Assessing semantic similarity is a fundamental requirement for many AI applications. Crisp ontology (CO) is one of the knowledge representation tools that can be used for this purpose. Thanks to the development of semantic web, CO-based similarity assessment has become a popular approach in recent years. However, in the presence of vague information, CO cannot consider uncertainty of relations between concepts. On the other hand, fuzzy ontology (FO) can effectively process uncertainty of concepts and their relations. This paper aims at proposing an approach for assessing concept similarity based on FO. The proposed approach incorporates fuzzy relation composition in combination with an edge counting approach to assess the similarity. Accordingly, proposed measure relies on taxonomical features of an ontology in combination with statistical features of concepts. Furthermore, an evaluation approach for the FO-based similarity measure named as FOSE is proposed. Considering social network data, proposed similarity measure is evaluated using FOSE. The evaluation results prove the dominance of proposed approach over its respective CO-based measure. © 2017 Wiley Periodicals, Inc.

1. INTRODUCTION

Similarity reasoning is the identification of syntactically different concepts that are semantically close. Assessing concept similarity is growing in importance within ontology engineering and, in particular, ontology merging and ontology alignment.\(^1\) An ontology is a knowledge representation mechanism, which is understandable by intelligent agents. It consists of a hierarchical description of concepts in a particular domain connected by taxonomic and non-taxonomic relations.\(^2\) It is employed in reasoning about domain concepts. Ontologies are the fundamental infrastructures in semantic web.\(^3\) With the rapid development of the semantic web, it is likely that the number of ontologies will greatly increase during the next few years, which leads to the arising demand for rapid and accurate assessing concept similarity.\(^4,5\) In this context, assessing concept similarity becomes more important in the presence

*Author to whom all correspondence should be addressed; e-mail: zarandi@aut.ac.ir

INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 32, 1167–1186 (2017) © 2017 Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com. DOI 10.1002/int.21895
of vague information. When some relations between domain concepts are vague or when there is uncertainty in defining a concept, these types of problems can be tackled with fuzzy information. Fuzzy logic theory was proposed by Zadeh\textsuperscript{7} and later applied successfully in various research areas. Fuzzy logic, as a powerful infrastructure in uncertainty management, was coupled with the ontology to originate fuzzy ontology (FO).\textsuperscript{13} FO is a generalization of crisp ontology (CO) where fuzzy relations exist between crisp concepts. FO have been successfully implemented in several application areas such as news summarization,\textsuperscript{14} diet recommendation,\textsuperscript{15,16} flight booking,\textsuperscript{17} information retrieval,\textsuperscript{18–25} reputation management,\textsuperscript{26,27} collision avoidance,\textsuperscript{28} and knowledge mobilization.\textsuperscript{29,30}

However, the literature on FO-based assessment of similarity is limited to the use of formal concept analysis (FCA). FCA is concerned with the formalization of concepts and conceptual thinking. A key limitation of FCA-based approach, however, is that it necessitates a particular type of world modeling, that is, concept-to-attributes, which may not be applied in all situations. In addition, considering the need for human intervention in concept-to-attributes database creation, FCA-based approach is semi-automatic rather than fully automatic and of exponential time order, \( O(2^N) \).\textsuperscript{32}

This article presents a measure for assessing concept similarity based on FO. With respect to other papers defined in the literature, the key concepts underlying the proposed approach are its independence of FCA and lower pre-required complexity. As a case study, this measure is then used to generate a similarity matrix of concept pairs in the context of social networks (SNs).

Furthermore, to evaluate the proposed semantic similarity measure, a new evaluation approach is proposed. The approach, named as FO-based similarity evaluation (FOSE), is the first data-driven evaluation approach of FO-based semantic similarity as far as this study is concerned. FOSE is then applied on our case study to evaluate proposed FO-based similarity measure.

The rest of this paper is organized as follows: Section 2 defines some notations and terminology used in the rest of this paper. Section 3 is devoted to the proposed FO-based assessing concept semantic similarity. In Section 4, the proposed approach for evaluating a FO-base similarity measure is introduced. A case study of deployment of proposed measure in a context of SNs is covered in Section 5 and finally comparison with literature work, concluding remarks, and future work studies are discussed in Section 6.

## 2. NOTATIONS AND PRELIMINARY DEFINITION

In this section, we formally define some terms and describe the notation used in this paper.

In formal terms, a fuzzy set can be defined as follows:

**Definition 1.** A fuzzy set \( S \) over the universe of discourse \( X \) is defined by its membership function \( \mu_S \), which maps \( S \) elements to a value between \( [0, 1] \) interval.\textsuperscript{17}

\[
\mu_S(x) : X \to [0, 1]
\]  

(1)
where $S$ is the fuzzy set and $\mu$ is the membership function. $\mu_s(x)$ represents the degree to which $x$ belongs to $S$. If $X$ is continuous, then $S$ can be rewritten as follows:

$$ S = \int_x \mu_S(x) / x $$

Additionally, $S$ can be organized into an ordered set of pairs as follows:

$$ S = \{(x, \mu_S(x)) | x \in X\} $$

**Definition 2.** If $X$ and $Y$ are any set and $R \subseteq X \times Y$, we call $R$ a binary relation from $X$ to $Y$ or a binary relation between $X$ and $Y$.

**Definition 3.** Let $X$ and $Y$ be two universes of discourse. A fuzzy binary relation $R(x, y)$ is a function of the product space of $X \times Y$, which assigns each pair $(x, y)$ its grade of membership $\mu_R(x, y)$ in $R$.

$$ R(x, y) = \{((x, y), \mu_R(x, y)) | (x, y) \in X \times Y\} $$

The value of $\mu_R(x, y)$ lies in $[0, 1]$ interval and represents the strength of the relationship between $x$ and $y$ where larger values denote stronger relations.

