Journal of Web Engineering, Vol. 10, No.1 (2011) 047-069

A NOVEL MULTI-ASPECT CONSISTENCY MEASUREMENT FOR ONTOLOGIES

ZHAO LU *

Department of Computer Science and Technology, East China Normal University, Shanghai, China
zlu@cs.ecnu.edu.cn

ZOLTÁN MIKLÓS

EPFL- I&C-LSIR, BC 142, Station 14, CH-1015 Lausanne, Switzerland
zoltan.miklos@epfl.ch

LIANG HE SONGMEI CAI JUNZHONG GU

Department of Computer Science and Technology, East China Normal University, Shanghai, China
lhe@cs.ecnu.edu.cn smcai@ica.stc.sh.cn jzgu@cs.ecnu.edu.cn

Received March 26, 2010
Revised January 7, 2011

Web developers have started to integrate semantic information to their systems increasingly often. The semantic metadata embedded with the resources is typically linked to ontologies or taxonomies. Meta information can bring a number of advantages for users. However, the ontologies might contain some errors or could be partially inconsistent. Therefore it is important to evaluate the quality of ontologies at various levels. Existing evaluation methods either investigate whether the ontologies are "fit for purpose", or focus on evaluating ontology consistencies from a single aspect. In this study, we focus on ontology consistency evaluation methods, which consider lexical, taxonomic and syntactic aspects at the same time. We propose new measures, which capture several essential aspects simultaneously. We demonstrate the effectiveness of the new measure through a case study and an extensive set of experiments.

Key words: Ontology, Semantic Relatedness, Ontology Consistency, WordNet
Communicated by: D. Lowe & O. Pastor

1 Introduction

Ontologies are part of the W3C standards stack for the Semantic Web, in which they are used to specify standard conceptual vocabularies for exchanging data across systems, for providing services for answering queries, for publishing reusable knowledge bases, and for offering services to facilitate interoperability between multiple, heterogeneous systems and databases [12]. Generally, representational primitives of an ontology are typically concepts, attributes (or properties), and relationships (or relations among concepts). In recent years, ontologies are used more and more in various Web-based

* Corresponding author

applications. However, in the course of representing domain knowledge by ontology, one inevitably introduces inconsistencies especially in the ontology construction and evolution process. These inconsistencies include structural inconsistency, logical inconsistency and user-defined inconsistency [14].

The need for an ontology evaluation methodology has become pressing as early as 1994 and since then the need has been greatly growing [34]. At present, special efforts are devoted for developing novel ways to measure and to evaluate the quality of ontologies, either qualitatively or quantitatively. Lewen [24] presents an open rating system-based approach for the evaluation. The core of this open rating system, which is partially implemented in Knowledge Zone, was extended with topic-specific trust to provide more accurate personalized ontology rankings. Gomez-Perez [9] presents an evaluation technique based on numerous criteria, including consistency, completeness, conciseness, expandability, and sensitiveness. Brewster *et al.* [4] suggested a method by decomposing ontologies into concepts and relationships in order to evaluate its fitness for conceptualizing particular sets of natural language texts, which is the corpus. These kinds of evaluations are based on statistical approaches.

One of the key issues in the ontology evolution and ontology matching [2, 28] is the problem of keeping the ontologies consistent. Generally, the consistency of an ontology can be evaluated at five levels: *Lexical level*, *Taxonomy level*, *Syntactic level*, *Context or application level*, and *Structure, architecture and design level*. Till now, most of the existing research focuses on only one of these levels. To the best of our knowledge, no comprehensive or general approach has been proposed that would concern the evaluation of multi-level consistencies in ontologies.

