Conference Paper

# Applied Ontologies Formation Based on Subject Area Texts

**O. L. Golitsyna and A. V. Zaitseva**

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe shosse, 31, Moscow, 115409, Russia

## Abstract

The problems of the formation of applied conceptual systems based on ontologies constructed automatically from the texts of the subject area documents are considered. Algorithms of operations on ontologies using the thesaurus as a general conceptual basis, unifying the terminology of the subject area, are proposed. Experiments with ontology collection obtained from the texts of design documentation showed that the semantic similarity of the resulting concepts of the system can be increased through the use of thesaurus links.

**Keywords:** ontologies, thesaurus, operations on ontologies, graph theory, semantic similarity, Neo4j, Java

Corresponding Author:
O. L. Golitsyna

Received: 22 July 2018
Accepted: 9 September 2018
Published: 8 October 2018

Publishing services provided by Knowledge E



Selection and Peer-review under the responsibility of the Breakthrough Directions of Scientific Research at MEPhI Conference Committee.



## 1. Introduction

Currently, a great variety of conceptual systems has been created reflecting certain aspects or levels of abstraction of the subject area. From the point of view of terminology, conceptual systems can be very similar but at the same time describe different sides of subject area. In this regard, it is important to solve the following tasks:

1. to find common in conceptual systems, to build a basis of concepts of the subject area;
2. to build a common conceptual system through the union of particular;
3. from the general to go to the particular, that is, to find knowledge reflecting a certain aspect of the subject area; and
4. to build the required level of abstraction.

Solving these problems (e.g., at the level of creating specialized operations) will provide the user with a set of tools for moving from universal conceptual systems

to highly specialized ones, for forming ontologies based on the results of information retrieval, for analyzing subject areas, and so on.

As one of the sources for extracting and forming concepts and links between them considering full-text documents of the subject area, for example, included in the documentary database. A separate document can represent ontology as a semantic search pattern. Thus, the aforementioned tasks can be solved by defining the structure of the ontology, the set of operations and the axioms of performing these operations within a given structure.

Next is presented an approach to the formation of applied ontologies based on the usage of union and intersection operations for ontological graphs constructed from the subject area texts.

## 2. Methods of Describing Ontologies and Operations on Them

An information description of a domain can be represented as a set of objects in this domain and a set of relationships between objects. Objects and relationships can be described by an array of characteristic properties.

In most ontology definitions, all (or some) of these sets are present, considered that the graph is the most constructive structural representation of the ontology. In this case, it is possible to talk about the application graph theoretical operations to ontologies. In [3, 4, 6, 7], the operation of combining (merging) ontologies through the union of graphs is considered, moreover, in [3, 7] said about the operation of intersection.

However, representing an ontology in the form of a graph, where the concepts of the subject area are vertices and the edges are relations between them so the performance of graph operations will require the formulation of rules for comparing vertices and edges.

It has been proposed in [1] to compare concepts at the level of coincidence of sign descriptions. The conclusion about the similarity or difference of ontologies is made on the basis of the results of the intersection of ontologies through the adjacency matrix. This approach can be used only on the assumptions that the vocabulary of the natural language is normalized, there is no synonymy and there is a single description for each real object.

In fact, the problems of synonymy, homonymy and other features of natural language require to consider not only lexicographic equality but also semantic proximity

when comparing concepts. In [2–4, 6], for example, the semantic proximity is associated with the calculation of common semantic features. Among the semantic features, we also consider the characteristic properties of vertices and edges. In [4], additionally, it has been proposed to describe semantics with the help of a meta-graph, combining semantically close concepts and relations in meta-vertices and meta-edges, respectively. [5] proposes to add an ontological graph with a directed tree that defines hierarchical relations between concepts.

It should be noted that with all the expediency of representing ontologies in the form of graphs for the approximation of operations on ontologies to graph theoretical operations, there is important problem of defining criteria for the similarity of concepts. This problem can be solved by introducing measures of semantic similarity describing the context using additional linguistic support in the form of dictionaries, thesaurus and taxonomies as well as combinations of these funds.

In this article, the algorithms for the formation of applied ontologies using graph theoretical operations have been proposed.

