The Knowledge Increase Estimation Framework for Integration of Ontology Instances' Relations

Adrianna Kozierkiewicz(✉) and Marcin Pietranik

Faculty of Computer Science and Management, Wroclaw University of Science
and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{adrianna.kozierkiewicz,marcin.pietranik}@pwr.edu.pl

Abstract. The previous authors' research showed that it is not only possible, but also profitable to estimate a potential growth of a level of knowledge that appears during an integration of ontologies. Such estimation can be done before the eventual integration procedure (or at least during such) which makes it even more valuable, because it allows to decide if a particular integration should be performed in the first place. Until now, authors of this paper prepared a formal framework that can be used to estimate the knowledge increase on the level of concepts, instances and relations between concepts. This paper is devoted to the level of relations between instances.

Keywords: Ontology · Integration · Knowledge increase

1 Introduction

With a growing number of available ontologies, their diversity and sizes, an efficient method of performing their integration is invaluable. However, no matter how efficient and well designed the integration algorithm, with large ontologies that are required to be merged, a method of estimating whether or not such integration will result in profitable outcomes can become useful.

We understand the integration itself as a task defined as follows: *for given n ontologies $O_1, O_2, ..., O_n$ one should determine an ontology O^* which is the best representation of given input ontologies.* Assuming that $\tilde{O}$ is a set containing all possible ontologies of interest this task can be achieved using a function σ with a signature $2^{\tilde{O}} \to \tilde{O}$ that accepts as an input any subset of $\tilde{O}$ and returns an outcome of the integration, denoted as O^*. Obviously, if only one ontology is given as an input, it is also returned as an output. Formally, $\sigma(O_1) = O_1$ for a selected $O_1 \in \tilde{O}$.

Our research are not focus on an issue of creating algorithms of the ontology integration (so a materialisation of the aforementioned function σ). However, in order to preserve the integrity of the paper, we have developed an integration algorithm that will be used to illustrate our ideas. As a foundation of the ontology integration algorithm, we tested and developed a procedure taken from the

© Springer Nature Switzerland AG 2018
A. Lupeikiene et al. (Eds.): DB&IS 2018, CCIS 838, pp. 172–186, 2018.
https://doi.org/10.1007/978-3-319-97571-9_15

literature [14]. Thus, we can concentrate on determining a formal framework that could be used to estimate a potential growth of knowledge that can be gained when ontologies are merged.

Going to the merits, by *"growth of knowledge"* we understand an indicator that shows whether or not the integration ends with a profitable outcome in the potential growth of knowledge point of view. For example, when two input structures contain the same knowledge, then nothing is gained from their integration. On contrary, if contents of input ontologies are entirely separate, then their merging can be invaluable.

Formally the task of creating a tool that allows such estimation can be defined as follows: *for given ontologies $O_1, O_2, ..., O_n$ from the set $\tilde{O}$ and an integration function σ one should determine a function Δ with a signature $\Delta : 2^{\tilde{O}} \to R$ that represents the increase of knowledge between input ontologies and a result of their merging $\sigma(O_1, O_2, ..., O_n) = O^*$.* Obviously, such approach can be decomposed into several subtasks that cover a wide array of ontology elements, namely: concepts, relations among them, their hierarchy, their instances and relations connecting instances.

In our previous publications [7–9], we developed a set of methods that allow to estimate the growth of knowledge during the integration of concepts (using a function Δ_C) and their hierarchy (using a function Δ_H), relations (using a function $\Delta_{R,C}$), and instances (using a function Δ_I). The missing piece is the estimation of knowledge that can be achieved through the integration of ontologies on a level of relations that described connections between concepts' instances. This level focuses on expressing how particular instances of the defined concepts interact with each other. In contrast to the level of concepts' relations (that we have covered in [9]) that is used to describe possible connections that may occur (e.g. *a man can be a husband of a woman*) the level of instances' relations focus of particular connections (e.g. *Joe is a husband of Jane*). Merging such statements may entail several difficulties concerning properties of relations such as their reflexivity, asymmetry or transitivity, that didn't appear when the integration of concepts relations has been considered.

For clarity, we assume that we deal only with the integration of only two ontologies. Therefore, the main contribution of the following paper can be formally defined as follows: For given two ontologies O_1 and O_2 that both contain sets representing relations between their instances denoted as R^{C_1} and R^{C_2} one should determine a function Δ_R with a signature $\Delta_R : \tilde{O} \times \tilde{O} \to [-1, \infty)$ representing an estimation of a potential growth of knowledge that can appear during the integration of O_1 and O_2 on the level of instances' relations.

The article is organised as follows. The next section contains a short description of a similar research found in the literature. In Sect. 3 the basic notions used throughout the paper are given. Section 4 formally describes a sought function Δ_R in terms of postulates that it must comply to. Eventually it ends with an algorithm that we have developed to calculate its values. It is then statistically analysed in Sect. 5. The paper ends with Sect. 6 which provides some conclusions and a brief overview of our upcoming work.

2 Related Works

Estimating the effectiveness of the ontology integration has not been widely investigated. In the literature it is possible to find some measures which allow to calculate how efficient the integration process is, however none of this research take into the consideration the potential growth of knowledge as the result of merging two or more ontologies.