**Definition 4.** Let $C$ be a set of concepts, $R$ a set of taxonomical relations between concept pairs, $A$ is a set of axioms, and $D$ the domain. An ontology can be formally represented as

$$ O = (C, R, A, D) $$

Typically, an ontology is illustrated as a directed acyclic graph (DAG) or a hierarchy, in which nodes correspond to concepts and edges represent relationships between pairs of concepts. In some ontologies, there is only one relationship between nodes, whereas in more general case, there exist more than one relationship between nodes. The most common type of ontology relation is the taxonomical “is-a” relation, which indicates the similarity of concept pairs. In a hierarchy corresponding to an ontology, there is a node specified as the root. The root is the starting node. A path is a sequence of adjacent (via the edges) nodes in the hierarchy. The name of each node at an intermediate level is associated with a parent, one or more child nodes and one or more sibling nodes. Parent node is the node one level higher in the hierarchy. Inversely, if a node is a parent of another node, the node is called a child of the parent. Consequently, a node may have several parent nodes, and vice versa. Sibling nodes share the same parent. The depth of a node is the length of the path to its root. Let us use $parent(c)$, $sibling(c)$, and $depth(c)$ operator to demonstrate the parent, sibling set, and depth of a node $c$ in a
hierarchy, respectively. A node that is connected to all lower-level nodes is demonstrated by \textit{ancestor}(c). Given two nodes in an ontology, they must share a set of common ancestor nodes, and the one with the highest depth is typically referred to as the lowest common ancestor (LCA) of the two nodes. Discarding the direction of the edges in an ontology, there exists at least one path between every pair of two nodes. Among all possible paths between two crisp concept $c_1$ and $c_2$ in an ontology, the one passing their LCA is the shortest path ($sp$) between two concepts, that is, $sp(c_1, c_2) = |\text{path}(c_1, \text{LCA}(c_1, c_2))| + |\text{path}(\text{LCA}(c_1, c_2), c_2)|$, where $|\text{path}(c_i, c_j)|$ counts the number of edges (relations) in the path from $c_i$ to $c_j$.

\textbf{DEFINITION 5}. A FO is defined by means of fuzzy relations characterized by a membership function. Considering Equation 5, this type of ontology is formulated as follows:

$$\text{FO} = (C, R_f, A, D)$$

(6)

$$R_f = \{(c_i, c_j), \mu_R(c_i, c_j)) \in C \times C, \ \mu_R(c_i, c_j) \in [0,1]\}$$

(7)

where, $R_f$ is a (binary) fuzzy relation over two countable crisp set of concepts and the membership function $\mu_R(c_i, c_j)$ denotes the strength of $R_f$ relations between $c_i$ and $c_j$. Despite a CO, in a FO, $R_f$ values lies in [0,1] interval rather than binary (0 or 1) values, where higher values of $\mu_R$ denotes stronger relation between its corresponding concepts.

A pictorial example of a FO is demonstrated in Figure 1 where crisp concepts are connected to each other by fuzzy “is-a” relations and a fuzzy degree of membership, $\mu(x)$ in [0,1] interval is assigned to each relation.
Let $R_1(x, y), (x, y) \in X \times Y$ and $R_2(y, z), (y, z) \in Y \times Z$ be two fuzzy relations. The max-product composition of two fuzzy relations $R_1$ and $R_2$ is denoted as $R_1 \circ R_2(x, z)$ and defined as follows:

$$R_1 \circ R_2 (x, z) = \max_y [\mu_{R_1} (x, y) \cdot \mu_{R_2} (x, y)]$$

(8)

A dataset, $X = \{x_1, x_2, \ldots, x_N\}$ is a set of $N$ objects or data points represented as feature vectors in a F-space. A distance matrix of a dataset, $\text{dist}(X)$, is a matrix demonstrating the distance between each pairs of $X$ elements. Having an ontology $O = (C, R, D)$ of $N$ concepts, a distance matrix is corresponded to $C$, which determines the distance of concept pairs.

Semantic similarity (distance) computes the similarity between concepts that need not be lexically similar. Semantic distance can be inferred from web data. Statistical analysis of web files for a set of concepts $C$ results in a distance matrix denoted as $\text{web\_dist}(C)$. Another approach for the assessment of concept semantic distance is based on an ontology.

**Definition 6.** Having a CO of $N$ concepts as $\text{CO} = (C, R, D)$, the semantic distance of concept pairs can be calculated based on ontology relations, $R$. Accordingly, the result is a distance matrix of concept pairs with the size $N \times N$ which is denoted as $\text{CO\_dist}$. However, if the source ontology is a FO, the distance measure would be a FO-based distance measure, denoted as $\text{FO\_dist}$.

Having two distance matrix of objects (concepts), their correlation and variation can be evaluated by Pearson correlation, relative root squared error (RRSE), and relative absolute error (RAE) as follows:

$$\text{Pearson correlation} (A, B) = \frac{\sum_{a \in A, b \in B} (a - \bar{a})(b - \bar{b})}{\sqrt{\sum_{a \in A, b \in B} (a - \bar{a})^2 (b - \bar{b})^2}}$$

(9)

$$\text{Relative root squared error} (A, B) = \text{RRSE} (A, B) = \sqrt{\frac{\sum_{a \in A, b \in B} (a - b)^2}{\sum_{a \in A} (a - \bar{a})^2}}$$

(10)

$$\text{Relative absolute error} (A, B) = \text{RAE} (A, B) = \frac{\sum_{a \in A, b \in B} |a - b|}{\sum_{b \in B} |b - \bar{b}|}$$

(11)

Pearson correlation measures the strength of a relation between two variables $A$ and $B$. $a$ and $b$ denote different values of variables $A$ and $B$, respectively. $\bar{a}$ and $\bar{b}$ specify the average value of variables $A$ and $B$, respectively. The higher value of Pearson correlation indicates a more dominant relation between two variables. The two other criteria, RAE and RRSE, measure relative variation between a target variable, $A$, and its predicted value, $B$. RRSE calculates the root of squared error.
between target variable $A$ and its predicted value $B$. RAE on the other hand calculates the absolute value of this variation. The smaller RAE and RRSE indicate a better prediction. Here $|x|$ refers to the absolute value of $x$.

In this context, the problem we want to solve is the following: having a set of concepts of web data, assess their semantic similarity based on FO and evaluate the results.