The definition of ontology from [8] is accepted as a basis:

$O = < S_f, S_c, S_t, \equiv >$, where

1. $S_f$ is a functional system consisting of concepts and relations derived directly from the texts of documents of the subject area;
2. $S_c$ is a conceptual system – the conceptual basis of the subject area, for example, a thesaurus can act;
3. $S_t$ is terminological system that reflects the properties of natural language at the level of terms related to equivalence or inclusion;
4. $\equiv$ is the operation of comparing the elements of different systems at the sign level, which provide their identity in the functional, system of concepts, and terminological systems.

Using a common conceptual basis in the implementation of operations on ontologies will increase the semantic coherence of the resulting conceptual system through the hierarchical and associative links between thesaurus descriptors.

# 3. The Operations of Union and Intersection on Ontologies

## 3.1. Structural representation of ontologies

According to [8], from the structural point of view, functional and conceptual levels of ontology are represented in the form of labeled, directed, weighted multigraphs $MG_f = < V_f, X_f >$ and $MG_c = < V_c, X_c >$.

In the context of the solution to the task – the formation of applied ontologies based on operations on ontological graphs built on the subject domain texts, $V_f$ is a set of concepts derived from texts; $X_f$ is a set of functional level links (relations) revealed between concepts; $V_c$ is a set of descriptors of the thesaurus; $X_c$ is a set of thesaurus links. Each link of the thesaurus (an arc in a graph) is set out by a triple: $x_k^c = \left(v_i^c, v_j^c, w_k^c\right)$, where $v_i^c$ is the vertex of the curve's start; $v_j^c$ is the vertex of the curve's end, $w_k^c$ is the weight of the curve (identifier of the relation). The set of curve's weights of the thesaurus graph $W_c = \{RT,NT,BT\}$, where *RT* (Related Term) is a link to descriptor association, *BT* (Broader Term) is a link to the upstream descriptor, *NT* (Narrower Term) is a link to the downstream descriptor.

## 3.2. Semantic similarity matrix

In accordance with [8], operations are performed on ontologies, reduced to one conceptual and terminological basis, that is, result of binary operations on ontologies $O_1 = \left\langle S_f^1, S_c, S_t, \equiv \right\rangle$ and $O_2 = \left\langle S_f^2, S_c, S_t, \equiv \right\rangle$, is an ontology $O_{op} = \left\langle S_f^1 op S_f^2, S_c, S_t, \equiv \right\rangle$.

The use of the thesaurus as a general concept basis allows us to introduce and use the measure of semantic similarity for the vertices of ontologies the descriptions of which coincide with the descriptors of the thesaurus.

In [10], a measure of the semantic similarity of descriptors $d_1$ and $d_2$, considering hierarchical and associative thesaurus links:

$S(d_1, d_2) = \alpha S_H(d_1, d_2) + \beta S_A(d_1, d_2)$, where

$S_H(d_1, d_2)$ and $S_A(d_1, d_2)$ is the measures of semantic similarity of hierarchical and associative thesaurus links relatively, α and β is coefficients that determine the link's weight (α + β = 1).

When calculating $S_H(d_1, d_2)$, all ancestors of descriptors in the hierarchy are considered (considering multiple inheritance):

$$S_H\left(d_1, d_2\right) = \frac{\left|UC\left(d_1, H^{d_1}\right) \cap UC\left(d_2, H^{d_2}\right)\right|}{\left|UC\left(d_1, H^{d_1}\right) \cup UC\left(d_2, H^{d_2}\right)\right|}$$

The set $UC(d_i, H^{d_i})$ contains a descriptor $d_i$, and all his ancestors in hierarchical web, viz. – $H^{d_i}$:

$$UC\left(d_i,\ H^{d_i}\right) = \left\{ d_j \in H^{d_i} | \ \exists m = \left(d_i, x_{i_1},\ x_{i_2} \ldots, x_{i_k}, d_j\right) \cup \left(d_i = d_j\right) \right\},$$

where *m* is a route connecting $d_i$ and $d_j$ through higher-level descriptors $x_{i_1}$, $x_{i_2}$,..., $x_{i_k}$.

