Assessment of ontological structures semantic similarity based on a modified cuckoo search algorithm

Y Kravchenko and V Bova
Computer-aided design department, Southern Federal University, Rostov-on-Don, Russian Federation

E-mail: ykravchenko@sfedu.ru, vvbova@sfedu.ru

Abstract. Enhancing the effectiveness of knowledge (information) classification and integration based on an assessment of semantic similarity of ontological structures is an urgent scientific problem. Evaluation of equivalent semantic similarity requires considerable computational resources. Consequently, a similarity of predicates of all ontologies’ concepts shall be verified in the exhaustive search. Exact methods do not allow to find a solution in polynomial time, since this problem is trans-computational and requires the use of random search approaches with decentralized control. The paper proposed a modification of the algorithm, inspired by the behavior of cuckoos in the process of nesting parasitism. The cuckoo search optimization algorithm increases the stochasticity of obtained quasi-optimal solutions in comparison with other bioinspired algorithms and increase the search speed. A comparative analysis of the results quality showed that solutions, obtained by the cuckoos’ search, outperform the results of other bioinspired approaches.

1. Introduction
Nowadays, the amount of generated and processed information is increasing exponentially. This fact inevitably leads to regression of effectiveness and relevance of the available data mining methods. Considering such metrics as Petabytes and Exabytes of information, it becomes necessary to upgrade the existing data mining tools. “Big Data” technologies, which had got such name in 2008, allow us to process a large scope of unstructured data by means of unprecedented hardware parallelization of information processes. This kind of solution can be good and effective but technically complicated and very expensive.

The methods of bioinspired search using parallel computational procedures are effectively applied to solve the Big Data problems. Bioinspired algorithms including swarm intelligence algorithms represent the compromise between randomization and local search, being able to generate the diverse solutions in the global search area on the one hand and focus the search in local search on the other. The key problem in wide using of bioinspired algorithms in terms of Big Data problems is the difference between the models of parallelism on the algorithm, iteration or solution levels [1-4]. Each model has its own benefits that should be considered while choosing the proper method for solving the needed task. The development of modified and hybrid bioinspired algorithms and building the ensembles of bioinspired algorithms on the basis of promising boosting and bagging procedures can provide the development of methods of combining the procedures of parallelizing bioinspired algorithms and improve the effectiveness of solving the optimization tasks using artificial intelligence methods in terms of Big Data processing.
To provide multiple and flexible usages and to avoid inconsistency in the data sets used to prepare for classification and integration of information, we suggest using ontological structures as data models. This can help to simplify the search of hidden dependencies and regularities by means of creating the metadata levels. We can distinguish semantic similarity as one of the metrics used to estimate the closeness between the investigated data elements.

2. Assessment of concepts semantic similarity

Similarities and differences of ontologies, used in various intellectual information systems (IIS) for knowledge acquisition and management, is determined by an analysis of semantic dependencies between existing classes of concepts. Semantic dependencies are the basis for agreed changes of elements and set permissible states of intelligent information systems [5-7]. In terms of agreed changes of knowledge elements in the context of the subject area, ontologies’ similarities and differences are found out. Calculation of semantic similarity and semantic relationships between ontologies’ concepts are the basis for these actions. The knowledge integration remains consistency between ontologies of IIS for knowledge acquisition and management and the set of semantic dependencies.

Let dependency between ontologies be reduced to semantic dependencies between concepts, defined on the set of dimensions of ontological concepts P:

1. Equivalence \( g_1 \): \( mapping(P^1) = P^2 \), if \( sim(P^1, P^2) \geq b \), where \( b \) is a semantic similarity threshold at which the concept \( P^1 \) is reflected in the ontology \( O_2 \). When all concepts \( P^1_i \in O_1 \) coincide with all concepts \( P^2_j \in O_2 \) (equivalence of \( R^1 \) and \( R^2 \)), two concepts are equal;

2. Hierarchy \( g_2 \): \( mapping(P^1) = P^2, P^2 = \{P^{2i}\} \), if \( q < sim(P^1, P^{2i}) \geq b \), where \( q \) is a similarity threshold, indicating the absence of concepts’ equivalence;

3. Metalevel \( g_3 \): \( mapping(P^1) = P^2, P^1 = \{P^{i1}\} \), if \( q < sim(P^1, P^{i1}) \geq b \);

4. Division \( g_4 \): \( mapping(P^1) = P^2 \), if \( q < sim(P^1, P^{2i}) \geq b \). Non-zero intersection of concepts \( P^1 \) and \( P^2 \) \( (R^1 \cap R^2 \neq 0) \) points to common attributes [1,2]. This fact confirms the existence of the concept \( P \) which is an upper class for concepts \( P^1 \) and \( P^2 \);

