Semantic Similarity Calculation Method using Information Contents-based Edge Weighting

Sunghwan Jeong[1], Jun Hyeok Yim[2], Hyun Jung Lee[3], and Mye Sohn[1]*
[1]Sungkyunkwan University, 2066, Seobu-ro, Jangan-gu, Suwon 440-746, Korea
{s103119, myesohn}@skku.edu
[2]Soosan INT Co., 10, Bamgogae-ro 1-gil, Gangnam-gu, Seoul 06349, Korea
gogo1525@soosan.co.kr
[3]Yonsei Institute of Convergence Technology, 162-1, Songdo-dong, Yeonsu-gu, Incheon, Korea
hjlee5249@gmail.com

Abstract

In this paper, we propose Semantic Similarity calculation measurement using INformation contents on EdGEs of ontology (SSINEGE) which is a hybrid edge- and information contents-based methodology. SSINEGE is devised to solve the limitation of the applying the same weighted edges by edge-based similarity. So, SSINEGE adopts information-contents theory to calculate the varied weights of edges. The varied weighted edges by SSINEGE can also solve a problem with the same degree of similarity for all pairs of concepts that are sharing a same Least Common Subsumer (LCS). To minimize the overlapped information-contents on the weighted, SSINEGE adopts the conceptual path between concepts instead of depths of the ontology. To verify the superiority of SSINEGE, we compared SSINEGE with widely used four similarity measurements including Leacock and Chodorow. We conducted two kinds of evaluations: first is calculation of similarity using the varied edge-weighting and second is for the discriminative capability using conceptual distances between comparative concepts. To verify the superiority of SSINEGE, we compared the calculated similarities of SSINEGE with Leacock and Chodorow. As the results, we verified that the calculated similarity of SSINEGE is significantly increased than the other comparatives.

Keywords: Hybrid semantic similarity, Ontology, Edge-based semantic similarity, Information Contents-based semantic similarity

1 Introduction

A lot of research has been widely conducted to compute and compare similarities between concepts or words, but there is big one limitation related to the word-sense disambiguation that may occur in areas such as automatic translation of documents, natural language processing, information integration, and recommendation [2]. Research related to the calculation of similarity between words is categorized in two: a lexical-based similarity calculation method and latent semantic analysis [1], [7], [10]. In the former method, the calculated similarity is varied depending on the defined words in the corpus. This kind of method calculates the similarity that is founded on the analysis of the shared words between two comparatives. The latter tries to calculate the degree of the similarity by latent relations between the comparative words using multivariate statistical analysis, singular value decomposition (SVD), or cosine similarity. Although the lexical-based similarity calculation method is easy to understand and comprehend, there are still some limitations, because the similarity of two words is calculated by only

Journal of Internet Services and Information Security (JISIS), volume: 7, number: 1 (February), pp. 40-53

*Corresponding author: : Department of Industrial Engineering, Sungkyunkwan University, 2066, Seobu-ro, Jangan-gu, Suwon 440-746, Korea. Tel: +82-31-290-7605

Table 1: **The limitations of traditional measurements for the semantic similarity**

Limitations	**Descriptions**
Edge-based	· The distances (weights) of all edges are the same in the ontology
From an algorithmical perspective	· All pairs of concept with the same LCS have the same degree of similarity · Insufficient consideration of semantic factors
From an ontological perspective	· Most of the developed ontologies have insufficient consideration of the properties other than hypernym/hyponym [3] · Need to additional considerations for semantic elements such as attributes, relationships, and annotations for the ontology

the number of the words in the corpus. So, it is difficult to identify semantic associations between the comparative words, because it is difficult to find the semantic associations by lexical similarity alone, while it is possible to derive conceptual associations.

One of the ways to overcome the limitations is to adopt ontology for calculation of the semantic similarity. Ontology is defined as a kind of dictionary to organize the types, properties, and interrelationships of concepts (words) that can be placed at a particular domain of discourse [19]. These domain-specific ontologies help that machines can understand and process the web documents. In addition, it can ensure the interoperability among web services [9]. The advantage of ontology-based similarity calculation is the extraction of the semantic similarity such as hypernym-hyponym, synonym-antonym, and/or disjoint relation between two words [15]. Research on the ontology-based similarity calculation is classified into four-category: edge-based, feature-based, information content-based, and hybrid method of two or more methods [4], [11], [12], [13]. However, there are following limitations. First, in the case of the edge-based semantic similarity, the edges are only used to determine whether concepts are connected or not. It means that there is no consideration of the semantic distance depending on the context. It can always calculate a same degree of similarity in any context. For instance, if a concept "Wine" is linked into two sub-concepts like 'Wine Body' and 'Region' by a property 'rdfs:subClassOf,' then the weights of edges on the two pairs (Wine, Wine Body) and (Wine, Region) are always a same as 1.0, even if it can be expected that there are some intuitive differences as conceptual distances. Second, even if ontology includes a variety of domain-knowledge and their relationships among domain-specific concepts, it has a limitation to use restricted information like the number of edges or information-contents of Least Common Subsumer (LCS). As the result, if the same LCS is applied, then the calculated semantic similarities will be the same. Last but not least, it is true that the domain-specific knowledge have limitations to be expressed by ontologies, such as name, types, properties, and interrelationships of the concepts that are associated with a particular domain of discourse. In this light, the problem on ontology-based semantic similarity is caused by the implied structural limitations on the ontology. Therefore, it is necessary to know the inherent limitations on the ontology and apply a new type semantic similarity to ontology. Table 1 summarizes the illustrated limitations in some measurements as follows.

