UDC 004.94 https://doi.org/10.15407/pp2021.02.024

O. Zakharova

DEFINING DEGREE OF SEMANTIC SIMILARITY USING DESCRIPTION LOGIC TOOLS

The purpose of this study is to determine effective approaches to define the value of semantic similarity of information. The special functions to determine quantitative indicators of a degree of semantic similarity of the information allow ranking the found information on its semantic proximity to search request/template. Forming such measures should take into account many aspects from the meanings of the matched concepts to the specifics of the business-task in which it is done. A combination of semantic and structural approaches is appropriate when constructing the similarity functions. This allows to do descriptions of the concepts more detail, and the impact of syntactic matching can be significantly reduced by using more expressive descriptive logics to represent information and by moving the attention to semantic properties.

The focus of this research is in the methods for evaluating similarity of concepts. Values of similarity between individuals and between a concept and an individual are defined by finding the most specific concept for individual(s) and evaluating the similarity between the appropriate concepts. Using some of defined measures is demonstrated on a geometry ontology application.

Key words: semantic similarity of information, a similarity value, least concept subsumer, the most specific concept, the most specific is-a ancestor, similarity measures, features-based models, semantic-network based models, information content based models, existential concepts similarity.

Introduction

The task of discovery of concepts that are semantically similar and evaluating a degree of their similarity is very important both for resolving applied problems (discovery of semantic web services, effective semantic search of information, data categorization, etc.), and for more general problems in the information technologies area, as, for example, integration of onto- logies/knowledge, information search, etc. There are a lot of approaches that try to resolve the problems of finding similarity by methods of text analysis or using special vocabularies as Wordnet [1], for example. As a rule, in such approaches only atomic concepts are considered, but more complex ones are out of the question. In addition, the cases of identifying similarities between individuals and between an individual and a concept are omitted. Also, note, that measures of information similarity should be based on semantics because the purely syntactic ap- proach is too weak to ensure that standard in- ferences are executed, especially if expressive descriptive logics (for example, ***ALC***) are considered as a language for knowledge re- presentation. It is clear that algorithms and functions of similarity measures must be effective. If they are too complex, they can't provide the desired result in a reasonable period of time and become commonly used.

Last time many studies have appeared that emphasize the feasibility of using ontologies and based on them functions of semantic similarity to compare concepts and / or individuals that can be obtained through the integration of heterogeneous sources of information [2,3,4,5].

The main purpose of this study is the analysis of methods, models and approaches for creating quantity indicators that evaluate a similarity degree of knowledges represented with descriptive logic (DL) tools, their classification and application.

Types and levels of similarity determination

Information/knowledge similarity may be considered and defined on different levels. Namely, we can identify:

1) **Conceptual level** – determination of similarity between concepts;

2) **Knowledge level** – determination of the similarity between instances of concepts;

3) **Mixed level** – determination of the similarity between an instance and a concept.

Similarity measures, as a rule, use the basic Set Theory and they are based on objects commonality. In particular, the basic criteria to determine such measures can be formulated as follows: the value of similarity between objects is not only a result of their common features but it is, also, a result of their differences. This criteria corresponds to the theoretical-informational definition of si- milarity. The objects, in this case, are concepts and instances of concepts.

We consider approaches for defining similarity measures and corresponding models for evaluating on the each level. But, first, we introduce some definitions used by the most existing models.

Basic concepts and definitions

Definition 1. **LCS (Least Concept Subsumer)** [23, 24] – the least common subsumer of concepts. Let L is DL. A description of the concept E in DL $\mathcal{L}$ is a LCS of concepts' descriptions $C_1, \cdots, C_n$ in $\mathcal{L}$ (shortly LCS($C_1, \cdots, C_n$)), if:

1) $C_i \sqsubseteq E$ for $i = 1, \cdots, n$ and

2) E is the least description of $\mathcal{L}$-concept that meets the first condition, so as, if E_0 is description of L-concept such that $C_i \sqsubseteq E_0$ for all $i = 1, \cdots, n$, then $E \sqsubseteq E_0$.

At once, it should be noted that LCS doesn't exist for everyone DL used to represent knowledge, but if LCS exists, it is unique to the point of equivalence. All measures that will be discussed below are based on DL $\mathcal{ALC}$. As shown in [6], LCS always exists for $\mathcal{ALC}$ DL and it is defined by the concepts' disjunction. In the case if the logic doesn't support the disjunction operator, LCS is calculated by selecting general concept names in its descriptions (within the concepts of the universum and existential constraints for the same role), not taking into account TBox as a whole [6]. But, in this case the result of LCS evaluation may be very common. Based on these considerations, LCS is calculated relative to TBox, on the basis of which the concepts are defined [7].

