Ontology-Based Probabilistic Estimation for Assessing Semantic Similarity of Land Use/Land Cover Classification Systems

Xiran Zhou 1,2, Xiao Xie 3,*, Yong Xue 1,2 and Bing Xue 3

Abstract: To accurately and formally represent the historical trajectory and present the current situation of land use/land cover (LULC), numerous types of classification standards for LULC have been developed by different nations, institutes, organizations, etc.; however, these land cover classification systems and legends generate polysemy and ambiguity in integration and sharing. The approaches for dealing with semantic heterogeneity have been developed in terms of semantic similarity. Generally speaking, these approaches lack domain ontologies, which might be a significant barrier to implementing these approaches in terms of semantic similarity assessment. In this paper, we propose an ontological approach to assess the similarity of the domain of LULC classification systems and standards. We develop domain ontologies to explicitly define the descriptions and codes of different LULC classification systems and standards as semantic information, and formally organize this semantic information as rules for logical reasoning. Then, we utilize a Bayes algorithm to create a conditional probabilistic model for computing the semantic similarity of terms in two separate LULC land cover classification systems. The experiment shows that semantic similarity can be effectively measured by integrating a probabilistic model based on the content of ontology.

Keywords: ontology; semantic similarity; land use/land cover classification systems

1. Introduction

Mapping land cover (LULC) provides important support for representing the historical trajectory and present situation of earth observation [1,2], land management [3], pattern analysis [4], settlement monitoring [5], landscape planning [6], etc. These LULC classification maps are available at multiple spatial and temporal scales generated by numerous types of classification standards for LULC. Currently, tens of LULC classification systems have been developed by different nations, institutes, and organizations, such as the NLCD1992 and the NLCD2006 developed by USGS (U.S. Geological Survey), the C-CAP developed by NOAA (National Oceanic and Atmospheric Administration), the Land Cover classification systems, legends developed by the UN (United Nations), and Chinese Current Land Use Classification.

These land cover classification systems and legends generate two significant challenges in integration and sharing: (1) polysemy: a land parcel might be defined as different LULC types by various LULC classification systems; (2) ambiguity: the same term of LULC might be defined differently according to various LULC classification systems. Polysemy and ambiguity belong to semantic heterogeneity [7], which focuses on addressing the confusion of expression in natural language processing. Li and Ling divided the semantic heterogeneity in terms of LULC classification systems and standards into three major factors [8]. (1) Confounding conflicts: the same definition or concept represents
diverse meanings. For example, the notion “commercial/industrial” belongs to the category “Commercial/Industrial” “Transportation” in NLCD1992 but belongs to the category “Developed High Intensity” in NLCD2006. (2) Scaling and unit conflicts: the same definition is represented at different scales and units. For example, the term “Low Density” in NLCD1992 and NLCD2006 is defined differently. (3) Naming conflicts: one word has multiple meanings, or one meaning can be expressed by using multiple words. For example, the “perennial” of NLCD 2006 and “long-term” of NLCD 1992 represent the same meaning.

To address the semantic heterogeneities, a number of works have proposed approaches regarding semantic harmonization to integrate multi-source information and features into a consistent one. Since the psychological study shows that similar features can attract more attention than different ones [9], semantic harmonization mainly focuses on semantic similarity to deal with semantic heterogeneity. Some previous works have used metadata to define the characteristics of the relationship of LULC types; however, the work proposed by Comber, Fisher, and Wadsworth [10] claimed that the metadata could not explicitly describe the meaning of LULC information. To deal with this challenge, a number of semantic harmonization regarding LULC focuses on statistical learning-based semantic similarity assessment, such as conceptual spaces [11], semantic metrics [12], integrating post-classification and semantic metrics [13], regression integrated correlation matrix [14], etc. Moreover, the user–machine interactive approach [15] and expert-enhanced system [16] have been developed to facilitate understanding the semantics for assessing semantic similarity.

Assessing the semantic similarity of various LULC terms requires the consideration of the explicit meanings of domain knowledge and the hidden expressions/relationships between a term and its neighboring terms. Thus, the domain ontologies could be a significant barrier to the implementation of those approaches in terms of semantic similarity assessment. For example, although the statistical model performs well on measuring the similarity of “high intensity” between high-intensity residential (NLCD 2006) and high-intensity developed (NLCD 1992), it cannot measure the relevance between developed and residential. The ontology can semantically define and formally represent the domain knowledge based on a hierarchical taxonomy, including classes, instances, attributes, and relationships. For semantic similarity measuring, previous works claimed that an ontology could systematically organize the domain knowledge and explicitly discover the relevance and correlations among domain individuals [17,18].

Until now, the state-of-the-art ontology-based semantic similarity assessment for language recognition and knowledge modeling consists of edge-based similarity measuring, feature-based similarity measuring, information content-based similarity measuring, and gloss-based similarity measuring [17,19–21]. Edge-based approaches are simple and easy to compute, but they cannot satisfy the demand for precision and accuracy of semantic similarity measures. Moreover, although the IC-based approaches successfully handle many applications regarding semantic similarity measures, informativeness or content are difficult to obtain from the limited volume information of LULC classification systems and standards. When the features are inadequate, feature-based approaches cannot accurately distinguish the small difference. The implementation of the gloss-based similarity method requires massive text information stored in a word base such as WordNet and Wiktionary; however, to our knowledge, the word base is still unreported in terms of LULC land cover classification and mapping. Thus, gloss-based similarity measuring might not be appropriate for measuring the semantic similarity of LULC classification systems and standards.

To accurately assess the semantics similarity of LULC classification systems with a limited amount of text information, we propose an ontology-enhanced probabilistic approach to enhance the semantic similarity measuring regarding the domain of LULC classification systems and standards. The remainder of this paper is organized as follows: Section 2 discusses the works relevant to ontology-based semantic similarity assessment; Section 3 presents our proposed methods for measuring semantic similarity, which includes an on-
tology named LuLcSys-Ontology for a formal representation of LULC, and a probabilistic model for semantic similarity based on LuLcSys-Ontology; Section 4 shows our semantic similarity assessment by using other approaches and our proposed one; Section 5 concludes our work, details our contributions to the literature, and predicts several prospective relevant research fields.