So, in this paper, we propose a new similarity calculation measurement, named by Semantic Similarity calculation measurement using Information-contents on EdGEs of ontology (SSINEGE, pronounced by "synergy"). It is developed to overcome the illustrated limitations.

SSINEGE is basically developed as the edge-based semantic similarity calculation using the information-contents theory to overcome the applying of the same weighted edges. Therefore, SSINEGE calculates the varied weighted edges depending on the context. To do this, SSINEGE adopts the information-contents theory to derive the weights of edges on the ontology. In this light, the SSINEGE is a hybrid approach of edge-based and information contents-based similarity calculation methods. However, the applied information contents-based similarity calculation has a limitation as ever, because all pairs of concepts with a same LCS have a same degree of similarity. SSINEGE tries to solve the problem by applying the varied weighted edges as a semantic similarity. Finally, it is necessary to minimize the distortion of semantic similarity by the weighted edge. To do this, SSINEGE computes the semantic similarity depending on the conceptual distance between comparative concepts and the shared LCS.

This paper is organized as follows. In Section 2, we summarize some related works to the ontology-based semantic similarity measurements. Section 3 describes the detailed descriptions about the development procedure of SSINEGE. Next, Section 4 describes experimental results to show the superiority of SSINEGE using the developed two kinds of foods ontology. Finally, Section 5 presents the conclusions and further research.

2 Literature Review

2.1 Ontology-based Semantic Similarity

Ontology-based semantic similarity calculation is a method that can consider semantic information (e.g., semantic distance, properties, and/or relations) as well as syntactical similarity between two concepts to be compared [12]. As mentioned state, ontology-based semantic similarity calculation method is classified into four-category such as edge-based, feature-based, information contents-based, and hybrid approach. The features, advantages, and disadvantages of these methods are summarized as follows.

Edge-based method is the most basic and intuitive method to calculate the semantic similarity between two concepts. It tries to find the path connecting two concepts to be compared. Then, it calculates the semantic similarity by measuring the number of edges that make up this path [8], [13], [20]. At this time, the semantic similarity between the two concepts becomes smaller as the number of edges constituting the path increases. This method has the advantages that it is simple to calculate the similarity and easy to understand the conceptual weight (distance). However, there is a problem that is based on the same weight (distance) of all the edges on the ontology as an implicit assumption. In other words, it implies a false premise that the similarity of two concepts connected by one edge is always the same.

Feature-based similarity method has been emerged to solve the problem as was pointed because it has been a fundamental problem of the edge-based similarity method. The feature-based similarity method computes the semantic similarity by taking into account the semantic factors such as relationships, properties, and annotations among concepts in the ontology [12], [16], [18]. At this time, the more two comparative concepts usually share the same semantic elements, the greater the similarity. The feature-based similarity method has an advantage that the properties of the concept can be considered to calculate the semantic similarity by the edge-based method. However, it is difficult to apply because ontologies do not properly reflect the properties of the concepts.

To overcome the limitations of the edge-based method, some researchers used information contents theory to advance the calculation of the semantic similarity [11], [14]. This method computes the semantic similarity using the share information by the two concepts, i.e., information contents of the LCS of two concepts to be compared. In case of information-contents method, the accuracy is higher than the edge-based method as well as the calculation is easy. However, it also has a limitation that all pairs of concept with the same LCS have the same degree of similarity. In recent, to overcome the limitation, research has been carried out simultaneous considering not only information contents of the LCS but also information contents of the concepts to be compared [11]. However, this method does not yet completely solve the problem of the information contents-based semantic similarity method.

2.2 Edge Weighting

To reduce the overlapping use of information is caused by the fact that the edges connecting the concepts have the same weight in the ontology, research have been carried out to assign weights to the edges [4], [6]. These research assumes that the conceptual distances are not actually the same between all pairs of concepts in the ontology. In other word, if an edge has a higher weight than the other, then the edge conceptually has shorter distance between two concepts. On the contrary, the shorter distance between concepts means greater semantic similarity than the comparatives. Ge and Qiu [4] proposes a measure in which the weight of the edge decreases as the depth of the ontology deepens. Kwon *et al.* [6] considers the type of semantic relationship as well as depth of the ontology. This method is based on WordNet ontology and weights are differently given according to the kinds of properties such as is-a, part-of, and has-part. However, these methods perform experiments to determine parameters that are required to calculate the weights of the edges. The parameters determined by the experiments have a disadvantage because it is difficult to obtain objectivity. In addition, if we assign only one weight to one property, then it is impossible to calculate the semantic similarity under consideration of the structural characteristics of ontology like the hypernym-hyponym relationship between concepts.