Taking into account the TBox, LCS definition can be reformulated as follows.

Definition 2. Let $\mathcal{L}1$ and $\mathcal{L}2$ are descriptive logics such as $\mathcal{L}1$ is sub-logic of $\mathcal{L}2$, so $\mathcal{L}1$ includes less constructors which are used to build expressions. For given TBox of $\mathcal{L}2$ logic $\mathcal{T}$, $\mathcal{L}1(\mathcal{T})$ is set of concept descriptions that can include concepts defined in $\mathcal{T}$. $C_1, \ldots, C_n$ are concept descriptions from $\mathcal{L}1(\mathcal{T})$, so LCS($C_1, \ldots, C_n$) in $\mathcal{L}1(\mathcal{T})$ w.r.t. TBox $\mathcal{T}$ is the description of the most specific $\mathcal{L}1(\mathcal{T})$ concept that includes $C_1, \ldots, C_n$ on TBox $\mathcal{T}$. In particular, it is such description of $\mathcal{L}1(\mathcal{T})$- concept D, as:

1) $C_i \sqsubseteq_{\mathcal{T}} D$ for $i = 1, \cdots, n$ and

2) If E is a description $\mathcal{L}1(\mathcal{T})$- concept such as $C_i \sqsubseteq_{\mathcal{T}} E$ for all $i = 1, \cdots, n$, $D \sqsubseteq E$.

If LCS for TBox doesn't exist (for example, in the case of cyclic TBox), its approximation is calculated. It is named Good Common Subsumer (GCS) [25] w.r.t. TBox and it exists for general TBox. GCS is calculated by defining the least conjunction of concepts and their objections which can include a conjunction of concept names of top level for each considered concept and the same conjunction of concepts constituting the rank of existential and universal constraints on the same role. GCS is the most specific covering than LCS calculated unrelated to Tbox. But, in a general case, it includes (or it is equivalent) LCS calculated w.r.t. TBox [7].

MSA (Most Specific is-a Ancestor)

[8] – the most specific ancestor in the hierarchy of the taxonomy. It is defined as binary relation on concepts taxonomy, but semantically it is similar to LCS. Both calculate the most specific generalization of input concepts (w.r.t. the operator of subsume). Their difference is next. MSA works on a taxonomy of concepts and returns one concept, which contains two original concepts (there is their is-a ancestor) and it does not include anyone else what meets the same requirements. LCS is a description that covers input concepts, and, as a result, returns all concepts included in it. If concepts only related by generic relations (TBox is a taxonomy) then LCS is reduced to one ancestor and LCS(C1,C2)=MSA(C1,C2).

MSC is the most specific concept. It is unary relation on a set of individuals of ABox.

Definition 3. [25] Let given ABox $\mathcal{A}$ and α is an individual from this ABox, then the most specific concept for *a* w.r.t. ABox $\mathcal{A}$ is a concept C, denoted as $C = \mathcal{MSC}_{\mathcal{A}}(a)$, such that $\mathcal{A} \models C(a)$, and $\forall D$ such that $\mathcal{A} \models D(a)$, $C \sqsubseteq D$ (where $\models$ is the inference operator).

At once, it should be noted that in the general case of acyclic ABox in the expressive DL $\mathcal{ALC}$ cannot be expressed by the final description of the concept [2], it is possible to obtain only its approximation. So, the existence of the most specific concept for an individual of ABox is not guaranteed, or it is difficult to calculate, and the approximation is limited by some depth of a set. A maximal depth of the approximation, as defined in [20], corresponds to the depth of ABox. In this case, we can define the most specific concept MSC(α) or its approximation MSC*(α) for any instance α of ABox.

Defining a semantic similarity of concepts

Today, a lot of researches exist that try to transform semantic relations between concepts into some quantitative indicators. It is clear, that the principles of formation of such measures are affected, first of all, by the essence of the compared concepts, and the business problem for the solution of which the similarity functions are chosen or determined. The most of existing studies use a semantic approach in conjunction with a structural one, which compares the descriptions of the concepts under consideration. Certainly, this allows to significantly detail the description, and the influence of syntactic matching can be reduced by using more expressive DLs to represent information and by moving the focus to semantic properties of concepts.

In establishing the degree of semantic correspondence between the concepts of the same ontology, the similarity function, in fact, is a mapping $\boldsymbol{S}$: $\boldsymbol{L}(\boldsymbol{T})\times \boldsymbol{L}(\boldsymbol{T})\rightarrow$Y, where $\boldsymbol{T}$ is TBox of this ontology represented in DL $\boldsymbol{L}$, and Y is real value, which quantifies the degree of similarity. In measures that are based on ratio Y ∈ [0,1] but another measure models also exist.

