Combined Method for Integration of Heterogeneous Ontology Models for Big Data Processing and Analysis

Viktor Kureychik and Alexandra Semenova(✉)

Autonomous Federal State Institution of Higher Education,
Southern Federal University, Rostov, Russia
kur@tgn.sfedu.ru, alexaforum@rambler.ru

Abstract. In the given paper a combined method for integration of heterogeneous ontology for big data processing and analysis is proposed. This allows perform semantic search through heterogeneous information resources, represented by different ontologies. The fundamental difference of the proposed approach is that it allows obtaining optimal weights on the basis of which the optimal alignment of ontologies is carried out. Performed calculations validate the productivity of the proposed method.

Keywords: Big data · Ontology · Ontology alignment · Swarm intelligence · Multiobjective optimization

1 Introduction

Recent advances in scientific and technical areas, including computer and information technologies, have led to the fact that one of the main trends in the modern science development is a significant increase of experimental data volumes and the associated problems of storage and processing. It is clear that further successful development of research projects is possible only if the scientific community will learn to process and analyze extra-large amounts of data, and extract from them new knowledge. Formalization of unstructured data is one of the solutions for solving the problem of big data processing. So we may apply the formal ontology - a modern paradigm of computing resources that describe the knowledge of the world and subject areas.

Many Russian and foreign researchers investigated the problem of application of ontologies to the processing and analysis of big data. Nevertheless, the growth of unstructured information flows, the need to improve the quality of its analysis and processing in information systems requires the development of new methods for effective processing of big data from various domains. Shared ontologies accumulation is seen as a mechanism of unlimited accumulation of knowledge about the world. Currently the problem of comparing and matching ontologies at the level of alignment, i.e. finding semantic correspondences between the elements of two independently developed ontologies, is not solved yet. The problem of ontology alignment is to find such a structure and permissible parameters that provide the optimal values by one or more quality criteria.

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_30

The purpose of this paper is to analyze the use of the ontological approach for big data processing and development of method for integration heterogeneous ontological models based on an evolutionary approach.

In this paper we propose a combined method for ontology alignment based on semantic similarity of concepts and multi-objective optimization of similarity weights. Modification of this method is the application of swarm intelligence algorithm for finding the weighting factors. The main advantages of the proposed approach are: finding the key concepts, eliminating of the subjectivity of their descriptions and dependence from the point of view of ontology developers. Generalized operation of concepts comparison along with the parsing and sorting algorithm will improve the quality of ontology alignment procedure. Therefore the interaction of heterogeneous information systems is provided. The fundamental difference of the proposed approach is that it allows obtaining optimal weights on the basis of which the optimal alignment of ontologies is carried out. Performed calculations validate the productivity of the proposed method.

2 The Problem of Unstructured Information Integration

The concept «BigData» refers to the data sets of extremely large volume and complexity that standard tools are not able to carry out their capture, storage, management and processing within a reasonable time for practice. Big data is characterized by parameters such as [1]:

- volume: big data doesn't sample; it just observes and tracks what happens;
- velocity: big data is often available in real-time;
- variety: big data draws from text, images, audio, video; plus it completes missing pieces through data fusion;
- validity: the property of being genuine, a true reflection of attitudes, behavior, or characteristics.

Unstructured information (NI) is information that either does not have a predetermined data structure or not organized in the given order [1]. Ontology is a formal explicit description of classes (concepts) in the domain, the properties and attributes of each concept (slots), and the restrictions imposed on slots (facets) [3]. Domain structure (ontology) development is the first step to bring the NI to a structured form. Each individual domain is only a subset of unstructured data set, so for the best possible coverage of the data and, as a consequence, a more complete analysis is necessary to allocate the maximum possible number of different domains to be analyzed [4].

Nowadays the heterogeneous information systems accumulate a considerable amount of knowledge. While integration of these systems a problem of classifying and structural representation of knowledge from different domains appears [5]. Different contexts of ontologies created by different communities are reflected in the different approach to the concepts of the specification that has become one of the causes of heterogeneity. As a result, the semantics of the concepts in the contexts described by different ontologies may be similar in different approaches of description of their structure: the structure, constraints and level of detail.