So, we introduce a new semantic similarity calculation.

3 Detailed Descriptions for SSINEGE Measurement

The applied basic philosophy to the design of the SSINEGE measurement is to minimize the illustrated disadvantages and to maximize the advantages of the existing semantic similarity measures. As mentioned state, SSINEGE as a new similarity measurement can solve two problems such as all edges in the ontologies have the same weight and all concepts under a particular LCS have the same similarity. To solve the first problem, information contents theory is applied to calculate varied differently weighted edges on all edges. So, there is no chance that the similarity of the two concepts becomes equal by sharing a particular LCS. Using the varied weighted edges, the second problem is naturally solved.

3.1 The Edge Weighting based on Information Contents Theory

In this section, we discuss the first problem, namely how to calculate the varied weights of the edges in the ontology. To do so, by the adopted information contents theory, we calculate the information contents of the concepts using ”probability of an occurrence of a specific event.” According to information contents theory, the contents of information of a specific event is defined as follows [17].

Definition 1 (*Information contents of event e*) Information contents of an event e ($ic(e)$) is simply represented as

$$ic(e) = -\log p(e) \tag{1}$$

, where $p(e)$is an occurrence probability of an event e. At this time, $ic(e)$ is represented as bit.

According to Eq. (1), the higher there are occurrence probability of an event, the less information contents of the event there is, vice versa. To calculate the information contents of the concepts using the information contents theory, let us project the events and their occurrence probabilities as the concepts and the search probabilities of the concepts of ontology, respectively. Generally, the developed ontology has a top node (thing), so general concepts are closely located at the top node and the more specific concepts are closely located at the leaf node (instance). Furthermore, the semantic relation (is-a) is represented by edges between two concepts.

In the case of the edge-based similarity calculation, the most widely used concept is the LCS which is based on the least common super concept of two concepts to be compared. It is general that most edge-based similarity calculation measurements use the number of edges from top concept to the LCS and/or the number of edges from the LCS to the concepts to be compared. The LCS is dynamically determined depending on the two concepts to be compared. At this time, the closer LCS is to the top concept, the smaller the number of searched edges. On the other hand, the closer LCS is to the leaf concept, the smaller the number of connected edges. So, the number of edges to be searched will be increased. In other words, in the case of the edge-based similarity calculation, the closer the concepts to the top concept, the greater probability that the edges connected to the concepts will be searched. Conversely, the closer the node is to the leaf node, the lower the probability that the edges connected to the concepts will be searched. Based on the above description, we define the information contents-based search probability of a particular edge on the ontology as follows.

Definition 2 (*Search probability of edge* e_{ij}) Search probability of an edge e_{ij} between node c_i and c_j is simply represented as

$$p(e_{ij}) = \frac{n_{e_{ij}} + 1.0}{N}, \quad i \neq j, i < j \tag{2}$$

, where N is the total number of edges on the ontology, and $n_{e_{ij}}$ is the number of relevant edges below c_j. Furthermore, c_i is super node of c_j. If c_j is leaf node, $n_{e_{ij}} = 0$ and $pr(e_{ij}) = 0.0$. To prevent that the search probability of the edges become zero, we calibrate using a value “1.0.”

The information contents a specific edge is defined as follow.

Definition 3 (*Information contents of edge* e_{ij}) Information contents of an edge e_{ij} is simply represented as

$$ic(e_{ij}) = -\log p(e_{ij}), \quad i = 1, 2, \cdots, n \tag{3}$$

, where $ic(e_{ij}) > 0$.

According to Eq. (3), the larger search probability of a specific edge, which is the closer to the top concept, and has the smaller information contents of the edge. Conversely, the lower probability of the searching edge that is the edge closer to the end concept, and has the larger information contents. In this way, we can derive the relationship between information contents of the edges and the number of edges under a certain concept. We can interpret that information contents at a particular edge is less or more because the number of edges below the particular edge is large or small. Using the information contents of the edge, the conceptual distance between the concepts is as follows.

Definition 4 (*Conceptual distance of edge* e_{ij}) Conceptual distance of an edge e_{ij} $cd()e_{ij}$)is simply represented as

$$cd(e_{ij}) = \frac{1}{ic(e_{ij})+1.0}, \quad \textit{where}\ 0.0 < cd(e_{ij}) < 1.0 \tag{4}$$

From Eq. (4), it can be seen that there is an inverse relation between the conceptual distance of the edge and the weight of the edge. Furthermore, the distance of the edge should have a value between 0.0 and 1.0. To do so, we compensate the denominator using value “1.0.”