In general, the task is more complex. If matching concepts from different ontologies with TBoxes $\boldsymbol{T}1$ and $\boldsymbol{T}2$ of DLs $\boldsymbol{L}1$ and $\boldsymbol{L}2$, respectively, it is needed to build the mapping $\boldsymbol{S}$: $\boldsymbol{L}1(\boldsymbol{T}1)\times \boldsymbol{L}2(\boldsymbol{T}2)\rightarrow$Y.

In any case, the similarity function must have the following properties:

1) let E is a set of items (objects of the same or different ontologies), for which the value of similarity should be determined, then function $\boldsymbol{S}$ is defined on the set E × E;

2) the function $\boldsymbol{S}$ is positively defined, so $\boldsymbol{S}$(C,D)≥0;

3) ∀C,D: $\boldsymbol{S}$(C,D)≤ $\boldsymbol{S}$(C,C).

Defining the similarity function, it is necessary to understand that concept similarity may be considered both in terms of the degree of their commonality and the degree of their difference, and the similarity function should have a positive correlation with the value of commonality between the concepts and negative correlation with an indicator of the difference between them. It is clear this indicator depends on many factors, namely: the specifics of the content that is studied, the expressiveness and homogeneity of the languages of representation of on- tologies, and so on. But the key question in determining a similarity function is “how to calculate the value of commonality (difference) of concepts”, which, in turn, is related to the question “how we collect an investigated information”. It is unlikely that the similarity indicator can be considered as an absolute value, but it should provide the possibility of a reliable ranking of concepts by similarity values. As the main approaches to the defining such function can be distinguished:

1) defining similarity as function of a path-distance between taxons in hierarchy which underlies this ontology [10, 11, 12];

2) evaluating a feature–based semantic similarity [13];

3) defining a value of similarity by information content [14,15];

4) existential similarity of concepts.

The first approach may be applied only based on one ontology, so its usage may be appropriate only if the evaluation is performed on the basis of a single source of information, and the matched information items are concepts of a same ontology or an integrated ontology of information sources. Another approach for calculating semantic similarity uses both general and discrimination features between concepts and/or individuals. Methods of third group are based on Theory of information. They determine measure of similarity between two concepts in the hierarchy from the point of view of the amount of information transmitted directly by the super-concept, which includes the matched concepts. We may name all measures

that are based on features of concepts as the measures of intentional similarity. Under the *existential similarity of concepts* we will understand the degree of their closeness by the sets of instances that they include.

In the case of matching concepts of different, probably heterogeneous, ontologies listed approaches works only when certain conditions and restrictions are carried out. First, formal representation of these ontologies should support inference engines such as subsume. (Note that subsume engine is supported by basic DL as, for example, $\mathcal{ALC}$). Second, applying the calculation approaches are based on using a general ontology, and local concepts in different ontologies should inherit the structure of des- cription from their general ontology. In [16] some approaches for matching such concepts from different ontologies by their individuals are proposed. Namely, it is made an assumption that when the restrictions are performed the criteria of matching two concepts may be intersection of sets of their individuals. To match descriptions of the concepts that may be united in the general ontology, three main approaches are applied:

• filtering based on path-distance between concepts in the general ontology;

• defining measures based on matching graph that establish one-to-one correspondence between elements of the concept descriptions;

• defining probabilistic measures that give the correspondence in terms of the joint distribution of concepts.

Also, if computation of similarity values is performed for concepts belonging to different ontologies, it is necessary to take into account a difference between formalization levels of specification of these ontologies. Particularly, in [17] a similarity function determines classes of similar entities by matching using synonym sets, semantic neighborhood, and discriminating features that are classified into parts, functions, and attributes. In [9] another approaches are presented. It is aimed at finding common features among concepts or statements.

Listed groups of approaches for similarity computation are based on appropriate models.

The most common evaluation models include:

· feature-based models;

· semantic-network based models;

· models based on information content.

In **feature-based models** concept C is characterized by set of its features, denoted ftrs(C). In [18] two groups of measures for such model are proposed:

1) contrast model where the similarity between two concepts C and D is defined by the linear function

$contra(C, D) = \theta f(ftrs(C) \cap ftrs(D)) - \alpha f(ftrs(C) \setminus ftrs(D)) - \beta f(ftrs(D) \setminus ftrs(C))$, where \ is operation of sets difference, α, β and θ non negative constants, and f(.) expresses a number of features in the set;

2) a normalized model of the ratio where similarity is defined as quotient of the sets:

$$sim(C, D) =_{def} \frac{f\big(ftrs(C) \cap ftrs(D)\big)}{f\big(ftrs(C) \cap ftrs(D)\big) + \alpha f\big(ftrs(C) \setminus ftrs(D)\big) + \beta f\big(ftrs(D) \setminus ftrs(C)\big)}$$