Linguistic approach for ontology integration involves the creation of a formal ontology of the upper level, the interaction of which with other ontologies is implemented on the basis of linguistic relations. Linguistic relations of such ontological models such as *synonym_of* (synonym), *hyponym_of* (gipernim), *overlap_of* (overlap) and other linguistic relationships allow formally implement the mapping of terms [6]. The disadvantage of this approach is that the linguistic relationships are not always adequately reflect the semantics because of the ambiguity of the language variables.

Ontology integration approach based on shared vocabulary allows to build an integrated model of the different domains of knowledge due to the fact that almost any notion of a vocabulary can be associated with any other term. However, in this case the integration of ontologies typically performed with some ontology developer limits and tips. This approach is implemented by viewing the two ontologies, finding in them synonyms, as well as by conflicts resolution and the creation of a third ontology.

Integration of heterogeneous ontologies may be performed based on alignment of instances: the semantic relationships between two classes of heterogeneous ontologies are merged on the basis of the intersection of sets of instances. Typically, ontology classes described multiple instances that allow better define the semantics of a class. Therefore, the association of ontologies based on instances is more effective [7].

The main disadvantage of the majority of unstructured data fusion methods is the need to engage an expert to confirm the correctness of the detection of the similarities and differences of semantic concepts. Thus, ontological approach provides a new level of information integration. For semantically correct interconnection of heterogeneous information systems it is necessary to compare ontology and to find out their differences and similarities. This problem is solved by semantic similarity techniques of concepts of ontonologies.

3 Ontology Integration Based on Semantic Similarity

An approach for integrating unstructured data based on a comparison of the results of concepts, their attributes and relationships between concepts on the level of ontology alignment is suggested [8]. Each concept of the domain ontology is defined as a unit of knowledge and identified by a name and a type. We define concept as [8]

$$C_i = (N_i, T_i), \tag{1}$$

where

- N_i – a unique name (identifier) of i-th concept;
- T_i – a type of i-th concept.

Let's $C = \{C_i | i = 1, 2, \ldots, n\}$ be a set of concepts and $R = \{R_1, R_2, R_3\}$ a set of relations between concepts. At that,

- $R1$– relation of inheritance (relation of «class-subclass»), $R1(C1,C2)$, where $C1$ – is a superclass of $C2$;
- $R2$ – relation of aggregation (relation of «whole-part»), $R2(C1,A')$: attributes of concept $C1$ are included in a set of attributes of all concepts A';

- $R3$ – relation of association (semantic relations), having transitive relation.

Let's consider the following expression of formal ontology [9]:

$$\text{OHT} = (C, P, R, A), \tag{2}$$

where

- C – denotes concept (or classes) set for a specific domain;
- P – set of concepts attributes. Property is a component of the relation p(c,v,f), where c ∈ C – ontology concept, v – property value, associated with c and f defines restrictions for facets in v. One of restriction is a type, capacity, and range.
- R = {r | r ⊆ C × C × R t} – set of binary relations between concepts in C. There is the following variety of relation types: 1:1, 1:many, many:many. The basic set of relations are: synonymOFF, kindOFF, partOFF, instanceOFF, propertyOFF.
- A – axioms' set. Axiom is a rule that specify cause-and-effect relationship.

The problem of heterogeneous ontology combination is formulated as follows: given two regular ontologies create a third regular ontology, which is the concept of the input ontologies, as well as additional restrictions and relationships, if they are required. Building an ontology mapping O_1 on O_2 ontology is to find for each concept of ontology O_1 similar to it concept of ontology O_2.

Different ontologies may have overlapping sets of attributes, relations and concepts. The resulting ontology, maintaining the specifications in such a way as to include all the possible relations between concepts and did not contain equivalent (duplicate) concepts is developed on the basis of multiple source ontologies. So mappings on the same concepts of ontologies match. The resulting ontology defines the concepts of compliance and interpretation of the rules that can successfully establish their interaction. The purpose of the integration of unstructured data is to maintain compliance of the set of ontologies to the defined set of semantic relations. Semantic relations defined on the ontology O is taken as z-predicate set on O'. If there is a semantic relation z in ontology O, we write $z\ (O)$.