According to Eq. (3) and Eq. (4), it can be seen that the weight of certain edges and the weighted edge-based distance depends on how many edges are associated with the certain edge. As a result, it is not necessary to consider that all concepts under a specific LCS have the same similarity even though SSINEGE is applied to calculate the LCS-based similarity.

3.2 Dynamic Weighted edge-based similarity calculation measurement

The similarity calculation method measures the LCS to utilize the depth from the LCS to the top concept and the depth from the LCS to the comparative concepts. Furthermore, the measurement only the number of edges as depths and does not consider the dynamically varied weights of the edges at all. Therefore, the measurements usually uses the just total depth which is calculated by adding the number of edge from LCS to a top node and to comparative nodes, because they do not consider the weighted edge for the similarity calculations. However, SSINEGE derived the weights of edges by the consideration of the number of lower edges on the ontology. This means that the information contents of the upper edges may contain information contents of the lower edges. As a result, some of the weighted edges which are used to calculate the information contents of the upper edge are also considered to the calculation of the information contents of the lower edges, so the distortion of the similarity value may be occurred on the weighted edge-based.

To minimize the overlapping use of information, SSINEGE computes the semantic similarity without consideration of the weighted edges from the LCS to the top concept. Instead of, SSINEGE generates the conceptual path to connect the two comparative concepts and LCS. The algorithm is shown in Figure 1 to find the conceptual path between two concepts.

The semantic distance between two concepts is derived as the sum of the weights of the edges constituting the conceptual path. The conceptual distance is defined as follow.

Definition 5 (*SSINEGE-based conceptual distance between c_p and c_q*) *SSINEGE*-based conceptual distance between concept c_p and c_q $(CD(c_p,c_q))$ is simply represented as

$$CD(c_q,c_q) = \sum_{i,j \in cp(c_p,c_q)} cd(e_{ij}) \tag{5}$$

, where $cp(c_p,c_q)$ represents the conceptual path between c_p and c_q on ontology.

Furthermore, SSINEGE-based semantic similarity is calculated as follow.

Algorithm of the conceptual path finding between concept c_p and c_q on ontology

```
$c_p$ $c_q$: two concepts for a semantic similarity calculation

Begin ConceptualPathFindingProcess
    $k = 0$
    $stack \leftarrow c_p$
    $pathstack[k] \leftarrow c_p$
    While (notEmpty($stack$)) {
        $temp \leftarrow stack.top$
        If ($temp == c_q$)
            Then $temppath \leftarrow$ Find $\exists pathstack.end == c_q$
                $cp(c_p, c_q) \leftarrow temppath$
                Break( )
        EndIf
        For all $i$
            If ($c_i == parentnodeoftemp \,\|\, c_i == childnodeoftemp$)
                Then $stack \leftarrow c_i$
                    $temppath \leftarrow$ Find $\exists pathstack.end == c_i$
                Add $c_i totemppath$
                $pathstack[k] \leftarrow temppath$
                $k++$
            EndIF
        EndFor
    }
    Return $cp(c_p, c_q)$
End Process
```

Figure 1: **Conceptual Path Finding Algorithm**

Definition 6 (*SSINEGE-based semantic similarity between* c_p *and* c_q) Semantic similarity between concept c_p and c_q simply represented as

$$sim(c_q, c_q) = 1 - \frac{CD(c_q, c_q)}{n(cp(c_p, c_q))} \tag{6}$$

, where $n(cp(c_p, c_q))$ is the total number of edges on conceptual path between c_p and c_q.

Through the above discussion, we have proved that SSINEGE solves the previously illustrated problems such as all of the weights of the edges as the similarities of concepts under a particular LCS are the same.

Figure 2 shows the algorithm for calculating the semantic similarity between two concepts using SSINEGE.

```
Algorithm for SSINEGE measurement-based Semantic Similarity Calculation

$O$: ontology
$c_i$: $i^{th}$ concept on $O$
$c_j$: concept link with $c_j$
$e_{ij}$: edge between $c_i$ and $c_j$
$E$: set of edge weight $e_{ij}$

Begin SSINEGE Similarity Calculation Process
    Load $O$
    $E \leftarrow$ EdgeWeightCalculationFunction($O$)
    Input($c_p$, $c_q$)
    $cp(c_p, c_q) \leftarrow$ ConceptualPathFindingProcess ($c_p$, $c_q$)
    $CD(c_p, c_q) \leftarrow \sum_{i,j \in cp(c_p, c_q)} cd(e_{ij})$
    $sim(c_p, c_q) \leftarrow 1 - \frac{CD(c_p, c_q)}{n(cp(c_p, c_q))}$
    Return $sim(c_p, c_q)$
End Process

Begin EdgeWeightCalculationFunction
    $temp \leftarrow c_0$
    $stack \leftarrow temp$
    While (notEmpty($stack$)){
        $temp \leftarrow stack.top$
        While ($temp.childnode != null$) {
            $stack \leftarrow \forall temp.childnode$
            $e_{ij} \leftarrow \sum_{ij \in temp}(temp.childnode)$
        }
    }
    For all $i$
        $e_{i,temp} \leftarrow e_{ij}$
        $stack \leftarrow e_{i,temp}$
        While (notEmpty($stack$)) {
            $temp \leftarrow e_{ij}$
            For all $j$
                If ($e_{i,temp} == e_{temp,j}$)
                    Then $e_{ij} = e_{i,temp} + e_{temp,j}$
                        $stack \leftarrow e_{i,temp}$
                EndIf
            EndFor
        }
    EndFor
    Return $E$
End Function
```

