Use of Similarity of Wiki Pages as an Instrument of Domain Representation for Semantic Retrieval

Julia Rogushina
Institute of Software Systems
of National Academy of Sciences of Ukraine
Kyiv, Ukraine
Email: ladamandraka2010@gmail.com

***Abstract*—We propose an ontology-based approach to the Web semantic search that uses thesaurus representation of user task. Domain ontologies are considered as a source of semantic markup of the Wiki resources pertinent to retrieval domain. We use practical aspects of ontological approach to organization of Wiki-based information resources. Ontological model of Wiki resource formalizes the knowledge base structure and explicitly represents its main features. Domain ontologies, Wiki resources and task thesauri are generated independently by different applications but are used in general technological chain of user-oriented semantic retrieval. Open information environment is considered as an external data base with great volumes of heterogeneous and semi-structured information.**

Wiki ontologies are considered as the basis for establishing a semantic similarity between domain concepts pertinent to user task. Such Wiki-ontology elements as classes, property values of class instances and relations between them are used as parameters for the quantitative assessment of semantic similarity.

***Keywords*—semantic search, domain ontology, task thesaurus, semantic Wiki**

I. Introduction

The development of information technologies shows the tendencies to transition from traditional means of data processing to semantic computer systems oriented on work into the open environment. Now is the time for integration of the traditional information technologies with achievements of artificial intelligence. For example, theoretical principles and practical means of semantic computer systems development are designed by the Open Semantic Technology for Intelligent Systems Design (OSTIS Technology) [1]. But other important problem deals with interoperability of knowledge created and used by intelligent systems of different developers.

The growth in the use of the Web brings with it an increase in the number of interconnections among information systems and resources supporting the various aspects of human activities. Such interconnections have to be carefully prescribed to ensure interoperability. Standards of the Semantic Web [2] provides the universal ontology-based means of knowledge representation. Increasingly, the kinds of information structures being standardized today are much more complex than they were even a decade ago. However every practical task needs in specific methods of their use. In this work we consider an important component of open intelligent systems that deals with information retrieval on semantic level.

Open intelligent systems need in information retrieval tools and methods that search user-oriented information into the Web open sources. Such retrieval requires a formalized model of the search domain and description of user needs and interests in this process.

Ontologies are widely used to describe domains, but this causes a number of problems.

- Creating ontologies is a complex process that requires the involvement of a knowledge engineer and domain expert.
- The domain ontology is usually quite complex and contains a lot of unnecessary information for the specific task pertinent to user query.
- Processing an ontology and its matching with other information resources (such as unstructured natural-language texts) is a long and complicated process that requires the use of other background knowledge (e.g. linguistic knowledge bases). Therefore, it is advisable to use simpler information structures to formalize domain knowledge in information retrieval tasks.

We consider in this work the use of task thesauri as special cases of ontologies. Task thesaurus T is based on domain ontology O and consists of the ontological concepts (classes and individuals) joined by the semantic similarity to the user task in domain described by this ontology O.

II. Problem definition

An analysis of research works in the sphere of distributed information technologies shows that many intelligent tasks need in external sources of background knowledge from the Web, corporative and local networks, data warehouses etc. However, the problem of extracting such knowledge in the general case is extremely complex, and one of its components is semantic search that applies knowledge about user and user current task for selection of pertinent information resources.

To ensure the effective use of ontologies for semantic search by the various intelligent applications and to simplify the knowkedge processing process we propose to generate simplified ontology-based such information structures as thesauri. Every task thesaurus contains only such part of the domain knowledge that is needed to search for information that is pertinent to the user's current task. Thesaurus is a representation of semantically similar (in the local sense of current task) domain concepts related to this task.

This approach requires to justify the ways of ontology knowledge representation by means of thesaurus, to develop an algorithm for generating of such thesaurus based on the domain ontology and the description of the task. In addition, we need in development of methods for processing of this task thesaurus in applied retrieval systems and justification of their effectiveness of the received results for various types of such systems. It is also important to determine what information resources are used for creation of domain ontologies and what restrictions are imposed on the ontologies created in this way. In particular, ontologies that are generated by semantically marked Wiki resources often contain enough knowledge to carry out semantic search, but their processing is much simpler due to restrictions on their structure.