5. Uncertainty \( g_5 \): \( mapping(P^1) = 0, \exists P^1, \forall P^2 \in O_2, sim(P^1, P^{2i}) \leq q \).

The ontological model of the knowledge integration based on the reflection of \( O_1 \) on \( O_2 \) denotes as follows:

\[
Z = \langle O, H^0, G, U, mapping \rangle
\]  

where \( O = \langle P, V, R, C \rangle \) is an ontology of integrated information system; \( H^0 \) is an integration information system with the ontology \( O, G = \{g_1, g_2, g_3, g_4, g_5\} \) is a set of semantic dependencies; \( U: F^0 \rightarrow F^0 \) is reflection, at which \( \forall H^0 \in F^0, \forall g \in G \) is carried out \( s(U(H^0)) \); \( F^0 \) is a set of heterogeneous information systems with common ontology \( O \). \( U \) is a changing of the information system in the form of reflection.

3. Problem statement

The objective function of the given problem can be set as follows:

\[
f(R^1, R^2) \rightarrow \max
\]

where \( R^1 \) and \( R^2 \) are sets of ontologies’ concepts \( O_1 \) and \( O_2 \). Each set of concepts contains specific attributes, belonging to each concept of two ontologies [8,9].

\[
R^1 = \{R_{11}^1, R_{12}^1, \ldots; R_{i1}^1, \ldots; R_{n1}^1\}
\]

where \( i = 1, n \), \( n \) is a number of concepts in the \( O_1 \).

\[
R^2 = \{R_{12}^2, R_{22}^2, \ldots; R_{i2}^2, \ldots; R_{m2}^2\}
\]

where \( i = 1, m \), \( m \) is a number of concepts in the \( O_2 \).

The assessment of equivalent semantic similarity is as follows:
where \( d > 0 \), \( d \) is a threshold value, the equality or excess of which means the presence of equivalent semantic similarity between predicates belonging to the \( P_i^1 \) and \( P_i^2 \). The higher \( d \), the more significant the equivalent semantic similarity between concepts.

Let us use an abstract multilevel ontological architecture of semantic connections in interdisciplinary space of knowledge. The architecture can be set as a graph model. Connections between graph vertices (concepts) are represented by relations \( C^1 \) and \( C^2 \) for ontologies \( O_1 \) and \( O_2 \).

Graph vertices contain attribute values of comparative concepts, i.e. the subset \( R_1^1; R_2^1; \ldots ; R_i^1; \ldots ; R_N^1 \) for the ontology \( O_1 \) and the subset \( R_1^2; R_2^2; \ldots ; R_j^2; \ldots ; R_M^2 \) for the ontology \( O_2 \). In addition, it should be noted that a dimension of attributes subsets for different concepts may vary. In order to address this discrepancy, for each attribute’s subsets let set the maximum possible dimension (MAX) and fill empty fields in concrete subsets with the NULL value (an attribute is absent). Let us set the threshold value \( d \). A population of attributes with number \( a \) and \( b \) (\( a \in [1: N], b \in [1: M] \)) is generated in a random way. All attributes pairs are called a sample. Each pair are set by “upper” and “lower” subsets. During the search process the “upper” subset is a constant and the “lower” subset is a search space (Table 1). For indication of the attribute’s status let us introduce a variable \( \alpha \), which are “1” for constant and “0” for search space. Changing the values of this variable means changing the status of subsets.

### Table 1. Pair of search subsets and its statuses.

| «upper» subset | \( R_{a1}^1 \) | \( R_{a2}^1 \) | \( \ldots \) | \( R_{MAX}^1 \) | \( \alpha = 1 \) |
|---------------|----------------|----------------|-------------|----------------|------------|
| «lower» subset | \( R_{b1}^2 \) | \( R_{b2}^2 \) | \( \ldots \) | \( R_{MAX}^2 \) | \( \alpha = 0 \) |

Maximum similarity (equivalence) between sets of attributes indicates high equivalent semantic similarity of concepts [8,9]. The next section describes the modified cuckoo search algorithm.