Figure 2: **Semantic Similarity Calculation Algorithm using SSINEGE Measurement**

4 Experiments and Performance Evaluation

To demonstrate the superiority of the proposed SSINEGE, we developed the ”food ontology” with 1,000 concepts and the ”foods ontology” with 5,000 concepts. The reason for developing two ontologies is to understand the effect of the number of concepts included in the ontology on semantic similarity. To develop the ontologies, we used an open-source ontology editor, named by protégé. The developed foods ontology has property like ‘$rdfs : subClassOf'$ which indicates the hypernym/hyponym relation between concepts. The top level classes of ”foods ontology” composed of $< owl : Classrdf : ID =$ ”$Fruit$”$/ >$, $< owl : Classrdf : ID =$ ”$Vegetable$”$/ >$, $< owl : Classrdf : ID =$ ”$Meat$”$/ >$, and $< owl : Classrdf : ID =$ ”$ProcessedFood$”$/ >$.

The experiment was carried out as following. For the experiment, eight pairs of concepts are selected for the comparisons. The semantic similarities of the eight pairs of concepts are calculated by SSINEGE. In addition, four other measurements are selected from the existing research results [8], [11], [14], [20]. For the experiment, the four types of similarity measurements are the most representative ontology-based semantic similarity calculations. Among them, the similarity values of [8] proved to be the most similar to the intuitive similarity of people [5]. Needless to say, we conducted experiments to comparing of the semantic similarities between concepts on two ontologies.

4.1 Relevance of Semantic Similarity

As mentioned state, among the edge-based semantic similarity measurements, [8] is the measurement that yields the most similar results to human intuition. Due to the nature of the similarity calculation measurement, it is not easy to make a direct comparison. So, based on the results of [8], we perform a relative comparison of the similarity values of SSINEGE measure and the three measures except [8]. The selected concepts for the comparison are as follows: $SoftPersimmon(SO)$, $SweetPersimmon(SW)$, $Pear(PE)$, $SoybeanPaste(SP)$,$RedPepperPaste(RE)$, $Nut(NU)$, $Walnut(WA)$, $Strawberry(ST)$, $Raspberry(RA)$, $PorkBelly(PB)$, $PorkNeckBelly(PN)$, $Fruit(FR)$, $ProcessedFood(PF)$, and $Ham(HA)$. To compute the semantic similarity, we generate eight pairs of concepts selected from 14 concepts. The results of similarity calculation are shown in Figure 3 and Figure 4. From the results, SSINEGE describes the most similar pattern to the results of [8].

More specifically, to prove the excellence of SSINEGE, we directly compared the similarity of SSINEGE with the similarity of [8]. The results are depicted in Figure 5 and Figure 6.

As depicted in Figure 5 and Figure 6, it can be seen that there are large deviations for some pairs of concepts between the semantic similarities of [8] and SSINEGE. To find the cause of the deviation, we performed additional experiments.

4.2 Discriminative Capability

In order to verify the superiority of SSINEGE measure, we conduct to analyze the results of experiment 1 in detail. In experiment 1, we performed the similarity comparison after generating the five pair of concepts such as (SO,SW), (FR,PF), (SO,PE), (FR,HA), and (NU,WA) from the food ontology with 5,000 concepts. Intuitively, the semantic similarity of (SO,SW) should be larger than that of (FR,PF) However, the calculated semantic similarities are $sim(SO,SW) = 0.75$ and $sim(FR,PF) = 0.35$ between the pairs using SSINEGE, while the both semantic similarities using [8] of them are equally ”0.75.” In a similar way, we compare the semantic similarity for the pairs of (SO,PE) and (FR,HA). Similar to

