The Combined Method of Semantic Similarity Estimation of Problem Oriented Knowledge on the Basis of Evolutionary Procedures

V.V. Bova, E.V. Nuzhnov, and V.V. Kureichik(✉)

Southern Federal University, Rostov-on-Don, Russia
vvbova@yandex.ru, {nev,vkur}@sfedu.ru

Abstract. In the article authors proposed the method of problem-oriented knowledge elements search and similarity estimation in subject area ontology given in a form of semantic net. The knowledge relevance is estimated by closeness to a certain similarity estimation measure between concepts included in integrated ontology elements meta-descriptions of intellectual information systems interdisciplinary software environment. To calculate knowledge elements semantic closeness and coherence authors developed a combined model of semantic similarity estimation involving a set of interpreted measure of taxonomical and associative dependences represented in meta-descriptions. The methodology is based on relative position of ontology graph concepts in common hierarchy and on measures of similarity between properties in high-dimensional attribute space. Authors developed an algorithm to calculate parameters values of semantic similarity estimation model on the basis of evolutionary procedures and genetic optimum search. The proposed algorithm is based on the usage of evolutionary processes of reproduction, crossover, mutation and natural selection analogues. To analyze the developed method a set of experiments was carried out. The obtained data shows theoretical significance and prospects of such method and allows us to determine optimal values of algorithm parameters.

Keywords: Semantic similarity · Ontology · Semantic network · Semantic search · Semantic metadata · Genetic algorithms · Genetic operators

1 Introduction

Due to constant growth of information flows in different interdisciplinary technical, economical and social intelligent information systems (IIS), the development of new ways of distributed sources information representation, formalization, systematization, integration and search are relevant today [1–4]. One of the main functions of modern IIS involves semantic search of problem-oriented knowledge elements of a distributed and, thus, heterogeneous representation character.

In this paper the term 'semantic search' is considered as information search which provides comparison and similarity estimation of information objects on semantic level i.e. with the use of knowledge. Existing mechanisms of semantic search [5–7] are based on methods and approaches of knowledge subject area ontological conceptualization.

R. Silhavy et al. (eds.), *Artificial Intelligence Trends in Intelligent Systems*,
Advances in Intelligent Systems and Computing 573, DOI 10.1007/978-3-319-57261-1_8

This paper deals with the method of knowledge bases search, where metadata are formed on the basis of corresponding subject area ontologies represented in the form of semantic net. The knowledge relevance is estimated by closeness to a certain evaluation metric of similarity between concepts included in ontology elements semantic meta-description.

To calculate measures of semantic closeness and coherence between problem-oriented knowledge elements authors propose a combined model of semantic similarity estimation involving a set of interpreted taxonomical and associative meta-descriptions dependences of knowledge elements represented in ontology [8–10].

The algorithm of semantic similarity estimation is based on evolutionary procedures and genetic optimum search operators, which allows us to exclude uninformative or insignificant knowledge elements descriptions, and to manage speed of learning with the use of similarity threshold value assignment [11–13].

2 Problem Statement and Subject Area Analysis

The absence of a 'gold standard' for semantic similarity measures is a well-known problem. Many researchers are focused on the development of semantic similarity estimation and comparison methods used for a wide range of information search problems [2, 7–10].

In this paper factors defining selection of semantic similarity measures applied for formal representation (profile) of the user's query are proposed as follows:

- to select criteria composing the similarity measure: taxonomical relations between concepts – characteristics of ontological structures (the path length, hierarchy depth, etc.) and associative (horizontal) relations defining asymmetrical semantic similarity measure.
- to select criteria importance degree – importance coefficients in hybrid measure of computational semantics [14].

With the use of ontology in works [2, 10] authors proposed the method of special metadata type creation – meta-descriptions including sets of simple proposition statements of the form of 'subject *(s)*–predicate *(p)*–object *(o)*' which are referred as triplets *(t)* and represent main semantics of described knowledge elements. It is noted that such meta-descriptions are important sources of information for search implementation. With the use of meta-descriptions it is possible to significantly improve the search mechanisms functionality. Similarity estimation is called semantic similarity estimation if and only if it is determined on the basis of meta descriptions and query semantics [9].