2. Related Works

2.1. Edge-Based Similarity Measuring

Edge-based similarity measuring aims to calculate the links or depth between the terms in a conceptual hierarchy. The approach to compute the link and depth of a path is shown as follows:

\[
\begin{align*}
\text{link} & = \min \{ \text{len}(\text{path}(a,b)) \} \\
\text{depth}(a) & = \min \{ \text{len}(\text{path}(a,r)) \}
\end{align*}
\]

(1)

where \( \text{path}(a,b) \) are the set that includes all paths between two separate terms \( a \) and \( b \), \( \text{len}(\text{path}(a,b)) \) is the set that includes the length of each path between \( a \) and \( b \). \( r \) is the root of a hierarchical taxonomy that includes both \( a \) and \( b \).

Other extensive works on edge-based similarity measuring include the approaches proposed by Li, Bandar, and McLean [22] and Al-Mubaid and Nguyen [23]. The edge-based similarity measure is straightforward and requires low-cost computing; however, it might be ineffective to deal with the semantic similarity assessment for a hierarchical taxonomy with a complex structure. Additionally, the path and depth of a term vary according to different ontologies, which means that this term might be measured as different. Finally, it cannot represent the hidden information in ontologies.

2.2. Information Content (IC)-Based Similarity Measuring

The IC-based similarity measuring assesses the semantic similarity based on the informativeness of the concept [24]. Assuming a concept as \( a \), \( p(a) \) is the probability of observing this concept, the informativeness of this concept (IC\((a)\)) is shown as follows:

\[
\text{IC}(a) = -\log(p(a))
\]

(2)

Resnik [24] and the following methods designed an approach to measure the semantic similarity between two concepts based on the informativeness, which is shown as follows:

\[
\text{sim}(a,b) = \arg\max_c (\text{IC}(c)) = \text{IC}(\text{COM}(c)), c \in \text{Sub}(a,b)
\]

(3)

where \( a \) and \( b \) are two independent concepts, \( \text{Sub}(a,b) \) denotes the set of all concepts that contains concepts \( a \) and \( b \). Depending on Equation (3), the subsequent studies on IC-based similarity measures include two focuses [19]: Corpora-based IC computation method and intrinsic IC computation method.

The corpora-based IC computation method computes the content of IC by using external information. Otherwise, the intrinsic IC computation method focuses on utilizing the knowledge included in ontology is more popular. Related applications include measuring IC from a conceptual hierarchy with optimized depth calculation [25], measuring IC from a conceptual hierarchy without depth calculation [26], and measuring IC from a conceptual hierarchy via a setting weights mechanism [27,28].

In general, an IC-based similarity measure relies on massive well-prepared data to discover the heterogeneous meanings of each term. In comparison to the volume of training data from semantic bases such as WordNet, the number of terms in the state-of-the-art LULC classification systems and standards is inadequate for generating an accurate measuring result. Moreover, although intrinsic IC computation methods can derive knowledge from ontology without the support of massive external information, the hierarchical taxonomy in an ontology might be very complex for this method.
2.3. Feature-Based Similarity Measuring

Feature-based similarity measuring focuses on the similarity between the properties of two concepts, which is based on the set theory proposed by Tversky [29]:

\[ im(a, b) = \frac{|d(a) \cap d(b)|}{|d(a) \cup d(b)| + \mu \times \frac{|d(a)/d(b)|}{|d(b)/d(a)|} + (1 - \mu) \times \frac{|d(b)/d(a)|}{|d(a)/d(b)|}} \]  

where \( d(a) \) and \( d(b) \) are the descriptions for concept \( a \) and \( b \), respectively, \( \mu \) is the weight, \( d(a)/d(b) \) denotes that the descriptions belong to \( a \) but not \( b \), and \( d(b)/d(a) \) denotes that the descriptions belong to \( b \) but not \( a \).

Since the hierarchical taxonomies in an ontology have been becoming more and more complex, the investigation on semantic similarity has concentrated on the similarity of features rather than terms [30]. Rodriguez and Egenhofer [31] proposed a feature-based semantic similarity with regard to the relationships between terms.

\[ sim(a, b) = \mu_1 \times sim_s(A, B) + \mu_f \times sim_f(A, B) + \mu_n \times sim_n(A, B) \]  

where \( A \) and \( B \) are the corresponding set of terms \( a \) and \( b \), respectively. \( sim_s() \), \( sim_f() \), and \( sim_n() \) are the synsets, features, and neighbor concepts, and \( \mu_s \), \( \mu_f \), and \( \mu_n \) are the weights for these three concepts, respectively. More details of computing \( sim_s() \), \( sim_f() \), and \( sim_n() \) can be found in Reference [31]. Other feature-based similarity measures include X-similarity [32], integrating information-theoretical domain [33], using taxonomical features [34], measuring similarity without pre-defined ontology [35], matching concepts from diverse ontologies [36], etc.

Appropriate weighting refers to the most significant limitation of the feature-based similarity measure. In general, a feature-based similarity measure might assign an appropriate weight for each feature by a trial-and-error procedure. Moreover, a feature-based similarity measure assigns a weight for each independent term; however, the terms in various LULC classification systems and standards might have overlapped features, making it difficult to determine an appropriate weight.

3. Methodology

3.1. Formal Representation of LULC

3.1.1. LuLcSys-Ontology

Based on Protege software [37], we developed a domain ontology named LuLcSys-Ontology to semantically define and formally organize the information extracted from LULC classification systems and standards. Figure 1 illustrates the conceptual model of LuLcSys-Ontology, which includes five components: Classes, Instances, Properties, and Restrictions. Instances includes the individuals that belong to a class item defined in Classes. The items in Properties refer to relationships, and the items in Restrictions refer to the precondition and context of relationships. More details are provided as follows.

- **Category** and **Code** are two subclasses of **Classes**: the instances in these two subclasses are from the names and legends of LULC categories;
- **Features** is the third subclasses of **Classes**: the instances in **Features** are from the descriptions of each LULC category;
- **Annotation properties, Data properties, and Object properties** are three subclasses of **Properties**. **Annotation properties** defines the meta-information of ontologies. **Data properties** defines the relationship between two objects. **Data properties** defines the relationship between an object and the range or value of its feature. All properties are predefined by OGW standards and W3C Semantic Web Standard;
- The items in **Restriction** are predefined by OGW standards and W3C Semantic Web Standard.
Figure 1. Conceptual model of LuLcSys-Ontology.

