Ontology similarity assessment based on lexical and structural model features extraction.

MARIUSZ CHMIELEWSKI, MAŁGORZATA PACIORKOWSKA, MACIEJ KIEDROWICZ
Cybernetics Faculty
Military University of Technology
gen. S. Kaliski Street 2, Warsaw
POLAND
mchmielewski@wat.edu.pl, malgorzata.paciorkowska@wat.edu.pl, mkiedrowicz@wat.edu.pl

Abstract: This work discusses a method for extracting quantitative measures from terminology and instance base components of semantic models. The method introduces a multi-criteria analysis while comparing ontology models and assessing semantic similarity. The article presents theoretical overview of the method and tool implementing evaluation schemes supplemented with practical examples. The capabilities of the method can be used for semantic pattern recognition within knowledge bases, which can be utilised by analytical tools especially in the security domain (criminal threat, financial fraud detection, etc.). The specificity of security applications requires methods dedicated for analysis of hidden, indirect, comprehensive and versatile data. Structural analysis method and its implementation in form of ETOSE plugin for Protégé OWL environment delivers process-based approach for evaluating instance bases. The mechanisms has been designed to operate as a data flow interceptor, collecting the data and transforming them into instances expressed in a specific domain ontology (set of ontology modules). Presented quantitative approach has been applied in terrorist threat assessment, financial fraud identification tasks where certain templates of behaviour and associations can be described. The method and tool utilize structural and lexicon comparison of compared ontologies in order to deliver multi-criteria evaluation of concepts, relationships and indirectly implemented axioms.

Key-Words: ETOSE, semantic structural analysis, semantic similarity, ontology, quantitative analysis, similarity measurement, OWL2, Protege

1 Introduction and main concepts

The demand for security applications and services is rising especially in the age of terrorism, cybersecurity threats and overwhelming financial and tax frauds. Recently developed tools are mainly aimed at processing large amounts of data rarely addressing hidden relationships between analysed facts. Semantic analysis methods can provide valuable extensions for such kind of analysis. Such methods can deliver terminology correspondences which may extend knowledge in the system by inferring new facts about hidden (in direct) associations between data instances such as people, organisations, events and more. This paper describes a developed method of semantic model analysis utilising structural and lexical measures which can be used for semantic association assessment as well as (and most importantly) for semantic similarity evacuation. The task of semantic similarity evaluation is a critical functionality while searching for patterns within knowledgebase. In such task the patterns can be expressed on the terminology or instance base level and are applied on migrated into instance base data. For the semantic similarity method formulation a set of definitions of semantic models and extracted multi-graph structures need to be provided. The method itself is based on weighted-graph analysis utilising structure related measures for which a semantic interpretations has been provided. To validate designed method a software environment ETOSE has been developed utilising mechanisms of Protégé OWL 5.0 platform. The developed method utilises multi-criteria

approach in order to define the decision maker preference model and weighted approach for semantic similarity definitions. The method has been applied as a part of indirect association analysis in crisis management and in particular criminal and terrorist threat analysis.

2 Measures of the structural similarity

Structural semantic measures have already been used in [9][10][11]. Measures are defined between selected elements of the ontology. The structural analysis of the ontology is based on four selected measures. These measures define the structural similarity of ontologies

Hyperlink Induced Topic Search (HITS) [1] is measure which is based on an algorithm originally designed for evaluating websites for website search process. The main idea implemented in the algorithm is that the popularity (authority) of a page depends on the number of other pages that link to it (and their ranking). Similarly the algorithm works for the ontology. For a given concept searches relationship which relate to him. The more relationship to a given concept, the higher its his value.

Page Rank [2] is an algorithm used by Google Search to rank website in their search engine results. PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The more pages refers to the page this page quality is higher. Page Rank has found use in ontology. The more concepts are referred to a given concept, the higher the value of structural similarity. A concept that has high value is an important part of the ontology. To calculate the PageRank value for a given concept (p_i) are being used: the set of concepts which links to p_i ($M(p_i)$), the damping factor (d) [2], the total number of the concepts (N) and the number of references on concept p_j ($L(p_j)$).

$$m^s_{PR}(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{m^s_{PR}(p_j)}{L(p_j)} \quad (1)$$

The semantic Structural Betweenness Measure [1] is a measure primarily used for identification of importance, centrality of the concept in terms of shortest paths traversing through the concept. The importance of the node is determined on the basis of the amount of occurrences of a concept in the shortest path between any two concepts.

$$\vec{m}^S_{SNBM}(v_i) = \sum_{\substack{v_i, v_j, v_k \in V \\ v_i \neq v_j \neq v_k}} \frac{card\left(\underset{\min}{Paths}_{v_j, v_k}(v_i) \right)}{card\left(\underset{\min}{Paths}(v_j, v_k) \right)} \quad (2)$$

Structural Node Closeness Measure (SNCM) [1] which determines the average distance between, a concept and another concept in the semantic graph. Such rank can assess the most central concept in terms of structural relation connectivity in domain model.

$$\vec{m}^S_{SNCM}(v_i) = \frac{1}{\sum_{v_j \in V \setminus \{v_j\}} \min_{Paths(v_i, v_j)} \left\{ \left| path(v_i, v_j) \right| \right\}} \quad (3)$$