Thus, to determine a degree of semantic similarity between knowledge elements it is proposed to introduce the measure of distance between their meta-descriptions. The measure represents the combination of several measures of distance between two vertexes (concepts or attributes) of shortest weighted path between ontology graph vertexes [15–17].

It is suggested that all concepts required to compare are located in the united ontology and, thus, in the united taxonomy [15]. If ontologies are separated, they should be united before the analysis [3].

In this paper problem statement authors use following descriptions of ontology components: the ontology O represents the sign system O = <C, E, R, T>, where C denotes a set of concepts (knowledge elements); E denotes a set of concepts examples; T denotes a set of predicates – relation types; R denotes a set of relations assigning following relation types between entities: taxonomical, attributive, quantitative, logical, etc.

Let us introduce following rules and constraints:

(1) On the basis of subject area ontology O, semantic meta-descriptions $m(c_i) = \{t_1, t_2, \ldots, t_{n(i)}\}$ are created for each knowledge elements $C = \{ci\}$, where $n(i)$ is a number of triplets in logical representation of a concept c_i; t_i denotes RDF-triplets having a form of tuples <s_i, p_i, o_i>, where s_i and o_i are included in the union of C_i and E_i, and p_i is include in R.

(2) Each query q created by the user from the set of queries Q consists of the set of triplets $q = \{t_1, t_2, \ldots, t_{n(q)}\}$, where $n(q)$ denotes a number of triplets included in the query q.

The assigned problem involves finding a weight function w, which determines the importance of any triplet $t \in T$ (where T denotes a set of possible triplets) when describing knowledge elements c_i from the query q: $0 \leq w(t, c_i) \leq 1$, where где $t \in T$, $c_i \in C$, $0 \leq w(t, q) \leq 1$, where $t \in T$, $q \in Q$.

For each query q it is required to determine a subset RES of the set of knowledge elements C, which includes relevant concepts for the assigned query q – the result set. C$_i$ is considered as relevant to the query q, if and only if the semantic similarity estimation between them exceeds a certain threshold value of semantic closeness. Therewith, to estimate the similarity between knowledge elements and the query authors propose to use their semantic meta-descriptions [10].

3 The Combined Method of Semantic Similarity Estimation

The key moment in semantic search problem-solving includes the development of semantic similarity quantitative estimations. Existing methods of computational semantics can be subdivided in several categories: measures based on hierarchical structures – methods of conceptual taxonomical closeness estimation using different metrics of finding the length of the shortest path between subject area ontology graph vertexes [2, 16–18]; measures using non-hierarchical relations – methods of relational closeness estimation [5–7]; measures using attribute values [8–10].

The main problem of most measures based on ontological structures is symmetry. Expert analysis shows that similarity measure is not always symmetrical for both hierarchical and attributive relations [5, 6, 8–10]. The relevant problem is semantic similarity estimation between ontological elements that are not related hierarchically, but have concrete problem-specific (horizontal or associative) relation.

Thus, the most promising measures today are hybrid measures, which combine several methods considering ontology structures and relation semantics. This allows us to calculate semantic similarity estimations between ontology elements (concepts, examples, relations – predicates). Similarity estimations are referred as elementary estimations, and similarities between triplets are determined on their basis [2, 5]. Further, similarities estimations between triplets are used to determine similarity between meta-descriptions.

To determine semantic similarity between triplets of queries meta-descriptions M_q and triplets of concepts set M_c let us introduce metrics of distance between ontology nodes on the basis of taxonomy and concepts characteristics, and metrics of density and information value of concepts related thematically. Then, the modified similarity measure can be represented as follows:

$$SIM\left(M_q, M_c\right) = \sum\nolimits_{i=1}^{n} w(t,q)_i Sim^i(c_1, c_i), \tag{1}$$

Where Sim^i is a similarity measure based on a certain criterion, weight $w(t,q)_i$ determines the relative importance of query triplets criterion, weight summary equals to 1, n denotes a number of criteria.