The details of LuLcSys-Ontology are shown in Table 1.

Table 1. Details on components of LuLcSys-Ontology.

| Component | Triple Relationship | Content |
|-----------|---------------------|---------|
| Classes   | “subject” or “object” in the triple | Three subclasses: Categories: the categories or classes of LULC classification systems and standards. Codes: the codes corresponding to categories or classes. Features: the characteristics of each category or class. |
| Instances | “subject” or “object” in the triple | The terms or notations derived from the textural descriptions of LULC classification systems and standards. |
| Properties| “predicate” in the triple | Three types of properties: the details are shown in Table 2. Restrictions define the validity of a property under specific conditions. |
| Restrictions| “predicate” in the triple | The characteristics of properties. |
| Function terms| “predicate” in the triple | The characteristics of properties. |

Moreover, each item in Instances should belong to at least one class in Classes. In LuLcSys-Ontology, properties are defined by the W3C Standards, including RDFS (Resource Description Framework Schema) and OWL (Ontology Web Language), and predefined by LuLcSys-Ontology. Since the Annotation property mainly represents the meta-information of ontology, we focus on data property-based triple (subject–data property–object), and the object property-based triple (subject–object property–object). In some cases, the data property-based triple might be incorporated into an object property-based triple.

Table 2 lists the details of Properties, Restrictions, and Function terms. The properties that start with lulcsys:, rdf:, and owl: show that this property is defined by LuLcSys-Ontology, RDFS, and OWL, respectively. The items of Restrictions and Function terms are defined by the W3C Semantic Web Standard.
Table 2. Properties, Restrictions, and Function terms in LuLcSys-Ontology.

| Property Types          | Properties                                                                 |
|-------------------------|---------------------------------------------------------------------------|
| Annotation properties   | rdfs: seeAlso, rdfs:LisdefinedBy, owl:priorVersion,                        |
|                         | owl:versionInfo, owl:deprecatedclass,                                     |
|                         | owl:deprecatedproperty                                                     |
| Data properties         | lulcsys:hasCode, lulcsys:hasMaxCover, lulcsys:hasMinCover, lulcsys:Area,  |
|                         | lulcsys:hasPerimeter, lulcsys:hasTall, Height                             |
|                         | lulcsys:hasFeatures, lulcsys:isDominatedBy, lulcsys:isUsedFor,            |
|                         | lulcsys:isUsedBy, lulcsys:isPlantedFor, lulcsys:isPlantedBy,              |
|                         | lulcsys:isSaturated/CoveredBy, lulcsys:isOnlyIn,                          |
|                         | lulcsys:isRemoved/ModifiedBy, lulcsys:isReplacedBy,                       |
|                         | lulcsys:isAffectedBy, lulcsys:isResultFrom,                               |
|                         | LuLcSys:hasDuration, rdf:type, rdfs:subClassOf,                           |
|                         | rdfs:subPropertyOf, rdf:member, owl:equivalentClass,                     |
|                         | owl:equivalentProperty, owl:sameAs, owl:differentFrom,                    |
|                         | owl:AllDifferent, owl:distinctMembers                                      |
| Object properties       |                                                                           |
|                         |                                                                           |
|                         |                                                                           |

| Restriction types       | Restrictions                                                               |
|-------------------------|---------------------------------------------------------------------------|
| Data/Object restrictions| Some (existential), Only (universal), Min (min cardinality),               |
|                         | Max (max cardinality), Exactly (exact cardinality)                         |
|                         | allValueFrom, someValueFrom                                                |
| Property restrictions   | inverseOf, TransitiveProperty, SymmetricProperty, FunctionalProperty,      |
|                         | InverseFunctionalProperty                                                  |
|                         |                                                                           |

Function terms

Functional, Inverse functional, Transitive, Symmetric, Asymmetric, Reflexive, Irreflexive

Based on the W3C Semantic Web Standard [38,39], all relationships in LuLcSys-Ontology were created as a triple relationship: “subject–predicate–object”. Taking three categories of NCLD 2006 (Deciduous Forest, Evergreen Forest, and Mixed Forest) as the example, Figure 2 shows the transformation from the descriptions of these three categories into the semantic information of LuLcSys-Ontology.

![Figure 2](image)

Figure 2. An example of the transformation from category descriptions to semantic information: (A) textural descriptions of three categories in NCLD 2006; (B) the corresponding semantic information modeled by triple relationships in LuLcSys-Ontology.
Figure 2A shows the descriptions of three categories involving Deciduous Forest, Evergreen Forest, and Mixed Forest. Figure 2B shows the semantics explicitly defined in LuLcSys-Ontology. We label various components in different colors. The orange texts refer to Classes, the italic black texts are Properties, the red texts are Property restrictions, the green texts are instances that are defined by Object properties, and the blue texts are the instances that are defined by Data properties. Based on these components, all descriptions are organized as triple relationships—as shown in Figure 2B.

Moreover, Figure 3 shows the partial structure of the LuLcSys-Ontology developed for NLCD 1992, including three classes of NLCD_1992: Categories, Codes, and Features. The yellow rectangles refer to the subclasses of these three classes, and the purple rectangles refer to the instances. All properties are represented by arrow lines. When an arrow line connects two rectangles, the rectangle that connects to the starting point of the arrow line refers to the “object” in the triple relationship, and the other rectangle refers to the “subject” in the triple relationship.