Initially heterogeneous ontologies are not associated between each other. Therefore we need to find semantically similar elements of ontologies. For the numerical evaluation of semantic similarity of ontology concepts an approach based on the results of studies of A.F. Tuzovskiy was chosen [10]. In the proposed method similarity measure consists of three components: attributive, taxonomic and relational measures. This method has been adapted for the calculation of the semantic similarity of two heterogeneous ontologies. Modification of this method is an application of particle swarm algorithm for finding the weights. The definition of lexical component is calculated as the ratio of the intersection of sets of words (synonyms) in terms of their association.

Let's $S^T(c_i, c_j), S^R(c_i, c_j), S^A(c_i, c_j)$ be a semantic measure of two concepts based on their position, attribute concept attribute value. Weights t, allow to control commutation process of semantic similarity of two concepts.

To estimate lexical similarity of two concepts $S^T(c_i, c_j)$ sets of concept terms are $\text{PL}_p(c_i)$ and $\text{PL}_p(c_j)$ compared, common and different components are found [8]:

$$S^T(c_i, c_j) = \begin{cases} 1, \text{если } c_i = c_j \\ \frac{|PL_p(c_i) \cap PL_p(c_j)|}{|PL_p(c_i) \cup PL_p(c_j)|}, \quad \text{если } c_i \neq c_j \end{cases} \tag{3}$$

where $PL_p(c_i) = \{L_i \in L | P_c(c_i) = L_i\}$ – is a set of lexical terms of a concept c_i.

To estimate relation similarity it is assumed that if two concepts have similar relation with the third concept they are more similar than two concepts having different relations. Let's assume that [8]

$$C_r(c_i) = \{c_j \in C | R_1(c_i, c_j) \vee R_2(c_i, c_j) \vee R_3(c_i, c_j) \vee c_j = c_i\} \tag{4}$$

− is a set containing concepts with relations R_1, R_2, R_3.

Define association relation of concepts such as [8]

$$R_A(c_j) = \{c_i : c_i \in C_r(c_j)\}. \tag{5}$$

Calculate the sum lexical similarity values for the set of concepts (c_j) and $R(c_i)$.

$$S_{RA}(R_A(c_i), R_A(c_j)) = \sum\nolimits_{c_i \in R_A(c_i), c_j \in R_A(c_j)} S^T(c_i, c_j) \tag{6}$$

Relation similarity measure $S^R(c_i, c_j)$ allows to evaluate similarity of two concepts based on concept similarity from a set (c_i) [8].

$$S^R(c_i, c_j) = \begin{cases} 1, \quad \text{если } c_i = c_j \\ \frac{S_{RA}(R_A(c_i), R_A(c_j))}{|R_A(c_j) \cup R_A(c_i)|} \quad if\ c_i \neq c_j \end{cases}$$

Compare attributes of two concepts. A set of attributes pertaining to a concept:

$$A^{Ci} = \{A_k^{Ci}, k \in [1 \ldots n_1]\}, \tag{7}$$

where n_1 – number of attributes in a concept c_i.

$$A^{Cj} = \left\{A_k^{Cj}, k \in [1 \ldots n_2]\right\}, \tag{8}$$

where n_2 − number of attributes in a concept c_j.

Attributive similarity measure $S^A(c_i, c_j)$ of concepts c_i and c_j is calculated by matching of common attributes: $A^{C_i} \cap A^{C_j}$. Attributive similarity measure $S^A(c_i, c_j)$ satisfy axioms of independence and resolvability, and is defined by the expression

$$S^A(c_i, c_j) = \frac{|A^{C_i} \cap A^{C_j}|}{|A^{C_i} \cup A^{C_j}|}, \tag{9}$$

where A^{C_i} – is a set of attributes of a concept c_i;

A^{C_j} – is a set of attributes of a concept c_j.