3. A NEW SEMANTIC SIMILARITY MEASURE BASED ON FO

Semantic similarity computes the similarity between concepts that need not be lexically similar.\(^{38}\) Ontology-based analysis has been a popular approach in recent years for semantic similarity assessment. However, they mainly utilizes CO structure.\(^{34}\) Despite the information richness of FO, it has been considered to a limited extent in the literature for semantic similarity measurement. A FO is more informative in comparison with a CO. In addition to providing taxonomical relations of concepts, it provides information on the strength of the relation between concept pairs. Accordingly, similarity assessment based on FO could provide results that are closer to reality. The input for similarity assessment based on FO is a FO of concepts connected by fuzzy relations. If a FO is not available, a CO can be converted to a FO using the algorithm proposed by Ref. 39. Their approach generates a FO from a CO by means of a distance matrix of concepts inferred from the web, $\text{dist}_{\text{web}}(C)$. This approach maps a two-dimensional $\text{dist}_{\text{web}}$ matrix to a nested ontology structure to generate a FO.

3.1. Assessing Concept Similarity Based on FO

This section aims at proposing a novel approach for semantic similarity assessment based on FO. Before proceeding, let us define some propositions, which will be used later in this section:

- Two completely similar objects give the maximum similarity, $(\text{sim}(x, y) = 1)$, whereas the least similar pairs give the minimum value $(\text{sim}(x, y) = 0)$. That is, similarity is the complement of dissimilarity in the range of $[0, 1]$ and so one can be easily derived from the other as follows:\(^{40}\)

$$\text{dist}(x, y) = 1 - \text{sim}(x, y) \quad (12)$$

- In a FO $= (C, R_f, D)$, the fuzzy value of taxonomical “is-a” relation between concept pairs demonstrates their level of belonging to each other. Accordingly, the fuzzy value of taxonomical “is-a” relations existing in a FO can be considered as the similarity of concepts, s. t.,

$$\forall c_i, c_j \in C, \exists R(c_i, c_j), R \in R_f : \text{sim}(c_i, c_j) = \mu_R(c_i, c_j) \quad (13)$$

Now let us assume a FO $= (C, R_f, A, D)$, consisting of $N$ concepts indexed as $c_i, 1 \leq i \leq N$, which are connected to each other by fuzzy relations indexed as
\( \mu_R(c_i, c_j) : 1 \leq i, j \leq N \), where \( \mu_R(c_i, c_j) \) denotes the strength of “is-a” relation between \( c_i \) to \( c_j \). Assuming \( c_i \) as the node with lower depth in comparison with \( c_j \), s. t. \( \text{depth}(c_i) \leq \text{depth}(c_j) \), then relations between \( c_i \) and \( c_j \) lie in one of the following two categories:

1. \( c_i \) is the parent or an ancestor of \( c_j \).
2. \( c_i \) and \( c_j \) are neither parent nor ancestors but are connected to each other via common ancestors.

According to this partitioning, our proposed approach for semantic distance assessment of any concept pairs \((c_i, c_j)\) in a FO is defined by Algorithm 1.

**Algorithm 1. calcDistance\((c_i, c_j, FO)\)**

1: if \((c_i = \text{ancestor}(c_j))\)  
2: \( \text{itr} = c_j \)  
3: \( \text{sim}(c_i, c_j) = 1 \)  
4: while \((\text{itr}! = c_i)\)  
5: \( \text{sim}(c_i, c_j) = \text{sim}(c_i, c_j) \times \mu_R(\text{itr}, \text{parent}(\text{itr})) \)  
6: \( \text{itr} = \text{parent}(\text{itr}) \)  
7: end while  
8: \( \text{dist}(c_i, c_j) = 1 - \text{sim}(c_i, c_j) \)  
9: else  
10: \( \text{anc} = \text{LCA}(c_i, c_j) \)  
11: \( \text{dist}(c_i, c_j) = \text{calcDistance}(c_i, \text{anc}, \text{FO}) + \text{calcDistance}(\text{anc}, c_j, \text{FO}) \)  
12: end if  
13: return

If \( c_i \) is an ancestor of \( c_j \): As mentioned earlier in Equation 13, for concepts directly connected to each other by “is-a” relation, the fuzzy value of their relation, that is, \( \mu_R(c_i, c_j) \), is set as their similarity. However in case that there are some intermediate relations on the path from \( c_i \) to \( c_j \), all these relations are composed using fuzzy max-product composition as defined in Equation 8 in order to generate the final value for similarity of \( c_i \) to \( c_j \). Afterward, considering Equation 15, their distance equals their similarity subtracted from 1.

More precisely, let us assume that there are \( k \) intermediate concepts between \( c_i \) and \( c_j \). The similarity of \( c_i \) to \( c_j \) is calculated by max-product composition (Equation 8) of all intermediate fuzzy relations on the path from \( c_i \) to \( c_j \) as follows:

\[
\text{FO}_{\text{sim}}(c_i, c_j) = R(c_j, \text{parent}(c_j)) \circ R(\text{parent}(c_j), \text{parent}(\text{parent}(c_j))) \circ \ldots \circ R(\text{child}(c_i), c_i) = \prod_{l=1}^{k} \mu_R(l, \text{parent}(l)) \tag{14}
\]

and accordingly their distance is equal to:

\[
\text{FO}_{\text{dist}}(c_i, c_j) = 1 - \prod_{l=1}^{k} \mu_R(l, \text{parent}(l)) \tag{15}
\]
As a pictorial example, consider Figure 2, which represents a FO fragment. The similarity of \( c_1 \) and \( c_{10} \) is computed by max-product composition of fuzzy relations on the path from \( c_1 \) to \( c_{10} \) s. t.

\[
\text{FO}_{\text{dist}}(c_1, c_{10}) = 1 - \left( R(c_{10}, c_7) \odot R(c_7, c_3) \odot R(c_3, c_1) \right)
\]

\[
= 1 - (0.7 \times 0.9 \times 0.4) = 0.75
\]

If \( c_i \) and \( c_j \) are connected via a common ancestor(s), which means that the path from \( c_i \) to \( c_j \) passes from their common ancestor(s), the pass containing their LCA is the shortest path between \( c_i \) and \( c_j \). Subsequently, their distance is set as the sum of the distance of each to their LCA. To calculate the distance of each to their LCA, the method defined in previous section is used, s. t. the max product composition of relations on the path from each node to their LCA is calculated. As an example consider Figure 2 again. The similarity of \( c_8 \) and \( c_{10} \) is calculated by consideration of the similarity of each to \( c_3 \), which is their LCA.