Precision and recall with respect to a reference mapping are the most popular methods of measuring the quality of ontology matching [12,14] which can be formalised as a problem of an information distance metric like in [15] or in [6]. Some modifications of these measures have been proposed by [5]. Euzenat also provided a semantics for alignments based on the semantics of ontologies and has designed semantic precision and recall measures. The definition of these measures are independent from the semantics of ontologies. Such approach requires the use of logical reasoning, where both correspondences and ontologies are considered.

In [6] authors have demonstrated a novel measure named link weight that uses semantic characteristics of two entities and Google page count to calculate an information distance similarity between them. The proposed measure has been used to align ontologies semantically. In [4] authors have evaluated a wide range of string similarity metrics, along with string preprocessing strategies such as removing stop words and considering synonyms, on different types of ontologies. Additionally, the most efficient metrics for merging process have been pointed out. In [11] some ontology metrics used to measure distance between ontologies' semantics, rather than ontological structures are proposed. Those cohesion metrics have been: Number of Ontology Partitions (NOP), Number of Minimally Inconsistent Subsets (NMIS) and Average Value of Axiom Inconsistencies (AVAI).

In [13] authors introduced quality measures that are based on the notion of mapping incoherence that can be used without a reference mapping. This measure provides a strict upper bound for the precision of a mapping and can therefore be used as a guideline for estimating the performance of matching systems.

Ceusters [3] developed a metric, which is designed to allow assessment of the degree to which the integration of two ontologies yields improvements over either of the input ontologies. Authors noticed the fact, that input ontologies can contain some mistakes. However this paper contains only theoretical considerations about some factors which can be used to assess the adequateness of both the original ontologies and the results of matching or merging.

Some papers are devoted to measuring the quality of a single ontology. In [2], a tool called Ontology Auditor has been described. Authors have developed a suite of metrics that assess the syntactic, semantic, pragmatic, and social aspects of the concerned topic. Authors have designed many metrics like: overall quality, syntactic quality, lawfulness, richness, semantic quality, interpretability, consistency, clarity, pragmatic quality, comprehensiveness, accuracy, relevance, social quality, authority, history and described each of them.

OntoQa [17] is another model that analyses ontology schemas and their populations and describes them through a well defined set of metrics. Authors defined two categories of metrics. The schema metrics address the design of the ontology which indicate the richness, width, depth, and inheritance of an ontology schema. The instance metrics were grouped into two categories: knowledge base metrics and class metrics. The former describe the knowledge base as a whole. The latter which describe the way each class that is defined in the schema is being utilised in the knowledge base.

The ROMEO [19] methodology identifies requirement that an ontology must satisfy and maps the requirements to evaluation measures like consistency, conciseness, completeness, coverage, correctness, clarity, expandability, minimal ontological commitment. Similarity functions (ontology evaluation approaches) has been developed in other systems like: OntoClean [18] or OntoMetric [10] however, the mentioned solution evaluate the quality of the single or set of ontologies and does not consider the integration them.

All of the described metrics have many disadvantages. Many of them are calculated in separation with each other and none of the metric can give a big picture of the performed integration. Some measures are extracted from information retrieval field and do not consider the expressive structure of ontologies, while other assess only a single ontology and it cannot be applied in the estimation of the quality of the ontology integration process.

In this paper we propose measures that do not have flaws described above. It is devoted to the estimation of the potential increase of knowledge. The proposed method allows to decide about the profitability of the eventual integration process, which is a continuation of our previous research presented in [7–9]. Until now we have developed methods of the estimating the knowledge increase during the integration of ontologies on the level of concepts, instances and relations and hierarchies of concepts. In this article we focus on the integration of relations that occur between instances.

3 Basic Notions

We define a real world using a pair *(A, V)*, in which A is a set of attributes that can be used to describe objects taken from some topic and V denotes a set of valuations of these attributes. Formally, if V_a denotes a domain of an attribute a, a following condition is met: $V = \bigcup_{a \in A} V_a$. An ontology is a tuple:

$$O = (C, H, R^C, I, R^I) \tag{1}$$

where C is a set of concepts, H is concepts' hierarchy, R^C is a set of relations between concepts $R^C = \{r_1^C, r_2^C, ..., r_n^C\}$, $n \in N$, $r_i \subset C \times C$ for $i \in [1, n]$, I denotes a set of instances' identifiers and $R^I = \{r_1^I, r_2^I, ..., r_n^I\}$ symbolises a set of relations between concepts' instances. Every relation from the set R^C has a complementary relation from the set R^I. In other words, a relation $r_j^C \in R^C$ is a set containing descriptions of potential connections that may occur between instances of concepts from the set C, while $r_j^I \in R^I$ contains definitions of actually materialised connections. For example, the set R^C may contain relations *is_husband* or *is_wife* and in one can find R^I statements that *Dale is a husband of Laura or that Jane is a wife of David*. Obviously, $|R^C| = |R^I|$.