Similarity measure S (c_i, c_j) of concept c_i of ontology O and concept c_j of ontology O' is defined

$$S(c_i, c_j) = t \cdot S^T(c_i, c_j) + r \cdot S^r(c_i, c_j) + \alpha \cdot S^A(c_i, c_j) \tag{10}$$

where t, r, a – are the coefficients, defining importance of similarity measures $S^T(c_i, c_j), S^R(c_i, c_j), S^A(c_i, c_j)$, respectively,

$$t, r, \alpha \in [0; 1], t + r + \alpha = 1, S(c_i, c_j) \in [0; 1]. \tag{11}$$

$$\begin{cases} S(c_i, c_j) = 1, & \textit{if concepts are equivalent}, \\ S(c_i, c_j) = 0, & \textit{if concepts are different}. \end{cases}$$

Heterogeneous ontology integration problems belong to a class of NP-hard optimization problems, and can be solved by evolutionary algorithms.

4 Multi-objective Optimization of Similarity Weights Calculation

Consider the modified swarm intelligence method for ontology alignment, using multi-objective optimization approach [11]. Algorithm for optimization of similarity weights calculation by particle swarm intelligence is depicted on Fig. 1 [12].

In this work we propose to apply PSO calculation based on multi-objective optimization. Multi-objective optimization or parallel programming is the process of simultaneous optimization of two or more conflicting objective functions in a given domain. Multi-criteria optimization task is formulated as follows [13]:

$$\min_{\vec{x}} \{f_1(\vec{x}), f_2(\vec{x}), \ldots, f_k(\vec{x})\}, \quad \vec{x} \in S \tag{12}$$

where $f_i : R^n \to R$ – is a $k(k \geq 2)$ of objective functions. Solution vectors $\vec{x} = (x_1, x_2, \ldots, x_n)^T$ belong to a non-empty domain set S.

Multi-objective optimization task is to find a vector of target variables satisfying cash constraints and optimizing the function of the vector whose elements correspond to the objective function. These functions form a mathematical description of the satisfactory test [14].

Consider the set of data, wherein the data lines are different similarity coefficients and columns - the relations between two different ontologies. For subsequent combining these similarity coefficients in one metric optimum weights were obtained. The proposed approach is possible to find the set of weights that meet the criteria of similarity which allows obtaining an optimal alignment. In the process of evaluation of swarm intelligence generalized function was calculated f_{integ}:

$$f_{\text{integ}}(O1_i, O2_i) = \sum_{k=1}^{7} w_k \times F_k(salign_{ij}), \tag{13}$$

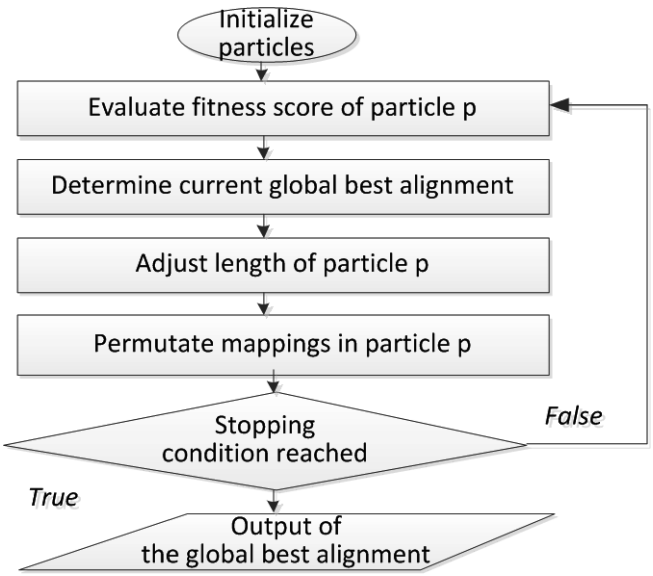

Fig. 1. Algorithm for optimization of similarity weights by particle swarm intelligence

where $\sum_{k=1}^{7} w_k = 1$.

If $f_{integ}(O1_i, O2_i)$ exceeds a threshold then $salign_{ij}$ is a valid alignment. In such a way all valid alignments are defined. Subsequently, using these valid alignments and reference alignments objective functions are calculated.

The method consists of the following stages.

(A) Initialization. The population is called a swarm, and it is composed of m number of appropriate solutions or particles. Each particle has n positions or cells comprising n weighting coefficients corresponding to n different similarity measures. Initially, for each cell of particles the value from 0 to 1 is selected randomly. Once selected primary swarms, calculated the corresponding values of fitness. The initial velocity of the particles of each cell is zero. The inputs to the proposed method are a swarm of 50 and weighting factors c_1 and c_2. The threshold value is chosen to be 0.5. The algorithm performed within 30 iterations.