To calculate Sim^i let us introduce the modification of asymmetrical similarity measure [5] considering all types of semantic relations R appropriate for triplet components similarity estimation. In suggested modification graph edges are assigned with a certain weigh coefficients depending on passing direction. This is based on the assumption that a child is more similar to a parent rather than opposite way.

1. For the relation 'parent-child' (*is-a*) two coefficients g and s are assigned, which represents similarity in direction of generalization and detailing.
2. For the relation *instanceOf* (connects concepts and concepts examples) two parameters $\delta, \gamma \in [0, 1]$ are assigned, which represent similarity between the example and the concept and between the concept and the example.
3. Similarity coefficients assigned for the relation *sameAs* (synonyms) and *invertOf* (antonyms) equals to 1 and -1 respectively.
4. For other semantic relations r_i we assign weight coefficient ω, which represents semantic similarity in accordance with these relations.

Let us consider $D = \{c_1, \ldots, c_n\}$ as the path between entities c_1 and c_n (which can be concepts, examples or predicates). The path D has following characteristics:

1. $s(D)$ is a number of edges in detailing direction;
2. $g(D)$ is a number of edges in generalization direction;
3. $ic(D)$ is a number of edges from the example to the concept;
4. $ci(D)$ is a number of edges from the concept to the example;
5. *inv(D)* is a number of inverse relation edges;
6. $oth(D)$ is a number of other relation edges.

The estimation of similarity between entities c_1 and c_2 in terms of criterion i and the path D is determined by the following formula:

$$Sim^i(c_1, c_2) = \max\nolimits_{j=1,..,m}\left\{\left(\left|(-1)^{inv\left(d_j\right)} s^{s\left(d_j\right)} * g^{g\left(d_j\right)} * \delta^{ic\left(d_j\right)} * \gamma^{ci\left(d_j\right)} * \omega^{oth\left(d_j\right)}\right|\right)\right\},) \tag{2}$$

where $d^1, \ldots, d^m$ denotes paths between vertexes c_1 and c_2.

To determine the density and information value of thematically related elements and their meta descriptions let us define the concept weight on the basis of occurrence degree. It is considered that the query concept weight depends on a number of meta descriptions concepts related with it *m(c_i)* which is represented by triplets *m*(c_i) = {t_1, t_2, …, $t_{n(i)}$}, where *n*(*i*) is a number of triplets in concept logical representation c_i [10].

$$w(t,q) = 1 + \ln\left(\varphi_{t,c_i}\left(1 + \sum\nolimits_{\mathrm{c_i \in C}} \varphi_{\mathrm{t,c_i}} SIM(c_1, c_i)\right)\right), \tag{3}$$

Where $\varphi_{\mathrm{t,c_i}}$ is a coefficient of occurrence degree of query triplet *q* in meta description *m(c_i)*, the coefficient is assigned in the algorithm, $SIM(c_1, c_i)$ is a measure of semantic similarity between meta descriptions of concept vertexes in *C*.

4 Genetic Algorithm of Semantic Similarity Estimation

To improve the effectiveness of semantic similarity estimation and to determine semantically prioritized knowledge objects for the purpose of their representation in search model authors propose to use genetic algorithm (GA) which allows us to find suboptimal solutions in polynomial time effectively [19, 20]. GA is an heuristic search algorithm used for optimization and modeling problem solving by means of random selection, combination and variation of searched parameters with the use of mechanisms analogous to natural selection [20]. The generalized structure of genetic search is shown on the Fig. 1.