Finally, the distance of each node to itself is set as 0. Consequently, the proposed distance measure can be summarized as follows:

\[
\text{FO}_{\text{dist}}(c_i, c_j) = \begin{cases} 
1 - \prod_{k=c_i} \mu_{R(k, \text{parent}(k))} & , c_i = \text{ancestor}(c_j) \\
1 - \prod_{k_1=c_i, LCA(c_i, c_j)} \mu_{R(k_1, \text{parent}(k_1))} & , c_i \neq \text{ancestor}(c_j) 
\end{cases}
\]

\[
+ \left( 1 - \prod_{k_2=LCA(c_i, c_j), c_j} \mu_{R(k_2, \text{parent}(k_2))} \right) 
\]

\[
(16)
\]
3.2. Examples

To illustrate the behavior of our approach, let us consider following portions of two FOs as depicted in Figures 1 and 2 and calculate semantic similarity of some of their concepts:

Example 1. Consider the FO of Figure 1, which consists of six concepts connected by five fuzzy relations. According to the proposed method the similarity between “Computer Science” and “Regression Analysis” and the distance between “Reinforcement Learning” and “Data Mining” is defined as follows:

\[
\text{FO}_{\text{sim}}(\text{Computer Science, Regression Analysis}) = 0.5 \times 0.7 = 0.35.
\]

\[
\text{FO}_{\text{dist}}(\text{Data Mining, Reinforcement Learning})
\]

\[
= (1 - 0.6) + (1 - 0.7 \times 0.8) = 0.4 + 0.44 = 0.84.
\]

Example 2. Consider the FO segment of Figure 2, which consists of eight concepts connected by seven fuzzy relations. According to the proposed method, the distance of assumed \( c_3, c_7, c_{10}, \) and \( c_8 \) concepts are calculated as follows:

\[
\text{FO}_{\text{dist}}(c_3, c_7) = 1 - \text{FO}_{\text{sim}}(c_3, c_7) = 1 - \mu_R(c_3, c_7) = 1 - 0.9 = 0.1.
\]

\[
\text{FO}_{\text{dist}}(c_{10}, c_3) = 1 - \text{FO}_{\text{sim}}(c_{10}, c_3) = 1 - \mu_R(c_{10}, c_7) \times \mu_R(c_7, c_3)
\]

\[
= 1 - 0.36 = 0.64.
\]

\[
\text{FO}_{\text{dist}}(c_8, c_{10})
\]

\[
= (1 - \text{FO}_{\text{sim}}(c_8, c_3)) + 1 - \text{FO}_{\text{sim}}(c_{10}, c_3)
\]

\[
= (1 - \mu_R(c_8, c_6) \times \mu_R(c_6, c_3)) + (1 - \mu_R(c_{10}, c_7) \times \mu_R(c_7, c_3))
\]

\[
= (1 - 0.8 \times 0.9) + (1 - 0.4 \times 0.9) = 0.92.
\]

3.3. Properties of the Proposed Measure

In order to show the validity of the presented measure, we have studied the properties that a distance measure must fulfill. It is important to note that the fulfillment of those properties is a requirement if the measure is used in conjunction with some reasoning techniques. A distance function must satisfy three properties: positivity, minimality, and symmetry as stated by:

\[
\text{PROPOSITION 1. The function } \text{FO}_{\text{dist}} \text{ fulfills the properties of}
\]

\[
\forall a, b \in \text{FO, } \text{FO}_{\text{dist}}(a, b) \geq 0 \quad \text{(positivity)}
\]

\[
\forall a \in \text{FO, } \text{FO}_{\text{dist}}(a, a) = 0 \quad \text{(minimality)}
\]

\[
\forall a, b \in \text{FO, } \text{FO}_{\text{dist}}(a, b) = \text{FO}_{\text{dist}}(b, a) \quad \text{(symmetry)}
\]

International Journal of Intelligent Systems DOI 10.1002/int
Proof.

Positiveness: Calculation of fuzzy distance between concepts A and B consists of three steps. First fuzzy value of the shortest path from each node to their LCA is calculated then these values are subtracted from 1 and afterward are summed up. The first part, calculation of distance of each node to their LCA, is composed of multiplication of some fuzzy values. Accordingly, the result lies in [0 1] interval. Afterward, the result is subtracted from one. Since the result lies in [0 1] interval its subtraction form 1 is always positive and then two positive numbers are added which results in a new positive number.

Minimality: As mentioned earlier, the distance of each node to itself is set to zero.

Symmetry: Concepts A and B are evaluated according to the fuzzy values of ontology relations on the path to their LCA. The order of calculation does not differ and all the operations performed on these values are commutative (product, max, and addition), accordingly the measure accomplishes the symmetry property.

4. A NEW APPROACH FOR EVALUATING FO-BASED SIMILARITY MEASURE

The literature on evaluation of ontology-based similarity measures is limited to consideration of CO. This section first investigates some criteria that can be considered in evaluation of a FO-based similarity measure. Next, a novel evaluation approach named as FOSE is introduced. To the best of our knowledge, FOSE is the first data-driven evaluation approach for FO-based similarity measures.

4.1. Possible Evaluation Criterion for a FO-Based Similarity Measure

To evaluate a CO-based dissimilarity measure, a common approach is to compare it with web data. However, to evaluate the power of fuzziness in a model, it is common to compare it against its equivalent crisp model. Accordingly, in our evaluation in addition to consideration of web-based distance of data, web_dist, we consider its equivalence CO-based distance measure, CO_dist.