The set of concept evaluating structural measures is defined as:

$$M_s(c) = \{m^s_i\}_{i=1..\overline{M_s(c)}} \quad (4)$$

The set of concept evaluating lexical measures is defined as:

$$M_l(c) = \{m^l_i\}_{i=1..\overline{M_l(c)}} \quad (5)$$

The vector of the structural similarity measures assessing a given concept (c_i) is defined as:

$$m_s(c_i) = < m^s_{w.hub}, m^s_{w.auth}, m^s_{PR}, m^s_{SNBM}, m^s_{SNCM} > \quad (6)$$

The vector of the lexical similarity for a concept is defined as:

$$m_l(c_i) = < m^l_{SM}, m^l_{Jaro}, m^l_{Jaccard}, m^l_{JaroWinkler}, m^l_{Sørensen-Dice} > \quad (7)$$

The aggregated measures for a single concept is calculated according to the following formula [3]:

$$\overline{m}(c_i) = \omega^s \cdot \sum_{k=1}^{\overline{M_s(c)}} w^s_k \cdot m^s_k + \omega^l \cdot \sum_{j=1}^{\overline{M_l(c)}} w^l_j \cdot m^l_j \quad (8)$$

Therefore, the distance between two compared concepts (multi-criteria) vector for the structural measure is defined as [3]:

$$d(c_i, c_j) = | \overline{m}(c_i) - \overline{m}(c_j) | \quad (9)$$

3 Measures of the lexical similarity

The lexical analysis is based on six chosen lexical similarity measures. All lexical similarity measures returns a degree of similarity as a value between 0 and 1. The value of zero means that the compared

strings are different and they have not got any common parts. The value of one shows that compared strings are the same. Values out of range [0,1] indicates an occurrence of the calculation error.

The Levenshtein distance (edit distance) [4] is a method for weighting the difference between two strings Li,Lj. The method allows to calculate the minimal number of changes needed to apply to two compared strings to made them the same. These strings can have different lengths. The result of the method is a number which defines number of changes (inserting a character, deleting a character, substituting a character) required to transform one string into another using a dynamic programming algorithm.

$$ed(L_i,L_j)_{(a,b)}=\begin{cases}\max(a,b) \\ \min\begin{cases}ed_{L_i,L_j}(a-1,b)+1 \\ ed_{L_i,L_j}(a,b-1)+1 \\ ed_{L_i,L_j}(a-1,b-1)+1_{(L_{ia}\neq L_{jb})}\end{cases}\end{cases} \quad (10)$$

The String Matching (SM) [4] is a lexical similarity measure for comparing similarity of strings (Li, Lj). The Levenshtein distance is being calculated in relation to the length of the shortest string of the two compared.

$$m^l_{SM}(L_i,L_j)=\max(0,\frac{\min(|L_i|,|L_j|)-ed(L_i,L_j)}{\min(|L_i|,|L_j|)} \quad (11)$$

The Jaro Distance (Jaro) [5] is a measure which compares two strings (Li, Lj) and returns the degree of similarity based on the number charcters in the same order in both strings (Li', Lj'). The measure is based on Levensthein's distance. The number of matching (but in different sequence order) characters divided by 2 defines the number of transpositions ($T_{L_i',L_j'}$).

$$m^l_{Jaro}(L_i,L_j)=\frac{1}{3}(\frac{|L_i'|}{|L_i|}+\frac{|L_j'|}{|L_j|}+\frac{|L_i'|-T_{L_i',L_j'}}{|L_i|}) \quad (12)$$

The Jaro-Winkler Distance [6] is a modified Jaro Distance measure. The modification introduced by W.E.Winkler allows to increase the value of similarity for strings whose Jaro similarity measure has exceeded a value 0.7. The measure uses the number of common characters counted from the beginning of one string up to the fourth character of the compared strings (b). The Jaro-Winkler measure includes a constant scaling factor (p), which is empirically determined and should not exceed 0.25. If the factor is greater than 0.25, the result of the Jaro-Winkler Distnace can become larger than 1.

$$m^l_{JaroWinkler}(L_i,L_j)=m^l_{Jaro}(L_i,L_j)+b\cdot p\cdot(1-m^l_{Jaro}(L_i,L_j)) \quad (13)$$

The Jaccard [5] is a lexical similarity measure with similar two string Li, Lj and treats them as multisets od words. It is defined as the quotient of the common part of the collections compared to the sum of these collections.

$$m^l_{Jaccard}(L_i,L_j)=\frac{|L_i\cap L_j|}{|L_i\cup L_j|}=\frac{|L_i\cap L_j|}{|L_i|-|L_j|+|L_i\cap L_j|} \quad (14)$$

The Sørensen-Dice [8] (also called Sørensen index or Dice coefficient) is a measure used to compare two sets. The Sørensen-Dice is used to calculate the measure of similarity between two sequences.

$$m^l_{SorensenDice}(L_i,L_j)=\frac{2|L_i\cap L_j|}{|L_i|+|L_j|} \quad (15)$$

4 Semantic Similarity Assessment and a tool to support evaluation

Semantic Similarity Assessment is performed in IV stages:

I. Translate ontology and instance base to set of corresponding structures emphasising the aim of analysis
II. Evaluate Lexicon similarity measures and evaluate Structural similarity for terminology or instance data layers
III. Review of evaluation results to identify corresponding (similar) concepts or instances in semantic model – recognise the case for semantic similarity
   Case 1: models are similar in terms of lexicons
   Case 2: models are similar in terms of their structure
IV. Knowledge engineer criteria weights configuration and aggregated similarity measure assessment

Final stage – manually confirm or correct similarity correspondences in semantic models verifying instance data transformation schemes

The Environment for Theoretical Ontologies Similarity Evaluation (ETOSE) is a plugin for the Protégé 5.0 program. The plugin has been implemented as Java standalone application. Detailed description of the ETOSE plugin is in [3].