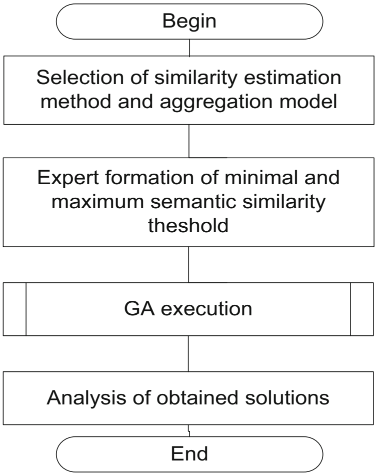
Begin
Selection of similarity estimation method and aggregation model
Expert formation of minimal and maximum semantic similarity theshold
GA execution
Analysis of obtained solutions
End

Fig. 1. The generalized architecture of genetic search

To determine optimal coefficients values authors defined the GA objective function with the use of similarity estimation maximization method:

$$\mathrm{F} = \max\left(SIM\left(M_q, M_c\right)\right). \tag{4}$$

The GA of model parameters values calculation for the purpose of semantic similarity estimation is shown on the Fig. 2.

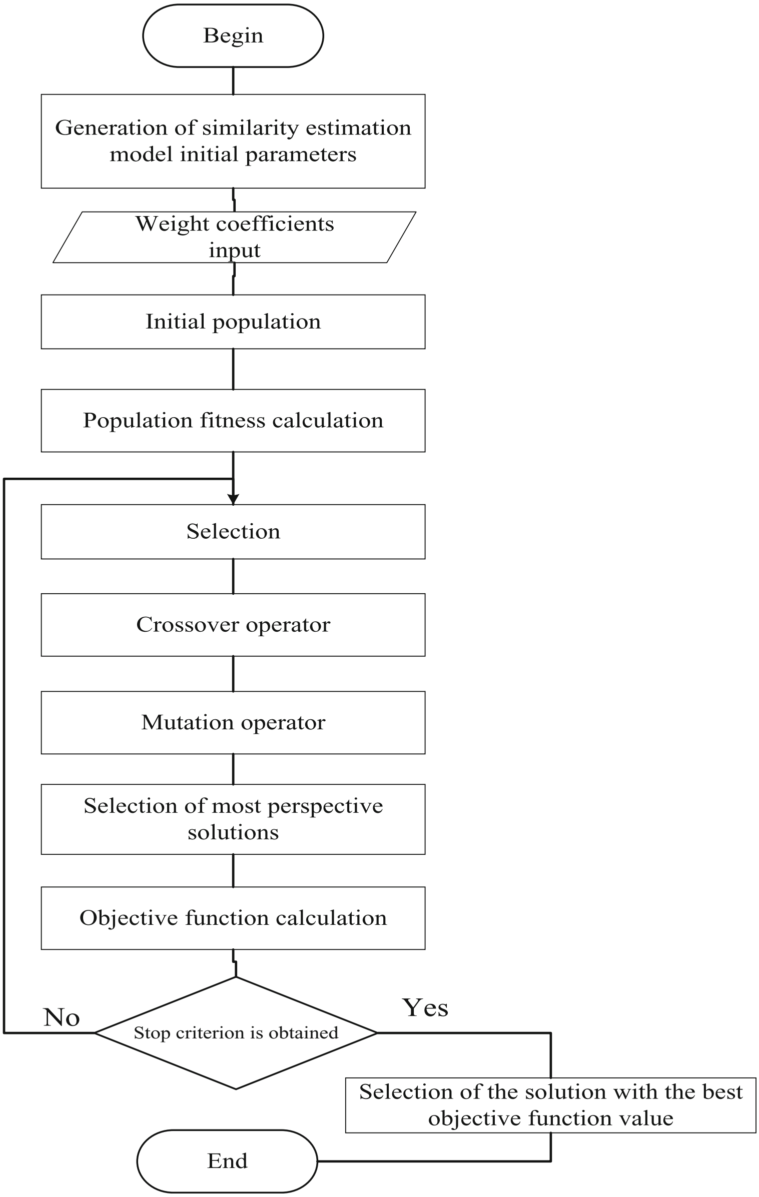
Begin
Generation of similarity estimation model initial parameters
Weight coefficients input
Initial population
Population fitness calculation
Selection
Crossover operator
Mutation operator
Selection of most perspective solutions
Objective function calculation
Stop criterion is obtained
No
Yes
Selection of the solution with the best objective function value
End

Fig. 2. The genetic algorithm of similarity estimation

The first step of the GA is to generate initial parameters of estimation model elements (population size and chromosome length) and to input values of weight coefficient and probability of crossover and mutation operators. Then, the initial population is to be formed on the basis of available learning data from the set C = {c_i} which semantic meta descriptions m(c_i) = {t_1, t_2, …, t_n(i)} was created for. Each chromosome element (gene) is a triplet in logical representation of the concept c_i. To estimate the fitness of each chromosome authors propose to calculate objective function value (4).