CO_dist: To have an acceptable evaluation, the underlying logic for calculation of CO-based measure, must be similar to proposed FO-based similarity measure. Rada proposed an approach for CO-based similarity measure which utilizes the concepts of shortest path between concepts and consideration of their LCA in the same way as our proposed FO-based similarity measure. Accordingly, in our evaluation we consider Rada measure for calculation of CO-based distance measure which is defined as follows:

\[
CO_{dist}(c_i, c_j) = |path(c_i, LCA(c_i, c_j))| + |path(LCA(c_i, c_j), c_j)|
\]  

(17)

web_dist: In order to evaluate distance based on web data, terms (concepts) co-occurrences in web files is considered, which is a common literature approach. Co-occurrence refers to the number of times two specific concepts have been appeared concurrently in a same file. This criterion is used to evaluate web-based distance of concepts.

FO_dist: Proposed FO-based distance measure (Equation 16) is used to calculate concepts distances based on FO.

Accordingly, there are three distance matrix of concept pairs one is web-based distance matrix, web_dist, one is distance matrix based on CO, CO_dist, and one is...
the distance matrix based on FO, FO_dist. The three distance matrix are put aside as follows to compare the results of proposed FO-based similarity measure.

1. Evaluation of the correlation between web_dist and CO_dist, and compare it with the correlation between web_dist and FO_dist.
2. Evaluation of the RAE between web_dist and CO_dist, and compare it with the RAE between web_dist and FO_dist.
3. Evaluation of the RRSE between web_dist and CO_dist, and compare it with the RRSE between web_dist and FO_dist.

The correlation, RAE, and RRSE are calculated as in Equations 9, 10 and 11, respectively.

4.2. The Proposed Approach for FOSE

According to previous section, evaluation of a FO-based similarity measure requires to compare it against its corresponding CO-based similarity measure. The criterion for comparison could be their RAE (Equation 11) to a standard web-derived distance matrix which both are built upon on. This evaluation approach clarifies how well FO has measured similarity in comparison with CO.

Considering the proposed FO-based similarity measure introduced in Section 3 and its respective CO-based similarity measure introduced in this section, a new evaluation criterion for FOSE method is proposed noted as FOSE. The criterion is ratio based. RAE of FO-based similarity matrix and web_dist matrix is set as its numerator and RAE of CO-based similarity matrix and web_dist matrix is set as its denominator. Accordingly, FOFOSE criterion is defined as follows:

\[
\text{FOSE}(\text{web\_dist}, \text{FO\_dist}, \text{CO\_dist}) = \frac{\sum_{f \in \text{FO\_dist}, w \in \text{web\_dist}} \text{RAE}(f, w)}{\sum_{c \in \text{CO\_dist}, w \in \text{web\_dist}} \text{RAE}(c, w)}
\]

where web_dist matrix is the web-based distance matrix as defined in Section 4.2. FO_dist is the distance based on FO using proposed method (Equation 16), CO_dist is the distance based on CO using Equation 17 and RAE(A,B) equals the RAE of point to point elements of A and B which is calculated as in Equation 11. FOSE considers three distance matrices to evaluate a FO-based distance measure: the distance matrix generated from web-data, (web_dist), the distance matrix based on CO, (CO_dist), and the distance matrix based on FO, (FO_dist). Comparing the RAE between web_dist and CO_dist with that of web_dist and FO_dist, FOSE evaluates a FO-based similarity measure. Thus, the FOSE value less than 1 indicates the superiority of FO over CO. The range of FOSE possible values and their meaning is demonstrated in Table I.

5. CASE STUDY

In this case study, we focus on the effectiveness of the FO-based semantic similarity measure in the context of SN. In our context, we focus on LinkedIn,
Table I. Different Values of Proposed FO-Based Similarity Evaluation (FOSE) Criterion and their Meaning.

| FOSE value | Interpretation of FOSE value                              |
|------------|-----------------------------------------------------------|
| FOSE < 1   | FO_dist is advantageous over CO_dist.                    |
| FOSE = 1   | FO_dist and CO_dist are of comparable ability in measuring similarity. |
| FOSE > 1   | CO_dist is advantageous over FO_dist.                    |

which is a notable example of business-oriented social networking site, founded in December 2002 and launched in May 2003. It is essentially used for professional networking. In June 2012, LinkedIn reports more than 175 million registered users in more than 200 countries and territories. Indeed, 80% of all jobs in United States filled through LinkedIn. Members can create customizable profiles that detail employment history, business accomplishments, and professional competencies in their area of expertise. Consequently, they may develop contacts, find jobs, and answer questions.

5.1. FO Development

In LinkedIn SN, each person fills out his/her own profile with a set of skills that defines his/her areas of expertise. A screen shot of a LinkedIn's profiles skills section is depicted in Figure 3. Subsequently, by crawling profile data of 130 unique users, various skills were collected from their LinkedIn profiles.

In order to create an ontology of skills for the SN, individual skills must be clustered as the groundwork for ontology construction. A common metric for web data clustering, is terms co-occurrences. Co-occurrence refers to the number of times two individual terms have been used within the same text file. Mapping this definition to our context is equivalent to the number of times two terms have been used concurrently as an individual skill. Calculation of this metric resulted in 28,536 couples of co-occurred terms. By the accomplishment of this stage, dataset specifications were extracted as outlined in Table II and the web-based distance matrix is constructed based on concept co-occurrences.

To create a CO of terms, agglomerative hierarchical clustering (AHC) is applied on the skill set. AHC forms the clusters from “bottom-up” and is a common literature approach for learning of an ontology from a data set. Having generated the CO, the corresponding CO-based distance matrix is generated using Equation 17.

5.2. Assessing FO-Based Concept Similarity

Having a CO, FO is generated by the algorithm, which maps a two-dimensional distance matrix of concepts to hierarchical ontology structure. For the generated FO, distance of concept pairs is calculated by the proposed approach (Equation 16) and FO-based distance matrix is determined.
Figure 3. A screenshot of LinkedIn’s profiles skills section.