Concepts taken from the set C are defined as quadrupoles $c = (id^c, A^c, V^c, I^c)$, where id^c is an identifier of a concept c, A^c is a set of its attributes, V^c is a set attributes domains (formally: $V^c = \bigcup_{a \in A^c} V_a$) and I^c is a set of particular concepts' instances. We can write $a \in c$ which denotes the fact that the attribute a belongs to the concept's c set of attributes A^c. An ontology is called *(A, V)-based* if the conditions $\forall_{c \in C} A^c \subseteq A$ and $\forall_{c \in C} V^c \subseteq V$ are both met. As aforementioned in Sect. 1, a set of all (A, V)-based ontologies is denoted as $\tilde{O}$.

Given a concept c, we define its' instances as a tuple $i = (id^i, v_c^i)$. id^i is an identifier and v_c^i is a function with a signature: $v_c^i : A^c \to V^c$. Using a consensus theory [14], the function v_c^i can be interpreted as a tuple of type A^c.

A set of instances from the Eq. 1 is defined below:

$$I = \bigcup_{c \in C} \{id^i | (id^i, v_c^i) \in I^c\} \tag{2}$$

We write $i \in c$ to express that an instance with an identifier i belongs to a concept c.

In order to simplify operations on sets, we define an auxiliary notion of a set *Ins(c)* containing identifiers of instances assigned to concept c. Formally:

$$Ins(c) = \{id^i | (id^i, v_c^i) \in I^c\} \tag{3}$$

Complementary, Ins^{-1} denotes a helper function that designates concepts containing a given instance's identifier. It has a signature $Ins^{-1} : I \to 2^C$ and is defined as follows:

$$Ins^{-1}(i) = \{c | c \in C \wedge i \in c\} \tag{4}$$

Relations from the set R^C acquire semantics using L_s^R which is a sublanguage of the sentence calculus. This is accomplished using a function $S_R : R^C \to L_s^R$. Such approach allows to define criteria for relationships between relations:

- *equivalency* between relations r and *r'* (denoted as $r \equiv r'$) appears only if a sentence $S_R(r) \iff S_R(r')$ is a tautology
- a relation *r'* is more general than the relation r (denoted as $r' \leftarrow r$) if $S_R(r) \implies S_R(r')$ is a tautology
- *contradiction* between relations r and *r'* (denoted as $r \sim r'$) is true only if a sentence $\neg(S_R(r) \wedge S_R(r'))$ is a tautology.

As mentioned earlier, relations from the set R^C define what concepts instances can be connected with each other, while R^I defines what actually is connected. To denote this requirement, we use the same index of relations taken from both sets - a relation $r_j^I \in R_I$ contains instance pairs that are mutually connected by a relation $r_j^C \in R^C$. Below, we define a set of formal criteria that these sets must meet:

1. $r_j^I \subseteq \bigcup_{(c_1,c_2)\in r_j^C} (Ins(c_1) \times Ind(c_2))$
2. $(i_1, i_2) \in r_j^I \implies \exists(c_1, c_2) \in r_j^C : (c_1 \in Ins^{-1}(i_1)) \wedge (c_2 \in Ins^{-1}(i_2))$ which describes that two instances may be connected by some relation only if there is a relation connecting concepts they belong to
3. $(i_1, i_2) \in r_j^I \implies \neg\exists r_k^I \in R^I : ((i_1, i_2) \in r_k^I) \wedge (r_j^C \sim r_k^C)$ which concerns a situation in which two instances cannot be connected by two contradicting relations (e.g. John cannot be simultaneously a husband and a brother of Jane)
4. $(i_1, i_2) \in r_j^I \wedge \exists r_k^I \in R^I : r_k^C \leftarrow r_j^C \implies (i_1, i_2) \in r_k^I$ which denotes that if two instances are in a relation and there exists a more general relation, then they are also connected by it (e.g. if John is a father of David, then he is obviously also his parent).

Relations connecting concepts are used to define what types of object can interact with each other. On this level it is not necessary to define specific properties of relations using their semantics originating from L_s^R. On the level of instances relations, that actually materialise relations the actual properties of relations come into play.

For example, defining that *a man* can *be a brother* or *be a husband* of *a woman* is quite simple on the level of concepts. On the level of instances it is crucial to also express that *John* cannot simultaneously be a husband and a brother of *Jane*. Merging two sets (that are used to define relations according to the Eq. 1) which contain pairwise excluding knowledge would lead to inner inconsistencies within an ontology that is a result of the integration. Therefore, in the L_s^R we distinguish two elements *is_asymmetric* and *is_transitive*, that for some selected relation r can be used to describe its following properties:

- $(S_R(r) \implies is\_asymmetric) \iff \forall\,(a, b) \in r : \neg\exists\,(b, a) \in r$
- $(S_R(r) \implies is\_transitive) \iff \forall\,(a, b, c) \in C : (a, b) \wedge (b, c)\,\exists\,(a, c) \in r$

To simplify the notation, in subsequent parts of the paper, we will use predicates $is\_asymmetric(r)$ and $is\_transitive(r)$.