In this study, which is based on our previous works [21, 27, 28], we develop a new measurement method for evaluating taxonomy consistencies in ontologies at three levels: *Lexical level*, *Taxonomy level* and *Syntactic level*. The corresponding metric is called S-Measure. We compute the Taxonomy Consistency Score (denoted as Stax) of an ontology to describe its degree of consistency, using a novel semantic relatedness measure and two algorithms. To measure the semantic relatedness between two words in WordNet, a new semantic relatedness measure is proposed in this study, called Semantic Path Weight Measure (or in short, p-Measure). We designed some experiments to demonstrate both the validity and effectiveness of S-Measure and p-Measure, see Section 6 for details.

To summarize, this paper makes the following contributions:

- We propose S-Measure to measure the taxonomy consistency of ontologies. The new ontology consistency measure does not only evaluate the consistency of ontologies at multiple aspects but it also weights different types of errors.

- We propose p-Measure to measure semantic relatedness, which considers (1) three relationship types, which are *hh* (hypemym/hyponym), *hm* (holonym/meronym) and *sa* (synonym/antonym), and (2) the strength of relations among words. We demonstrate the effectiveness of p-Measure by comparing it to seven other popular semantic relatedness measures.

- We show that the principle of S-Measure is reasonable through an analytical evaluation. We also use experiments to show that our ontology measure performs better than Lexical F-measure when measuring multi-aspect consistencies.

The rest of this paper is organized as follows. Section 2 highlights related works. Section 3 presents S-Measure including several related definitions. Section 4 describes p-Measure for pairs of words in detail. Section 5 describes two Algorithms to compute Stax. Section 6 presents experimental evaluations of p-Measure, and analytical and experimental evaluations of S-Measure. Section 7 concludes the paper finally.

2 Related Work

2.1 Ontology Evaluation

Various approaches concerned with ontology evaluation depend on the type of ontology and the evaluation purpose. Generally speaking, these approaches can be classified into four categories [3]:

- By comparing ontologies with a set of Golden Standard. Here the Golden Standard refers to an existing ontology or other representations of the same problem domain.
- By plugging ontologies into some special applications and measuring the quality of results.
- Through comparing ontologies to a set of unstructured or informal data (e.g., text documents) which describe the same problem.
- An ontology evaluation is conducted through assessing the fitness value according to a set of predefined criteria, standards or requirements.

The first three approaches are domain-oriented, and the fourth relies on some predefined rules. None of the four categories take the taxonomic relationships of ontologies into consideration.

According to literatures, several different ontology evaluation levels can be concluded as follows:

- *Lexical level*, which measures the ontology quality by comparing words (lexical entries) of ontologies with a set of words which represents a problem domain such as in paper [4, 29, 35].
- *Taxonomy level*, which considers the hierarchical connections among concepts of an ontology using "is-a" relations or other semantic relations such as in [13].
- *Syntactic level*, which considers syntactic requirements of an ontology describe language [11].
- *Context or application level*, which considers context of ontologies, such as an ontology which references or is referenced by the one being evaluated, or the application it is intended for [6, 34].
- *Structure, architecture and design level*, which takes into account principles and criteria considered during the ontology construction process [26].

To assure a good quality, it is highly desirable that ontologies to be evaluated through qualitative and quantitative methods. The former approaches evaluate different ontologies with regard to an application, and the latter can be used to evaluate the quality of a single ontology. As an important activity of ontology evaluation, inconsistency detection is responsible for checking the degree of consistency of an ontology with respect to a predefined ontology consistency conditions. Its goal is to find all parts in ontologies that do not meet some consistency conditions. Hasse & Stojanovic [15] propose three different types of ontology consistency concepts; namely syntactical consistency, logical consistency and structural consistency.

Guarino & Welty [13] have argued that some ontologies contain inadequate taxonomic relationships so they proposed the OntoClean methodology. The main function of the OntoClean is the formal properties evaluation by means of a predefined ideal taxonomical structure of meta-properties, such as essence, identity, unity, and dependence. The evaluation is dictated by the constraints imposed on the different meta-properties. It is not convincing though that the meta-properties schema introduced is usable by knowledge engineers or domain experts, as different knowledge engineers tend to describe the same concept with significantly different sets of meta-properties.