The chromosomes selection is carried out by determined method with the use of elitist strategy and partial substitution the least fitted chromosomes by the best fitted ones in terms of saving population size [19]. To generate new specimen set for each pair of selected parent chromosomes it is required to use crossover and mutation operators with pre-assigned probability. The crossover is carried out in a random way with the probability *Pc*. Crossing point is determined in random way within the assigned interval.

The mutation procedure is carried out with the child population obtained as the result of crossover and involves change of gene value by means of randomly selected number from the interval [0, 1] with the probability *Pm*.

The selection of the most perspective solutions is carried out on the basis of probabilities *Pv*, calculated for each population individual with the use of proportional selection [11–13]. After calculation the each chromosome fitness using the formula (4) and the selection of the best one it is required to decide whether to continue the evolutionary procedure of next generation creation or to end the learning procedure. The higher the objective function value is, the higher is the chromosome fitness. The GA work stops under one of following conditions:

(1) if the function *F* obtained expected value;
(2) if the assigned number of iterations (generations) does not improve already obtained valued of the *F*;
(3) if the time allotted for the problem solution is up.

Premature stop of the GA work can occur in case of population degeneration, which means the reduction of chromosomes diversity. The extreme form of degeneration is the condition, when all individuals have identical chromosomes [11].

As a result of the process of artificial evolution including the selection, the crossover, the mutation, the chromosome selection the quality of solutions in population gradually improves.

5 Experimental Research

Computational experiments were carried out for the purpose of the developed GA effectiveness research. To estimate the developed algorithm authors made a comparative analysis with the algorithm based on similarity measure TF-IDF. In the context of the measure each triplet is considered as an individual concept. To estimate the similarity the method uses cosine measure [5, 18] and the MKNN (Mutual KNN) algorithm providing the method of similarity estimation between nearest neighbor (concept triplets and relations) [16].

To carry out experimental research of the developed algorithm effectiveness authors designed software allowing us to implement the iterative procedure of semantic similarity estimation method parameters setting. Experimental research results allowed us to determine the dependence of algorithm execution time on input parameters of the similarity model: n is a number of chromosomes in population; the chromosome is represented as $m(c_i) = \{t_1, t_2, \ldots, t_n(i)\}$, where $n(i)$ is a number of triplets in logical representation of the concept c_i.

The dependences of developed algorithm execution time and TF-IDF and MKNN algorithms on number of similarity estimation model input data are shown on the Fig. 3. Time complexity of the developed algorithm is $O(n^2)$.