III. Task thesauri and their features

Wikipedia defines a thesaurus as a reference work that lists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms), in contrast to a dictionary, which provides definitions for words, and generally lists them in alphabetical order [3]. It is important that thesaurus as opposed to dictionary does not contain all word synonyms and their definitions.

Such definition does not use ontological approach but reflects the main characteristic of thesauri deal with it orientation on some particular task.

In the context of information retrieval, thesaurus is a form of controlled vocabulary that seeks to dictate semantic manifestations of metadata in the indexing of content objects. Its use is aimed to minimize semantic ambiguity by ensuring uniformity and consistency in the storage and retrieval. In this meaning thesaurus has to contain at least three elements: - list of words that correspond to domain terms, – the hierarchical relations between these words (e.g. parent/broader term; synonym, etc.), - a set of rules for thesaurus usage.

Thesaurus can be used for domain representation. If thesaurus represents ontological concepts as terms and uses ontological relations to link these concepts then we can consider such thesaurus as a special case of domain ontology oriented on analyses of natural language texts. Thesaurus contains only ontological terms (classes and instances) but does not describe all semantics of relations between them.

Some methods of thesauri generation use ontologies as a source of domain knowledge and integrate it with the current task description. Task thesaurus is a thesaurus that is generated automatically on base of the domain ontology selected by user and the NL description of particular task that is interesting for this user [4]. A simple task thesaurus is a special case of task thesaurus based on the terms of a single domain ontology. A composite task thesaurus is a task thesaurus that is based on the terms of two or more domain ontologies by operations on simple task thesauri of these ontologies.

Generation of the simple task thesaurus uses as input data two parameters:

- domain ontology selected by user;
- task description – the natural language text defines the current problem.

The text of task description contains elements related to the ontology concepts.

The process of simple thesaurus constructing contains two main steps:

- Step 1. Automated generation of the subset of ontology concepts correlates with fragments of task description:
  Substep 1.1 User explicitly and manually selects task-pertinent terms from the automatically generated list of classes and instance X. In the simplest cases, the construction of the thesaurus may be completed in this step, but it requires a lot of efforts from the user.
  Substep 1.2 Thesaurus is expanded with the help of various methods for processing of natural-language applied to task description (linguistic analysis, statistical processing, semantic markup analysis) that allow to detect NL fragments related to terms from O.
- Step 2. Expansion of the simple thesaurus by other ontology concepts according to the set of conditions that can use all elements of the O ontology.

Linguistic knowledge bases (KB) can be used for thesaurus construction. We can apply specific domain-oriented linguistic KBs that accumulate a large amount of lexical information. Such information is not universal and depends either from domain and natural language used in task definition. Therefore we cannot use Text Mining systems oriented on processing English texts. We apply direct updating of the domain lexical ontology by users and export linguistic knowledge from relevant vocabularies and knowledge bases, as well as from semantically marked Ukrainian texts.

In many cases, information about properties of ontological classes and individuals, their allowed values and their relations with other terms is appropriate in thesaurus constructing. Such information can be processed for refining the initially formed thesaurus in accordance with explicitly formulated user conditions. These conditions are defined by the specific nature of the task, but are not derived from its description and can be considered as meta-rules of retrieved information.

Complex task thesauri are generated from the built earlier task thesauri (simple or complex) with the help of set theory operations such as sum of sets, intersection of sets etc.

IV. Existing approaches to similarity measures

Similarity is a fundamental and widely used concept. Many researchers analyze the principles and measures of semantic similarity of domain concepts. In the cognitive domain, similarity is treated as a property characterized by human perception but use of this property in information systems requires the quantitative evaluations.

Now many similarity measures are used in various applications, such as information content, mutual information [5], Dice coefficient [6], distance-based measurements [7] etc. In [8] similarity defined in terms of information theory is applicable if domain has a probabilistic model. The similarity measure is derived from a set of assumptions about similarity and is not defined directly by some formula.

The similarity of concepts is also related to their content. One of the key factors in the similarity of the two concepts is the degree of information sharing in the taxonomy. The edge-counting method takes this into account indirectly. The information content of concept can be quantified by the logarithmic function of probability of concept use. Thus, the level of concept abstraction (i.e., its place in taxonomy) causes the less informational content. If there is a unique upper concept in taxonomy then its information content is 0. This quantitative characterization of information provides a new way of measuring semantic similarity based on the extension of concepts.