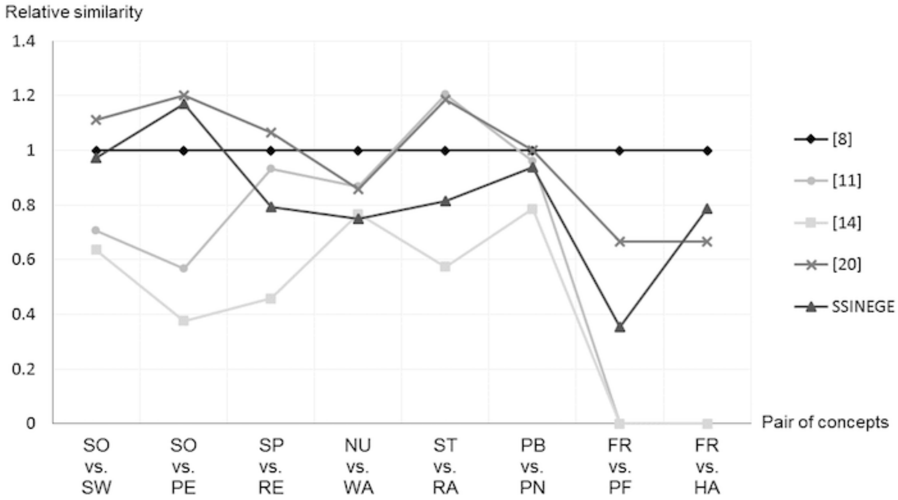


Figure 3: **Comparison of semantic similarity (No. of concepts: 1,000)**

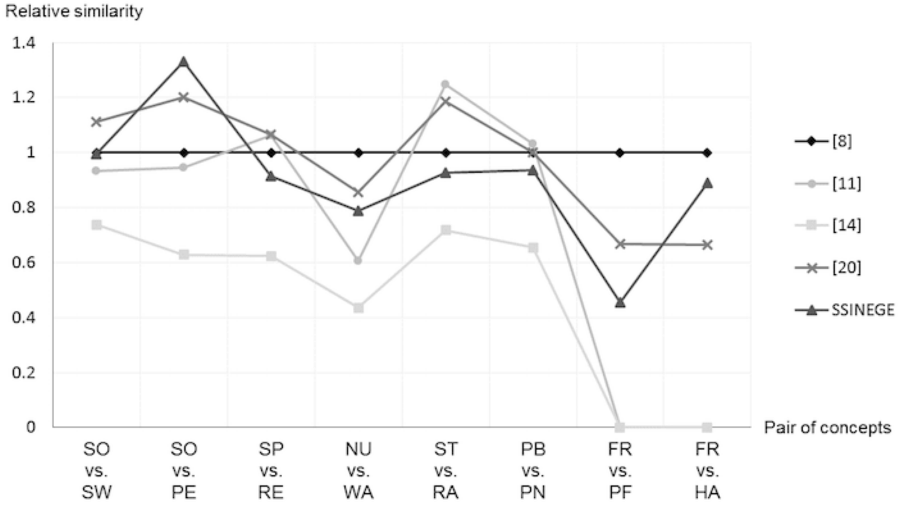


Figure 4: **Comparison of semantic similarity (No. of concepts: 5,000)**

the previous case, the semantic similarities by [8] are equally "0.5," while the semantic similarities by SSINEGE are $sim(SO,PE) = 0.67$ and $sim(FR,HA) = 0.45$. Although the semantic similarity of "*FR*" and "*HA*" that are belonging to the different categories should be smaller than those of "*SO*" and "*PE*" belonging to the same category, named fruit, [8] seems to have no ability to discriminate it. From the experiments, the results are depicted in Figure 7.

Through the experiments, it proved that the calculated similarity by SSINEGE is more intuitive than [8]. It means that the result is the closest to people's thought. In addition, we have demonstrated that SSINEGE has the ability to more sensitively distinguish semantic differences in concepts than [8].

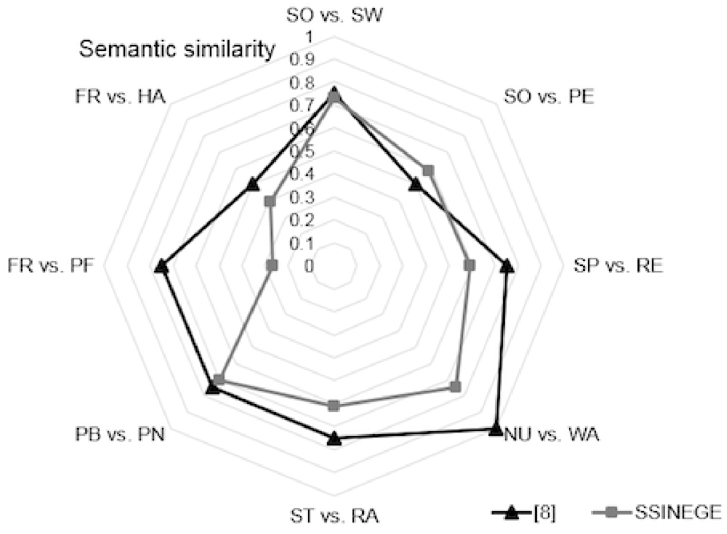


Figure 5: **Similarity comparison between SSINEGE measurement and [8] (No. of concepts: 1,000)**