If suppose that similarity function is symmetric then $\alpha = \beta = 0{,}5$. Assuming that the function f is distributive on intersected sets then $sim(C, D)$ may be transformed as follows:

$$sim(C, D) =_{def} \frac{2f(ftrs(C) \cap ftrs(D))}{f\big(ftrs(C)\big) + f\big(ftrs(D)\big)}$$

In the **semantic-network based models** a reference information is given in the semantic-network form that includes nodes-concepts and, at least, is-a edges (sometimes it contains more complex relations as in WordNet). It is an example of the case when similarity computation is based on measures of path-distance between concepts in the network. As concepts are in the taxonomy (linked by generic relations) so the similarity value between two concepts is computed by calculating edges on the path from considered concepts to their closer ancestor. If the entities are divided by only some connections then they are rated as similar. The more connections they share, the less similar they are [8, 19, 12, 20]. So, to evaluate similarity of

concepts C and D it is found the most specific is-a ancestor E = MSA(C,D) of C and D and a similarity measure is computed as the sum of path distances from C to E and from E to D. More advanced estimates may take into account the depth of the concept MSA(C, D), the density of the edges at the path nodes, and the weight of the edges.

In the **models based on information content** the information pr(C) about the probability that entity is described by the concept is used as well as semantic network. This probability, as usual, is estimated based on an initial particular task.

The value of information content is measured based on the probability pr(C) as IC(C) $=def-\log pr$(C). In [21] it is proposed the measure of the similarity of the concepts C and D based on probabilistic estimation of their MSA:

$$sim(C,D) =def\ IC(MSA(C,D)) =def-\log pr(MSA(C,D)).$$

In [22] it is proposed the measure of the path distance in the network based on their information content. It takes into account such factors as the depth and density of the edges of the path between the concepts is:

$$dist(C,D) =\_def\ IC(C)+IC(D)-2IC(MSA(C,D))$$

In [18] it is proposed the similarity measure that defined by the ratio:

$$sim(C,D) =\_def\ (2IC(MSA(C,D)))/(IC(C)+IC(D))$$

Defining similarity values for DL descriptions of concepts

All metrics above are defined for atomic concepts. But these measures may be reformulated for complex DL concepts. Note, we suppose that the concept descriptions are represented in basic DL which support only operation of intersection of concepts. Any description of a complex concept may be normalized, namely, decomposed in such way that it will contain only atomic concepts.Usually, this is done simply by substituting the concept descriptions in the definition instead of non-atomic concepts. Denote as *nf(C)* the set of atomic concepts which is in the normal form of the concept C. Note that $C \sqsubseteq D$ (where $\sqsubseteq$ - structural subsumption), if $nf(D) \subseteq nf(C)$.

Taking into account given structural description of the concept, the measures above can be reformulated as follows.

For feature-based model, we will consider features of the concept as atomic concepts and a complexconcept as conjunction of these atomic concepts. Considering the peculiarities of the intersection and the difference of the sets of atomic properties, the similarity measures, under the conditions of their symmetry, can be determined as follows:

$$contra(C,D) =def\ f(lcs(C,D)) - 0{,}5 * f(diff(C,D)) - 0{,}5 * f(diff(D,C))$$

$$sim(C,D) =_{def} \frac{2 * f((C,D))}{2 * f(lsc(C,D)) + f(diff(C,D) + f(diff(D,C))}$$

Note, the function f is a counter of properties, possibly weighted, in these measures.

Now, let consider the model of the semantic network. When the network is an hierarchy and the concept C has is-a ancestor: U1, U2, U3, ..., Un-1, Un, introduce the concept C* such that C := C*⊓ U1⊓ U2⊓ U3⊓...⊓Un-1⊓ Un. In the result T-Box the defined concept has the same hierarchy as initial nodes in the semantic network. Moreover, if the source network U1, U2, . . . Un = ⊤ is the path to the root of is-a hierarchy then the normal form of the concept U1 in DL is nf(U1) = U_(1)^*⊓ U_(2)^*⊓...⊓U_(n-1)^*. Other words, if the network is a tree then concept's cardinality is the normal form of the concept C - |nf(C)| and it is equal to the distance of the path from node U1 to the root. Paths from C and D to the root are intersected in E= MSA(C,D), that is the same as LCS(C,D) on the subsumption hierarhy. Then the distance between the concepts C and D may be defined as follows:

dist(C,D) =_def |nf(C)|+|nf(D)|-2|nf(LCS(C,D))|,* where |X| is the cardinality of the concept X.