Table II. Dataset Specifications.

| Specification                          | Value               |
|----------------------------------------|---------------------|
| Number of profiles                     | 130 persons         |
| Number of ontology leaf node           | 1,098 skills        |
| Number of skills for an individual person | Between 6 and 51 skills |
| Number co-occurred couples of terms    | 28,536              |
5.3. Evaluation of FO-Based Similarity Measure

At this step, the FO-based distance measure, FO\_dist, is compared against the CO-based distance measure, CO\_dist, using correlation (Equation 9), RRSE (Equation 10) and RAE (Equation 11), and FOSE criterion. The overall evaluation results are summarized in Figures 4–7.

Table III outlines the four dataset’s specifications that were considered in this experiment. All datasets are generated from skillsets of LinkedIn profiles (Table II). As is demonstrated, datasets are sorted in ascending order based on their size.

Figure 4 demonstrates the correlation between FO\_dist and web-based distance matrix, web\_dist, with that of CO\_dist and web\_dist for all datasets. According to
**Figure 6.** RRSE of FO-based semantic similarity in comparison with respective CO-based semantic similarity.

**Figure 7.** Fuzzy ontology-based similarity evaluation (FOSE).

**Table III.** Datasets specifications.

| Dataset number | Number of user profiles | Number of unique concepts | Number of co-occurred concept couples |
|---------------|-------------------------|----------------------------|--------------------------------------|
| Dataset 1     | 7                       | 6                          | 11                                   |
| Dataset 2     | 14                      | 218                        | 19,527                               |
| Dataset 3     | 80                      | 701                        | 19,846                               |
| Dataset 4     | 214                     | 1,638                      | 50,123                               |

*International Journal of Intelligent Systems* DOI 10.1002/int
Figure 4, FO has a higher correlation with the initial web-based distance matrix compared with CO. Considering Table III, the larger datasets result in higher superiority of FO in comparison with CO for the conceptual modeling of the real world. This shows the potential of this approach for larger datasets as is necessary for SN analysis.

In Figures 5 and 6, the comparison is performed between FO-based similarity measure and its corresponding CO-based similarity measure by means of RAE and RRSE criteria. Figure 5 demonstrates the RAE between FO_dist and CO_dist with web_dist. As illustrated for all datasets, RAE of FO-based measure is lower than CO-based similarity, which underlines its dominance in semantic similarity assessment.

In the same way, Figure 6 illustrates the RRSE between each of FO_dist and CO_dist with web_dist. Results indicates FO-based measure superiority over CO-based measure especially when the size of datasets increases.

Eventually, the proposed evaluation criterion, FOSE, is calculated for the proposed FO-based similarity measure and summarized in Figure 7. Considering Table I, values less than 1 and close to 0 indicates the FO-based similarity measure superiority over CO-based similarity measure. Subsequently, the decreasing trend of FOSE values by increasing datasets size and the near to zero, (0.07), value for the largest dataset indicates proposed method perfection.

6. CONCLUSION AND FUTURE WORK

In this section, first the works of other authors, whose proposals are close to ours are reviewed and compared with our proposal. Then the paper is concluded and directions for future works are outlined.

6.1. Related Work

In this section, we first review the literature on CO-based semantic similarity measures and afterward consider the literature on a more narrowed topic of FO-based similarity measures.

Common approaches for CO-based similarity measures can be classified to edge-counting, information-based, and feature-based approaches.\textsuperscript{50,51}

Edge-counting approaches refer to the group of methods that utilize ontology structure to assess the similarity. Rada\textsuperscript{45} considered the length of the shortest path between ontology nodes. The longer the path, the more semantically far the concepts are considered. Later on, Refs. 52–54 integrated some other features of ontology structure like its depth to improve the accuracy of this similarity measure.

Features-based approaches consider the degree of overlapping between sets of ontological features. Concepts are considered as sets of features and their evaluation is estimated based on the number of similar features they have. In this way, common features tend to increase similarity and non-common ones tend to diminish it. This approach is originated from Tversky model\textsuperscript{55} and has been applied in molecular biology,\textsuperscript{56} adaptive e-learning\textsuperscript{57} and ontology merging\textsuperscript{58} and alignment.\textsuperscript{59}

Information-based approaches calculates statistical specification of concepts based on a corpus or other data sources.\textsuperscript{60} Some of common statistical specifications in the literature
are concept frequencies, which refers to the number of occurrences in the corpus and co-occurrences, which considers simultaneous occurrence of two concepts in the same file.

**FO-based similarity assessment** is mainly obtained by fuzzy FCA (FFCA). FFCA-based approaches lie in feature-based approaches since they consider common features of concepts in the lattice to assess their similarity. FFCA has been used in various domains. For instance, used it in combination with WordNet for similarity assessment and used an ontology obtained by FCA to assess similarity for products for information retrieval. A highlighted work of this domain is Ref. 61, which has used FFCA method in combination with information content theory to assess similarity. Despite its popularity, FCA-based approach is of high computational complexity. Its pre-requirement is to build formal concept lattices, which is a complex task of $O(2^N)$ time order.

The proposed measure assess similarity based on the shortest path between concepts which is an edge-counting approach in combination with fuzzy grades of membership which is an information-based approach. Accordingly, proposed measure lies in a hybrid category as illustrated in Table IV.

### 6.2. Summary and Future Work

Assessing concept semantic similarity is a critical module in many applications of artificial intelligence. This paper proposed a novel approach for concept similarity assessment which is based on FO. The proposed approach incorporates fuzzy relation composition in combination with an edge counting approach to assess the similarity. Accordingly, proposed measure relies on taxonomical features of an ontology in combination with its fuzzy grades of membership. Consequently, the approach lies in the hybrid category of assessing similarity that utilizes both taxonomical and information content metrics.

Differently to the literature on FO-based semantic similarity measures, the proposed approach does not utilize FCA. Considering the limited world modeling of FCA and its high computational complexity, the proposed approach has lower pre-required complexity and higher flexibility.