3.1.2. Rules Building

In comparison to a spatial database, the key advantage of an ontology is the capability of discovering hidden knowledge through rule-based reasoning supported by triple relationships. In this paper, we built reasoning rules by SWRL (semantic web rules language) [39], which is defined by the W3C Semantic Web Standard. Assume the triple relationship (subject–predicate–object) in ontology as $P(Sub, Obj)$, where $Sub$, $P()$ and $Obj$ denotes subject, property, and object, respectively. Additionally, $Sub_{new}$, $P_{new}()$, and $Obj_{new}$ respectively, denotes the new subject, property, and object after reasoning based on $P(Sub, Obj)$. The basic structure of SWRL in this paper is as follows:

$$P(Sub, Obj) \Rightarrow P_{new}(Sub_{new}, Obj_{new})$$

(6)

Then, based on the data properties and object properties, we develop two types of rules: the rule of data property-based triple, and the rule of object property-based triple.
Assuming object property and data property as $oP()$ and $dP()$, based on Equation (6), we have the rule based on object property-based triples and data property-based triples as follows:

$$
[oP_i(\text{?}_i, \text{?}_{11}) \cap dP_i(\text{?}_i, \text{?}_{12})] \land \ldots \land [oP_i(\text{?}_i, \text{?}_{11}) \cap dP_i(\text{?}_i, \text{?}_{12})] = > oP_{new}(\text{?}_{new}, \text{?}_{new})
$$

(7)

where $i$ is the total number of object property-based triples. $oP_{new}(\text{?}_{new}, \text{?}_{new})$ denotes a new object property-based triple. According to Equation (6), this new triple is also the result of logical reasoning.

We present an example of the reasoning based on Deciduous Forest in Figure 2 and in Table 3. Assuming we have a tree called “target_tree”; then, we have two data property-based triples and two object-property-based triples:

**Table 3. Examples of semantic modeling for LULC classification systems and standards.**

| Parameters | Meaning | Parameters | Meaning |
|------------|---------|------------|---------|
| Constructed materials account for 80 to100 percent of the cover. | Impervious surfaces account for 80% to 100% of the total cover. |
| Object property ($oP_i()$) | hasCover | Object property ($oP_i()$) | hasCover |
| Data ($D_i$) | 80%-100% | Data ($D_i$) | 80%-100% |
| Data property ($dP_i()$) | noLessThan | Data property ($dP_i()$) | noLessThan |

**Data property-based triple 1:** Trees lulcsys:hasHeight (owl:minCardinality) 5.

**Data property-based triple 2:** Trees lulcsys:hasShedsPercentage (owl:minCardinality) 75%.

**Object property-based triple 1:** Deciduous Forest lulcsys:isDominatedBy Trees.

**Object property-based triple 1:** target_tree rdf:isInstanceOf Deciduous Forest.

Based on these three triples, we can deduce a hidden relationship being unsupported by a spatial database:

target_tree rdf:isInstanceOf Deciduous Forest.

### 3.2. Probabilistic Reasoning Embedded Ontology-Based Semantic Similarity Measuring

As mentioned previously, feature-based measuring is limited to accurately weighting each feature without massive training samples. Thus, semantically modeling the features, rather than quantitatively weighting, would be an alternative solution. We integrate the probabilistic model (Bayes) and the feature-based measuring method to assess semantic similarity. Based on the object property-based triples and data property-based triples in LuLcSys-Ontology, we create the Bayes-based conditional probabilities to assess the semantic similarity.

For separate terms (subjects) $S_1$ and $S_2$ in two LULC classification systems and standards, we assume that the object property-based triple and data property-based triple of $S_1$ are $P(S_1, O_1)$ and $P(S_1, D_1)$, respectively. Similarly, for $S_2$, we assume its object
property-based triple and data property-based triple as \( P(S_2, O_2) \) and \( P(S_2, D_2) \). Moreover, the common features of objects and data between \( S_1 \) and \( S_2 \) are \( O_c \) and \( D_c \), \( O_c \subseteq O_1 \cap O_2 \) and \( D_c \subseteq D_1 \cap D_2 \). The semantic similarity of \( S_1 \) and \( S_2 \) \((\text{sim}(S_1, S_2))\) is measured by the following expression:

\[
\text{sim}(S_1, S_2) = \Pr(S_1, S_2) = \Pr(S_c)
\]

(8)

In Equation (8), we transform the semantic similarity of \( S_1 \) and \( S_2 \) to the probability of observing that they are similar, which is denoted as \( \Pr(S_1, S_2) \). The similarity is measured based on their common features of object \( (O_c) \) and common features of data \( (D_c) \), which is represented by \( \Pr(S_c) \). \( \Pr(S_c) \) is obtained by the following expression.

\[
\Pr(S_c) = \begin{cases} 
\exists o \Pr(S_c, O_c) : \Pr(S_c | O_c) = \frac{\Pr(S_c, O_c)}{\Pr(O_c)} \\
\exists o \Pr(S_c, O_c) \land d \Pr(S_c, D_c) : \Pr(S_c | (D_c | O_c)) = \frac{\Pr(S_c, (D_c | O_c))}{\Pr(D_c | O_c)}
\end{cases}
\]

(9)

In Equation (9), \( \Pr(S_c | O_c) \) refers to the probability of observing \( S_1 \) and \( S_2 \) are similar based on \( O_c \). \( \Pr(S_c | (D_c | O_c)) \) is the probability of observing \( S_1 \) and \( S_2 \) are similar based on \( D_c \). The following table shows an example that explains the parameters in Equation (9).

4. Experiments

The datasets for the experiment include three major regional LULC classification systems and standards: NLCD1992 and NLCD 2011/2006/2011 from USGS, and NOAA Regional Land Cover Classification Scheme from NOAA. The first experiment assesses the semantic similarity between NLCD 1992 and NLCD 2011/2006/2011. Considering that the difference between NLCD 2011/2006/2011 and NOAA Regional Land Cover Classification Scheme has attracted much attention, the second experiment focuses on assessing the semantic similarity of these two land cover classification systems and legends. The classes of these land cover classification systems and legends are listed in Table 4.

According to the categories and descriptions of NLCD 1992, NLCD 2011/2006/2011, and NOAA Regional Land Cover Classification Scheme, we develop three separate LuLcSys-Ontologies for these land cover classification systems and legends: NLCD92_Ontology for NLCD 1992, NLCD11_Ontology for NLCD 2011/2006/2011, and NOAA_Ontology for NOAA Regional Land Cover Classification Scheme. Then, we compute the semantic similarity based on the triples of each two ontologies: NLCD92_Ontology and NLCD11_Ontology, and NLCD11_Ontology and NOAA_Ontology. The computing method includes three existing ontology-based approaches: edge-based measures [23], feature-based measures [26], information content-based measures [25], and our proposed approach.