### 4. Modified cuckoo search algorithm

Let us describe a modified algorithm inspired by cuckoos’ behavior in the process of brood parasitism for assessment of equivalent semantic similarity. In contrast to a canonical cuckoo search algorithm [5-9], let us suppose that in each nest in addition to an original owner’s egg, there is another one that have a certain degree of similarity with the original egg. According to the suggested coding mechanism, the original egg corresponds to the element of the “upper” subset \( R_{a}^1, \alpha \in [1: N], \alpha = 1 \), and the planted egg corresponds to the element of the “lower” subset \( R_{b}^2, b \in [1: M], \alpha = 0 \). The similarity of these subsets is the initial fitness function that needs to be maximized. The cuckoo’s goal is to find such “egg” that \( R_{y}^1, y \in [1: M], y \neq b \), where \( y \) is a random vertex number from the \( O_2 \), that is not included in the nest population. And in this, the degree of similarity with the egg \( R_{a}^1 \) will be greater than \( R_{b}^2 \). Having achieved this goal, the cuckoo increases the probability of survival of the planted egg, since the owner of the nest will not be able to recognize the difference and will not throw out the egg. Let us represent a step-by-step description of the algorithm.

1. Let take the maximization function (2) as the objective function of the given problem. To calculate the function \( f(R^1, R^2) \) values, expression (5) is used.
2. Let \( R_1^1; R_2^1; \ldots ; R_i^1; \ldots ; R_N^1 \) for the ontology \( O_1 \) and \( R_1^2; R_2^2; \ldots ; R_j^2; \ldots ; R_M^2 \) for the ontology \( O_2 \) are initial data for search and denote expression (4) and (5).
3. To provide an unified dimension of attributes’ subsets, it is introduced the maximum possible dimension with the value MAX. Empty fields in subset are filled a NULL value, which means “no attribute”.
4. Let us set \( d \) as a threshold value, equality or excess of which by the objective function \( f(R^1, R^2) \) value means the presence of equivalent semantic similarity between concepts \( P_i^1 \) and \( P_i^2 \).
5. Let us set \( t = 1, t \in [1: T] \) as an iteration counter, where \( T \) is a maximum number iteration.
6. A new population of nests $s_i, i \in [1:S]$ is initialized, $S$ is a set of host bird’s nest. Pairs are formed from attributes of $O_1$ and $O_2$ ontologies. Coordinates vectors of nests $X_i$ are defined. The probability to discover duof alien eggs in the nest can be set as $p_i \in (0:1)$.

7. On the basis of vertexes from the ontologies $O_2$, not included in the nests’ population, a new cuckoos’ population $c_j, j \in [1:S]$ is generated with coordinates $X_j$.

8. If the number of iteration $t \leq T$, then for each new nest the objective function $f(X_i) = f(R^2_a \cap R^2_\gamma)$ is calculated, otherwise – the algorithm is terminated.

9. A random cuckoo $c_j$ flies in a random nest $s_n$, while changing the subset $R^2_a$ at the subset $R^2_\gamma$, $\forall y, b, y \neq b$.

10. A new objective function value is represented as $f'(X'_i) = f(R^1_a \cap R^2_\gamma)$.

11. If $f'(X'_i) > f(X_i)$, then $f(X_i) = f'(X'_i)$, and $X_i = X'_i, X'_j = X'_i, X_j = X'_j$. Thus, a cuckoo lays the egg in a host bird nest, otherwise – a cuckoo flies to the other host bird nest.

12. Predetermined worst $k$ nests, $k \in [1:S/2]$ are deleted from the population with a probability $p_i$. In this, cuckoo’s coordinates relevant to attributes $R^2_\gamma$ are saved. Next, in accordance with paragraphs 6 and 7 it is generated the same number new pairs of nests (i=k) and required number of new cuckoos $j = S - k$. The value of iteration counter is increased $t + 1$ and go to paragraph 8.

It is important to note that initial coordinates of cuckoos and nests are random, evenly distributed in space vertices of $O_1$ and $O_2$ ontologies.

The cuckoo makes its movements in the form of Levy flights [7], implemented according to the following canonical formula:

$$X'_j = X_j + V \otimes L_{|X|}(\lambda),$$

where $V = (v_y, z \in [1:|X|])$ is a step size which related with the vector $X; L_{|X|}(\lambda) - (|X| \times 1)$ is a random vector of independent real random numbers which relevant to Levy distribution [7].

Levi flights have a high probability of small and sufficient probability of large steps. So, the equation (6) supports agent trajectory as a set of a small number of large (global) flights and a significant number of small (local) ones. If the agent does not immediately fall into the coordinates of the empty nest, then he moves to the nearest possible one.

The equation (6) defines $(|X| \times 1)$-dimensional stochastic process represented by a Markov chain. Here, new agent’s coordinates depend on current coordinates and the transition probability. According to the authors, a random walk (6) is more effective than other algorithms [7]. The nature of the birds’ flight during the food search is close to the Levi flights.