To calculate $S_A(d_1, d_2)$ in [9], the concept of a 'semantic double' for an associative descriptor is introduced in spite of an association tie has the symmetry property; however, while the relationship type is not set strictly, the exact semantic profile of the symmetry cannot be determined. To compile the set of semes, we should consider the association tie as being single direction (outgoing) and compile the 'semantic double' of the descriptor.

$$S_A\left(d_1, d_2\right) = \frac{\left|UC\left(d_1,\ A^{d_1}\right) \cap UC\left(d_2,\ A^{d_2}\right)\right|}{\left|UC\left(d_1,\ A^{d_1}\right) \cup UC\left(d_2,\ A^{d_2}\right)\right|}.$$

The set of semantic factors $UC(d_i,\ A^{d_i})$ for associative links is constructed as follows:

$$UC\left(d_i,\ A^{d_i}\right) = \left\{ d_j \in A^{d_i} | \ \exists a = \left(d_i, \widetilde{d_j}\right) \cup d_j \cup \left(d_i = d_j\right) \right\},$$

where *a* is the association tie linking $d_i$ and $\widetilde{d_j}$ ($\widetilde{d_j}$ is the 'semantic double' of descriptor $d_j$).

For the calculations in this article, the values of the coefficients of the additive measure from [10] have been uses: α = 0.65, β = 0.35. In practice, such measure values determine that a hierarchical relationship is two times preferable to an associative one.

When implementing operations on ontologies at the level of functional systems, the measure of semantic similarity can be considered as some assessment of the correspondence of the contexts of the use of concepts: if in one ontology, a concept semantically close to the concept in another ontology is encountered, then it is possible to talk about a possible common thematic context. However, it should be noted that the use of measures of semantic similarity is always accompanied by the problem of establishing its threshold value.

Consider the task of charting the semantic intersection of patterns $D_1 = \left(d_1^1, d_2^1, \ldots, d_n^1\right)$ *and* $D_2 = \left(d_1^2, d_2^2, \ldots, d_m^2\right)$. For all pairs of descriptors in the thesaurus (including the combined thesaurus), a semantic similarity matrix may be calculated measuring $n \times m$: $W = (w_{ij})$, where $w_{ij} = S(a_i, b_j),\ i = \overline{1..m},\ j = \overline{1..n}$.

In this case, the sufficient value of the proposed semantic similarity measure in order to include a pair of descriptors within the semantic intersection can be determined not

by the set fixed threshold, but rather by using the local context defined by the descriptors of each pattern, that is, the semantic intersection comprises those descriptors for which the condition of coinciding maximums holds:

$$\max_{j=\overline{1,n}} \left(w_{ij}\right) = \max_{i=\overline{1,m}} \left(w_{ij}\right) .$$

## 3.3. Algorithms of operations of union and intersection for ontologies

Let the multigraphs A = $<V_a, X_a>$ and $B = <V_b, X_b>$ be functional systems of ontologies $O_1$ and $O_2$, respectively and multigraph $T = <V_t, X_t>$ be a general conceptual basis (thesaurus).

Consider the algorithm of the operation of combining two ontologies $O_1$ and $O_2$.

1. The set of intersections of the vertices of the multigraphs A and B is calculated on the basis of equality of sign descriptions: $V_a \cap V_b = V_{ab}$.
2. Forms the sets $(V_a \setminus V_{ab}) \cap V_t$ and $(V_b \setminus V_{ab}) \cap V_t$. If at least one of the sets is empty, go to step 6.
3. For thesaurus's descriptors from sets $(V_a \setminus V_{ab}) \cap V_t$ and $(V_b \setminus V_{ab}) \cap V_t$, a matrix of semantic proximity is calculated.
4. Formed pairs of descriptors for which the condition of coincidence of maxima is fulfilled.
5. For the pairs of descriptors on step 4, routes are created in the multigraph $T$. A set of descriptors of the thesaurus included in the constructed routes $D_t \subset V_t$ are formed.
6. The set of vertices of the functional multigraph of the resultant ontology is the union of the sets $V_a \setminus V_{ab}$, $V_b \setminus V_{ab}$, $V_{ab}$and $D_t$. The set $D_t$ can be empty.
7. The set of arcs of the functional multigraph of the resulting ontology is constructed as a union of sets $X_a$, $X_b$ and routes from step 5 if the set of routes is not empty.