CleOn evaluates taxonomic relationships in ontologies using paths from concept nodes to the root node [33]. CleOn has extracted the paths from WordNet 2.1 to get the path information they require, and only hypemym relationships defined in WordNet 2.1 are considered. However, except the hypemym/hyponym relationships, other relationships, such as the holonym/meronym and synonym/antonym relationships, are also defined in WordNet 2.1.

Gangemi *et al.* [8] suggested an evaluation framework to integrate different approaches for ontology evaluation and validation by means of a formal model (i.e., meta-ontology), that is called O^2. They identify three main types of measures for evaluation: structural measures, functional measures and usability-profiling measures.

2.2 Semantic Relatedness Measure

Measures of semantic relatedness or semantic similarity are used in various applications such as word sense disambiguation, information extraction and retrieval, and automatic correction of word errors in text. Semantic similarity typically shows a synonymy relation between two words, while there are other kinds of relations contained in the notion of semantic relatedness, e.g., metonymy, antonym, functional association, and other “non-classical" relations. For example, the relation between a car and an engine is a *part-whole* relation, the relation between “good” concept and “bad” concept is an *antonym* relation, and intuitively there is a kind of relation between “snow” concept and “ski” concept, but this kind of relation sometimes is hard to qualify. In this study, we use semantic relatedness rather than semantic similarity based on three observations: (1) Semantic similarity is one kind of semantic relatedness. (2) In general, computational linguistic applications benefit more from calculated semantic relatedness rather than calculated semantic similarity. (3) When we evaluate an ontology, except synonym relations, other kinds of relations contained in a domain ontology should be considered also.

There are three kinds of measuring relatedness approaches: dictionary-based approaches, approaches based on Roget-structured thesauri and approaches using WordNet and other semantic networks. The approaches using WordNet are predominant. As a lexical hierarchical system, WordNet 3.0, which is produced by Miller *et al.* from Princeton University in the 1990s [19], is currently one of the most popular and the largest online dictionary. For brevity, in the following sections we use WordNet for WordNet 3.0.

In WordNet, nodes represent concepts and edges represent relations between concepts. Nodes at deeper levels are more informative and specific than nodes that are nearer to the root. The backbone of the noun network is the subsumption hierarchy (e.g. hyponym/hypemym). WordNet supports multiple inheritance between nodes and therefore has the greatest number of relations implemented. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (or simply synsets), each

expressing a distinct concept in WordNet. A synset can be viewed as a concept evoked by one or more senses of words. There are 117,597 synset nodes and each WordNet synset has a corresponding node. Synsets are interlinked by means of conceptual-semantic and lexical relationships. For each sense of a word, several other types of semantic relations are supplied besides the traditional "is-a" and "part-of" relations, and these semantic relations are not systematically and formally defined.

Until now, several WordNet-based approaches have been proposed to compute the semantic relatedness (or similarity) between concepts. These approaches can be classified into three main categories: edge-based, information content-based and hybrid approaches.

Hirst & St-Onge [16] suggested an edge-based measuring approach: for two WordNet concepts c_1 and c_2 ($c_1 \neq c_2$), there is,

$$rel_{hso}(c_1, c_2) = C - len(c_1, c_2) - k \times turns(c_1, c_2)$$

where C and k are constants (in practice, they let $C = 8$ and $k = 1$), and $turns(c_1,c_2)$ is the number of times the path between c_1 and c_2 changes direction. Thus, the longer is the path and the more changes of the direction it contains, the lower is the weight. This method considers all relations, patterns and the number of changes of direction in a path. The main drawback of this measure is that it considers that each edge of each type represents the same information content. However, apart the measure, only few works have been made on semantic relatedness measures using heterogeneous relations.

Other kinds of measuring approaches are the information based approaches and the integrated approaches, such as the Resnik's approach based on information theory [32], the combined approach discussed by Jiang & Conrath [20], and Lin's universal similarity measure approach proposed based on there intuitions concerned commonality between two concepts [25].