Table 5 shows the result of the semantic similarity assessment between NLCD 1992 and NLCD 2011/2006/2011. By comparing the textural descriptions of these two LULC classification systems and standards, both polysemy and ambiguity can be observed. In other words, no two classes are exactly the same, although they are defined as the same term. Based on the path and depth of each two terms in ontologies, PDBM cannot effectively assess the semantic similarities between most of the classes in NLCD 1992 and NLCD 2001/2006/2011. Meanwhile, we can observe that information content-based measures (ICBM) cannot assess the semantic similarities of some classes in these two LULC classification systems and standards. When there exists a limited volume of common features between two classes, the informativeness of their seminaries is challenging to assess; however, ICBM performs well on distinguishing some small differences between the two classes. For example, although the four classes of NLCD 1992 involving Row Crops, Small Grains, Fallow, and Orchards/Vineyards/Other are similar to the class of NLCD 2001/2006/2011 named Cultivated Crops, the similarities between each of these four classes of NLCD 1992 and Cultivated Crops are different. ICBM can produce more accurate results than feature-based measures (FBM) in measuring this semantic similarity. Moreover, many results by FBM are closer to the results of our proposed approach; however, FBM struggles to assess the small differences between two classes. For example, the
semantic similarity of Grasslands/Herbaceous and Sedge/Herbaceous is not the same as the semantic similarity of Grasslands/Herbaceous and Lichens and Moss, because Lichens and Moss are specifically defined for the landscape of Alaska; however, FBM produces the same similarity result. Thus, without the support of a conditional probabilistic model, ICBM and FBM are limited in measuring the semantic similarity of LULC classification systems and standards based on ontology.

Table 4. Illustration on LULC classification systems and standards used for experiment.

| NLCD 1992 Class                        | NLCD 2001/2006/2011 Class                                | NOAA Regional Land Cover Classification Scheme          |
|----------------------------------------|----------------------------------------------------------|--------------------------------------------------------|
| Open Water                             | Open Water                                               | Developed, High Intensity                              |
| Perennial Ice/Snow                     | Perennial Ice/Snow                                       | Developed, High Intensity                              |
| Low Intensity Residential              | Developed, Open Space                                    | Developed, Medium Intensity                            |
| High Intensity Residential             | Developed, Low Intensity                                 | Developed, Low Intensity                              |
| Commercial/Industrial/Transportation    | Developed, Medium Intensity                              | Developed, Open Space                                  |
| Bare Rock/Sand/Clay                    | Developed High Intensity                                  | Cultivated Crops                                       |
| Quarries/Strip Mines/Gravel Pits       | Barren Land (Rock/Sand/Clay)                             | Pasture/Hay                                            |
| Transitional                           | Deciduous Forest                                         | Grassland/Herbaceous                                   |
| Deciduous Forest                       | Evergreen Forest                                         | Deciduous Forest                                       |
| Evergreen Forest                       | Mixed Forest                                             | Evergreen Forest                                       |
| Mixed Forest                           | Dwarf Scrub                                               | Mixed Forest                                           |
| Shrubland                              | Shrub/Scrub                                              | Scrub/Shrub                                            |
| Orchards/Vineyards/Other               | Grassland/Herbaceous                                     | Barren Land                                            |
| Grasslands/Herbaceous                  | Sedge/Herbaceous                                         | Tundra                                                 |
| Pasture/Hay                            | Lichens                                                  | Perennial Ice/Snow                                     |
| Row Crops                              | Moss                                                     | Palustrine Forested Wetland                            |
| Small Grains                           | Pasture/Hay                                              | Palustrine Scrub/Shrub Wetland                         |
| Shrubland                              | Shrub/Scrub                                              | Palustrine Emergent Wetland (Persistent)               |
| Orchards/Vineyards/Other               | Grassland/Herbaceous                                     | Barren Land                                            |
| Grasslands/Herbaceous                  | Sedge/Herbaceous                                         | Tundra                                                 |
| Pasture/Hay                            | Lichens                                                  | Perennial Ice/Snow                                     |
| Row Crops                              | Moss                                                     | Palustrine Forested Wetland                            |
| Small Grains                           | Pasture/Hay                                              | Palustrine Scrub/Wetland                                |

Table 5. Results of semantic similarity measuring between NLCD 1992 and NLCD 2011/2006/2011.

| NLCD 1992 Class                        | NLCD 2001/2006/2011 Class                                | PDBM *    | FBM **   | ICBM *** | Our **** |
|----------------------------------------|----------------------------------------------------------|-----------|-----------|-----------|-----------|
| Open Water                             | Open water                                               | 1         | 0.67      | 0.29      | 0.75      |
| Perennial Ice/Snow                     | Perennial Ice/Snow                                       | 1         | 0.5       | 0.69      | 0.75      |
| Low-Intensity Residential              | Developed, Low Intensity                                 | 1         | 0.48      | 0.59      | 0.68      |
| Low-Intensity Residential              | Developed, Medium Intensity                              | 1         | 0.52      | 0.59      | 0.70      |
| High-Intensity Residential             | Developed High Intensity                                 | 1         | 0.75      | 0.81      | 0.82      |
| Commercial/Industrial/Transportation    | Developed High Intensity                                 | 1         | 0.4       | 0.52      | 0.67      |
| Bare Rock/Sand/Clay                    | Barren Land (Rock/Sand/Clay)                             | 1         | 0.5       | 1         | 0.56      |
| Quarries/Strip Mines/Gravel Pits       | Developed, Low Intensity                                 | 0         | 0.25      | 0         | 0.25      |
| Quarries/Strip Mines/Gravel Pits       | Developed, Medium Intensity                              | 0         | 0.25      | 0         | 0.25      |
| Quarries/Strip Mines/Gravel Pits       | Developed High Intensity                                 | 0         | 0.25      | 0         | 0.25      |
| Quarries/Strip Mines/Gravel Pits       | Developed, Open Space                                    | 0         | 0.25      | 0         | 0.25      |
| Transitional                           | Developed, Low Intensity                                 | 0         | 0.2       | 0         | 0.2       |
| Transitional                           | Developed, Medium Intensity                              | 0         | 0.2       | 0         | 0.2       |
| Transitional                           | Developed High Intensity                                 | 0         | 0.2       | 0         | 0.33      |
Table 5. Cont.