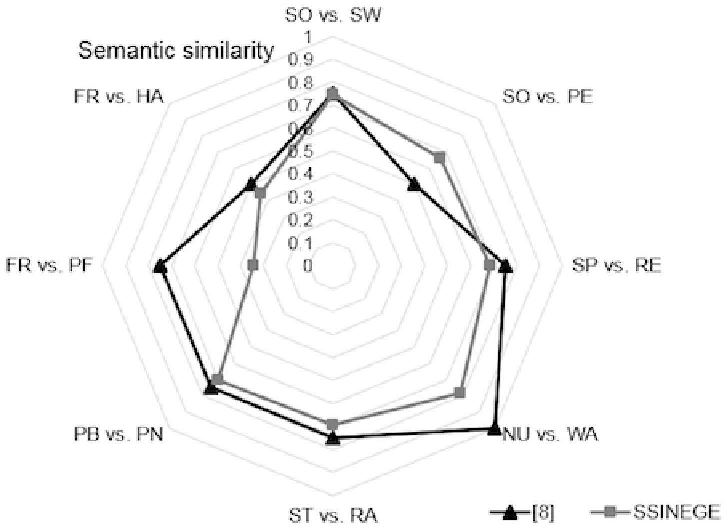


Figure 6: **Similarity comparison between SSINEGE measurement and [8] (No. of concepts: 5,000)**

5 Conclusion and Further Research

In this paper, we propose Semantic Similarity calculation measurement using INformation contents on EdGEs of ontology (SSINEGE) as a hybrid edge- and information contents-based methodology. SSINEGE measurement is developed to overcome the limitations such as all edges in the ontologies have the same weight and all concepts under a particular LCS have the same similarity. To solve the problems, SSINEGE measurement adopts the information contents theory to derive the varied weights of on the ontology. Additionaly, to minimize the overlapped use of information, SSINEGE measurement generates

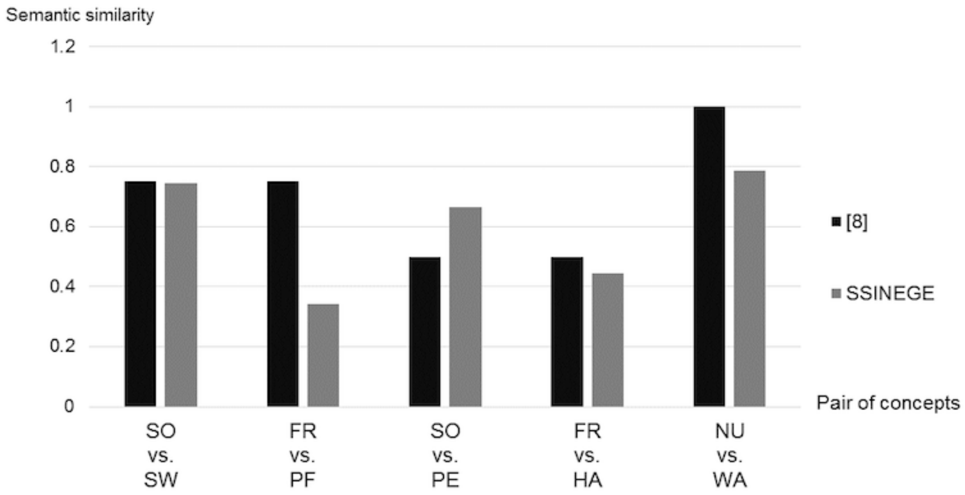


Figure 7: **Discrimination capability Comparison between SSINEGE measure and [8]**

the conceptual path between the comparative concepts and the LCS for the semantic similarity. To verify the superiority of the proposed SSINEGE, we performed a relative comparison of the similarity values between SSINEGE and the three measurements [11], [14], [20] except for Leacock and Chodoro [8], because Leacock and Chodoro calculates the most similar to the intuitive similarity of people. As the result, we found that the calculated similarity by SSINEGE is similar to [8]. In addition, it proved that the discrimination capability of SSINEGE is better than [8].

In future studies, this research can be extended to several directions. It is needed to be performed by additional experimentations on SSINEGE against various ontologies. It requires detailed analysis to prove the superiority of performance of SSINEGE through applying the ontologies to various domain, sizes and depths. In addition, the test is needed to be done on public ontology including WordNet to evaluate the performance of SSINEGE and to find the best conceptual path on the ontology.

Acknowledgments

This research was partially supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2016R1D1A1B03932110), partially supported by the IT R&D program of KEIT (No. 1005-0810, Development of Disability Independent Accessibility Enhancement Technology for Input and Abnormality of Home Appliances).

References

[1] G. Bouma. Normalized (pointwise) mutual information in collocation extraction. In *Proc. of the Biennial GSCL Conference 2009, Potsdam, Germany*, volume 156, 2009.