The more information is shared by two concepts, the more similar they are, and the information co-shared by the two concepts is determined by the information content of the concepts included in them into taxonomy.

Some measures of similarity [9] take into account only the depth of the nodes of the terms. Although the similarity is calculated taking into account all the upper bounds for the two concepts, the information measure allows to identify the minimum upper bound, but no class is less informative than its superclasses.

Measures to determine the semantic similar concepts (SSC) on the basis of ontologies use various semantic features of these concepts – their properties (attributes and relations with other concepts), the relative position in ontological hierarchies. The SSC set is a fuzzy set of concepts with the semantic distance less than the selected threshold.

Similarity is an important and fundamental concept in AI and many other fields. Various proposals for similarity measures are heuristic in nature and tied to a particular domain or form of knowledge representation.

The most general definitions of similarity are based on three intuitive assumptions:

- the similarity between A and B depends directly on their commonality;
- the similarity between A and B depends inversely on the differences between them;
- the maximum similarity between A and B is reached if A and B are identical, no matter how much commonality they share.

The similarity of two objects is related to their commonality depends directly on number of their common features and depends inversely on number of their differences. Concept similarity can be defines by similarity of strings and words. Feature vectors are widely used for knowledge representation, especially in case- based reasoning and machine learning. They can be applied for representation of words. Weights of features is used to account the importance of various features for word similarity. Some special features are applicative for natural language words and non-applicative for arbitrary strings of characters.

The similarity measures suppose that words derived from the same root as some initial word have the better similarity rankings. Other similarity measures are based on the number of different trigrams in the matching strings and on proposed by user definition of similarity under the assumption that the probability of a trigram occurring in a word is independent of other trigrams in the word. Similarity measures between words correspond to their distribution in a text corpus.

Semantic similarity can be based on similarity between concepts in domain taxonomy (such as the WordNet or CYC). The semantic similarity between two classes characterize not the set of their individuals or subclasses classes. Instead, generic individuals of these classes are compared.

A problem with similarity measures is that each of them is tied to a particular application or assumes a particular domain model. For example, distance-based measures of concept similarity assume that the domain is represented in a network. Another problem with the similarity measures is that their underlying assumptions are often not explicitly stated. Without knowing those assumptions, it is impossible to make theoretical arguments for or against any particular measure.

Methods aimed at SSC finding in different ontologies can be used to analyze the semantic similarity between the domain concepts. The assessment of similarity of concepts may be based on their positions in the hierarchy of classes with defined similarity: if the subclasses and superclass of these concepts are similar, then the same concepts are also similar.

The following parameters (features) can be considered into quantified similarity assessment of the two ontological classes:

- similarity assessing of their direct superclasses;
- similarity assessing of all their superclasses;
- similarity assessing of subclasses of concepts;
- similarity assessing of instances of classes.

Semantic similarity is a special case of semantic affinity. For the individual case of ontology, where the only relation between concepts is applied - the hierarchical relation of type IS-A, - taxonomy - the similarity of the two terms can be estimated by the distance between the concepts into the taxonomy.

The semantic distance between the concepts depends on the length of the shortest path between the nodes and the overall specificity of the two nodes. The shorter the path from one node to another, the more similar they are. If there are several paths between elements then the length of the shortest path is used [10]. This length is determined by the number of nodes (or edges) in the shortest path between two corresponding nodes of the taxonomy [11], taking into account the depth of the taxonomic hierarchy.

However, this approach is compounded by the notion that all taxonomy edges correspond to homogeneous distances. Unfortunately, in practice the homogeneity of distance in taxonomy is not supported.

In real taxonomies, there is great variability of distances covered by a single taxonomic relation, especially if some subsets of taxonomies (such as biological categories) are much denser than other ones. For example, WordNet [12] contains direct links between either fairly similar concepts or relatively distant ones. Therefore, it is advisable to take into account the semantics of relations between concepts for different taxonomic relationships and to consider the number of instances in subclasses.

In [13] a measure of semantic similarity is based on domain taxonomy that take advantage of taxonomic similarity in resolving syntactic and semantic ambiguity.