A hybrid measure is proposed by Hong-Minh & Smith [17] for measuring semantic similarity by adding depth factor and link strength factor,

$$sim(c_1, c_2) = \begin{cases} \max\limits_{c \in Sup(c_1, c_2)} (IC(c \mid p) \times f_c(d)), & c_1 \neq c_2, \\ 1, & c_1 = c_2. \end{cases}$$

here $f_c(d)$ is consider as an exponential-growth function (i.e., the Equ.(4) in Section 4), and there is $d=max(depth(c_1),depth(c_2))$. The strength of a link is defined to be the conditional probability of encountering a child node c_i, given an instance of its parent node p. However, the proposed definition for the strength of relation between nodes does not take into account the relation type, that is an important factor for the distance approaches.

More recently, Mazuel & Sabouret [22] have proposed an integrated measure using heterogeneous relations: hierarchical links and non-hierarchical links,

$$rel(c_1, c_2) = 2 \times IC_{\max} - dist(c_1, c_2)$$

where

$$dist(c_1, c_2) = \min_{\{p \in \pi(c_1, c_2) \mid HSO(p) = true\}} W(p)$$

and and $W(p)$ is the weight of the path $path(x,y)$. Links "Part-of", "Member-of "and "Substance-of" are described as non-hierarchical links.

In order to measure semantic relatedness using heterogeneous relations, Mazuel & Sabouret analyze two kinds of single-relation paths, i.e., hierarchical relation and the non-hierarchical relation, and mixed-relation path which can be factorized as an ordered set of n single-relation sub-paths. The main drawback of this measure is that the semantic relatedness of two words depends on not only the information content of nodes but also the types of all relations appearing in the path $path(x,y)$. For non-hierarchical relations, Mazuel & Sabouret considers a static weight factor based on edge-count methods. In hierarchical relations the weight of path is computed by information content of nodes, but information content should be computed firstly, otherwise the measure cannot be conducted.

The strength of a child link is proportional to the conditional probability of encountering an instance of the child concept c_i given an instance of its parent concept p, that is $P(c_i|p)$. Mazuel & Sabouret defines the link strength (LS) by:

$$LS(c_i, p) = -\log(P(c_i \mid p) = IC(c_i) - IC(p)$$

The LS states the difference of the information content values between a child concept and its parent concept. The weight (wt) for a child node c and its parent node p can be determined as follows:

$$wt(c, p) = \left(\beta + (1-\beta)\frac{\overline{E}}{E(p)}\right)\left(\frac{d(p)+1}{d(p)}\right)^{\alpha}\left[IC(c) - IC(p)\right]\ T(c, p)$$

here $T(c,p)$ is the link type factor, and $T(c,p) = 1$.

In WordNet, the *hh* link type is the most common, however other link types, such as *hm* and *sa,* should also be considered as they would have different effects in calculating the weight. To differentiate the weights of links connecting a node and all its child nodes, one needs to consider the link strength of each specific child link. This could be measured by the closeness between a specific child node and its parent node, against those of its siblings.

3 Backgrounds

In this study, we tackle the problem of measuring ontologies at Lexical level, Taxonomy level and Syntactic level at the same time. We view an ontology as a multi-hierarchical structure, where a node represents a concept, and an edge models the binary specialization relation (e.g., is-a, part-of) between two concepts. The multi-hierarchical structure implies that a concept can have more than one parent. When all the concepts have at most one parent, then the structure can be considered as a tree. However, the multi-hierarchical structure is a Directed Acyclic Graph (DAG), if at least one concept has more than one parent. A DAG is a directed graph with no directed cycles, that is, for a vertex *v,* there is no nonempty directed path that starts and ends on *v.* Moreover, DAGs can be considered to be a generalization of trees in which certain subtrees can be shared by different parts of the tree.