In our considerations we do not include relations that are symmetric. The reason why is a property of such relation - if two symmetric relations (that are in fact sets) are summed, the resulting set will also be symmetric. As a result, no conflicting statements about connected instances will emerge. The same situation occurs when integrating relations that are reflexive or irreflexive. Therefore, we also do not include them in our framework.

4 Integration of Ontology Instances' Relations

In this section, an algorithm for the ontology integration on instances' relations level will be presented. As an input, it requires to ontologies that are defined according to Eq. 1. As a result, it returns sets of integrated relations (for both levels of concepts and instance) and a final estimation of the knowledge increase on the level of instance gained during the conducted integration. As stated in Sect. 1, this estimation is a value of a function $\varDelta_R$ that for two ontologies $O_1 = (C_1, H_1, R^{C_1}, I_1, R^{C_1})$ and $O_2 = (C_2, H_2, R^{C_2}, I_2, R^{C_2})$ has a signature $\varDelta_R : R^{C_1} \times R^{C_2} \rightarrow [-1, \infty]$.

Assuming that both ontologies contain only one relation (formally: $R^{C_1} = \{r_1^C\}, R^{I_1} = \{r_1^I\}, R^{C_2} = \{r_2^C\}, R^{I_2} = \{r_2^I\}$), the function $\varDelta_R$ is described by the following postulates:

1. $\varDelta_R = -1 \iff (r_1^C \equiv r_2^C) \wedge is\_asymmetric(r_1^C) \wedge \forall_{(a,b) \in r_1^I} \exists (b,a) \in r_2^I$
2. $\varDelta_R = 0 \iff (r_1^C \equiv r_2^C) \wedge (r_1^I \cup r_2^I = r_1^I \cap r_2^I)$
3. $\varDelta_R = 1 \iff \neg((r_1^C \equiv r_2^C) \vee (r_1^C \leftarrow r_2^C) \vee (r_1^C \sim r_2^C))$
4. $\varDelta_R \in [1, \infty] \iff (r_1^C \equiv r_2^C) \wedge is\_transitive(r_1^C)$

The first postulate concerns an issue in which two relations are equivalent, but asymmetric. In such situation the resulting relation must also be asymmetric, but also cannot contain pairs of instances that interfere with the asymmetry. If all of the instances' pairs from two relations do so, then the $\varDelta_R$ must express that the integration not only do not increase the knowledge, but actually causes its loss.

The second postulate illustrate the repetitive knowledge in two ontologies. In such situation $\varDelta_R$ should be equal to 0, which expresses the situation where nothing is gained from the conducted integration.

The third postulate defines a situation in which two relations expresses two completely different interactions that may occur between instances. These interactions do not interfere with each other (which may result in knowledge loss), but do not entail the emergence of any new knowledge about instances' relations. This kind of synergy is expressed using the fourth postulate.

The eventual shape of the proposed method is presented on Algorithm 1. At first, in lines 1 to 3, the algorithm creates two empty sets for the results and initialises the knowledge increase estimator $\varDelta_R$. The backbone of the algorithm (line 4) is an iteration through a Cartesian product of two sets of concepts relations.

In each loop, the algorithm at first (in line 5) checks if the two relations that are currently analysed are equivalent. If this is the case, then the algorithm attaches to the final result a sum of processed relations (a single relation that contains elements of both inputs) and adds to the knowledge increase estimator $\varDelta_R$ an ordinary Jaccard's similarity between the two in order to express how the conducted integration enriches the overall knowledge. This is done in lines 6 to 10.

However, the equivalency requires to check if the resulting relation (created in lines 6 and 7) is asymmetric or transitive. The former property may entail potential loss of knowledge due to the fact that in the resulting ontology two instances cannot be connected symmetrically. If in the two input ontologies two instances are connected by the equivalent relations, but in the different order, this can cause that the knowledge coming from the fact that the two instances are connected becomes uncertain. To avoid such situation, the algorithm removes (in lines 16 and 17) both connections and decreases the value of the estimator δ_R.

In the next step (line 23), the algorithm checks the transitivity of the resulting relation created in lines 6 and 7. This situation may entail the emergence of a new knowledge. For example, for the relation *is family* a situation in which one of the input ontologies contains a pair *(John, Steven)* and the second contains a pair *(Steven, Wilson)* should result that after the integration its output should also include a pair *(John, Wilson)*. This knowledge has not been present in the input ontologies, but emerges thanks to the conducted integration. In other words, the integration of ontologies is a synergy, where its output is something more than a strict sum of the input. Therefore, the algorithm increase the knowledge estimator δ_R that can acquire values larger than 1, meaning that it is not a metric, but it may indicate that a new knowledge has been created.

The subsequent part of the algorithm (lines 31–36) copes with a situation in which one of the processed relations is more general than the other. In such situation, the algorithm adds to the resulting ontology both relations, but also expands the less specific relation with the elements of the second relation. This indicates that some of the knowledge is gained thanks to the integration, but some is actually repeated in both input ontologies. Therefore, the knowledge increase estimator δ_R is increased only by the relations that are not present in both relations.