| NLCD 1992 Class       | NLCD 2011/2006/2011 Class     | PDBM * | FBM ** | ICBM *** | Our **** |
|-----------------------|--------------------------------|--------|--------|----------|----------|
| Transitional          | Developed, Open Space          | 0      | 0.2    | 0        | 0.33     |
| Deciduous Forest      | Deciduous Forest               | 1      | 0.67   | 0.29     | 0.94     |
| Evergreen Forest      | Evergreen Forest              | 1      | 0.75   | 0.81     | 0.95     |
| Mixed Forest          | Mixed Forest                  | 1      | 0.75   | 0.29     | 0.92     |
| Shrubland             | Dwarf Scrub                    | 0.66   | 0.17   | 0.59     | 0.31     |
| Shrubland             | Shrub/Scrub                   | 1      | 0.43   | 0.59     | 0.79     |
| Orchards/Vineyards/Other | Cultivated Crops             | 0      | 0.25   | 0.13     | 0.65     |
| Grasslands/Herbaceous | Grassland/Herbaceous          | 1      | 0.75   | 0.81     | 0.86     |
| Grasslands/Herbaceous | Sedge/Herbaceous              | 0.66   | 0.2    | 0        | 0.61     |
| Grasslands/Herbaceous | Lichens                       | 0.66   | 0.2    | 0        | 0.25     |
| Grasslands/Herbaceous | Moss                          | 0.66   | 0.2    | 0        | 0.25     |
| Pasture/Hay           | Pasture/Hay                   | 1      | 0.67   | 0.16     | 0.83     |
| Row Crops             | Cultivated Crops              | 1      | 0.25   | 0.82     | 0.84     |
| Small Grains          | Cultivated Crops              | 1      | 0.25   | 0.37     | 0.65     |
| Fallow                | Cultivated Crops              | 1      | 0.25   | 0        | 0.33     |
| Urban/Recreational Grasses | Developed, Open Space | 1      | 0.75   | 0.81     | 0.95     |
| Woody Wetlands        | Woody Wetlands                | 1      | 0.75   | 0.69     | 0.97     |
| Emergent Herbaceous Wetlands | Emergent Herbaceous Wetlands | 1      | 0.75   | 0.69     | 0.97     |

* Path and depth-based measures: PDBM, ** Feature-based measure: FBM, *** Information content based measure: ICBM, **** Our proposed approach: Our approach.

Table 6 shows the result of the semantic similarity assessment between NLCD 2001/2006/2011 and NOAA Regional Land Cover Classification Scheme. The results include both polysemy and ambiguity. PDBM cannot effectively assess the semantic similarities for a majority of classes between NLCD 2001/2006/2011 and NOAA Regional Land Cover Classification Scheme. Without a manual interpretation, ICBM seems to have challenges to deal with measuring the semantic similarities of some classes (e.g., Barren Land (Rock/Sand/Clay) and Barren Land) between these two LULC classification systems and standards. Moreover, although FRM overperforms ICBM, it still cannot recognize the hidden differences. For example, the semantic similarity assessment of Palustrine Emergent Wetland (Persistent) and Emergent Herbaceous Wetlands, and Estuarine Emergent Wetland (Persistent) and Emergent Herbaceous Wetlands requires discovering the hidden relationship among Palustrine, Estuarine, and Emergent; however, this hidden relationship might not be explicit without the domain knowledge semantically organized by the conceptual hierarchy of the ontology.

Table 6. Results of semantic similarity measuring between and NLCD 2011/2006/2011 and NOAA Regional Land Cover Classification Scheme.

| NLCD 2001/2006/2011 Class | NOAA Class        | PDBM * | FBM ** | ICBM *** | Our **** |
|---------------------------|-------------------|--------|--------|----------|----------|
| Open water                | Open Water        | 1      | 1      | 1        | 1        |
| Perennial Ice/Snow        | Perennial Ice/Snow| 1      | 1      | 1        | 1        |
| Developed, Low Intensity  | Developed, Low Intensity | 1 | 1 | 1 | 1 |
| Developed, Medium Intensity | Developed, Medium Intensity | 1 | 1 | 1 | 1 |
| Developed High Intensity  | Developed, high intensity | 1 | 1 | 1 | 1 |
| Open space                | Open Space        | 1      | 0.75   | 0.59     | 0.75     |
| Barren Land (Rock/Sand/Clay) | Barren Land       | 1      | 0.5    | 0.45     | 0.67     |
| Barren Land (Rock/Sand/Clay) | Tundra           | 0.66   | 0.11   | 0        | 0.31     |
| Deciduous Forest          | Deciduous Forest  | 1      | 1      | 1        | 1        |
| Evergreen Forest          | Evergreen Forest  | 1      | 1      | 1        | 1        |
| Mixed Forest              | Mixed Forest      | 1      | 1      | 1        | 1        |
| Dwarf Scrub               | Scrub/Scrub       | 0.5    | 0.2    | 0.11     | 0.31     |
| Shrub/Scrub               | Scrub/Scrub       | 0.5    | 1      | 1        | 1        |
| Cultivated Crops          | Cultivated Crops  | 1      | 1      | 1        | 1        |
Table 6. Cont.