The Etose plugin:
- presents graphical visualization of selected ontologies
- calculates lexical measures described in the chapter 3
- calculates structural measures described in the chapter 2
- shows the structural similarity measures for the ontologies
- shows the lexical similarity matrix of two ontologies calculated for a measure chosen by the user
- shows notifications about actions performed by the user
- calculates the frequency of appearing of the given concept in the natural language based on laws: Zipf and Lotka
- exports calculated values of lexical and structural similarity measures to CSV files

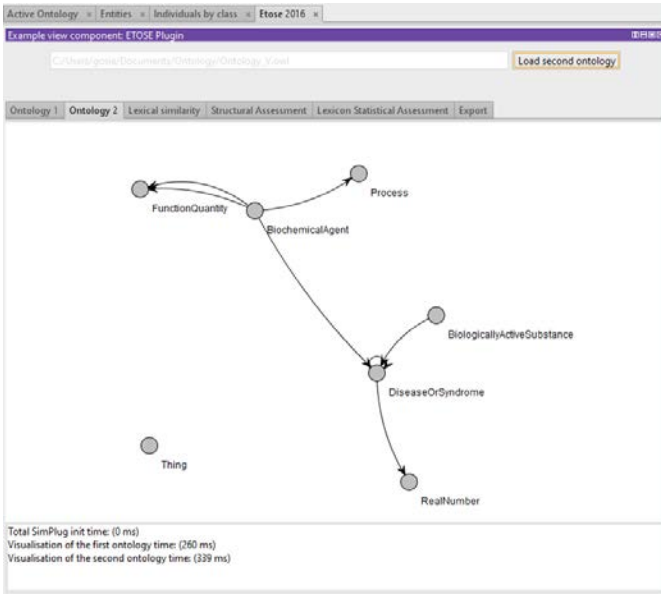

Fig. 1. ETOSE plugin ontology structure visualisation, integrated within Protégé 5.X OWL modeling environment

5 Tests

Semantic similarity determines the measure of structural similarity. Structural similarity is based on vertices and relations existing between vertices. Semantic models are represented by multigraphs on which concepts are vertices of the graph, and relationship between the vertices are edges of the graph. In order to demonstrate the quantitative approach capabilities a few ontology and instance base examples have been proposed, referencing crisis management domain.

1.1 Example 1 (similar structures, different lexicons)

In the first example two ontologies were compared. These ontologies are the same in terms structural, but differ in terms of lexical. In the first ontology named “terrorist attack en” is used terminology in English language. The second ontology named “terrorist attack pl” presents Polish based lexicon (concept’s labels) saving the initial mode structure.

.

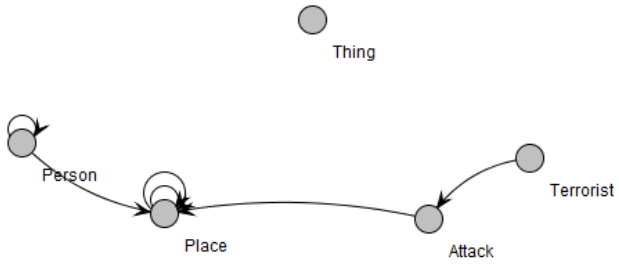

Fig. 2. Ontology labeled graph presenting „terrorist atack en” ontology in the ETOSE plugin. Coresponding structures are the same.

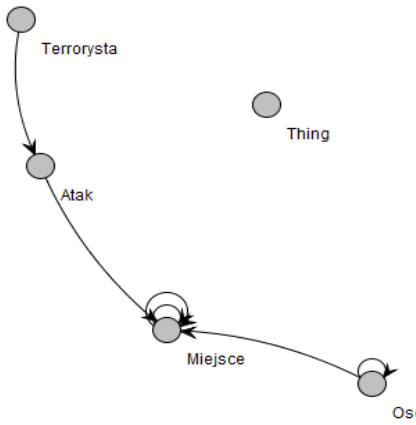

Fig. 3. Ontology labeled graph presenting „terrorist atack pl” ontology in the ETOSE plugin. Coresponding structures are the same.

Results for structural measures for the law enforcement domain ontology:

	HITS:Hub	HITS:Auth	PageRank	SNBM	SNCM	Result
Miejsce	0	0	0	0	0	0
Terrorysta	0	0	0	0	1	0,2
Osoba	0,7071	0	0	0	0,5	0,2414
Atak	0,7071	0	0	1	0,5	0,4414

Fig. 4. Values calculated for the structural measure of the „terrorist atack pl” ontology by the ETOSE plugin

Results for structural measures for the second ontology:

	HITS:Hub	HITS:Auth	PageRank	SNBM	SNCM	Result
Place	0	0	0	0	0	
Person	0,7071	0	0	0	0,5	0,241
Attack	0,7071	0	0	1	0,5	0,441
Terrorist	0	0	0	0	1	0,

Fig. 5. Values calculated for the structural measure of the „terrorist atack en ” ontology by the ETOSE plugin

According to the formulas given in Chapter **Błąd! Nie można odnaleźć źródła odwołania.**, the measure values for each concept of the ontology are calculated. On the basis of these measures, according to the formula , values for the general structural measure are calculated (“Result” column). These results were used to calculate the similarity of both structural ontologies. For this purpose, there was calculated absolute values of difference measures for the vertices.