[2] D. Cai, Y. Bai, S. Yu, N. YE, and X. REN. A context based word similarity computing method. *Journal of Chinese Information Processing*, 24(3), March 2010.

[3] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs. Swoogle: a search and metadata engine for the semantic web. In *Proc. of the 13th ACM International Conference on Information and Knowledge Management (CIKM'04), Washington D.C., USA*, pages 652–659. ACM, 2004.

[4] J. Ge and Y. Qiu. Concept similarity matching based on semantic distance. In *Proc. fo the 4th International Conference on Semantics, Knowledge and Grid (SKG'08), Beijing, China*, pages 380–383. IEEE, December 2008.

[5] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. Petrakis, and E. Milios. Information retrieval by semantic similarity. *International journal on semantic Web and information systems (IJSWIS)*, 2(3):55–73, 2006.

[6] J. Kwon, C.-J. Moon, S.-H. Park, and D.-K. Baik. Measuring semantic similarity based on weighting attributes of edge counting. In *Proc. of the 13th International Conference on AI, Simulation, and Planning in High Autonomy Systems (AIS'04), Jeju Island, Korea*, volume 3397 of *Lecture Notes in Computer Science*, pages 470–480. Springer Berlin Heidelberg, October 2005.

[7] T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240, April 1997.

[8] C. Leacock and M. Chodorow. Filling in a sparse training space for word sense identification, 1994.

[9] A. L. Lemos, F. Daniel, and B. Benatallah. Web service composition: a survey of techniques and tools. *ACM Computing Surveys (CSUR)*, 48(3):33:1–33:41, 2016.

[10] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proc. of the 5th annual international conference on Systems documentation (SIGDOC'86), Toronto, Ontario, Canada*, pages 24–26. ACM, 1986.

[11] D. Lin. An information-theoretic definition of similarity. In *Proc. of the 15th International Conference on Machine Learning (ICML'98), Madison, Wisconsin, USA*, volume 98, pages 296–304. Morgan Kaufmann Publishers Inc., July 1998.

[12] E. G. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou. X-similarity: computing semantic similarity between concepts from different ontologies. *Journal of Digital Information Management (JDIM)*, 4(4):233–237, 2006.

[13] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.

[14] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th International Joint Conference on Artificial intelligence (IJCAI'95), Montreal, Quebec, Canada*, volume 1, August 1995.

[15] R. Richardson, A. Smeaton, and J. Murphy. Using wordnet as a knowledge base for measuring semantic similarity between words. Technical Report CA-1294, School of Computer Applications, Dublin City University, 1994.

[16] M. A. Rodriguez and M. J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):442–456, 2003.

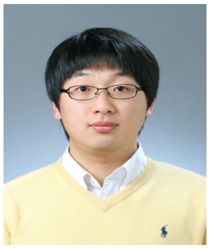
[17] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[18] A. Tversky. Features of similarity. *Psychological review*, 84(4):327–352, July 1977.

[19] Wikipedia. Ontology, January 2012. *https://en.wikipedia.org/wiki/Ontology* [Online; Accessed on January 31, 2017].

[20] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proc. of the 32nd Annual Meeting on Association for Computational Linguistics (ACL'94), Las Cruces, New Mexico, USA*, pages 133–138. Association for Computational Linguistics, June 1994.

Author Biography



Sunghwan Jeong is in Ph. D. course in the department of Industrial Engineering at Sungkyunkwan University. He received his bachelor's degree from s Sungkyunkwan University. His main interests are ontology, semantic web, web service, Web-of-Thing, and Cyber-Physical System.

Jun Hyeok Yim is a R&D Engineer in the SOOSAN INT. He received his BA and MS degrees from Sungkyunkwan University. His main interests are Network security and semantic web.

Hyun Jung Lee is a research professor of Yonsei Institute of Convertgence Technology at Yonsei University. She received PhD from Korea Advanced Institute of Science and technology (KAIST). Her long-term research theme is modeling and developing Intelligent Information Systems (IIS) and Electronic Commerce (EC) within management information systems. Her primary research interests are effective information processing to online market with rigorous application of a multiple levels-of-analysis perspective, customized product recommendations on online market, developing semantic web service applications in knowledge management, delivery and fulfillment issues on ubiquitous computing, cost reduction decision using information processing and research methodologies. She published several papers on Decision Support Systems, Soft Computing Journal, and so on.

Mye Sohn is a professor in the department of Systems Management Engineering at Sungkyunkwan University and the director of the centre for Woman in Science, Engineering and Technology in Gyeonggi. She received her MS and Ph. D. from the Korea Advanced Institute of Science and Technology (KAIST). Sohn's research has been focused on the semantic web, ontology, web service, context-aware computing, and Web-of-Things. She got the best paper awards in KIIS ('03), KMES ('11), and IMIS ('2014). She has published several papers in Communications of the ACM, Information Sciences, An International Journal of Mathematical and Computer Modelling, and so on.