   The algorithm for the operation of the intersection of ontologies $O_1$ and $O_2$includes the following steps:

   First, the aforementioned Steps 1–5.

6. The set of vertices of the functional multigraph of the resulting ontology is a union of sets $V_{ab}$ and $D_t$. The set $D_t$ can be empty.

7. The set of curve of the functional multigraph of the resulting ontology is constructed as the union of the sets Xa ∩ Xb and the routes of step 5 if the set of routes is not empty.

The result of the operations of union and intersection is the multigraph of the functional system and the thesaurus as a common conceptual basis.

## 4. Experimental Research

To implement operations on ontologies and conduct experimental research, a Java program was developed using the Neo4j graph database system. The choice of a graphical DBMS is determined by the structural representation of ontologies and thesaurus in the form of graphs. In addition, Neo4j offers built-in search functions for routes between concepts so it greatly simplifies the implementation of algorithms; besides, there is a web interface that allows to visualize graphs. Figure 1 shows the interface of the program and the web interface of the Neo4j DBMS.

The experiment was carried out on ontology, built on the texts of design documents of the domain 'Atomic Energy'. As a general conceptual basis, the thesaurus INIS IAEA is used.

Figure 2 shows a fragment of the result union operation combines the next ontologies: (1) 'Layout of the main equipment', (2) 'Strength and seismic stability', (3) 'Intracorporeal devices', (4) 'Technical solutions for the modernization of the reactor plant', (5) 'Sources of radiation'. At the functional level, the common nodes of several ontologies were united through the terms 'FA (fuel assembly)', 'ICD (intracorporeal devices)' and others.

The matrix of semantic similarity in the third stage of ontologies union when a fourth is added to the union of the three ontologies is shown in Table 1. In accordance with the condition of the coincidence of the maxima, the routes from the thesaurus for the pairs of terms 'fuels' –'fuel gas', 'safety' – 'reactor safety'.

A fragment of the thesaurus that illustrates the calculation of the measure of semantic similarity is shown in Figure 3. For the case of the descriptors 'fuels' and 'fuel gas', the values $S_H(d_1, d_2)$ and $S_A(d_1, d_2)$ are calculated as follows:

$|\cup C\ (\text{fuels}, H^{\text{fuels}})|$ = 1 is a root top in the thesaurus

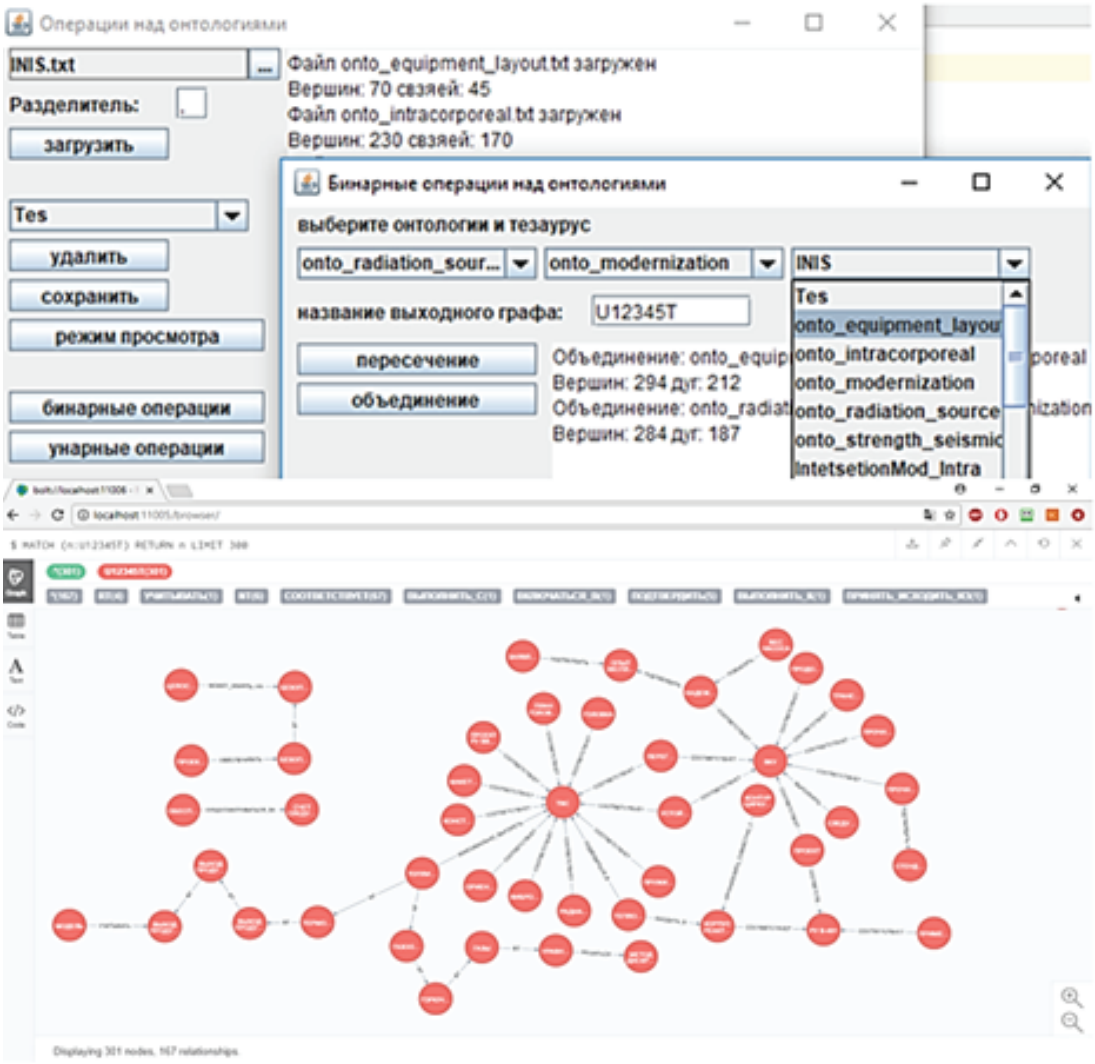


**Figure** 1: User interface windows and web interface Neo4j.

$|\cup C\ (\text{fuel gas},\ H^{\text{fuel gas}})| = 6$

$$\left|UC\left(\text{fuels},\ H^{\text{fuels}}\right) \cap UC\left(\text{fuel gas},\ H^{\text{fuel gas}}\right)\right| = 1$$

$$\left|UC\left(\text{fuels},\ H^{\text{fuels}}\right) \cup UC\left(\text{fuel gas},\ H^{\text{fuel gas}}\right)\right| = 6$$

$$S_H\left(\text{fuels, fuel gas}\right) = \frac{1}{6} = 0.167$$

$|\cup C\ (fuel\ gas, A^{\text{fuel gas}})| = 10$

$|\cup C\ (\text{fuels}, A^{\text{fuels}})| = 15$

$$\left|UC\left(\text{fuels},\ ^{\text{fuels}}\right) \cap UC\left(\text{fuel gas},\ ^{\text{fuel gas}}\right)\right| = 5$$

$$\left|UC\left(\text{fuels},\ ^{\text{fuels}}\right) \cup UC\left(\text{fuel gas},\ ^{\text{fuel gas}}\right)\right| = 25$$

$$S\,(\text{fuels, fuel gas}) = \frac{1}{5} = 0.2$$

$$S\,(\text{fuels, fuel gas}) = 0.65 * \frac{1}{6} + 0.35 * \frac{1}{5} = 0.178$$