	Miejsce	Terrorysta	Osoba	Atak
Place	0	0,2	0,2414	0,4414
Person	0,2414	0,0414	0	0,2
Attack	0,4414	0,2414	0,2	0
Terrorist	0,2	0	0,0414	0,2414

Values for the structural similarity of the „terrorist atack en” ontology and the „terrorist atack pl” ontology calculated on the basis of the values obtained with the ETOSE plugin

With the ETOSE plugin, results of structural similarity are measured for each ontology separately. It may be recalled that the scope of structural measures is the set of [0,1], where 1 indicates that the similarity of structures is negligible and 0 means that the structures are the same. To facilitate analysis of the figures above, there was utilized colours. The darker colour in the cell, the higher structural similarity obtained when comparing both ontologies. Received values of metrics allow to conclude that the highest structural similarity exists between vertices “Terrorysta” and “Terrorist” “Osoba” and “Person”, “Atak” and “Atack”, “Miejsce” and “Place”. The structural similarity between each pair of these vertices is zero, which means that the structure of vertices in each pair is the same. The vertices in pairs: “Miejsce” and “Atack”, “Atak” and “Place” have the lowest similarity measures between them. The difference in the structural similarity of these vertices is 0.4414.

Expected results were achieved, because vertices with the zero similarity value are structurally the same. Its just differ in names, which are in different languages. For vertices which are not related to each other, the similarity value is greater than 0.

1.2 Example 2 (two identical ontologies)

This example uses a single ontology, which means that in both compared ontologies is the same ontology. Ontology consists of three concepts and three roles. Below are results of all the lexical measures which are used by the ETOSE plugin. The last table (figure 30) presents values of lexical similarity, where each measure was taken with a weight of 0.2.

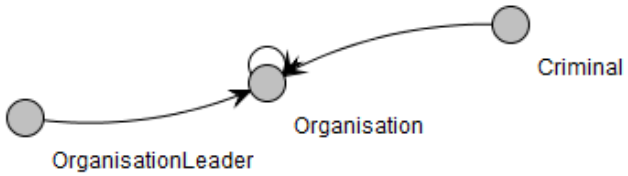

Fig. 6. Ontology labeled graph presenting first and second ontologies in the ETOSE plugin. The ontology is part of the “LawEnforcementDomain” ontology

	Organisation	OrganisationLeader	Criminal
Organisation	1	0,6667	0,3333
OrganisationLeader	0,6667	1	0,2381
Criminal	0,3333	0,2381	1

Fig. 7. Values calculated for the Jaccard measure by the ETOSE plugin

	Organisation	OrganisationLeader	Criminal
Organisation	1	0,8889	0,5278
OrganisationLeader	0,8889	1	0,4907
Criminal	0,5278	0,4907	1

Fig. 8. Values calculated for the Jaro measure by the ETOSE plugin

	Organisation	OrganisationLeader	Criminal
Organisation	1	0,9333	0,5278
OrganisationLeader	0,9333	1	0,4907
Criminal	0,5278	0,4907	1

Fig. 9. Values calculated for the Jaro-Winkler measure by the ETOSE plugin

	Organisation	OrganisationLeader	Criminal
Organisation	1	0,5	0
OrganisationLeader	0,5	1	0
Criminal	0	0	1

Fig. 10. Values calculated for the String Matching measure by the ETOSE plugin

	Organisation	OrganisationLeader	Criminal
Organisation	1	0,7857	0
OrganisationLeader	0,7857	1	0
Criminal	0	0	1

Fig. 11. Values calculated for the Sørensen-Dice measure by the ETOSE plugin

	Organisation	OrganisationLeader	Criminal
Organisation	1	0,7549	0,2778
OrganisationLeader	0,7549	1	0,2439
Criminal	0,2778	0,2439	1

Fig. 12. Values for the lexical similarity of the first ontology and the second ontology calculated on the basis of the values obtained with the ETOSE plugin

As expected, both ontologies have the same concepts, so there were received three values 1. In the example, can be observed differences in results between measures. It is an effect that various measures uses different features during calculation, what was described in chapter VI .

1.3 Example 3 (the secnod ontology is an extension of the first ontology)

The example presents two ontologies. The first ontology consists of four terms and three roles (object properties). The second ontology consists of same concepts and roles as the first ontology, but has been extended by two concepts and three roles.

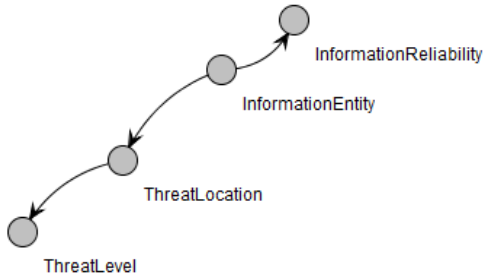

Fig. 13. Ontology labeled graph presenting first ontology in the ETOSE plugin. The ontology is part of the “LawEnforcementDomain” ontology.