Respectively, for information models:

*dist(C,D) =_def IC(C)+IC(D)-2*IC(lcs(C,D)),*

*sim(C,D) =_def (2*IC(lsc(C,D)))/(IC(C)+IC(D))*

Existentional measures of the concept similarity

In the existeontional approaches a similarity value is calculated by counting joint instances of the concept extensions [26] or by measuring a content variation between concepts [27, 28, 29].

As a rule, an ontology has a structure which is more complex that simple taxonomy. So, similarity measures that are based on distances in the taxonomy or based on usage of MSA can't be applied.

Note that the semantic relation of subsumption is based on canonic interpretation of ABox and assumption of unique namespace (UN A) of DL. It follows that the interpretation of the instances of ABox are themselves, and different individuals, corresponding to different objects of the business -area, have different names in the namespace. So, we will determine the similarity measure based on their extensions in the canonical interpretation of DL [25].

Let $\mathscr{L}$ is a set of concepts in DL $\mathcal{ALC}$, a $\mathcal{A}$ is ABox with the canonical interpretation $\mathcal{I}$. The semantic similarity of concepts $\mathcal{s}$ is a function: $\mathcal{s}$:$\mathcal{L}\times\mathcal{L}\rightarrow$[0,1], that is defined as:

$$s(C,D) = \frac{|I^{\mathcal{I}}|}{|C^{\mathcal{I}}|+|D^{\mathcal{I}}|-|I^{\mathcal{I}}|} * \max(|I^{\mathcal{I}}|/|C^{\mathcal{I}}|, |I^{\mathcal{I}}|/|D^{\mathcal{I}}|),$$

where $I=C\sqcap D$ and $(.)^{\mathcal{I}}$ is an extension of the concept in the interpretation $\mathcal{I}$.

The measure above may be justified as follows. If the concepts C and D are equivalent (both C⊑D and D⊑C are true) then $\mathcal{s}$=1. If the concepts are different at whole and intersection of their extensions is empty, then the similarity value is a minimal, so it is equal 0. In the case of non-empty intersection of the concepts the measure has a value in the rank from 0 to 1. So, this measure expresses a degree of the similarity of the concepts C and D reduced on the value of max(|I^I |/|C^I |,|I^I |/|D^I |). This value presents a difference of these concepts. It means that the similarity is considered as value weighted respectively the similarity degree (it is not absolute value). This measure corresponds to a rather strict semantic relation between the concepts, which is provided by the subsumption.

Measures of GCS-similarity of the concepts

The measures of GCS-similarity are determined based on the term of GCS-cover. They may be applied in the cases when other measures, namely, ones based on concept extensions intersections, information content or path distance, don't work. The measures based on GCS also use the term of a concept extension but the similarity value is defined as variation of number of instances in the concepts' extensions relatively the number of instances in the extension of their super-concept instead of counting common instances of the concepts. The common super-concept is defined by GCS of the concepts and the measure w.r.t. TBox $\mathcal{T}$ of $\mathcal{ALC}$ is formally determined as follows.

$\mathcal{T}$ is $\mathcal{ALC}$- TBox. $\mathcal{L}$ is descriptive logic that include $\mathcal{ALC}$. C and D are concept descriptions in $\mathcal{L}(\mathcal{T})$. Then a semantic similarity measure $\mathcal{S}$ is a function $\mathcal{S}$: $(\mathcal{T})\times(\mathcal{T}) \rightarrow[0,1]$ that determined as follows

$$\mathcal{S}(\mathrm{C},D) = \frac{min(|C^{\mathcal{J}}|,|D^{\mathcal{J}}|)}{|(GCS(C,D))^{\mathcal{J}}|} * \left(1 - \frac{\left|\left(GCS(C,D)\right)^{\mathcal{J}}\right|}{|\Delta^{\mathcal{J}}|} * \left(1 - \frac{min(|C^{\mathcal{J}}|,|D^{\mathcal{J}}|)}{\left|\left(GCS(C,D)\right)^{\mathcal{J}}\right|}\right)\right),$$

where $(.)^{\mathcal{J}}$ calculates a concept's extension w.r.t. interpretation $\mathcal{J}$ (canonical interpretation [2, 9]).

So, if two concepts are semantically similar they should have good common super-concept that is close to both concepts, namely, it is extension of super-concept which contains a lot of individuals that are common with initial concepts. In such case the value of the function is approaching 1. Vise versa, if the initial concepts are very different, then their GCS and their super-concept contains many instances that do not belong to the source concepts, i.e. the value of the similarity will approach 0. This measure doesn't require the intersection of the concepts and doesn't take into account the path distance between them. Moreover, to

avoid obtaining an incorrect value of similarity in the case when one concept is very similar to the super-concept and very different from another concept, which is compared, the minimal extension of concepts is considered in the measure's definition.