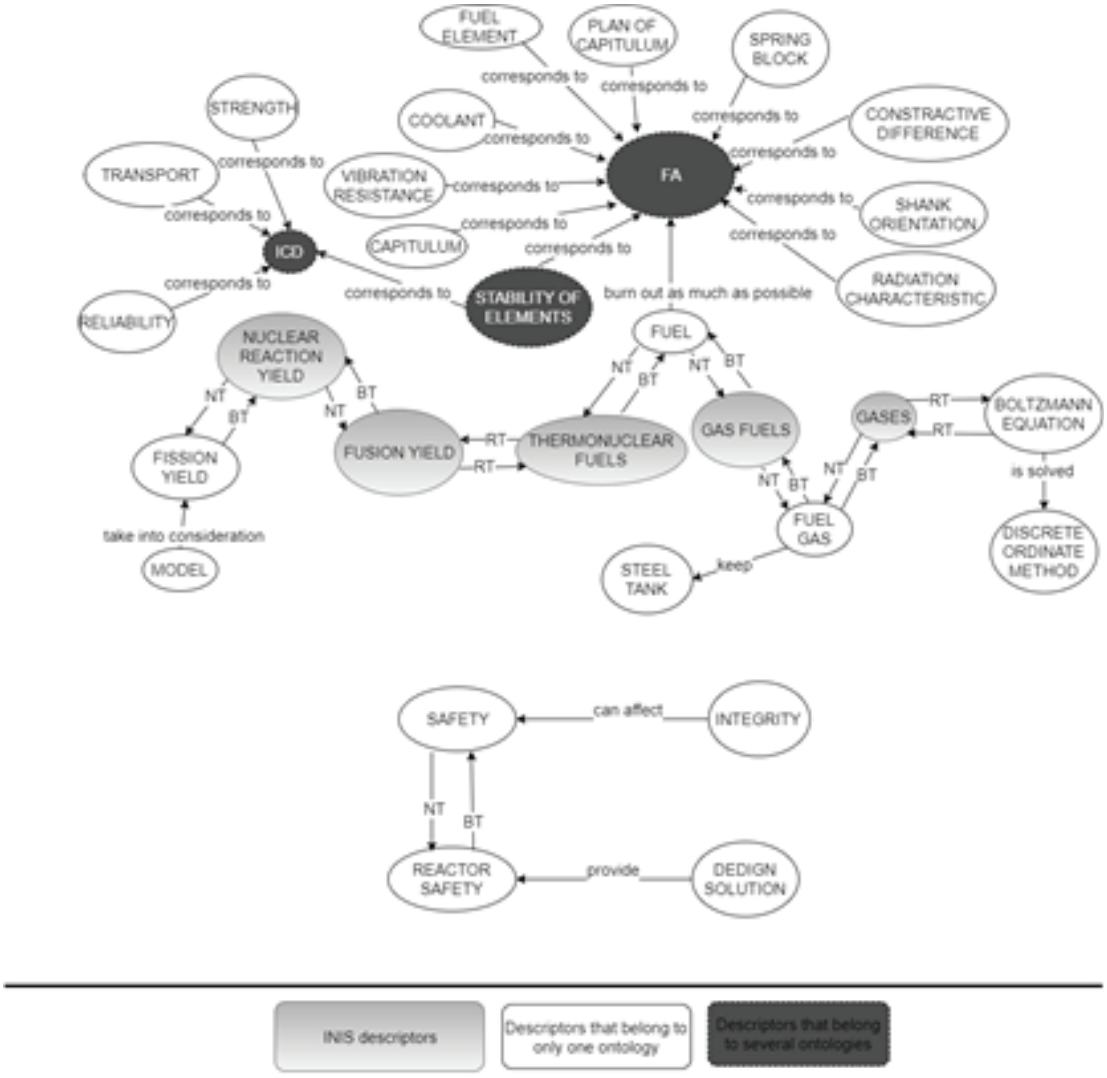


**Figure** 2: A fragment of the resulting graph combining ontologies.

The calculations of the connectivity components of multigraphs for the five ontologies showed the following values: 'Layout of the main equipment' – 62, 'Strength and seismic stability' – 25, 'Intracorporeal devices' – 81, 'Technical solutions for the modernization of the reactor plant' – 41, 'Sources of radiation' – 19. As a result of the union operation, a multigraph consisting of 186 connectivity components was formed at the functional level. After applying the thesaurus, the number of connectivity components was reduced to 170. Thus, using a common conceptual basis made it possible to increase the semantic similarity of the result.