Semantic similarity represents a special case of semantic relation between concepts. In [11] the assessment of similarity in semantic networks is defines with the help of taxonomic links. Although other types of links such as "part-of" can also be used for assessment of similarity [14].

All these researches use only some aspects of ontological representation of knowledge limited by:

- hierarchical relations – taxonomic relations between classes and relations between classes and their individuals;
- other types of relations which semantics influence on their weight for the similarity but does not used in logical inference;
- properties of class individuals and their properties that matched in process of similarity estimation but do not analyzed at the level of expressive possibilities. However all these ontological features can be represented by semantic Wiki resources. Therefore we propose to use such Wiki resources as a source of semantic similar concepts for other intelligent systems.

Now a lot of software supports Wiki technologies. One of the most widely used is MediaWiki. The basic functionality allows to create pages connected by hyperlinks, set their categories and publish their content with some structure elements etc. Semantic MediaWiki (SMW) extends semantically this Wiki engine by use of semantic properties of pages [15]. SMW definitely displays content with these annotations in the formal description using the OWL DL ontology language [16].

V. Semantic similarity of concepts into the Wiki resources

We approve the proposed above approach in development of semantic search and navigation means implemented into e-VUE – the portal version of the Great Ukrainian Encyclopedia (vue.gov.ua). This resource is based on ontological representation of knowledge base. To use a semantic Wiki resource as a distributed knowledge base we develop knowledge model of this resource represented by Wiki ontology [17]. This model provides semantic markup of typical information objects (IOs) by domain concepts [18].

Application of semantic similarity estimation for this IR provides the functional extension of Encyclopedia by new ways of content access and analysis on the semantic level.

One of the significant advantages of e-VUE as a semantic portal is the ability to find SSCs. Criteria of e-VUE concept similarity is based on the following assumptions:

- concepts that correspond to Wiki pages of the same set of categories are semantically closer than other e-VUE concepts;
- concepts corresponded to Wiki pages with the same or similar meanings of semantic properties are semantically closer than concepts corresponded

to Wiki pages with different values of semantic properties or those ones with not defined values;

- concepts defined as semantically similar by the both preceding criteria are more semantically similar than concepts similar by one of criteria.

e-VUE users can apply SSC search if they are unable to select correctly the concept category or if they enter concept name with errors. Similar concepts help to find the desired Wiki page. We propose to user retrieval of globally similar (by the full set of categories and values of semantic properties) and locally similar (by some subset of these features) IOs.

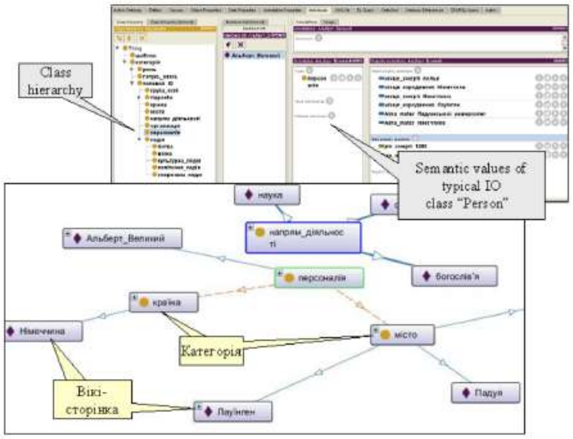


Figure 1. Wiki ontology defined the structure of e-VUE typical information objects (fragment).

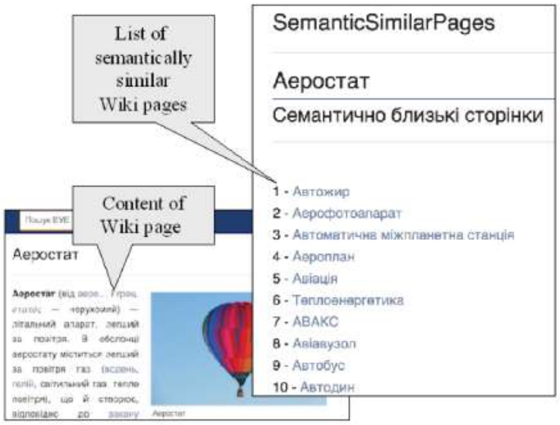
Wiki ontology is a basis for research of similarity between concepts of Wiki resource because the estimates of similarity are based on such elements of semantic markup as categories and properties of these concepts. We can process in these estimates only those characteristics of IOs that are explicitly represented by ontology elements (Fig. 1).