| NLCD 2001/2006/2011 Class | NOAA Class                      | PDBM * | FBM ** | ICBM *** | Our **** |
|---------------------------|--------------------------------|--------|--------|----------|----------|
| Grassland/Herbaceous      | Grassland/Herbaceous           | 1      | 1      | 1        | 1        |
| Sedge/Herbaceous          | Grassland/Herbaceous           | 0.66   | 0.33   | 0        | 0.61     |
| Lichens                   | Grassland/Herbaceous           | 0      | 0.25   | 0        | 0.25     |
| Moss                      | Grassland/Herbaceous           | 0      | 0.25   | 0        | 0.25     |
| Pasture/Hay               | Pasture/Hay                    | 1      | 1      | 1        | 1        |
| Woody Wetlands            | Palustrine Forested Wetland     | 0.66   | 0.6    | 0.59     | 0.83     |
| Woody Wetlands            | Palustrine Scrub/Shrub Wetland  | 0.66   | 0.4    | 0.33     | 0.67     |
| Woody Wetlands            | Estuarine Forested Wetland      | 0.66   | 0.6    | 0.59     | 0.83     |
| Woody Wetlands            | Estuarine Scrub/Shrub Wetland   | 0.66   | 0.4    | 0.33     | 0.67     |
| Emergent Herbaceous Wetlands | Palustrine Emergent Wetland  | 0.66   | 0.33    | 0.11   | 0.83    |
| Emergent Herbaceous Wetlands | Estuarine Emergent Wetland  | 0.66   | 0.33    | 0.11   | 0.83    |
| Unconsolidated Shore      |                                |        |        |          |          |

* Path and depth-based measures: PDBM, ** Feature-based measure: FBM, *** Information content based measure: ICBM, **** Our proposed approach: Our approach.

As we can see from Tables 5 and 6, using previous ontology-based semantic similarity for LULC classification systems and standards, the performance of existing approaches is ranked as: FBM > ICBM > PBM; however, the weaknesses of each approach prevent them from producing an accurate result of semantic similarity. By incorporating probabilistic models into FBM, our proposed approach can more accurately measure semantic similarity.

The result of semantic similarity measuring could be useful for a number of applications. First, the changes of LULC have been a significant research focus of remote sensing and land planning. Due to the fact that LULC maps within different periods were generated by various LULC classification systems, the changes of LULC based on those maps might not be available. The similarity degrees among these LULC classification systems can facilitate people quantitatively analyzing the changes of LULC in a more accurate way. Moreover, LULC classification systems are generated based on specific LULC conditions of different areas, countries, or regions. The semantic similarity of LULC classification systems of different places represents the characteristics of these places in terms of LULC to some extent.

5. Conclusions

The emergence of multi-type LULC classification systems and standards facilitates the generation of LULC classification maps and digital products; however, the heterogeneities of diverse LULC classification systems and standards impact the efficiency of using these products in land monitoring, management, and utilization. To address the heterogeneities, ontology-based approaches have been commonly exploited by information science. This paper integrates probabilistic models and ontologies to facilitate measuring semantic similarity of different LULC classification systems and standards.

In this paper, we developed domain ontologies to explicitly define the descriptions and code of different LULC classification systems and standards as semantic information and rules for logic reasoning. Based on the semantics and rules, we applied the Bayes algorithm to create a conditional probabilistic model for computing the semantic similarity of LULC categories in separate LULC classification systems and standards. The experiment shows that semantic similarity can be effectively measured by integrating a probabilistic model based on the content of ontology.

There are several possible extensions of this research that focus on integrating the content of different LULC classification systems and standards. To explicitly represent the hidden semantic information, the fusion of various domain ontologies for LULC classification systems and standards still needs to be investigated. Moreover, since the nature of LULC information inherits geographical context, geo-referenced information would be an aspect of the semantic similarity measuring. Based on discussions of the
feature-based approach and the IC-based approach, it might be useful to study integrating informativeness and features to assess the semantic similarity of LULC classification systems and standards.

**Author Contributions:** Conceptualization, X.Z. and Y.X.; methodology, X.Z. and X.X.; validation, X.X. and B.X.; data curation, X.Z.; writing—original draft preparation, X.Z. and Y.X.; writing—review and editing, X.Z., X.X. and B.X. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (140119001) and the National Natural Science Foundation of China (41701466, 41975041).

**Data Availability Statement:** The data used in this paper can be found via these links: https://www.mrlc.gov/data/legends/national-land-cover-database-2001-nlcd2001-legend; https://www.mrlc.gov/data/legends/national-land-cover-database-2006-nlcd2006-legend; https://www.mrlc.gov/data/legends/national-land-cover-database-2011-nlcd2011-legend; https://coast.noaa.gov/digitalcoast/training/ccap-land-cover-classifications.html.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Rounsevell, M.D.A.; Arneth, A.; Alexander, P.; Brown, D.G.; Ellis, E.; Finnigan, G.; Galvin, K.; Grigg, N. Towards decision-based global land use models for improved understanding of the Earth system. *Earth Syst. Dynam.* 2014, 5, 117–137. [CrossRef]

2. Perez-Hoyos, A.; Garcia-Haro, F.J.; Valcárcel, N. Incorporating Sub-Dominant Classes in the Accuracy Assessment of Large-Area Land Cover Products: Application to GlobCover, MODISLC, GLC2000 and CORINE in Spain. *IEEE J.-STARS* 2014, 7, 187–205. [CrossRef]

3. Jepsen, M.R.; Levin, G. Semantically based reclassification of Danish land-use and land-cover information. *Int. J. Geogr. Inf. Sci.* 2013, 27, 2375–2390. [CrossRef]

4. Netzel, P.; Stepinski, T.F. Pattern-Based Assessment of Land Cover Change on Continental Scale with Application to NLCD 2001–2006. *IEEE Trans. Geosci. Remote* 2015, 53, 1773–1781. [CrossRef]

5. Novack, T.; Kux, H.J.H. Urban land cover and land use classification of an informal settlement area using the open-source knowledge-based system InterIMAGE. *Health Risk Soc.* 2010, 55, 23–41.

6. Tomaselli, V.; Dimopoulos, P.; Marangi, C.; Kallimanis, A.S.; Adamo, M.; Tarantino, C.; Panitsa, M.; Terzi, M.; Veronica, G.; Lovergine, F.; et al. Translating land cover classifications to habitat taxonomies for landscape monitoring: A Mediterranean assessment. *Landsc. Ecol.* 2013, 28, 905–930. [CrossRef]

7. Aydinoglu, A.C.; Yomralioglu, T.; Ican, H.I.; Sesi, F.A. Managing land use/cover data harmonized to support land administration and environmental applications in turkey. *Sci. Res. Essays* 2010, 5, 275–284.