---

**Table IV.** Some representative proposals for ontology-based similarity assessment.

| Proposed approach | Edge-counting | Information-based | Feature-based | Fuzzy ontology | Crisp ontology |
|-------------------|---------------|-------------------|---------------|---------------|---------------|
| 45                | *             |                   |               |               |               |
| 64                | *             |                   |               |               |               |
| 65                |               | *                 | *             |               |               |
| 6                 |               |                   | *             |               |               |
| 44                |               |                   |               | *             |               |
| 50                |               |                   |               |               | *             |
| 51                |               |                   |               |               | *             |
| 61                |               | *                 |               |               |               |
| 56                |               |                   |               |               |               |
| Proposed approach | *             | *                 |               |               | *             |
Furthermore, an evaluation method for FO-based similarity measure, named as FOSE, was proposed, which determines the variation between proposed FO-based similarity measure with real world-data and compare it with that of CO-based similarity measure. As far as the present research is concerned, FOSE is the first data-driven evaluation approach for FO-based similarity measures.

To evaluate the proposed measure by means of FOSE, a case study of LinkedIn SN was considered. Experimental results reveal superiority of proposed FO-based similarity measure in comparison with CO-based measures with respect to its correlation to the real world data and error minimization. Our future work is concerned with assessing similarity of concepts in a context with higher degrees of uncertainty based on interval and general type-2 fuzzy ontologies.

References

1. Klein M. Combining and relating ontologies: an analysis of problems and solutions. Workshop Ontol Inf Sharing, IJCAI’01. Seattle, USA; 2001.
2. Opuszko M, Jena F. Using semantic web technologies for classification analysis in social networks. 15. Interuniv. Doktorandenseminar Wirtschaftsinformatik der Univ. Chemnitz, Dresden, Freiberg, Halle-Wittenberg, Jena und Leipzig; 2011. pp. 11–21.
3. Passin TB. Explorer’s guide to the semantic web. Greenwich, CT: Manning; 2004.
4. Zhao Y, Halang W. Rough concept lattice based ontology similarity measure. In: Proc 1st Int Conf Scalable Inf Syst; 2006. p 15.
5. Zhao Y, Wang X, Halang W. Ontology mapping based on rough formal concept analysis. In: Adv Int’l Conf Telecommun Int’l Conf Internet Web Applic Services (AICT-ICIW’06); 2006. p 180.
6. Formica A. Similarity reasoning for the semantic web based on fuzzy concept lattices: an informal approach. Inf Syst Front 2012;1–10.
7. Zadeh LA. The concept of a linguistic variable and its applications to approximate reasoning I. Inf Sci (NY) 1975;8:199–249.
8. Miyamoto S. Fuzzy sets in information retrieval and cluster analysis, Vol. 4. Springer Netherlands; 1990.
9. Crestani F, Pasi G. Soft computing in information retrieval: techniques and applications, Vol. 50. Physica-Verlag Heidelberg; 2000.
10. Bezdek J, Dubois D, Prade H. Fuzzy sets in approximate reasoning and information systems, Vol. 5. Springer US; 1999.
11. Lau RYK, Li Y, Xu Y. Mining fuzzy domain ontology from textual databases. In: IEEE/WIC/ACM Conf Web Intell; 2007. pp 156–162.
12. Han L, Chen G. A fuzzy clustering method of construction of ontology-based user profiles. Adv Eng Softw 2009;40(7):535–540.
13. Parry D. Fuzzification of a standard ontology to encourage reuse. IEEE Int Conf Inf Reuse Integr 2004 IRI 2004 Proc 2004; 2004. pp 582–587.
14. Lee C-S, Jian Z-W, Huang L-K. A fuzzy ontology and its application to news summarization. IEEE Trans Syst Man Cyber B 2005;35(5):859–880.
15. Lee C-S, Wang M-H, Hagras H. A type-2 fuzzy ontology and its application to personal diabetic-diet recommendation. IEEE Trans FUZZY Syst 2010;18(2).
16. Lee C-S, Wang M-H, Acamorpora G, Hsu C-Y, Hagras H. Diet assessment based on type-2 fuzzy ontology and fuzzy markup language. Int J Intel Syst 2010;25(12):1187–1216.
17. Bukhari AC, Kim Y. Integration of a secure type-2 fuzzy ontology with a multi-agent platform A proposal to automate the personalized flight ticket booking domain. Inf Sci (NY) 2012;198:24–27.
18. Calegari S, Sanchez E. A fuzzy ontology-approach to improve semantic information retrieval. URSW 2007;327.
19. Parry D. A fuzzy ontology for medical document retrieval. In: Second Workshop Austr Inf Security Data Min Web Intell Softw Int; 2004. pp 121–126.
20. De Maio C, Fenza G, Loia V, Senatore S. Hierarchical web resources retrieval by exploiting fuzzy formal concept analysis. Inf Process Manag 2012;48(3):399–418.
21. Remi S, Varghese SC. Domain ontology driven fuzzy semantic information retrieval. Procedia Comput Sci 2015;46(Icict 2014):676–681.
22. Gao H, Zhao J, Yin Q, Wang J. Ontology-based enterprise information retrieval model. In: IEEE Int Conf Grey Syst Intell Services. GSIS 2009; 2009. pp 1326–1330.
23. Balasubramaniam K. Hybrid fuzzy-ontology design using FCA based clustering for information retrieval in semantic web. Procedia Comput Sci 2015;50:135–142.
24. Hourali M, Montazer GA. An intelligent information retrieval approach based on two degrees of uncertainty fuzzy ontology. Adv Fuzzy Syst 2011;2011:7.
25. Zhai J, Shen L, Liang Y, Jiang JJ. Application of fuzzy ontology to information retrieval for electronic commerce. In: Int Symp Electron Commerce Sec Appl; 2008. pp 221–225.
26. Portmann E, Meier A, Cudré-Mauroux P, Pedrycz W. FORA—A fuzzy set based framework for online reputation management. Fuzzy Sets Syst 2015;269:90–114.
27. Portmann E. Fuzzy online reputation analysis framework. FORA Framew 2013;117–140.
28. Ali F, Kim EK, Kim Y-G. Type-2 fuzzy ontology-based semantic knowledge for collision avoidance of autonomous underwater vehicles. Inf Sci (NY) 2015;295:441–464.
29. Pakonen A, Tommila T, Hirvonen J. A fuzzy ontology based approach for mobilising industrial plant knowledge. In: IEEE Int Conf Emerg Technol Factory Automation; 2010.
30. Carlsson C, Mezei J, Brunelli M. Fuzzy ontology used for knowledge mobilization. Int J Intell Syst 2013;28(1):52–71.
31. Poelmans J, Elzinga P, Viane S, Dedene G. Formal concept analysis in knowledge discovery: a survey. no. 1999; 2009.
32. Cimiano P, Hotho A, Staab S. Learning concept hierarchies from text corpora using formal concept analysis. J Artif Intell Res(JAIR) 2005;24:305–339.
33. Kalhori MRN, Zarandi MHF, Torshizi AD. Utilizing fuzzy relations in interval type-2 possibilistic C-means clustering. In: IEEE-NAFIPS; 2014. pp 1–6.
34. Gan M, Dou X, Jiang R. From ontology to semantic similarity: calculation of ontology-based semantic similarity. Sci World J 2013;2013:793091.
35. Bahrinani V, Caragea D, Aljandal W, Hsu WH. Ontology engineering and feature construction for predicting friendship links in the live journal social network. In: 2nd SNA-KDD Workshop, Vol. 8; 2008.
36. Bobillo F, Straccia U. Fuzzy ontology representation using OWL 2. Int J Approx Reason 2011;52(7):1073–1094.
37. Zimmermann HJ. Fuzzy set theory-and its applications. Springer Netherlands; 2001.
38. Hliaoutakis A. Semantic similarity measures in MeSH ontology and their application to information retrieval on Medline, Master’s thesis; 2005.
39. Safaeipour H, Zarandi MHF, Bastani S. Crisp to fuzzy ontology conversion in the context of social networks: a new approach. In: NAFIPS; 2016.
40. Zadeh LA. Similarity relations and fuzzy orderings. Inf Sci (NY) 1971;3(2):177–200.
41. Harispe S, Sánchez D, Ranwez S, Janaqi S, Montmain J. A framework for unifying ontology-based semantic similarity measures: a study in the biomedical domain. J Biomed Inform 2014;48:38–53.
42. O’sullivan D, Smyth B, Wilson D. Understanding case based recommendation: a similarity knowledge perspective. Int J Artif Intell Tools 2005;14(01n02):215–232.
43. Euzenat J, Shvaiko P. Ontology matching, Springer-Verlag Berlin Heidelberg; 2007.
44. Akmal S, Shih LH, Batres R. Ontology-based similarity for product information retrieval. Comput Ind 2014;65(1):91–107.
45. Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 1989;19(1):17–30.
46. Ben Ahmed E, Nabli A, Gargouri F. Group extraction from professional social network using a new semi-supervised hierarchical clustering. Knowl Inf Syst 2013;1–19.