Defining the similarity measures at knowledge and mixed levels

Recall that similarity metrics of knowledge and mixed levels are measures for determining values of matching individuals and an individual and a concept, respectively. Determining measures involving individuals is based on the term of the Most Specific Concept (MSC). We can compute MSC or its appro¬xi-ma¬tion for each instance in ABox. These terms are equivalent in some cases.

Let a and b are two instances of ABox, A*=MSC*(a), B*=MSC*(b). Then, semantic similarity measures may be applied to the descriptions of the concepts A* and B*, and a result value will express the degree of similarity of corresponding individuals:

$$\forall a,b: s(a,b) = s(A^*,B^*) = s(MSC^*(a), MSC^*(b))$$

Likewise, the value of similarity between the descriptions of the concept C and the individual a may be calculated by determining the approximation of MSC of the instances and further applying the similarity measure to the concept C and the approximation MSC* of the instance a:

$$\forall a,C: s(a,C) = s(A^*,C) = s(MSC^*(a), C)$$

So, both measures are reduced to determining the similarity of the concept descriptions after preliminary approximation of instances. In this case, any of the above models can be used to calculate the value of similarity of concepts.

It should be noted that the complexity of the proposed methods depends on the complexity of the standard methods of inherits in DL.

Applying the similarity measures based on the DL ontology POGeometry (an example)

Consider the application of the measures of similarity of concept descriptions based on their extensions in canonical interpretation DL on an example of the domain ontology POGeometry.

TBox of the domain ontology POGeometry:
Coordinate, GeometricFigure
Vertex ⊑ has.XCoordinate
Vertex ⊑ has.YCoordinate
XCoordinate ⊑ Coordinate
YCoordinate ⊑ Coordinate
Coordinate ⊑ hasValue.NUMBER
Vector ⊑ 2has.Vertex
Vector ⊑ has.VectorLength
Vector ⊑ has.VectorAngle
VerterxLength ⊑ hasType.NUMBER
VerterxAngle ⊑ hasType.NUMBER
Height ⊑ hasType.NUMBER
EdgeLenth ⊑ hasType.NUMBER
…
Polygon ≡ GeometricFigure ⊓=has.Vertex⊓ =has.Vector
Circle ⊑ GeometricFigure
Quadrangle ≡ Polygon ⊓=4has.Vertex⊓ =4has.Vector
Triangle ≡ Polygon ⊓ =3has.Vertex =3has. Vector
Polygon ⊑ has.Vertex
Polygon ⊑ has.Vector
Triangle ⊑ =3has.Height
Square ⊑ hasType.NUMBER
GeometricFigure ⊑ has.Square
Circle ⊑ GeometricFigure
ABox:
Triangle(ABC), Triangle(XYZ), Triangle(A1B1C1), Triangle(B1C1D1), Triangle(A1C1D1), Triangle(A1B1D1), Triangle(X1X2X3), Triangle(X2X3X4), Triangle(X3X4X5), Triangle(X4X5X6), ..., Quadrangle(A1B1C1D1), Polygon(X1X2X3X4X5X6.), Circle(O1), Circle(O2)

Taking into account the definitions of the concepts Quadrangle and Triangle we may inherit the subsumption of the concepts Triangle ⊑ Polygon and Quadrangle ⊑ Polygon. So, all individuals of the concepts Triangle and Quadrangle are instances of the concept Polygon.

So, $|Polygon^{\mathcal{I}}|$=47, $|Triangle^{\mathcal{I}}|$=29, $|Quadrangle^{\mathcal{I}}|$=17.

Then, the similarity of the concepts Triangle and Polygon may be determined based on sets of their instances as follows:

Let I=Triangle⊓Polygon, then

$$\begin{aligned}&s(Polygon, Triangle)\\&= \frac{|I^{\mathcal{J}}|}{|Polygon^{\mathcal{J}}| + |Triangle^{\mathcal{J}}| - |I^{\mathcal{J}}|}\\&* \max\left(\frac{|I^{\mathcal{J}}|}{|Polygon^{\mathcal{J}}|}, \frac{|I^{\mathcal{J}}|}{|Triangle^{\mathcal{J}}|}\right)\\&= \frac{29}{47 + 29 - 29} * max\left(\frac{29}{47}, \frac{29}{29}\right) = \frac{29}{47}\\&= 0{,}62\end{aligned}$$

Taking into account that the interpretations of the concepts Triangle and Quadrangle have no intersection $|I^{\mathcal{J}}| = 0$, where $I = Triangle \sqcap Quadrangle$, their values of similarity by instances will also be equal 0. In this case, certainly, the feature-based similarity measures or similarity measures using the least common subsumer are more reliable.