5. Experiments

To confirm the effectiveness of the proposed approach, a software module has been developed. The software includes following functions: creation of the ontology, generation of concept’s attributes, assessment of semantic similarity on the basis of the modified cuckoo search algorithm. Attribute values are represented by average abstract data, received by a random way with a predetermined probability.

Carried out experimental research allowed to obtain a time complexity of the proposed modified algorithm. As a result, time complexity of the algorithm is represented by $O(n^2)$, where $n$ is a number of input parameters. The study of the suggested algorithm speed was conducted in comparison with a bacterial foraging optimization algorithm and a monkey search algorithm (Table 2).

The bacterial algorithm uses an abstract multi-level ontological architecture of semantic relations in the information knowledge space represented as a graph model. The parallelism of the calculations is provided by the cluster-based approach applied together with the bacterial colony of the dimension $S$, which is determined in accordance with the evaluated ontologies’ dimension and represent the number of the agents in the bacterial colony. To improve the effectiveness of search in the successful directions, the proposed algorithm uses greedy heuristics at the reproduction of the bacteria with high quality of the fitness function.
Table 2. Comparison of algorithms’ running time.

| Graph dimension | Modified cuckoo search algorithm | Bacterial foraging optimization algorithm | Monkey search algorithm |
|-----------------|----------------------------------|------------------------------------------|------------------------|
| 1000            | 40.90                            | 52.70                                    | 48.10                  |
| 2000            | 43.64                            | 56.57                                    | 52.14                  |
| 5000            | 47.72                            | 57.16                                    | 54.78                  |
| 10000           | 55.54                            | 65.71                                    | 62.05                  |
| 20000           | 64.80                            | 83.98                                    | 81.86                  |
| 50000           | 75.11                            | 102.35                                   | 101.34                 |

The suggested modified algorithm is more stochastic in comparison with the bacterial algorithm described before, which provides a more targeted and gradual search of the quasioptimal solutions in accordance with the selected vector of the bacteria’s swimming. The search is implemented step-by-step in the graphs of both first and second ontologies. Conversely, the monkey search algorithm finds the solution step-by-step in one ontology making the local jumps in it, while the second ontology is used to make the random global jumps only. Hence, the modified bacterial algorithm can provide better solutions while the modified monkey algorithm can give a higher speed of work.

The results of experiments have demonstrated the superiority of the suggested modified algorithm over considered analogous. The high speed of the modified cuckoo search algorithm is due to its stochasticity, and the required quality of solutions is due to the effectiveness of search procedures.

6. Conclusion

This paper presents the development of the modified optimization algorithm using the cuckoo search method for the semantic similarity assessment of ontological concepts. Unlike the canonical cuckoo search algorithm, the original coding of solutions is proposed. An important advantage of the cuckoo search algorithm is the small number of free parameters, which makes the algorithm more versatile and effective compared to other bioinspired optimization algorithms.

To estimate the quality of the proposed modified cuckoo search algorithm, experimental studies were conducted that confirmed the effectiveness of the development algorithm in comparison with the bacterial foraging optimization algorithm and the monkey search algorithm. According to the estimates, the time complexity of the developed algorithm does not exceed the polynomial one.

References

[1] Xiong C, Power R and Callan J 2017 Explicit semantic ranking for academic search via knowledge graph embedding Proc. of the 26th Int. conf. on WWW 1271–79
[2] Resnik P 1995 Using information content to evaluate semantic similarity in a taxonomy Proc. 14th Int. Joint Conf. on Artificial Intelligence (Montreal) p 448
[3] Ehrig M and Sure Y 2004 Ontology mapping – an integrated approach The Semantic Web: Research and Applications. Proc. 1st European Semantic Web Symposium vol 3053 (Berlin: Springer) pp 76–91
[4] Slota M and Leite J 2014 The rise and fall of semantic rule updates based on se-models J. TPLP 14 869–907
[5] Kureichik V, Kravchenko Y 2013 Bioinspired algorithm applied to solve the travelling salesman problem World App. Sc. J. 22 1789-97
[6] Kravchenko Y, Kursitsys I and Bova V 2018 Knowledge sifters in MDA technologies IOP Conf. Series: Journal of Physics: Conf. Series 1015 042024
[7] Sousa T, Silva A and Neves A 2004 Particle Swarm based Data Mining Algorithms for classification tasks J. Par. Comp. 30 767–83
[8] Miller G 1991 Contextual correlates of Semantic Similarity J. Language and cognitive processes 6 1–28
[9] Nguyen H 2006 New Semantic Similarity techniques of concepts applied in the biomedical domain and wordnet Thesis for the Degree Master of Science (Univ. of Houston-Clear Lake)