EVALUATION OF THE ONTOLOGICAL METHOD OF INTEGRABLE DATA INCONSISTENCY IDENTIFICATION

Alexander A. Kropotin, Yuliya V. Bidulya, Alexander G. Ivashko

Abstract

This paper evaluates the performances of the method of description logics implementation to identify semantic conflicts and inconsistency of integrable data in the context of relational databases integration presented by entity-relationship diagrams during information systems integration.

At first, we present a developed method method of description logics implementation to identify semantic conflicts and inconsistency of integrable data which based on original mathematical model and algorithms of description logics calculus applying.

We split evaluating into two sets. In the first set of experiments we compared how the classification and consistency checking of the ontology terminology could be applied for the identify semantic conflicts of integrable data. In the second set of experiments the consistency checking of the ontology was estimated for the identify semantic conflicts and inconsistency of integrable data.

Keywords: Ontology; description logics; database; ER-model

1. Introduction

The development of information technology has led to the emergence of a huge number of diverse information systems. As a rule these systems are designed to automate a business processes, to store the personal and reference data, and to generate different reports based on stored data [1]. It often occurs that the data in these systems duplicate each other, which creates big problems in their integration. That is why ensuring interoperability and the integration of information systems are the most urgent tasks [2].

Due to the fact that the volume of information processed is constantly growing, information technology specialists need to provide an effective data exchange between information systems avoiding the duplication and loss of information [1; 2; 3]. Increasingly, information systems developers are faced with many weak linked and disparate sources of information and they have to constantly perform routine actions to perform low-level operations with data in a heterogeneous environment. A special place is occupied by questions of data integration, which may not be consistent and lead to the emergence of structural and semantic conflicts. Traditional methods for solving this problem mainly based on linking between the elements of systems constructed manually, and are aimed at solving the problem for each specific
case [1; 2; 4]. In this case, a change in the structure of one of the systems may cause the disruption of the program complex and requires re-integration [1; 3]. Thus, traditional methods do not provide a sufficiently flexible solution for creating the data integration system [2; 4].

At the present time, there are many methods for solving this problem, but most effectively is to apply the technologies of the semantic web and related data [1; 2; 3; 4; 8; 9; 10; 14].

These technologies have special standards for machine representation of knowledge bases and software products that implement inference algorithms [5]. To solve this problem, we developed the method of applying the descriptive logics formalism to identify automatically conflicts of entity-relationship conceptual schemes of relational databases [7; 14; 15]. The purpose of this paper is to evaluate the developed method using computational experiments.

2. Method description

The developed method of applying the formalism of descriptive logics [13; 14; 15] is based on the previously developed mathematical model of the ontological representation of relational databases based on their semantic description and the method of applying the calculus of descriptive logics to identify the conflicts of these schemes.

As a mathematical model of the ontological representation of relational databases based on their semantic description, the ontological model of conceptual objects was previously proposed [7; 14; 15], which provides rules for describing of conceptual schemes of the entity-relationship relational databases as axioms and statements of descriptive logic. These rules are based on the syntactic similarity of relational databases and ontologies, are substantiated by their semantic difference and are aimed at ensuring the possibility of identifying semantic conflicts of metadata and inconsistencies of integrable data through checking the feasibility of their ontologies. In accordance with the ontological model rules for representing conceptual objects, the elements of the entity-relationship schemes of databases are represented as follows [7; 14; 15]:
- each set of entities and each set of values are represented as the atomic concept;
- each set of links and attribute are represented atomically as role;
- specific elements of such sets are represented as corresponding individuals.

The method of applying the calculus of descriptive logics for identifying conflicts of entity-relationship conceptual schemes of relational databases allows us to reduce the process of identifying of semantic conflicts of metadata and the inconsistency of data which occurs when their integration, to solving of algorithmic problems of consistency and classification of terminology, and the consistency of the ontology of these data [7; 14; 15]. It should be noted that under the consistency of ontology or terminology is understood as the existence of the given ontology or terminology model in which all concepts and roles are not empty.

3. Evaluating

The applicability of the ontological model of conceptual objects and the method for identifying semantic conflicts and the inconsistency of the integrable data were experimentally evaluated of the using 861 conceptual database schemas that represent 368 possible cases of conflicts: attribute-attribute, attribute-entity, entity-relationship, dependency, behavior, inheritance, key, data types, acceptable values, permitted values, and the relevance of values.

This evaluation included three stages:
1. Preparation of experimental data, represented by sets of conceptual schemas of the entity-relationship of relational databases, each of which represents different cases of occurrence of a corresponding conflict of two such schemes. Based on the proposed ontological model for representing conceptual objects and the algorithm for constructing terminological axioms and statements based on the semantic representation of the relational databases of each of such a set, ontologies are constructed according to the OWL 2.0 standard.
2. Conducting computational experiments, focused on the tabular algorithm tasks using the experimental data and revealing the conflicts that they simulate in accordance with the proposed way of applying the formalism of descriptive logics to identify semantic conflicts and data discrepancy of their integration. There are tasks performing:
   - the consistency checking and terminology classification to identify conflicts: attribute-attribute, attribute-entity, entity-relationship, behavior, inheritance and key;
   - the consistency of the ontology entirely to detect conflicts of the format of values, the presentation of values and the relevance of values;
   - the consistency of terminology and ontology consistency to identify dependency conflicts, data types, acceptable values, permitted values.
3. Evaluation of the computational experiment results, which are based on a comparison of the number of actual semantic conflicts of entity-relationship conceptual schemas, and ones revealed during the computational experiments provided for corresponding inconsistencies in terminological axioms and / or statements.
The essence of the computational experiments consists in solving algorithmic problems of ontology consistency, consistency and terminology classification on sets of experimental data, and in identified inconsistencies calculation.