International Journal of Intelligent Systems DOI 10.1002/int
47. Alsaleh S. Recommending people in social network using data mining. PhD diss., Queensland University of Technology; 2013.
48. Portmann E, Meier A. A fuzzy grassroots ontology for improving weblog extraction. JDIM 2010;8(4):276–284.
49. Alsaleh S, Nayak R, Xu Y. Grouping People in Social Networks Using a Weighted Multi-Constraints Clustering Method. 2012.
50. Sánchez D, Batet M, Isern D, Valls A. Ontology-based semantic similarity: a new feature-based approach. Expert Syst Appl 2012;39(9):7718–7728.
51. Hossein Zadeh PD, Reformat MZ. Assessment of semantic similarity of concepts defined in ontology. Inf Sci (NY) 2013;250:21–39.
52. Wu Z, Palmer M. Verbs semantics and lexical selection. In: Proc 32nd Annu Meet Assoc Comput Ling; 1994. pp 133–138.
53. Li Y, Bandar ZA, McLean D. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng 2003;15(4):871–882.
54. Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. WordNet An Electron Lex database 1998;49(2):265–283.
55. Tversky A. Features of similarity. Psychol Rev 1977;84(4):327.
56. Zhang S-B, Tang Q-R. Protein–protein interaction inference based on semantic similarity of gene ontology terms. J Theor Biol 2016;401:30–37.
57. Saleena B, Srivatsa SK. Using concept similarity in cross ontology for adaptive e-Learning systems. J King Saud Univ Comput Inf Sci 2015;27(1):1–12.
58. Chowdhury NA, Dou D. Improving the accuracy of ontology alignment through ensemble fuzzy clustering. In: Meersman R, Dillon T, Herrero P, editors. Uncertainty in ontology alignment. Berlin Heidelberg: Springer-Verlag; 2011. pp 826–833.
59. de Souza KXS, Davis J. Aligning ontologies and evaluating concept similarities. In: On the Move to Meaningful Internet Systems, 2004. pp 1012–1029.
60. Cross V, Yu X, Hu X. Unifying ontological similarity measures: a theoretical and empirical investigation. Int J Approx Reason 2013;54(7):861–875.
61. Safaeipour H, Fazel Zarandi MH, Turksen IB. Developing type-2 fuzzy FCA for similarity reasoning in the semantic web. In: NAFIPS, 2013.
62. Chen R-C, Bau C-T, Yeh C-J. Merging domain ontologies based on the WordNet system and fuzzy formal concept analysis techniques. Appl Soft Comput 2011;11(2):1908–1923.
63. Formica A. Concept similarity in formal concept analysis an information content approach. Knowledge-Based Syst 2008;21(1):80–87.
64. Wu Z, Palmer M. Verbs semantics and lexical selection. In: Proc 32nd Annu Meet Assoc Comput Ling; 1994. pp 133–138.
65. Belohlavek R, Outrata J, Vychodil V, Dvorák J, Qutrata J, rf J, Rák D. Fast factorization by similarity in formal concept analysis of data with fuzzy attributes. Int J Found Comput Sci 2007;73(1):1012–1022.