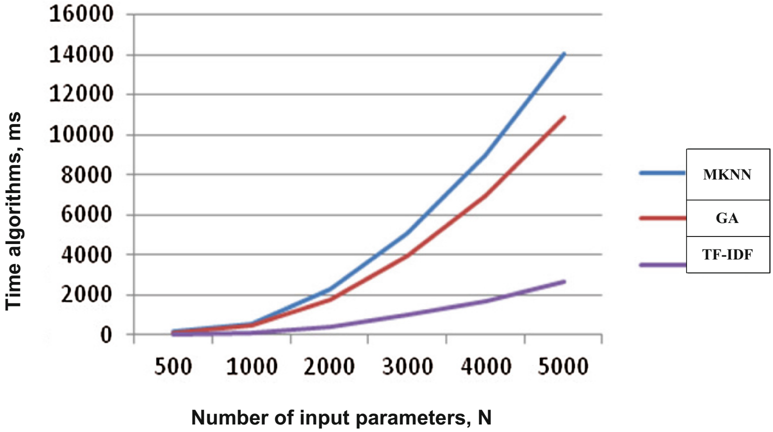

Fig. 3. The dependences of algorithms time on number of input parameters

Authors carried out a set of experiments in terms of completeness and accuracy of relevant concept extraction performed by the MKNN algorithms and the genetic algorithm with the use of two previously described similarity metrics and different number of ontology elements – triplets C_k represented by their meta descriptions. Results show almost linear growth of number of obtained concepts in dependence on the parameter k for both of algorithms (Fig. 4).

GA extracts more relevant concepts in terms of completeness and accuracy of the query. The reason is that the MKNN algorithm excludes pairs of elements that are not nearest neighbors [11].

During the GA search process rules of comparison of ontology elements triplets and query triplets are to be used. Performance accuracy depends on the quality of effective solutions obtained by the GA after each iteration, weighting results of criteria definition of knowledge elements similarity in ontology.

Proposed combined method represents the original mechanism of semantic search, which uses the GA to estimate similarity between ontology elements on the basis of the user's query description semantic metadata.

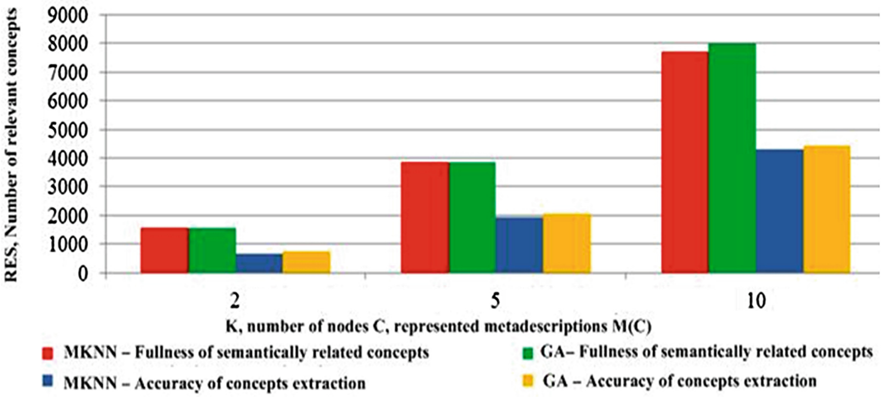

Fig. 4. Dependences of completeness and accuracy of concepts extraction

6 Conclusion

To determine measures of semantic closeness and coherence of problem-oriented knowledge elements authors developed the combined model of semantic similarity estimation which uses a set of interpreted taxonomical and associative dependences of meta descriptions represented in ontologies. The algorithm of semantic similarity estimation is based on evolutionary procedures and genetic optimum search operators which allows us to exclude non-informative and insignificant knowledge elements descriptions and to manage the speed of learning with the use of similarity threshold value assignment.

Acknowledgment. The study was performed by the grant from the Russian Science Foundation (project # 14-11-00242) in the Southern Federal University.

References

1. Bova, V.V., Kureichik, V.V., Zaruba, D.V.: Heuristic approach to model of corporate knowledge construction in information and analytical systems. In: 8th IEEE International Conference on Application of Information and Communication Technologies (AICT 2016), pp. 221–229. IEEE Press, Baku (2016)
2. Kravchenko, Y.A., Kuliev, E.V., Kursitys, I.O.: Information's semantic search, classification, structuring and integration objectives in the knowledge management context problems. In: 8th IEEE International Conference on Application of Information and Communication Technologies (AICT 2016), pp. 136–141. IEEE Press, Baku, Azerbaijan (2016)
3. Bova, V.V., Kureichik, V.V., Legebokov, A.A.: The integrated model of representation model of representation oriented knowledge in information systems. In: 8th IEEE International Conference on Application of Information and Communication Technologies (AICT 2014), pp. 111–115. IEEE Press, Astana (2014)

