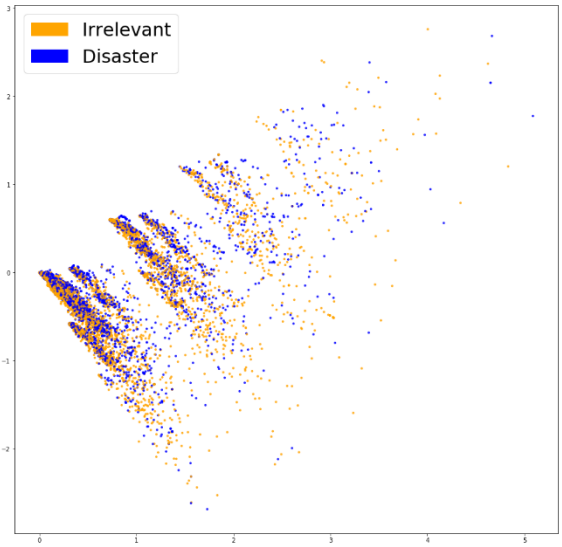

Fig. 3. Classification results for the “Bag of words” model.

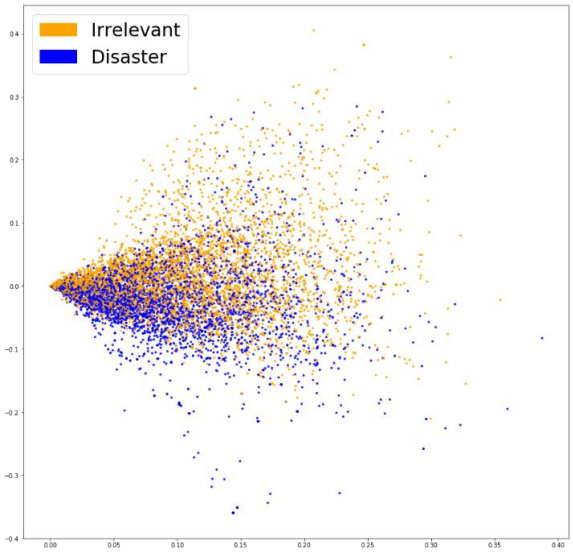

Fig. 4. Classification results for the TF-IDF model.

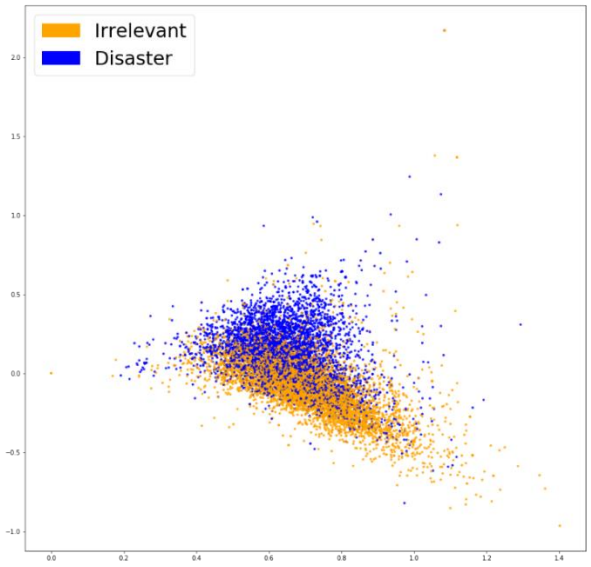

Fig. 5. Classification results for the "Word2Vec" model.

Fig. 6. Comparison of classification algorithms in accordance with the values of the evaluation function.

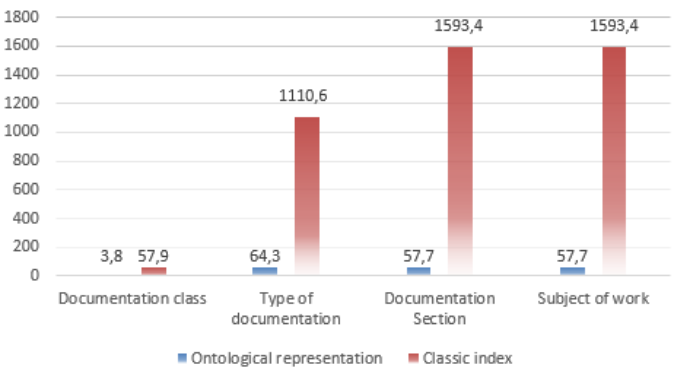

Fig. 7. Comparison of classification algorithms in accordance with the classification time (sec.).

As can be seen from the results of the experiments, the classification of ontological representations is faster (up to 27 times) relative to the classification time of classical indices.

The quality of classification of ontological representations in comparison with the results of classification of classical indices is slightly worse only when divided by the class of documentation. When dividing a multitude of documents by type of documentation, section of documentation and subject of work, the quality of classification of ontological representations is higher than that of classical indices.