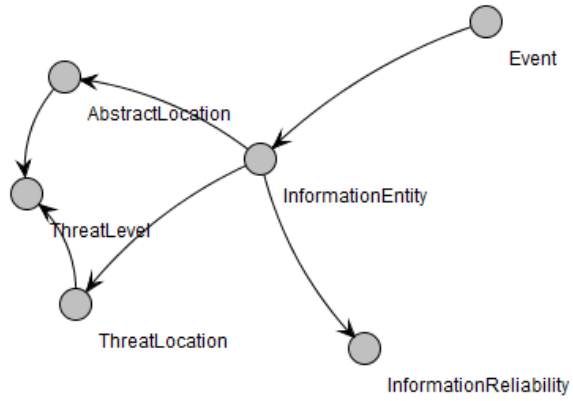

Fig. 14. Ontology labeled graph presenting second ontology in the ETOSE plugin. The ontology is part of the “LawEnforcementDomain” ontology.

Below are presented results of lexical and structural similarity calculated with the ETOSE plugin.

	InformationReliability	ThreatLocation	ThreatLevel	InformationEntity
ThreatLocation	0,3457	1	0,5627	0,4037
ThreatLevel	0,2858	0,5801	1	0,2526
AbstractLocation	0,3265	0,6618	0,2624	0,4092
Event	0,2238	0,2432	0,2857	0,2539
InformationEntity	0,694	0,4037	0,2526	1
InformationReliability	1	0,3575	0,2858	0,7048

Fig. 15. Values for the lexical similarity of the first ontology and the second ontology calculated on the basis of the values obtained with the ETOSE plugin

Values in the table are received from weighted average of lexical measures calculated in the ETOSE plugin. Each measure was taken with a weight of 0.2. There was received four values equal 1. It means that these concepts are structurally the same. This was expected, because the first ontology is a part of the second ontology. It can be seen that for concepts which were added to the second ontology (they are not present in the first onotology), the greatest lexical similarity exists between the concepts "AbstractLocation" and "ThreatLocation" and its value is 0.6618.

	InformationReliability	ThreatLocation	ThreatLevel	InformationEntity
ThreatLocation	0,0654	0,1318	0,4643	0,0741
ThreatLevel	0,2013	0,1349	0,1976	0,1926
AbstractLocation	0,0158	0,0822	0,4147	0,0245
Event	0,2013	0,1349	0,1976	0,1926
InformationEntity	0,2539	0,1875	0,145	0,2452
InformationReliability	0,5196	0,4532	0,1207	0,5109

Fig. 16. Values for the structural similarity of the first ontology and the second ontology calculated on the basis of the values obtained with the ETOSE plugin

Results obtained for structural measures confirm that the greatest structural similarity exists between concepts " InformationReliability " (first ontology) and "AbstractLocation" (second ontology). These are concepts that can be reached from one concepts of the graph (these are the concepts that specify the range of the property). Both concepts are possible only to reach but there are no edges from them to pass to next concepts.

	InformationReliability	ThreatLocation	ThreatLevel	InformationEntity
ThreatLocation	0,20555	0,5659	0,5135	0,2389
ThreatLevel	0,24355	0,3575	0,5988	0,2226
AbstractLocation	0,17115	0,372	0,33855	0,21685
Event	0,21255	0,18905	0,24165	0,22325
InformationEntity	0,47395	0,2956	0,1988	0,6226
InformationReliability	0,7598	0,40535	0,20325	0,60785

Fig. 17. Values for the aggregated measure of the first ontology and the second ontology calculated on the basis of the values obtained with the ETOSE plugin

To calculate the aggregated measures was used ω^s and ω^l equals 0,5. Also used the values obtained in Fig. 15 and Fig. 16. The greatest the aggregated measures value exists between the concepts "InformationReliability" (first ontology) and "InformationReliability" (second ontology) and its value is 0.7598. The smallest the aggregated measures value exists between the concepts "InformationReliability" (first ontology) and "AbstractLocation" (second ontology) and its value is 0.17115.

1.4 Example 4 (ontologies with a significant difference in the amount of concepts and roles)

The first ontology consists of six concepts and seven roles. The second ontology consists of three concepts and two roles. The first ontology has two times more concepts and roles than the second ontology. The number of concepts and roles does not directly affect lexical or structural similarity but the larger the ontologies are being compared, the greater is the probability that ontologies will be similar to each other.

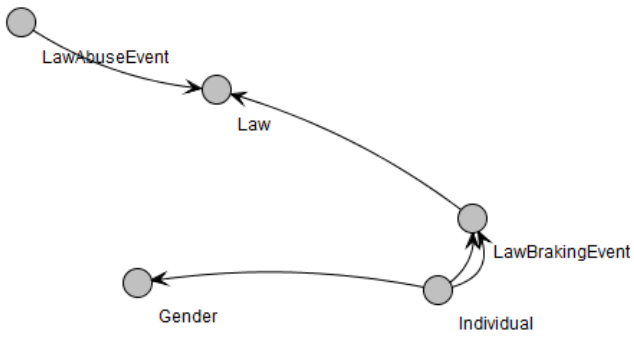

Fig. 18. Ontology labeled graph presenting first ontology in the ETOSE plugin. The ontology is part of the "LawEnforcementDomain" ontology