The next stage of the algorithm handles a situation when two relations are contradicting. Both relations are included in the final ontology, but this issue may result in knowledge decrease, due to the fact that two instances cannot be simultaneously connected by such relations. For example, *John* cannot be both *a brother* and *a husband* of *Jane*. If such pairs are found, then they are removed from the resulting ontology (lines 40 and 41) and the knowledge estimator Δ_R is decreased accordingly (line 42).

The last part of the algorithm (lines 44 to 49) covers the simple case when two relations express completely different knowledge concerning how instances interact with each other. In this situation, both relations are added to the final ontology and the estimator Δ_R is increased with a maximal value (equal to 1), because both input ontologies bring new knowledge to the final result.

Eventually, the algorithm returns created sets of relations and a mean knowledge increase (line 52) that has been acquired due to the conducted integration.

5 Evaluation of the Proposed Formula

Our research focuses on creating a new methodology of estimating a potential growth of knowledge during the ontology integration process. Therefore, there is no benchmark dataset that could be used to prove the correctness of our ideas. To verify our ideas, we used a statistical analysis of data obtained from questionnaire that contained 20 questions showing results of the integration of two ontologies on the instances' relation level[1]. The human judgment is the popular and approved methodology [16] for this kind of verification.

[1] https://goo.gl/iktxsK.

Algorithm 1. Ontology integration and knowledge increase estimation on relation level

```
Require: $O_1 = (C_1, H_1, R^{C_1}, I_1, R^{C_1}), O_2 = (C_2, H_2, R^{C_2}, I_2, R^{C_2})$;
$R^{C^*} = \phi$;
$R^{I^*} = \phi$;
$\Delta_R = 0$;
for $(r_i^{C_1}, r_j^{C_2}) \in R^{C_1} \times R^{C_2}$ do
    if $r_i^{C_1} \equiv r_j^{C_2}$ and $(r_i^{I_1} \cup r_j^{I_2}) \notin r^{I^*}$ then
        $r^{C^*} = r_i^{C_1} \cup r_j^{C_2}$;
        $r^{I^*} = r_i^{I_1} \cup r_j^{I_2}$;
        $R^{C^*} = R^{C^*} \cup \{r^{C^*}\}$;
        $R^{I^*} = R^{I^*} \cup \{r^{I^*}\}$;
        $\Delta_R = \Delta_R + 1 - \frac{|r_i^{I_1} \cap r_j^{I_2}|}{|r_i^{I_1} \cup r_j^{I_2}|}$;
        if $is\_assymetric(r^{C^*})$ then
            $count = 0$
            $s = |r^{I^*}|$
            for $(a, b) \in r^{I^*}$ do
                if $(b, a) \in r^{I^*}$ then
                    $r^{I^*} = r^{I^*} \setminus \{(b, a)\}$
                    $r^{I^*} = r^{I^*} \setminus \{(a, b)\}$
                    $count + = 2$
                end if
            end for
            $\Delta_R = \Delta_R - 2 \cdot \frac{count}{s}$
        end if
        if $is\_transitive(r^{C^*})$ then
            for $(a, b) \in r^{I^*}$ do
                if $\exists c \in C^* : (b, c) \in r^{I^*}$ then
                    $r^{I^*} = r^{I^*} \cup \{(a, c)\}$
                end if
            end for
            $\Delta_R = \Delta_R + \frac{|r^* \setminus (r_i^{I_1} \cup r_j^{I_2})|}{|r_i^{I_1} \cup r_j^{I_2}|}$
        end if
    else if $r_i^{C_1} \leftarrow r_j^{C_2}$ then
        $r^{C^*} = r^{C^*} \cup \{r_j^{C_2}\}$
        $r^{C^*} = r^{C^*} \cup \{r_i^{C_1} \cup r_j^{C_2}\}$
        $r^{I^*} = r^{I^*} \cup \{r_j^{I_2}\}$
        $r^{I^*} = r^{I^*} \cup \{r_i^{I_1} \cup r_j^{I_2}\}$
        $\Delta_R = \Delta_R + \frac{|r_i^{I_1} \cap r_j^{I_2}|}{|r_i^{I_1}|}$
    else if $r_i^{C_1} \sim r_j^{C_2}$ and $r_i^{I_1} \notin r^{I^*}$ then
        $r^{C^*} = r^{C^*} \cup \{r_i^{C_1}\}$
        $r^{C^*} = r^{C^*} \cup \{r_j^{C_2}\}$
        $r^{I^*} = r^{I^*} \cup \{r_i^{I_1} \setminus r_j^{I_2}\}$
        $r^{I^*} = r^{I^*} \cup \{r_j^{I_2} \setminus r_i^{I_1}\}$
        $\Delta_R = \Delta_R + \frac{|r_i^{I_1} \setminus r_j^{I_2}|}{|r_i^{I_1}|} + \frac{|r_j^{I_2} \setminus r_i^{I_1}|}{|r_j^{I_2}|} - \frac{|r_i^{I_1} \cap r_j^{I_2}|}{|r_i^{I_1} \cup r_j^{I_2}|}$
    else
        if $r_i^{I_1} \notin r^{I^*}$ then:
            $r^{C^*} = r^{C^*} \cup \{r_i^{C_1}\}$;
            $r^{C^*} = r^{C^*} \cup \{r_j^{C_2}\}$;
            $r^{I^*} = r^{I^*} \cup \{r_i^{I_1}\}$;
            $r^{I^*} = r^{I^*} \cup \{r_j^{I_2}\}$;
            $\Delta_R = \Delta_R + 1$;
        end if
    end if
end forreturn $R^{C^*}, R^{I^*}, \frac{\Delta_R}{|R^{I^*}|}$;
```