In accordance with the conflict classification of the conceptual schemes, the experimental data sets for the experimental evaluation were divided into two groups: the metadata conflict group and the data conflict group.

As a result of the experimental evaluation, on the sets of experimental data of the group of metadata conflicts, in all, 813 databases were processed and 328 cases of conflicts of their metadata from 350 possible were revealed. The results of the experimental evaluation on the sets of experimental data of the group of metadata conflicts are given in Table 1.

<table>
<thead>
<tr>
<th>Experimental data</th>
<th>Number of semantic conflicts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conflict type</strong></td>
<td><strong>Database number</strong></td>
</tr>
<tr>
<td>Attribute-attribute</td>
<td>68</td>
</tr>
<tr>
<td>Attribute-entity</td>
<td>40</td>
</tr>
<tr>
<td>Entity-relationship</td>
<td>132</td>
</tr>
<tr>
<td>Dependencies</td>
<td>20</td>
</tr>
<tr>
<td>Behavior</td>
<td>14</td>
</tr>
<tr>
<td>Inheritance</td>
<td>12</td>
</tr>
<tr>
<td>Key</td>
<td>24</td>
</tr>
<tr>
<td>Data types</td>
<td>424</td>
</tr>
<tr>
<td>Acceptable values</td>
<td>16</td>
</tr>
<tr>
<td>Permitted values</td>
<td>16</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>813</strong></td>
</tr>
</tbody>
</table>

Table 1. The results of the experiment on the sets of experimental data of the group of metadata conflicts.

As a result of the experimental evaluation, on the sets of experimental data of the data conflict group, in all, 48 databases were processed and 18 cases of conflicts emergence out of 18 possible. The results of the experimental evaluation on the sets of experimental data of the data conflict group are given in Table 2.

<table>
<thead>
<tr>
<th>Experimental data</th>
<th>Number of semantic conflicts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conflict type</strong></td>
<td><strong>Database number</strong></td>
</tr>
<tr>
<td>Format of values</td>
<td>16</td>
</tr>
<tr>
<td>Presentation of values</td>
<td>16</td>
</tr>
<tr>
<td>Relevance of values</td>
<td>16</td>
</tr>
<tr>
<td><strong>Beero</strong></td>
<td><strong>48</strong></td>
</tr>
</tbody>
</table>

Table 2. The results of the experiment on sets of experimental data of the data conflict group.

An experimental evaluation of the developed method of applying the descriptive log formalism to automate the process of identifying the conflicts of conceptual schemes of the entity-relationship of relational databases clearly showed that:

1. The developed model of the ontological representation of conceptual relational databases is oriented to ontological modeling of only the semantic component of such bases, excluding their structural representation.

2. The proposed method of using the descriptive calculus to identify the semantic conflicts of database schemas and data inconsistencies allows us to identify only the semantic inconsistency of such schemes, ignoring their structural differences. Such applicability of the method is due to the tasks and results of research into integration problems, which showed that in view of the semantic coherence, structural heterogeneity is solvable by standard data integration tools.

3. The proposed method cannot be applied for rigid typing of data and does not extend to conflicts formed by differences only in the cardinality constraints of equivalent link sets. This limitation of the method is due to various assumptions about the openness and closure of the world, which is accepted in ontologies and databases, respectively.
4. Conclusion

The paper presents the results of an experimental evaluation of the applicability of the developed method of using the descriptive log formalism to automate the process of identifying conflicts of conceptual schemes of the essence-relationship of relational databases.

The experimental estimation was carried out on 861 database on conflict detection: attribute-attribute, attribute-essence, entity-relationship, dependencies, behavior, data types, allowed and allowed values, value format, representation of values and relevance of values.

The results of the experiments showed that the proposed method of identifying semantic conflicts and inconsistency of integrable databases is based on solving algorithmic problems of consistency and classification of terminology, and the consistency of ontology, makes it possible to reveal 95% of cases of different occurrence of semantic inconsistency.

The applicability of the proposed method for identifying semantic conflicts and the inconsistency of integrable data that is impossible with rigid data typing is determined experimentally and does not extend to conflicts formed by differences in only the cardinality constraints of equivalent sets of constraints.

In the future work it is planned to develop a model for the ontological representation of business processes and its further comparison with the ontological representation of databases in order to identify the correspondence of the database to the business process and vice versa. This approach will avoid the occurrence of many conflicts and inaccuracies when designing a database for each particular business process.

5. References