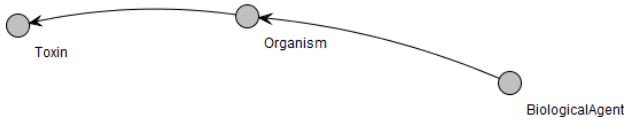

Fig. 19. Ontology labeled graph presenting second ontology in the ETOSE plugin. The ontology is part of the "WeaponsOfMassDestruction" ontology

	BiologicalAgent	Toxin	Organism
Law	0	0	0
Individual	0,087	0	0
Gender	0,1053	0	0
LawBrakingEvent	0,1429	0,1111	0
LawAbuseEvent	0,1538	0	0

Fig. 20. Values calculated for the Sørensen-Dice measure by the ETOSE plugin

In this case, the ontologies are not lexically similar. The largest, but still small, similarity exists between concepts "BiologicalAgent" and "LawAbuseEvent", which has a value 0.1538. This is due to the lexical similarity of concept's names parts ("Agent" - "Event").

	BiologicalAgent	Toxin	Organism
Law	0,3661	0,2232	0,5998
Individual	0,0097	0,1526	0,224
Gender	0,2999	0,157	0,5336
LawBrakingEvent	0,0924	0,2353	0,1413
LawAbuseEvent	0,2696	0,1267	0,5033

Fig. 21. Values for the structural similarity of the first ontology and the second ontology calculated on the basis of the values obtained with the ETOSE plugin

A similarity of 0.0097 exists between concepts "BiologicalAgent" and "Individual". This is due to the fact that both concepts specify the domain of the property (from these concepts it is possible to reach another concepts).

	BiologicalAgent	Toxin	Organism
Law	0,0118	0	0,2144
Individual	0,2896	0,2441	0,2333
Gender	0,0421	0,2022	0,2445
LawBrakingEvent	0,3635	0,0444	0,2244
LawAbuseEvent	0,3284	0,0118	0,2205

Fig. 22. Values for the lexical similarity of the first ontology and the second ontology calculated on the basis of the values obtained with the ETOSE plugin

1.5 Example 5 (structurally similar ontologies)

The example uses two ontologies that have similar construction for a few concepts. The first ontology consists of six concepts and seven roles. The second ontology consists of five concepts and five roles.

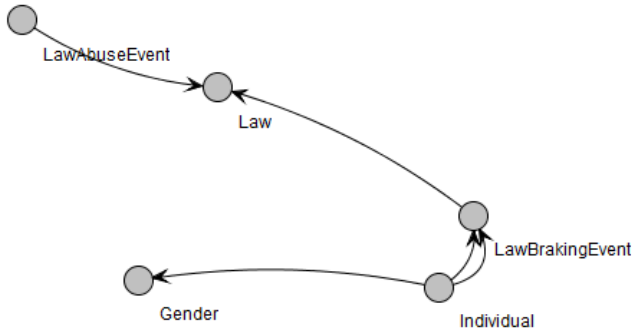

Fig. 23. Ontology labeled graph presenting first ontology in the ETOSE plugin. The ontology is part of the "LawEnforcementDomain" ontology

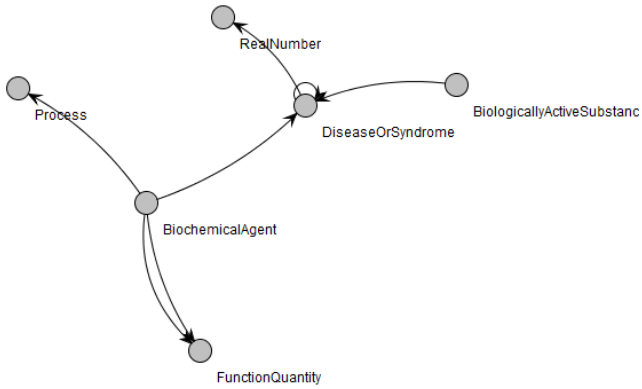

Fig. 24. Ontology labeled graph presenting second ontology in the ETOSE plugin. The ontology is part of the "WeaponsOfMassDestruction" ontology

Results of the calculated structural measures for the first and second ontologies are presented below.

	Law	Individual	Gender	LawBrakingEvent	LawAbuseEvent
BiochemicalAgent	0,3708	0,005	0,3046	0,0877	0,274
DiseaseOrSyndrome	0,3886	0,0128	0,3224	0,0699	0,292
RealNumber	0,0021	0,3737	0,0641	0,4564	0,094
Process	0,0628	0,313	0,0034	0,3957	0,033
FunctionQuantity	0,1312	0,2446	0,065	0,3273	0,034
BiologicallyActiveSubstance	0,2309	0,1449	0,1647	0,2276	0,134

Fig. 25. Values for the structural similarity of the first ontology and the second ontology calculated on the basis of the values obtained with the ETOSE plugin

Received results of structural similarity confirm that half of concepts in structural similarities have values less than 0.2.