The main aim of our experiments is demonstrated how the developed measure reflect a way people evaluate the knowledge increase. Prepared examples have covered all possible cases like: the integration of two equivalent relations, two contradicting relations, two asymmetric relations, two transitive relations, and situation where one relation is more general than other one. The instances' relations cannot be considered without a semantic meaning, therefore we didn't use any symbolic data. Additionally, we cannot use the real ontologies because the provided datasets are too big and it is very hard to process them manually by an expert. Some examples of prepared ontologies are presented in Figs. 1 and 2.

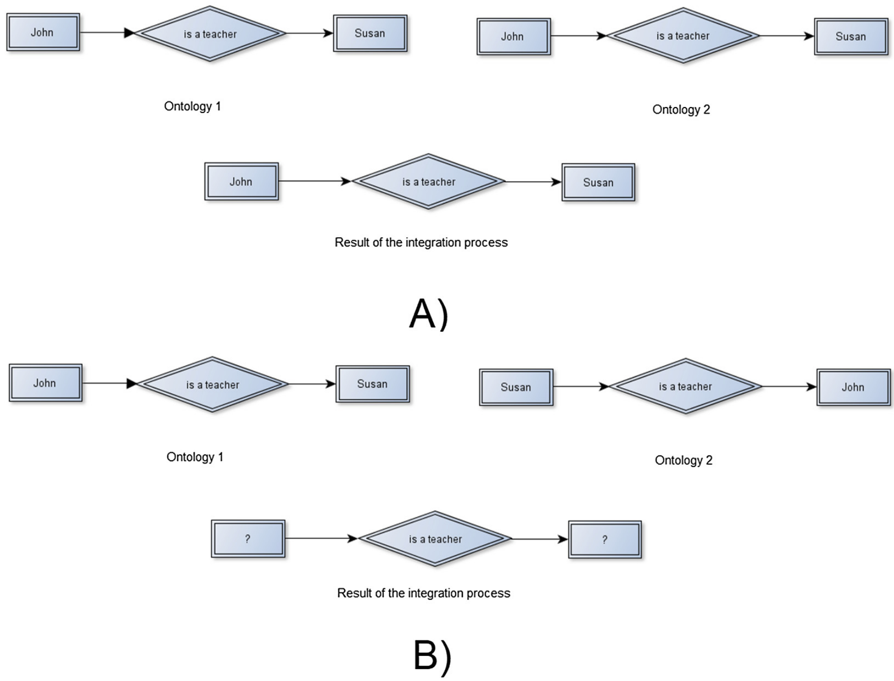

Fig. 1. Examples of knowledge increase during ontologies integration at instances' relations level; (A) $\Delta_R = 0$; (B) $\Delta_R = -100\%$.

The experiment has been divided in two parts. In the first one, we wanted to check the general trend. We have asked our responders about an overall opinion about knowledge change during the integration process. The responders could choose one from the three options: *the knowledge has grown, the knowledge remained the same, the knowledge has been lost (in other words - some semantic conflict occurred)*. In the second part the responders were asked to rate the level of knowledge change with the use of the scale range from -100% to ∞. This range corresponds with range $[-1, \infty]$, however the percentage values are more intuitive for human.