It should be noted that shown example is based on basic DL which use only the inter-section constructor, and TBox, in fact, is a taxonomy. So, LCS always exists for their con-cepts, and for any concepts C and D from this Tbox the statement *LCS(C,D)=MSA(C,D) is the true. Particularly, Polygon =LCS(Triangle, Quadrangle) = MSA(Triangle, Quadrangle)*.

The function of the similarity concepts based on LCS may be defined based on the path distances between concepts or based on intersections of extensions of corresponding concepts (their sets of individuals).

dist(Triangle,Quadrangle) =_def
|nf(Quadrangle)|+|nf(Triangle)|-
2|nf(lcs(Triangle,Quadrangle))|=*
|nf(Quadrangle)|+|nf(Triangle)|-
2|nf(Polygon)|=2+2-2*1=2*

Using feature-based model the similarity measure is:

$$s(Triangle, Quadrangle) =_{def} \frac{2f(ftrs(Triangle) \cap ftrs(Quadrangle))}{f\big(ftrs(Trangle)\big) + f\big(ftrs(Quadrangle)\big)} = \frac{2 * 1}{3 + 3} = \frac{1}{3}$$

Taking into account that, in this case, GCS=LCS=MSA the similarity value is:

$$\begin{aligned}&\mathcal{S}(Triangle, Quadrangle)\\&= \frac{\min(|Triangle^{\mathcal{J}}|, |Quadrangle^{\mathcal{J}}|)}{|(LCS(Triangle, Quadrangle))^{\mathcal{J}}|}\\&* \left(1 - \frac{\left|\big(LCS(Triangle, Quadrangle)\big)^{\mathcal{J}}\right|}{|\Delta^{\mathcal{J}}|}\right.\\&\left. * \left(1 - \frac{\min(|Triangle^{\mathcal{J}}|, |Quadrangle^{\mathcal{J}}|)}{\left|\big(LCS(Triangle, Quadrangle)\big)^{\mathcal{J}}\right|}\right)\right)\\&= \frac{\min(|Triangle^{\mathcal{J}}|, |Quadrangle^{\mathcal{J}}|)}{|Polygon^{\mathcal{J}}|}\\&* \left(1 - \frac{|Polygon^{\mathcal{J}}|}{|\Delta^{\mathcal{J}}|} * \left(1 - \frac{\min(|Triangle^{\mathcal{J}}|, |Quadrangle^{\mathcal{J}}|)}{|Polygon^{\mathcal{J}}|}\right)\right)\\&= \frac{17}{47} * \left(1 - \frac{47}{49} * \left(1 - \frac{17}{47}\right)\right) = \frac{17}{47} * \frac{19}{49} \approx 0{,}14\end{aligned}$$

Conclusions

In this paper the analysis of semantic similarity indicators, classified by approaches and estimation models is carried out. Described measures use semantic reasoning such as, for example, instances checking of given ABox (it means calculating the concepts extensions). The internal complexity of expressive DL languages, such as ALC, causes the non-effectiveness of structural approaches to reasoning, so the definition of similarity functions is based on the use of The Set Theory. This allows the use of numerical approaches at the symbolic level of representation of DL.

The estimation models and similarity measures on different estimation levels are analyzed in the article. The main is defining similarity between concepts (models of conceptual level). The tasks of calculating the values of similarity between individuals or between an individual and a concept reduced to finding MSC for individual(s) and estimating similarity of appropriate concepts.

The most of described measures are built based on basic DLs which support only intersection constructor. But described approaches may be applied for any DL that provides basic reasoning services, namely: instances checking and MSC (approximation).

Proposed similarity measures may be useful for resolving a lot of different problems of different types, particularly big data problems such as, for example, information retrieval in the context of terminological systems of knowledge representation, data classification and categorization, etc.