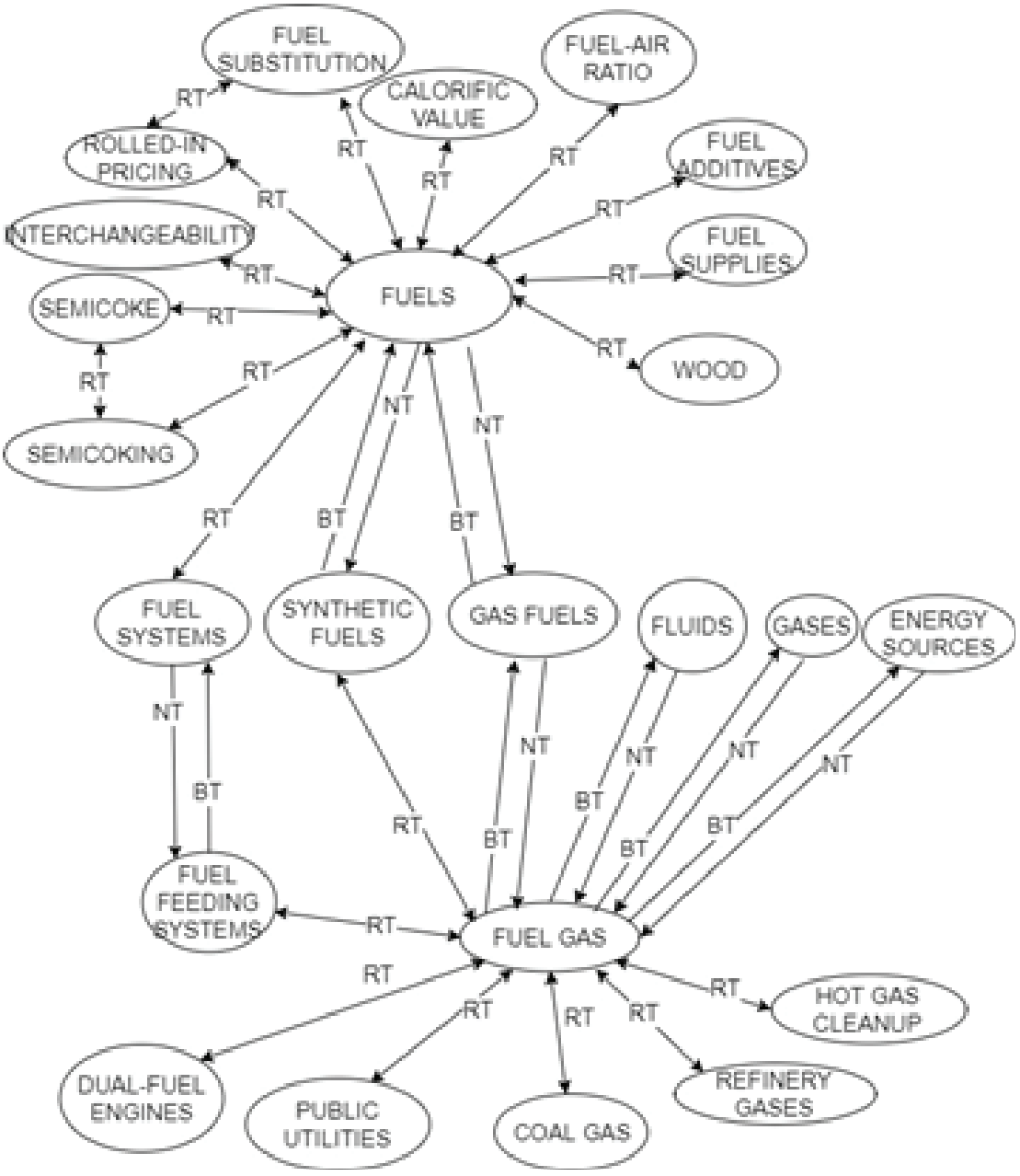


**Figure** 3: Fragment of the thesaurus graph for calculating the measure of semantic similarity between descriptors «fuels» and «fuel gas».

## 5. Summary

The article presents one of the approaches to the use of documentary databases for the formation of applied ontologies. Such ontologies can be constructed, for example, based on the results of factual or thematic searches using various conceptual systems.