	Law	Individual	Gender	LawBrakingEvent	LawAbuseEvent
BiochemicalAgent	0,0111	0,2843	0,2155	0,3698	0,2933
DiseaseOrSyndrome	0,0105	0,2032	0,2949	0,2041	0,2406
RealNumber	0,2078	0,2442	0,2792	0,2495	0,2535
Process	0	0	0,211	0,1813	0,1475
FunctionQuantity	0,0111	0,251	0,1734	0,245	0,2045
BiologicallyActiveSubstance	0,0069	0,2759	0,02	0,2631	0,235

Fig. 26. Values for the lexical similarity of the first ontology and the second ontology calculated on the basis of the values obtained with the ETOSE plugin

Two similarity measurements of ontologies were performed. In both tests, the same ontology was loaded as the first (example 4, example 5). Ontologies loaded as second in both examples were different from each other in terms of stucture and concept names.

Based on these measurements it can be concluded that more structural similarity exists between the ontologies compared in the Example 4 than in 5. In the Example 4, 50% of the comparison concepts were similar with the aggregated value of similarity less than 0.15. In the Example 5, this dependency correspond to only 27% of compared concepts.

What is more, it can be observed that higher lexical similarity exists between the ontologies in the Example 4 than in 5. In the Example 3, 77% of compared concepts were similar with the aggregated value of similarity greater than 0.2. In the Example 5, that dependence can be applied to only 67% compared concepts.

The calculation time of structural and lexical measures is dependent on the size of the selected ontologies. For ontologies up to 100 concepts and 100 roles, the time of computations is up to 1500 ms. The deviation is 27 ms. For ontologies that have about 2000 concepts and about 2000 roles, the time of measurements is up to 7000 ms. The deviation is 230 ms.

6 Ontologies used in tests

Proposed tests have been produced on ontologies of different sizes. Evaluated ontologies describe security risks domain. [13][14][15] The smallest model "law enforcement domain" ontology consists of 89 concepts and 40 properties. The largest "weapons of mass destruction" ontology consists of 1608 concepts and 209 properties. In order to test

the functionality of the ETOSE plugin there were performed tests on four selected ontologies („law enforcement domain", „mind swap terrorists", „weapons of mass destruction" and „terrorism" ontologies). A presentation of evaluation results has been prepared below - calculated structural and lexicon measures.

	Ontology	
	Law enforcement domain	Mind swap terrorists
Axiom	501	9732
Logical axiom count	332	2673
Declaration axioms count	152	1926
Class count	89	1608
Object property count	40	209
Data property count	18	103
Individual count	12	86
DL expressivity	ALCHOIF(D)	SHI(D)

Fig. 27. Metrics obtained in the Protégé appliaction for example ontologies

	Ontology	
	Terrorism	Weapons of mass desctruction
Axiom	9813	938
Logical axiom count	2671	504
Declaration axioms count	2011	207
Class count	1608	185
Object property count	209	17
Data property count	102	4
Individual count	86	162
DL expressivity	SHI(D)	AL(D)

Fig. 28. Metrics obtained in the Protégé appliaction for example ontologies

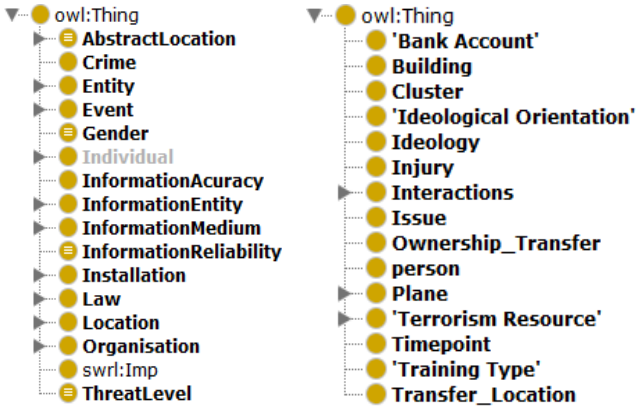

Fig. 29. (left) concepts taxonomy in the „law enforcement domain", (right) concepts taxonomy in the „mind swap terrorists" ontology - (Protégé environment)

Provided models descriptions describe the complexity of analysed and semantically compared models. Proposed quantitative approach can be successfully applied for assessing also model (terminology) cohesion identifying possible ontology modules.

Fig. 30. (left) concepts taxonomy in the „weapons of mass destruction", (right) concepts taxonomy in the „terrorism" ontology - (Protégé environment)

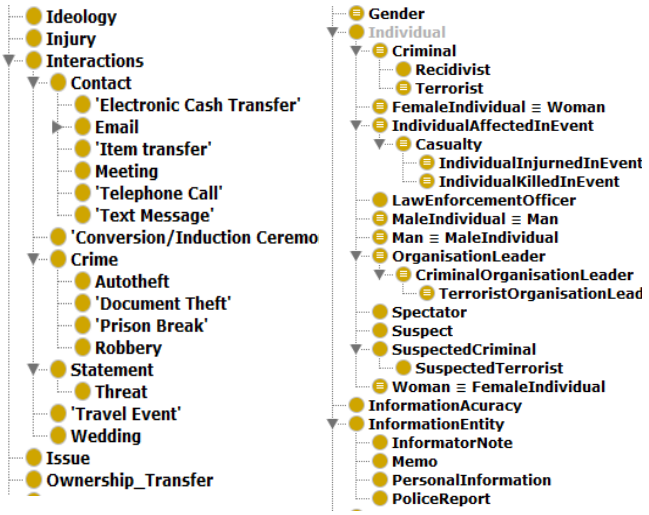

Fig. 31. (left) concepts taxonomy in the „terrorism", (right) concepts taxonomy in the „law enforcement domain " ontology - (Protégé environment)

7 Conclusions and future development

Proposed multi-criteria method delivers new means of ontology model similarity measurements. Based on structural analysis of the ontology and supplemented instance base, an analyst is able to evaluate and identify structural similarities which in case of semantic models may evidence about correspondences between models. A structure similarity for ontology might not suffice for adequate similarity assessment. Due to that fact a set of lexicon similarity measures have been propose complementing capability of similarity metrics. Further extensions of the method will include

assessment of ontology axioms (rules and DL constructors).