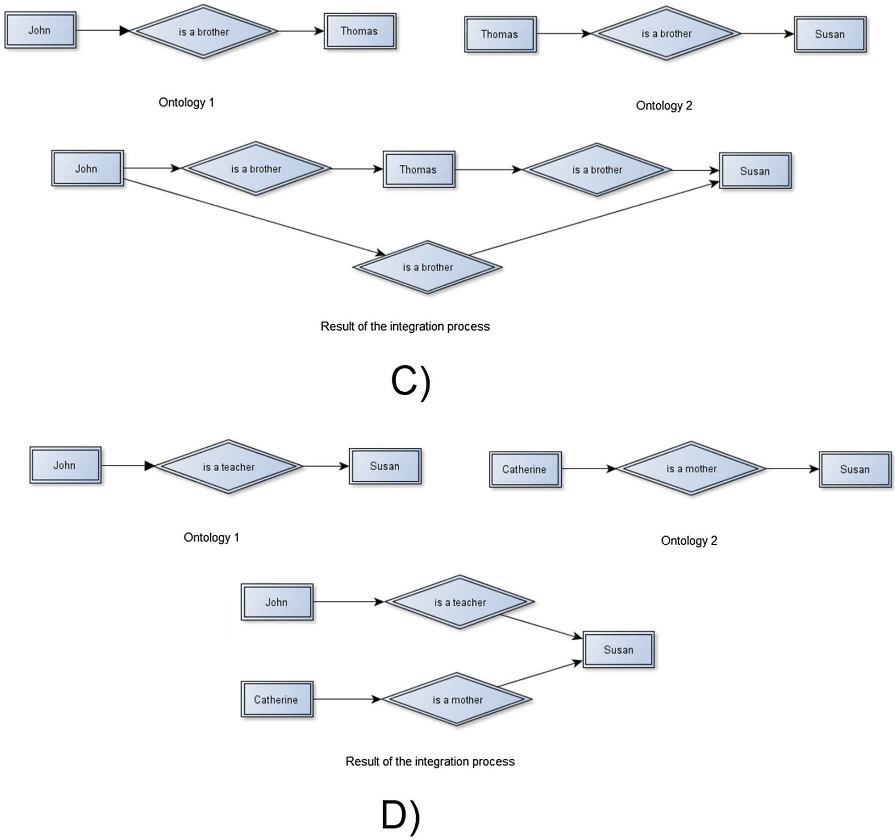

Fig. 2. Examples of knowledge increase during ontologies integration at instances' relations level; (C) $\Delta_R = 150\%$; (D) $\Delta_R = 100\%$.

Participants were not randomly selected - the survey was sent to a self-selected, biased population of people with technical science background. We decided to use that type of sampling because our responders needed to be familiar with issues related to databases, ontologies, instances and relations. The questionnaire was filled by 45 responders differing in age, sex, and educational status.

The obtained data have been pre-processed before a statistical analysis. The answers of our responders have been treated as experts' opinions and a median of their answers for each question has been calculated as the final estimation of the general trend and a value of knowledge increase estimation. In other words - based on the collected data, we have determined a consensus [14], as the final, common opinion. According to the literature it is possible to determine such consensus satisfying a 1- or 2-*optimality* criterion. For our purpose, we have chosen the 1-*optimality* postulate, which requires the result of the integration to be as near as possible to element of the input. The final summary of our analysis can

Table 1. The results of statistical analysis.

The type of experiment	The result	p-value
Cohen's Kappa coefficient-the general trend verification	$\kappa = 70.15\%$	0.000006
Correlation Coefficient test-the absolute agreement	0.579	0.0045
Correlation Coefficient test-the consistency	0.714	0.0045

be found in the Table 1. In the next subsection, some discussion about obtained results are presented.

5.1 The General Trend Verification

We had two samples to analyse. The first sample contained 20 elements, each determined as the consensus of expert's opinion referring to the general trend of the knowledge increase for the cases presented in the questionnaire. The second sample coming from the Algorithm 1.

All of the values are on nominal scale, therefore, the significance test for Cohen's Kappa coefficient has been used for the analysis. It was made with a significance level $\alpha = 0.05$. The agreement with a chance adjustment $\kappa = 70, 15\%$ is smaller than the one which is not adjusted for the chances of an agreement and the *p-value* is equal 0.000006. Such result proves a statistical agreement between these two samples on the assumed significance level. It means that people can notice a general trend referring to the knowledge increase in the given integration task.

5.2 The Value of Formula Verification

As in the previous section, we had two samples to analyse. The first calculated as the consensus of experts' opinions and the second was created automatically using the formula presented in Algorithm 1. However, in this part of our experiment, our samples contained a value of potential growth of knowledge which occurs during the ontology integration process. For this purpose, we have checked the normality distribution of both samples using the Shapiro-Wilk test. For both samples *p-value* was greater than the accepted significance level $\alpha = 0.05$, therefore, we couldn't reject the null hypothesis, which states that both samples come from a normal distribution.

For the further analysis, we selected the Intraclass Correlation Coefficient test (ICC) [1]. It measures the strength of an inter-judge reliability – the degree of their's assessments concordance. We had two samples for which we have tested the consistency and the absolute agreement. For both test *p-value* was around 0.0045 which allows us to claim that values obtained from the Algorithm 1 and those based on experts' opinions are statistically concordant in the analysed population. However, the absolute agreement is not very high and is 0.579. The consistency has been calculated as 0.714. The gathered results are presented in Fig. 3. It could be seems that the respondent's answers do not achieve extreme points on the scale. However, each point on the left side of the Fig. 3 is consensus of 45 responders answers calculated in simply way as a median.

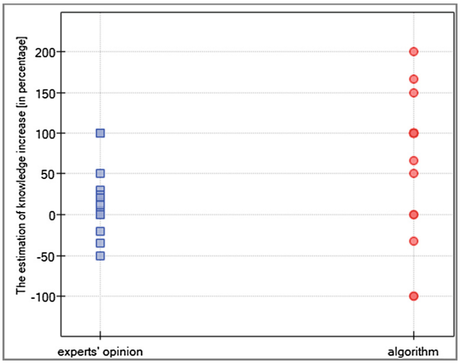

Fig. 3. The results of the experiment.

We can claim that the proposed method corresponds with a natural way people estimate the increase of knowledge. The integration process of instances' relations is not very intuitive, which makes it an interesting direction of upcoming research.

6 Future Works and Summary

The paper addresses the problem of the ontology integration on instances' relations level. Authors proposed the algorithm which allows to estimate a potential growth of knowledge during the merging process of two ontologies on instances' relations level.