(B) Objective function. The proposed approach works with multiple objective functions: the accuracy and recall of the search. Accuracy is the criterion of correct alignment found in the resulting alignment. Recall is the criterion of finding the right alignment found from a given reference alignment. The criterion of «accuracy» is calculated by the following formula:

$$P = \frac{|A| - |A \cap R|}{|A|} \tag{14}$$

The criterion of «recall» (quantitative parameter of the results of information retrieval, which is determined by dividing the amount granted as a result of the

search of relevant concepts to the total number of relevant concepts present in the ontological model) is calculated by the following formula

$$R = \frac{|A| - |A \cap R|}{|A|} \tag{15}$$

As proposed multi-objective Particle Swarm Optimization is implemented as minimization problem so first objective is computed as (1-precision) and second objective is computed as (1-recall).

(C) Next Generation Swarm is Produced by Evaluating the Position and Velocity. Each cell or position represents the weight (normalized value of the cell) with respect to the similarity measure. The cells inside the particles contain values from 0 to 1, and the speed of each gene is given zero values. Using the information obtained in the previous step, the position and velocity of each particle of each cluster are updated. Each particle keeps track of the best position it has reached, which is also called pbest. In terms of multi-criteria approach, the position is selected for pbest, whose adaptation of the particle dominates the other devices. And the best position among all particles called global best or gbest. When the particle moves to a new position at a rate that its position and velocity changes in accordance with Eqs. (16) and (17) [13]:

$$v_{ij}(t+1) = w \times v_{ij}(t) + c_1 \cdot r_1 \cdot \left(pbest_{ij}(t)x_{ij}(t)\right) + c_2 \cdot r_2 \cdot \left(gbest_{ij}(t) - x_{ij}(t)\right) \tag{16}$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \tag{17}$$

where t is a time stamp, j-th cluster of i-th particle. Velocity $v_{ij}(t+1)$ is calculated by usage of previous velocity $v_{ij}(t)$, pbest and gbest. Then a new position $x_{ij}(t+1)$ is obtained by adding new velocity with current position $x_{ij}(t)$. c_1 and c_2 are set to 2, r_1 and r_2 are random values from the range from 0 to 1.

After applying non-dominated sorting and crowding distance sorting to the archive, a Local Search is conducted for obtaining the better approximation of weights regarding optimal alignment. In the Local-Search algorithm, the best particle replaces the worst particle of the new generation.

5 Experimental Research

Experimental researches performed with different number of ontology entities have shown that the algorithm has polynomial time complexity $O(n^2)$. Diagram of time complexity of the algorithm is shown on Fig. 2.

We've compared the suggested approach with single objective optimization by accuracy and recall (Table 1) [7].

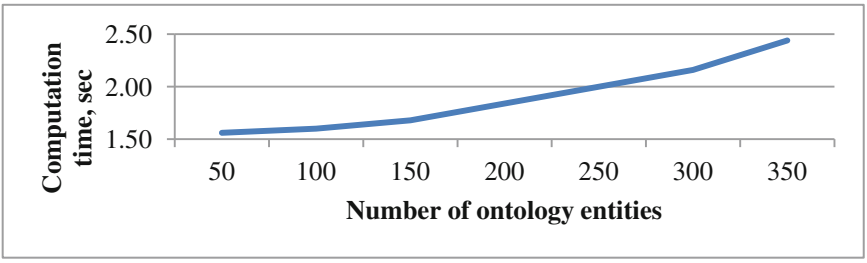

Fig. 2. Diagram of time complexity of the algorithm

Table 1. Results of experiments

Method	Accuracy	Recall	F-measure
Multi-objective optimization	0,8957	1,00	0,9873
Optimization by one objective function (accuracy)	0,7856	0,946	0,8583
Optimization by one objective function (recall)	0,3666	1,00	0,0103

Efficiency of the suggested approach for the criterion accuracy is 0,81428 (high). Again with respect to the f-measure the table shows that our method outperforms other single objective versions. Therefore, the proposed method is effective.