Conclusion

Thus, an ontological model of a text document as an electronic archive resource and an ontologically oriented algorithm for the classification of technical documents is proposed. As can be seen from the results of the experiments, the formation of the ontological presentation of each individual document in the archive can significantly increase the speed of automatic classification of documents (up to 27 times) while maintaining or slightly improving the quality of classification.

In future works, it is planned to introduce fuzzy elements in the ontological representation of project documents.

Acknowledgment

This paper has been approved within the framework of the federal target project "R&D for Priority Areas of the Russian Science-and-Technology Complex Development for 2014-2020", government contract No 05.604.21.0252 on the subject "The development and research of models, methods and algorithms for classifying large semistructured data based on hybridization of semantic-ontological analysis and machine learning".

References

[1] W. Yong, L. Liming and Q. Yongsheng, "Improvement of big data retrieval algorithm in the intelligent archives management," 12th IEEE International Conference on Electronic Measurement & Instruments (ICEMI), pp. 487-491, 2015. DOI: 10.1109/ICEMI.2015.7494245.

[2] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Rertieval," ACM Press, New York, 1999.

[3] K. Manning, P. Raghavan and H. Schütze, "Introduction to the Information Search," M.: LLC "I.D. Williams, 2011.

[4] F. Song and W. Bruce, "A general language model for information retrieval (poster abstract)," Research and Development in Information Retrieval, pp. 279-280, 1999.

[5] E.M. Voorhees, "Natural language processing and information retrieval," Information Extraction: Towards Scalable, Adaptable Systems, pp. 32-48, 1999.

[6] G. Salton, "Automatic Text Processing," Addison-Wesley Publishing Company, Inc., Reading, MA, 1989.

[7] L. Pandolfo, L. Pulina and M. Zielinski, "Towards an Ontology for Describing Archival Resources," 2017.

[8] L. Pandolfo, L. Pulina and G. Adorni, "A framework for automatic population of ontology-based digital libraries," Advances in Artificial Intelligence, pp. 406-417, 2016.

[9] S. R. Kruk and B. McDaniel, "Semantic digital libraries," Springer, 2009.

[10] Z. Yan, F. Scharffe and Y. Ding, "Semantic Search on Cross-Media Cultural Archives," Advances in Intelligent Web Mastering. Advances in Soft Computing, vol 43, pp. 375-380, 2007.

[11] Yu. Zagorulko, I.S. Kononenko and E.A. Sidorova, "Semantic approach to the analysis of documents based on the ontology of the subject area," International Conference on Computational Linguistics and Intellectual Technologies Dialogue, 2016.

[12] T.A. Gavrilova and V.F. Khoroshevsky, "Knowledge Base of Intelligent Systems," St. Petersburg: Peter, 2000.

[13] T. Schneider, A. Hashemi, M. Bennett, M. Brady, C. Casanave, H. Graves, M. Grüninger, N. Guarino, A. Levenchuk, E. Lucier, L. Obrst, S. Ray, R. Sriram, A. Vizedom, M. West, T. Whetzel and P. Yim, "Ontology for Big Systems," The Ontology Summit Communiqué. Applied Ontology, vol. 7, pp. 357-371, 2012. DOI: 10.3233/AO-2012-0111.

[14] J. Serrano-Guerrero, J.A. Olivas, J. de la Mata and P. Garces, "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, vol. I, pp. 503-508, 2005.

[15] Yu.V. Vizilter, V.S. Gorbatsevich and S.Y. Zheltov, "Structure-functional analysis and synthesis of deep convolutional neural networks," Computer Optics, vol. 43, no. 5, pp. 886-900, 2019. DOI: 10.18287/2412-6179-2019-43-5-886-900.

[16] N. Yarushkina, V. Moshkin and A. Filippov, "Development of a knowledge base based on context analysis of external information resources," Proceedings of the International conference Information Technology and Nanotechnology. Session Data Science, Samara, Russia, pp. 328-337, 2018.

[17] A. Namestnikov, A. Filippov and V. Avvakumova, "An Ontology-Based Model of Technical Documentation Fuzzy Structuring," 2nd International Workshop on Soft Computing Applications and Knowledge Discovery (SCAKD), 2016.

[18] V. Moshkin and N. Yarushkina, "Modified Knowledge Inference Method Based on Fuzzy Ontology and Base of Cases," Creativity in Intelligent Technologies and Data Science, pp. 96-108, 2019. DOI: 10.1007/978-3-030-29750-3_8.

[19] A. Filippov, V. Moshkin, A. Namestnikov, G. Guskov and M. Samokhvalov, "Approach to Translation of RDF/OWL-Ontology to the Graphic Knowledge Base of Intelligent Systems," Proceedings of the II International Scientific and Practical Conference Fuzzy Technologies in the Industry, Ulyanovsk, pp. 44-49, 2018.

[20] Yu. Radionova, "A method for constructing an evaluation function that deterines the effectiveness of automatic clustering algorithms," Automatio of control processes, no. 15, pp. 23-28, 2009.