The proposed method can be verified using several types of statistical tools used on the collected data, i.e. surveys, experimental studies and observations, among which a survey was selected for the stated purpose. The questionnaire containing 20 different scenarios has been prepared and presented to 45 volunteers. Based on the collected data, the consensus of experts' answers has been determined for each question. These results were compared with answers obtained by the execution of the proposed algorithm.

Statistical analysis (the Cohen Kappa method and ICC measure) allowed us to draw a conclusion that the created method of estimating a potential growth of knowledge is intuitive in a human experts way. The first experiment showed substantial agreement around 70%. It means that people can notice a general trend referring to the knowledge increase. The idea of relations of instances and their integration is not intuitive and it is hard to explain to people that are not familiar with the topic. However, the absolute agreement was nearly 58% in case of the comparison of the value of potential growth of knowledge. This agreement can be interpreted as fair.

In this paper, the method which estimates the objective growth of knowledge has been proposed. It means that our algorithm do not judge the disperse in the amounts of knowledge available in the input ontologies. In our upcoming publications, we would like to focus on subjective measures (from the point of view of the integrated ontologies) and conduct more experiments using real ontologies, which could bring more expressive conclusions.

Acknowledgments. This research project was supported by grant No. 2017/01/X/ST6/00491 from the National Science Centre, Poland

References

1. Bartko, J.J.: The intraclass correlation coefficient as a measure of reliability. Psychol. Rep. **19**(1), 3–11 (1966). https://doi.org/10.2466/pr0.1966.19.1.3
2. Burton-Jones, A., et al.: A semiotic metrics suite for assessing the quality of ontologies. Data Knowl. Eng. **55**(1), 84–102 (2005)
3. Ceusters W., Smith B.: Towards a realism-based metric for quality assurance in ontology matching. In: Proceedings of the 2006 Conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006), pp. 321–332. IOS Press (2006)
4. Cheatham, M., Hitzler, P.: String similarity metrics for ontology alignment. In: Alani, H., et al. (eds.) ISWC 2013. LNCS, vol. 8219, pp. 294–309. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41338-4_19
5. Euzenat, J.: Semantic precision and recall for ontology alignment evaluation. IJCAI **7**, 348–353 (2007)
6. Jiang, Y., Wang, X., Zheng, H.T.: A semantic similarity measure based on information distance for ontology alignment. Inf. Sci. **278**, 76–87 (2014)
7. Kozierkiewicz-Hetmańska, A., Pietranik, M.: The knowledge increase estimation framework for ontology integration on the concept level. J. Intell. Fuzzy Syst. **32**(2), 1161–1172 (2017). https://doi.org/10.3233/JIFS-169116
8. Kozierkiewicz-Hetmańska, A., Pietranik, M., Hnatkowska, B.: The knowledge increase estimation framework for ontology integration on the instance level. In: Nguyen, N.T., Tojo, S., Nguyen, L.M., Trawiński, B. (eds.) ACIIDS 2017. LNCS (LNAI), vol. 10191, pp. 3–12. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54472-4_1
9. Kozierkiewicz-Hetmańska, A., Pietranik, M.: The knowledge increase estimation framework for ontology integration on the relation level. In: Nguyen, N.T., Papadopoulos, G.A., Jędrzejowicz, P., Trawiński, B., Vossen, G. (eds.) ICCCI 2017. LNCS (LNAI), vol. 10448, pp. 44–53. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67074-4_5
10. Lozano-Tello, A., Gomez-Perez, A.: OntoMetric: a method to choose the appropriate ontology. J. Database Manag. **15**(2), 1–18 (2004)
11. Ma, Y., Jin, B., Feng, Y.: Semantic oriented ontology cohesion metrics for ontology-based systems. J. Syst. Softw. **83**(1), 143–152 (2010)
12. Maleszka, M., Nguyen, N.T.: A method for complex hierarchical data integration. Cybern. Syst. **42**(5), 358–378 (2011)
13. Meilicke, Ch., Stuckenschmidt, H.: Incoherence as a basis for measuring the quality of ontology mappings. In: Proceedings of the 3rd International Conference on Ontology Matching, vol. 431. CEUR-WS. org (2008)

14. Nguyen, N.T.: Advanced Methods for Inconsistent Knowledge Management. Springer, London (2008). https://doi.org/10.1007/978-1-84628-889-0
15. Pietranik, M., Nguyen, N.T.: A Multi-atrribute based framework for ontology aligning. Neurocomputing **146**, 276–290 (2014). https://doi.org/10.1016/j.neucom.2014.03.067
16. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 1, pp. 448–453 (1995)
17. Tartir, S., et al.: OntoQA: metric-based ontology quality analysis. http://lsdis.cs.uga.edu/library/download/OntoQA.pdf (2005). Accessed 22 Oct 2017
18. Welty, C., Guarino, N.: Supporting ontological analysis of taxonomic relationships. Data Knowl. Eng. **39**(1), 51–74 (2001)
19. Yu, J., Thom, J.A., Tam, A.: Requirements-oriented methodology for evaluating ontologies. Inf. Syst. **34**(8), 766–791 (2009)