Therefore the development of Wiki ontology defines the expressiveness of search procedure on base of Wiki resource marked by this ontology. Similarity can be defined by any subset of ontological classes and values of their properties but all other content of Wiki pages is not available for this analysis (these characteristics can be received by statistic analyses of from NL processing systems but they are over the consideration of this work).

According to the specifics of encyclopedic IR, it is impractical to search for pages that match all available parameters because some parameter groups are unique (for example, last name and year of birth) and some other ones dependent functionally on other parameters (although they have independent importance e.g. the name in the original language).

Therefore we realize the following examples of local SSPs retrived by:

- the fixed subset of categories of current page (Fig. 2);
- the values of the fixed subset of semantic properties of current page;
- the combination of categories and values of semantic properties of current page.



SemanticSimilarPages

Аеростат

Семантично близькі сторінки

1 - Автожир
2 - Аерофотоапарат
3 - Автоматична міжпланетна станція
4 - Аероплан
5 - Авіація
6 - Теплоенергетика
7 - АВАКС
8 - Авіавузол
9 - Автобус
10 - Автодин

Figure 2. Semantic similar concepts of e-VUE for concept "Aerostat".

It should be noted that built-in tools of Semantic MediaWiki don‘t support search for SSCs (local and global) and all of these requests are realized by special API queries that analyze code of the Wiki pages.

VI. Use of SSCs for information retrieval

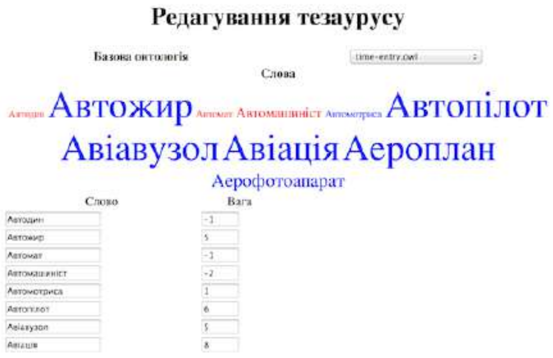
The set of SSCs can be considered as a thesaurus of a user's task for intelligent retrieval systems that support personified search of information pertinent to user needs. An example of such system is semantic retrieval system MAIPS based on ontological representation of background knowledge [19].

This system is oriented on users with stable informational interests into the Web. Ontologies and thesauri provie formalized conceptualization of subject domain pertinent to user tasks. The search procedure in MAIPS is personified by indexes of natural language text readability.

MAIPS uses OWL language for representation of domain ontologies and thesauri, it supports automated thesauri generation by natural language documents and set-theoretic operations on them. Task thesaurus in MAIPS is constructed directly by the user in order to display the specifics of the task which causes these information needs.

We propose the possibility to import this information from external Wiki resources where the set of thesaurus terms is generated as a group of semantically similar concepts. The most pertinent results user receives in situation if Wiki resource is matched semantically by terms of pertinent ontology.

User can improve this thesaurus on base the selected domain ontology by weights of concept importance for



Редагування тезаурусу

Базова онтологія: time-entry.owl

Слова

Слово	Вага
Автодин	-1
Автожир	5
Автомат	-1
Автомашиніст	-2
Автомотриса	1
Автопілот	6
Авіавузол	5
Авіація	8

Figure 3. Visualization of domain thesaurus in MAIPS.

task. MAIPS visualise thesaurus by the cloud of tags (Fig. 3). Users can also manually edit any previously created thesaurus by adding or deleting some terms. In addition, MAIPS realizes such set-theoretic operations on thesauri as union, intersection and complement.

VII. Conclusion

The main idea for the study is to ensure the integration of various intelligent systems that use domain knowledge represented by anthologies. In order to simplify the processing of such knowledge we propose to pass from ontologies to their special case – thesauri. Actuality of this problem is caused by development of intelligent applications based on the Semantic Web technologies [20]. Thesaurus of task contains only limited subset of domain concepts and their relations.

Such knowledge structures are more understandable for users, their creation and processing take less time and qualification. We demonstrate some methods of automatic generation of thesauri by appropriate ontologies and Wikis, and on example of MAIPS we show the usage of such thesauri as a source of domain knowledge for intelligent information retrieval.