6 Conclusion

The ontological approach for big data processing is considered. The approach for integrating unstructured data is based on comparison of the results of concepts, their attributes and relationships between concepts on the level of ontology alignment. Each concept of the domain ontology is defined as a unit of knowledge and identified by a name and a type. The purpose of the integration of unstructured data is to maintain compliance of the set of ontologies to the defined set of semantic relations. Heterogeneous ontology integration problems belong to a class of NP-hard optimization problems, and can be solved by evolutionary algorithms. In this work we propose to apply PSO calculation based on multi-objective optimization. Experimental researches performed with different number of ontology entities have shown that the algorithm has polynomial time complexity.

The main advantages of the proposed approach are: finding the key concepts, eliminating of the subjectivity of their descriptions and dependence from the point of view of ontology developers. Generalized operation of concepts comparison along with the parsing and sorting algorithm will improve the quality of ontology alignment procedure. Therefore the interaction of heterogeneous information systems is provided. The fundamental difference of the proposed approach is that it allows obtaining optimal weights on the basis of which the optimal alignment of ontologies is carried out. Performed calculations validate the productivity of the proposed method.

References

1. Analysis of unstructured data: applications of text analytics and sentiment mining. Elektronnyj resurs, data obrashcheniya aprel (2016). https://support.sas.com/resources/papers/proceedings14/1288-2014.pdf
2. Lapshin, V.: Ontologii v komp'yuternyh sistemah. Nauchnyj mir, Moskva (2010)
3. Gruber, T.R.: The role of common ontology in achieving sharable, reusable knowledge bases. In: Allen, J.A., Fikes, R., Sandewell, E. (eds.) Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning, pp. 601–602. Morgan Kaufmann, Cambridge (1991)
4. Ontologicheskie metody i sredstva obrabotki predmetnyh znanij: monografiya. Palagin, A. V., Kryvyj, S.L., Petrenko, N.G. - Lugansk: izd-vo VNU im. V. Dalya, 324 s (2012)
5. Kopajgorodskij, A.N. Primenenie ontologij v semanticheskih informaci-onnyh sistemah. Ontologiya proektirovaniya № 4(14), str.90–98 (2014)
6. Semenova A.V., Kurey, F. [chik V.M. Obzor metodov analiza i obrabotki lingvi-sticheskoj ehkspertnoj informacii. Informatika, vychislitel'naya tekhnika i inzhenernoe obrazovanie, № 1 (12). S. 25–77 (2015)
7. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
8. Bubareva, O.A.: Matematicheskaya model' processa integracii informacionnyh sistem na osnove ontologij. Bubareva, O.A., Popov, F.A. Sovremennye problemy nauki i obrazovaniya, № 2 (2012). http://www.science-education.ru/102-6030. Data obrashcheniya 19 April 2016
9. Semenova, A.V., Kureychik, V.M.: Domain ontology development for linguistic purposes. In: 9th International Conference on Application of Information and Communication Technologies (AICT) (2015)
10. Tuzovskij, A.F.: Metod obedineniya ontologij predmetnyh oblastej znanij. Izvestiya Tomskogo politekhnicheskogo universiteta, T 309, № 7, C. 138–141 (2006)
11. Bock, J.: Ontology alignment using biologically-inspired optimisation algorithms. Dissertation, Karlsruher Institut fur Technologie (KIT) Fakultut fur Wirtschaftswissenschaften (2012)
12. Semenova, A.V., Kureychik, V.M.: Application of swarm intelligence for domain ontology alignment. In: Abraham, A., Kovalev, S., Tarassov, V., Snášel, V. (eds.) Proceedings of the First International Scientific Conference "Intelligent Information Technologies for Industry" (IITI 2016). Advances in Intelligent Systems and Computing, vol. 450, pp. 261–270. Springer, Cham (2016)
13. Semenova, A.V., Kureychik, V.M.: Multi-objective particle swarm optimization for ontology alignment. In: 10th International Conference on Application of Information and Communication Technologies, pp. 141–147 (2016)
14. Gladkov, L.A., Kureychik, V.V, Kureychik V.M.: Bioinspirirovannye metody v optimizacii: monografiya. - M: Fizmatlit, S. 384 (2009)