8. Li, C.; Ling, T.W. OWL-Based Semantic Conflicts Detection and Resolution for Data Interoperability. Conceptual Modeling for Advanced Application Domains. In Proceedings of the ER 2004 Workshops CoMoGIS, COMWIM, ECDM, CoMoA, DGOV, and ECOMO, Shanghai, China, 8–12 November 2004.

9. Cave, K.R. The Feature Gate model of visual selection. *Psychol. Res.* 1999, 62, 182–194. [CrossRef]

10. Comber, A.; Fisher, P.; Wadsworth, R. You know what land cover is but does anyone else? An investigation into semantic and ontological confusion. *Int. J. Remote Sens.* 2005, 26, 223–228. [CrossRef]

11. Ahlqvist, O. Using uncertain conceptual spaces to translate between land cover categories. *Int. J. Geogr. Inf. Sci.* 2005, 19, 831–857. [CrossRef]

12. Ahlqvist, O. Using semantic similarity metrics to uncover category and land cover change. In *GeoSpatial Semantics*; Springer: Berlin, Germany, 2005; pp. 107–119.

13. Ahlqvist, O. Extending post-classification change detection using semantic similarity metrics to overcome class heterogeneity: A study of 1992 and 2001 US National Land Cover Database changes. *Remote Sens. Environ.* 2008, 112, 1226–1241. [CrossRef]

14. Pazúr, R.; Oťaheľ, J.; Maretta, M. The distribution of selected CORINE land cover classes in different natural landscapes in Slovakia: Methodological framework and applications. *Morav. Geogr. Rep.* 2015, 23, 45–56. [CrossRef]

15. Stepinski, T.F.; Cohen, J.P. Comparing semantically-blind and semantically-aware landscape similarity measures with application to query-by-content and regionalization. *Ecol. Inform.* 2014, 24, 69–77. [CrossRef]

16. Feranec, J.; Solin, L.; Kopecka, M.; Oťaheľ, J.; Kupkova, L.; Stych, P.; Bick, I.; Kolar, J.; Cerba, O.; Soukup, T.; et al. Analysis and expert assessment of the semantic similarity between land cover classes. *Prog. Phys. Geog.* 2014, 38, 301–327. [CrossRef]

17. Gan, M.; Dou, X.; Jiang, R. From ontology to semantic similarity: Calculation of ontology-based semantic similarity. *Sci. World J.* 2013, 2013, 793091. [CrossRef] [PubMed]

18. Batet, M.; Sánchez, D.; Valls, A.; Gibert, K. Semantic similarity estimation from multiple ontologies. *Appl. Intell.* 2013, 38, 29–44. [CrossRef]
19. Taieb, M.A.H.; Aouicha, M.B.; Hamadou, A.B. Ontology-based approach for measuring semantic similarity. *Eng. Appl. Artif. Intell.* 2014, 36, 238–261. [CrossRef]

20. Rodriguez, M.A.; Egenhofer, M.J. Comparing geospatial entity classes: An asymmetric and context-dependent similarity measure. *Int. J. Geogr. Inf. Sci.* 2004, 18, 229–256. [CrossRef]

21. Janowicz, K.; Kessler, C. The role of ontology in improving gazetteer interaction. *Int. J. Geogr. Inf. Sci.* 2008, 22, 1129–1157. [CrossRef]

22. Li, Y.; Bandar, Z.A.; McLean, D. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. Knowl. Data Eng.* 2003, 15, 871–882.

23. Al-Mubaid, H.; Nguyen, H.A. A cluster-based approach for semantic similarity in the biomedical domain. In Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, New York, NY, USA, 31 August–3 September 2006.

24. Resnik, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* 1999, 11, 95–130. [CrossRef]

25. Sebti, A.; Barfroush, A.A. A new word sense similarity measure in WordNet. In Proceedings of the International Multiconference on Computer Science and Information Technology, Wisła, Poland, 18–20 October 2008; pp. 369–373.

26. Sánchez, D.; Batet, M. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *J. Biomed. Inform.* 2011, 44, 749–759. [CrossRef]

27. Taieb, M.A.H.; Aouicha, M.B.; Hamadou, A.B. Computing semantic relatedness using Wikipedia features. *Knowl.-Based Syst.* 2013, 50, 260–278. [CrossRef]

28. Taieb, M.A.H.; Aouicha, M.B.; Hamadou, A.B. A new semantic relatedness measurement using WordNet features. *Knowl. Inform. Syst.* 2014, 41, 467–497.

29. Tversky, A. Features of similarity. *Psychol. Rev.* 1997, 84, 327–352. [CrossRef]

30. Cross, V.; Silwal, P.; Xi, C. Experiments Varying Semantic Similarity Measures and Reference Ontologies for Ontology Alignment. In *Proceedings of the Extended Semantic Web Conference, Montpellier, France, 26–30 May 2013*; Springer: Berlin/Heidelberg, Germany, 2013.

31. Petrakis, E.G.M.; Varelas, G.; Hliaoutakis, A.; Raftopoulou, P. X-Similarity: Computing semantic similarity between concepts from different ontologies. *IEEE Trans. Knowl. Data Eng.* 2003, 15, 442–456. [CrossRef]

32. Jiang, Y.; Zhang, X.; Tang, Y.; Nie, R. Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. *Inform. Process. Manage.* 2015, 51, 215–234. [CrossRef]

33. Solé-Ribalta, A.; Sánchez, D.; Batet, M.; Serratosa, F. Towards the estimation of feature-based semantic similarity using multiple ontologies. *Knowl. -Based Syst.* 2014, 55, 101–113. [CrossRef]

34. Gennari, J.H.; Musen, M.A.; Fergerson, R.W.; Grosso, W.E.; Crubézy, M.; Eriksson, H.; Noy NFTu, S.W. The evolution of Protégé: An environment for knowledge-based systems development. *Int. J. Hum. -Comput. Stud.* 2003, 58, 89–123. [CrossRef]

35. Hendler, J.; Lassila, O.; Berners-Lee, T. The semantic web. *Sci. Am.* 2001, 284, 34–43.

36. Horrocks, I.; Patel-Schneider, P.F.; Boley, H.; Tabet, S.; Grosof, B.; Dean, M. SWRL: A semantic web rule language combining OWL and RuleML. *W3C Membr. Submiss.* 2004, 21, 79.