References

1. Fellbaum, C. (Ed.). (1998). Wordnet: An Electronic Lexical Database. MA: MIT Press.
2. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P., eds.: The Description Logic Handbook. Cambridge University Press (2003)
3. Staab, S., Studer, R., eds.: Handbook on Ontologies. International Handbooks on Information Systems. Springer (2004)
4. Thompson, K., Langley, P.: Concept formation in structured domains. In Fisher, D., Pazzani, M., Langley, P., eds.: Concept Formation: Knowledge and Experience in Unsupervised Learning. Morgan Kaufmann (1991)
5. Haussler, D.: Learning conjuntive concepts in structural domains. Machine Learning (1989) 7–40
6. F. Baader, R. K¨usters, and R. Molitor. Computing least common subsumers in description logics with existential restrictions. In T. Dean, editor, Proceedings of the 16th International Joint Conference on Artificial Intelligence, pages 96–101. Morgan Kaufmann, 1999.
7. F. Baader, R. Sertkaya, and Y. Turhan. Computing least common subsumers w.r.t. a background terminology. In V. Haarslev and R. M¨oller, editors, Proceedings of Proceedings of the 2004 International Workshop on Description Logics (DL2004). CEUR-WS. org, 2004.
8. R. Rada, H. Milli, E. Bicknell, M. Blettner, ”Development and Application of a metric on Semantic Nets”, IEEE Trans. on Systems, Man, and Cybernetics, 19(1): 17-30 (1989)
9. Mantay, T.: Commonality-based ABox retrieval. Technical Report FBI-HH-M- 291/2000, Department of Computer Science, University of Hamburg, Germany (2000)
10. Collet, C., Huhns, M.N., Shen, W.M.: Resource integration using a large knowledge base in carnot. IEEE Computer 24 (1991) 55– 62
11. Fankhauser, P., Neuhold, E.J.: Knowledge based integration of heterogeneous databases. In Hsiao, D.K., Neuhold, E.J., Sacks-Davis, R., eds.: Proceedings of the IFIP WG 2.6 Database Semantics Conference on Interoperable Database Systems (DS-5). IFIP Transactions, North-Holland (1992)
12. Bright, M.W., Hurson, A.R., Pakzad, S.H.: Automated resolution of semantic heterogeneity in multidatabases. ACM Transaction on Database Systems 19 (1994) 212–253
13. Tversky, A.: Features od similarity. Psycological Review 84 (1997) 327–352
14. Jang, J., Conrath, D.: Semantic symilarity based on corpus statistic and lexical taxonomy. In: Proceedings of the International Conference on Computational Linguistics. (1997)
15. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11 (1999) 95–130

16. Weinstein, P., Bimingham, P.: Comparing concepts in differentiated ontologies. In: Proceedings of 12th Workshop on Knowledge Acquisition, Modelling, and Management. (1999)
17. Rodr´ıguez, M.A., Egenhofer, M.J.: Determining semantic similarity among entity classes from different ontologies. IEEE Transaction on Knowledge and Data Engineering 15 (2003) 442–456
18. A. Tversky, "Features of Similarity", Psychological Review 84(4): 327-352, 1977.
19. J. Lee, M. Kim, and Y. Lee. Information retrieval based on conceptual distance in is-a hierarchies. Journal of Documentation, 2(49):188–207, 1993.
20. D. Maynard, W. Peters, and Y. Li. Metrics for evaluation of ontology-based information extraction. In Proceeding of the EON 2006 Workshop, 2006.
21. P. Resnik, "Using Information Content to Evaluate Semantic Similarity", Proc. IJCAI 1995 : 448-453
22. G. Miller & W.G. Charles, "Contextual correlates of semantic similarity", Language and Cognitive Processes, 6, 1-28, 1991.
23. W. Cohen, A. Borgida, H. Hirsh: "Computing Least Common Subsumers in Description Logics", AAAI 1992: 754-760
24. R. Kusters & R. Molitor, "Computing Least Common Subsumers in ALEN", IJCAI 2001: 219-224
25. Claudia d'Amato, Steffen Staab, Nicola Fanizzi, F. Esposito: "Efficient Discovery of Services Specified in Description Logics Languages", SMRR 2007
26. C. d'Amato, N. Fanizzi, and F. Esposito. A semantic similarity measure for expressive description logics. In A. Pettorossi, editor, Proceedings of Convegno Italiano di Logica Computazionale, CILC05, Rome, Italy, 2005
27. C. d'Amato, N. Fanizzi, and F. Esposito. A dissimilarity measure for ALC concept descriptions. In Proc. of the 21st Annual ACM Symposium of Applied Computing, SAC2006, 2006.
28. P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, 11:95–130, 1999.
29. A. Borgida, T. Walsh, and H. Hirsh. Towards measuring similarity in description logics. In Horrocks, U. Sattler, and F. Wolter, editors, Proceedings of the 2005 International Workshop on Description Logics (DL2005), volume 147 of CEURWorkshop Proceedings. CEUR-WS. org, 2005.

Received: 23.04.2021

About the author:

Olga Zakharova,
PhD,
Researcher.
Number of scientific publications in Ukrainian journals – 31. http://orcid.org/0000-0002-9579-2973.

Affiliation:

Institute of Software Systems of National Academy of Sciences of Ukraine.
Ac. Glushkov Avenue, 40. Phone.: 526 5139.
E-mail: ozakharova68@gmail.com. Mob. phone: +38(068)594756