The ETOSE is a solution which meets the requirements for comparing similarities between two ontologies. It can be used in various situations, for example computer science, safety, security or daily life. Every topic which is able to be described with an ontology can be compared with the solution. It means that there is no restrictions on the subject the ontology nor the construction of the ontology. It provides solutions to calculate the similarity of lexical measures and structural measures, which enables multi-criteria analysis of the similarity between two ontologies. The plugin presents an overview through different approaches to the analysis of similarities between ontologies. Constructed software environment provides an overview of different types of similarity measures. The advantage of the solution comes from the combination and application of various semantic importance measures, both lexical and structural. Proposed process highlights differences between various methodologies of similarities between ontology structures. In addition, the plugin delivers parametrisation for multi-criteria semantic assessment and capabilities to adjust the parameters.

The ETOSE plugin architecture has been designed for further extensions, considering additional lexical and structural similarity measures as well as new approaches for aggregating the measures. The environment supports also orchestration of calculation methods producing a processing workflows [14][16] for ontology elements assessment [10] [11]. This would facilitate more detailed analysis of selected models of semantic similarities. On the other hand, another way of development can provide an comparison between more than two ontologies. It is also planned use the machine learning technique for semantic pattern recognition.

References:

[1] Chmielewski M., *Ontology-based indirect association assessment method using graph and logic reasoning technique,* Military University of Technology (2011)

[2] L. Page, S. Brin, R. Motwani, T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report. Stanford InfoLab, 1999

[3] M. Chmielewski, M. Paciorkowska, M. Kiedrowicz, A semantic similarity evaluation method and a tool utilised in security applications based on ontology structure and lexicon analysis, 21st International Conference on Circuits, Systems, Communications and Computers(CSCC 2017), AgiaPelagiaBeach, Crete Island,Greece, July 14-17, 2017

[4] A. Maedche, S. Staab, Measuring Similarity between Ontologies, *In Proceedings of the European Conference on Knowledge Engineering and Knowledge Management*, Germany (2002)

[5] W.W.Cohen, P.Ravikumar, S.E.Fienberg, *A comparison of string distance metric for name-matching tasks*, Carnegie Mellon University

[6] W.E. Winkler, *String comparator metrics and enhanced decision rules in the Fellegi-Sunter Model of record likage*, U.S. Bureau of the Census, Washington

[7] M.N.Jones, D.J.K.Mewhort, *Case-sesitive letter and bigram frequency counts from large-scale*, English corpora (2004)

[8] V.Thada, V.Jaglan, Comparison of Jaccard, Dice, *Cosine Similarity Coefficient to find best fitness value dor Web Retrieved Documents Usig Genetic Algorithm*, India

[9] M. Chmielewski, P. Stąpor, Protégé Based Environment for DL Knowledge Base Structural Analysis. Computational Collective Intelligence. Technologies and Applications. ICCCI 2011. Lecture Notes in Computer Science, 6922. Springer, Berlin, Heidelberg, (2011)

[10] Chmielewski M., Stąpor P. (2016) *Medical Data Unification Using Ontology-Based Semantic Model Structural Analysis*. In: Świątek J., Borzemski L., Grzech A., Wilimowska Z. (eds) Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part III. Advances in Intelligent Systems and Computing, vol 431. Springer, Cham

[11] Chmielewski M., Stąpor P. (2017) Money Laundering Analytics Based on Contextual Analysis. Application of Problem Solving Ontologies in Financial Fraud Identification

and Recognition. In: Borzemski L., Grzech A., Świątek J., Wilimowska Z. (eds) Information Systems Architecture and Technology: Proceedings of 37th International Conference on Information Systems Architecture and Technology – ISAT 2016 – Part I. Advances in Intelligent Systems and Computing, vol 521. Springer, Cham

[12] M. Kiedrowicz, "The importance of an integration platform within the organization", Zeszyty Naukowe, vol. 46, pp. 83-94, 2014.

[13] M. Kiedrowicz and J. Stanik, "Selected aspects of risk management in respect of security of the document lifecycle management system with multiple levels of sensitivity", (in:) Information Management in Practice, (eds) B.F. Kubiak and J. Maślankowski, pp. 231-249, 2015.

[14] M. Kiedrowicz, and R. Waszkowski, *"Business rules automation standards in business process management systems"*, (in:) Information Management in Practice, (eds) B.F. Kubiak and J. Maślankowski, pp. 187-200, 2015.

[15] M. Kiedrowicz, "Objects identification in the informations models used by information systems", GIS ODYSSEY 2016, pp. 129-136, 2016.

[16] M. Kiedrowicz, T. Nowicki, and R. Waszkowski, *"Business process data flow between automated and human tasks"*, 3rd International Conference on Social Science (ICSS 2016) December 9–11 2016, pp. 471-477, 2016.