References

[1] Golenkov V., Shunkevich D., Davydenko I., Grakova N. Principles of organization and automation of the semantic computer systems development, 2019.

[2] Ray S.R. Interoperability standards in the semantic web, J. Comput. Inf. Sci. Eng., 2002, Vol.2(1), pp.65-69.

[3] “Thesaurus,” Available at: https://en.wikipedia.org/wiki/Thesaurus (accessed 11.11.2019)

[4] Gladun A., Rogushina J. Use of Semantic Web Technologies and Multilinguistic Thesauri for Knowledge-Based Access to Biomedical Resources, International Journal of Intelligent Systems and Applications, 2012, Vol.1, pp.11-20.

[5] Hindle D. Noun classification from predicate-argument structures, ACL-90, 1990, pp. 268-275.

[6] Frakes W. B. , Baeza- Yates R. Information Retrieval, Data Structure and Algorithms, 1992.

[7] Lee J. H., Kim M. H., Lee Y. J. Information Retrieval Based on Conceptual Distance in is-a Hierarchies, 1989, Vol.49(2), pp.188-207.

[8] Lin D. An information-theoretic definition of similarity, ICML, 1998, Vol. 98, pp. 296-304.

[9] Wu Z., Palmer M. Verb semantics and lexical selection, 32nd Annual Meeting of the Association for Comput. Linguistics. Las Cruces, 1994, pp. 133-138.

[10] Rada R. , Bicknell E. Ranking documents with a thesaurus, JASIS, 1989, Vol.10(5), P.304-310.

[11] Rada R., Mili H., Bicknel E., Blettner M. Development and application of a metric on semantic nets, IEEE Transaction on Systems, Man, and Cybernetics, 1989, V.19(1), pp.17-30.

[12] Fellbaum C. WordNet, Theory and applications of ontology: computer applications, 2010, pp. 231-243.

[13] Resnik P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, Journal of Artificial Intelligence Research, 1999, Vol. 11, pp.95-130.

[14] Richardson R., Smeaton A.F., Murphy J. Using WordNet as a knowledge base for measuring semantic similarity between words, Working paper CA-1294, Dublin City University, School of Computer AppUcations, 1994.

[15] Rogushina J. Processing of Wiki Resource Semantics on Base of Ontological Analysis, Open semantic technologies for intelligent systems, 2018, pp.159-162.

[16] McGuinness D. L., Van Harmelen F. OWL web ontology language overview, W3C recommendation, 2004, 10(10).

[17] Rogushina J. Analysis of Automated Matching of the Semantic Wiki Resources with Elements of Domain Ontologies, International Journal of Mathematical Sciences and Computing (IJMSC), 2017, Vol.3, No.3, pp.50-58.

[18] Rogushina J. The Use of Ontological Knowledge for Semantic Search of Complex Information Objects, Open semantic technologies for intelligent systems, 2017, pp.127-132.

[19] Rogushina J. Use of the Ontological Model for Personification of the Semantic Search, International Journal of Mathematical Sciences and Computing (IJMSC), 2016, Vol. 2, No. 1, pp.1-15.

[20] Semantic Web, Available at: https://www.w3.org/standards/semanticweb/ (accessed 20.10.2019)

Использование подобия страниц Вики-ресурсов как инструмента представления предметной области для семантического поиска

Рогушина Ю.В.

Мы предлагаем основанный на онтологиях подход к семантическому поиску в Веб, использующий тезаурусное представление задачи пользователя. Онтологии домена рассматриваются как источник семантической разметки Вики-ресурсов домена.

Мы используем онтологии для формализации структуры базы знаний Вики-ресурсов, которая явно представляет ее основные функции. Онтологии, Вики-ресурсы и тезаурусы задач создаются независимо различными приложениями, но используются в общей технологической цепочке семантического поиска, ориентированного на пользователя. Открытая информационная среда рассматривается как внешняя база данных, содержащия большие объемы гетерогенной и частично структурированной информации.

Вики-онтологии рассматриваются как основа для установления семантического подобия между понятиями предметной области, которые относятся к задаче пользователя. В качестве параметров для количественной оценки семантического подобия используются элементы Вики-онтологии (классы, значения свойств экземпляров классов и отношения между ними).

Received 12.12.2019