The article presents one of the approaches to the use of documentary databases for the dynamic formation of applied ontologies. Such ontologies can be constructed,

Table 1: Semantic similarity matrix.

| | Fuel Gas | Demolition | Safety | Deformation |
|---|---|---|---|---|
| Security | 0.011 | 0 | 0.056 | 0.007 |
| Work | 0.008 | 0 | 0.031 | 0.001 |
| Neutron Fluence | 0 | 0 | 0.004 | 0 |
| Fuels | **0.178** | 0 | 0.026 | 0.008 |
| Space | 0.097 | 0 | 0.069 | 0.007 |
| Reactor Safety | 0.035 | 0 | **0.311** | 0.030 |
| Thickness | 0 | 0 | 0.007 | 0.004 |
| Butter | 0.001 | 0 | 0.005 | 0 |
| Stability | 0.006 | 0 | 0.163 | 0.012 |
| Reactor Operation | 0.007 | 0 | 0.055 | 0.040 |
| Natural Circulation | 0 | 0 | 0 | 0 |
| Reliability | 0.005 | 0 | 0.102 | 0.012 |
| Biological Shielding | 0.034 | 0 | 0.015 | 0.009 |
| Anchoring | 0 | 0 | 0 | 0 |
| Stabilization | 0.007 | 0 | 0.006 | 0.001 |
| Retention | 0.011 | 0 | 0.011 | 0.003 |
| Water | 0.029 | 0 | 0.007 | 0.017 |
| Emplacement | 0 | 0 | 0 | 0 |
| Shielding | 0.025 | 0 | 0.041 | 0.010 |
| Dimensions | 0.001 | 0 | 0.016 | 0.008 |
| Oscillations | 0.001 | 0 | 0.005 | 0.008 |
| Equipment | 0.008 | 0 | 0.094 | 0.007 |
| Reactor Shutdown | 0.033 | 0 | 0.086 | 0.023 |

for example, based on the results of factual or thematic searches involving various conceptual systems.

The representation of ontologies at the structural level in the form of multigraphs allows to use for sequential accumulation of knowledge about the subject area the operations of union and intersection by analogy with graph theoretical operations. The proposed approach to the implementation of operations provides an additional inclusion in the resulting ontology of hierarchical and associative links of thesaurus descriptors in order to increase the semantic connectivity.

Use in the operations of union and intersection the matrix of semantic similarity and the approach to formation of pairs of descriptors on the basis of the coincidence of the maxima allows to solve the problem of determining the threshold value for identifying semantic similarity.

## References

[1] Meenachi. N. M. and M. Sai Baba. (2017). Matrix rank-based ontology matching: An extension of string equality matching. *International Journal of Nuclear Knowledge Management (IJNKM)*, vol. 7, no. 1, pp. 1–11.

[2] Lyubchenko, V. V. and Kavitskaya, V. S. (2013). Method for determining the equivalence of classes of ontologies, in *Proceedings of the Odessa Polytechnic University*, vol. 2, no. 41, pp. 242–246.

[3] Biryukov, D. N. and Lomaco, A. G. (2015). Semantics of knowledge contexts in ontological modeling of conflict subject areas, in *Proceedings of SPIIRAS*, vol. 5, no. 42, pp. 155–179.

[4] Samokhvalov, E. N., Revukov, G. I., and Gapanyuk, Yu. E. (2015). Metagraphs for Information Systems Semantics and Pragmatics Definition. *Bulletin of MSTU*, vol. 1, pp. 83–89.

[5] Palagin, A., Kryvy, S., and Petrenko, N. (2015). Development, research and presentation of functions and operations on ontologies. *International Journal "Information Theories and Applications"*, vol. 22, no. 2, pp. 103–114.

[6] Novototskikh, D. V., Romanov, V. P., and Safonova, M. S. (2016). Dynamic structure of modern innovative enterprises. *Statistics and Economics*, vol. 13, no. 5, pp. 57–62.

[7] Kryvy, S. L. (2016). Formalized ontological models in scientific research. *Control Systems and Machines*, no. 3, pp. 4–15.

[8] Golitsyna, O. L., Maksimov, N. V., and Okropishina, O. V. (2012). The ontological approach to the identification of information in tasks of document retrieval. *Automatic Documentation and Mathematical Linguistics*, vol. 46, no. 3, pp. 125–132.

[9] Golitsyna, O. L., Maksimov, N. V., and Fedorova, V. A. (2016). On determining semantic similarity based on relationships of a combined thesaurus. *Automatic Documentation and Mathematical Linguistics*, vol. 50, no. 4, pp. 139–153.