4. Kuliev, E.V., Kravchenko, Y.A., Kulieva, N.V., Kureichik, V.V.: Problem-oriented knowledge processing on the basis of hybrid approach. In: Proceedings of IEEE East-West Design & Test Symposium (EWDTS 2016), pp. 510–513, Yerevan, Armenia (2016)
5. Nguen, B.F., Tuzovskii, A.F.: Overview of semantic search approaches. In: Proceedings of Tomsk State University of Control Systems and Radio Electronics, vol. 2, pp. 234–237 (2010)
6. Penin, T., Wang, H., Tran, T., Yu, Y.: Snippet generation for semantic web search engines. In: Domingue, J., Anutariya, C. (eds.) ASWC 2008. LNCS, vol. 5367, pp. 493–507. Springer, Heidelberg (2008). doi:10.1007/978-3-540-89704-0_34
7. Knappe, R.: Measures of semantic similarity and relatedness for use in ontology-based information retrieval. Ph.D. thesis. Roskilde University, p. 143 (2006)
8. Bova, V.V.: Conceptual model of knowledge representation in the construction of intelligent information systems. In: Proceedings of SFU, vol. 156, pp. 109–117. TTI SFU, Taganrog (2014)
9. Kryukov, K.V., Pankova, L.A., Pronina, V.A., Shipilina, L.B.: Measures of semantic similarity in ontologies. J. Manage. Problems **2**, 2–14 (2010)
10. Tuzovskiy, A.F.: Working with ontologies in the knowledge management system the organization. In: Abstracts of the Second International Conference on Cognitive Science (CogSci-2006), pp. 581–583. SPb: SPbGU (2006)
11. Bova, V., Zaporozhets, D., Kureichik, V.: Integration and processing of problem-oriented knowledge based on evolutionary procedures. In: Abraham, A., Kovalev, S., Tarassov, V., Snášel, V. (eds.) Proceedings of the First International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'16). AISC, vol. 450, pp. 239–249. Springer, Cham (2016). doi:10.1007/978-3-319-33609-1_21
12. Rodzin, S., Rodzina, L.: Theory of bioinspired search for optimal solutions and its application for the processing of problem-oriented knowledge. In: 8th IEEE International Conference on Application of Information and Communication Technologies (AICT 2014), pp. 142–147. IEEE Press, Astana (2014)
13. Bova, V.V., Legebokov, A.A., Gladkov, L.A.: Problem-oriented algorithms of solutions search based on the methods of swarm intelligence. J. World Appl. Sci. J. **27**(9), 1201–1205 (2013)
14. Castano, S., Ferrara, A., Montanelli, S., Racca, G.: Semantic information interoperability in open networked systems. In: Bouzeghoub, M., Goble, C., Kashyap, V., Spaccapietra, S. (eds.) ICSNW 2004. LNCS, vol. 3226, pp. 215–230. Springer, Heidelberg (2004). doi:10.1007/978-3-540-30145-5_13
15. Panchenko, A.: Technology of the automated thesaurus construction for Information Retrieval. J. Intell. Syst. Technol. **9**, 124–140 (2009)
16. Zhu, H., Zhong, J., Li, J., Yu, Y.: An approach for semantic search by matching RDF graphs. In: Proceedings LAIRS Conference, pp. 450–454 (2002)
17. Gladkov, L.A., Kravchenko, Y.A., Kureichik, V.V.: Evolutionary algorithm for extremal subsets comprehension in graphs. J. World Appl. Sci. J. **27**, 1212–1217 (2013)
18. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. Knowl. Data Eng. **15**(4), 871–882 (2003)
19. Bova, V.V., Kureichik, V.V., Zaruba, D.V.: Data and knowledge classification in intelligence informational systems by the evolutionary method. In: 6th International Conference on Cloud System and Big Data Engineering (Confluence), pp. 6–11, Noida, India (2016)
20. Zaporozhets, D.Y., Zaruba, D.V., Kureichik, V.V.: Hybrid bionic algorithms for solving problems of parametric optimization. J. World Appl. Sci. J. **23**, 1032–1036 (2013)