Figure 1 illustrates two possible structures for the ontology *Transport*, where Figure 1(a) contains a simple tree structure and Figure 1(b) has a DAG structure. As most ontologies have such a structure, we restrict our attention to these types of ontologies. Formally, we give the definition of ontology used in this study in Definition 1.

Definition 1 (*A multi-hierarchical structure of ontology*). An ontology $O := (C,root,\leq_c)$ is a multi-hierarchical structure, where C is a set of concept identifiers, $\{c_1,c_2,...,c_i,...,c_n\}$, a node represents a concept c, and the root is a designated root concept for the partial order $\leq_c$ on C. This partial order models the binary specialization relation between two concepts. The equation $\forall c \in C : c \leq_c root$ holds for this concept hierarchy.

Definition 2 (*Path*). A *path* betwee *n* two concepts c_1 and c_n is denoted as,

$$path(c_1, c_n) = \{c_1, e_{1,2}, c_2, e_{2,3}, ...c_k, e_{k,k+1}, c_{k+1}, ..., c_n\}$$

and $e_{k,k+1}$ is the link between two concepts c_k and c_{k+1} appearing in the path.

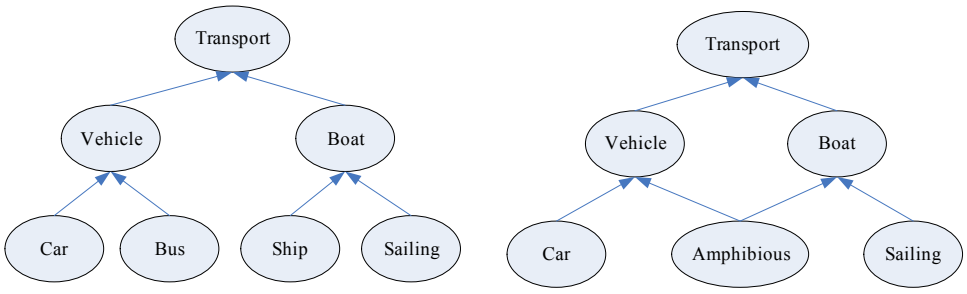

Figure 1 Two ontologies with a tree (a) and DAG (b) structure.

The *length* of a path is the number of edges that appear in the path. The *shortest path* from the concept c_1 to the concept c_n is the path with the minimum length, and the path length of the shortest path between two nodes is defined as the *distance* of the two nodes, denoted as $dist(c_i, c_j)$. The *depth* of a node c_i, denoted as $depth(c_i)$, is the shortest path from it to the global root concept *root.*

There are different kinds of links between two concepts in an ontology, such as *hh* (e.g., is-a), *hm* (e.g., part-of, member-of, substance-of) and *sa* (i.e., the two concepts are synonymous, or the two concepts are antonyms of each other), respectively. In most of the situations, each kind of link carries different amount of information, however, these is less research considering this. The type of link is a good factor to tackle this problem. In this study, we assign a *link type weight* for each link t in an ontology and denote it as w_t.

In order to determine whether the relationships between a father-node and all its child-nodes are reasonable or not, we introduce the notion of *relationship reasonable threshold* (Th) in Definition 3. Considering both the limitation of the depth of each node in the ontology O and the human cognitive limitations, the *relation reasonable threshold* is defined as follows:

Definition 3 (*Relationship reasonable threshold*). Given an ontology *O,* for an concept c in it, there is $\exists c \in C : c \leq_c root$, then the relationship reasonable threshold of the concept c, $Th(c)$, which satisfies:

$$Th(c) = (w_{\max})^n \tag{1}$$

where $w_{\max}$ is the biggest value among all link type weights appearing the $path(root,c)$.

For the parameter n in Equ(1), there are two considerations: (1) Since there are maximum 16 words in a path of the noun hierarchy of WordNet, then two words are unrelated if there are more than eight nodes contained in the shortest path between them. (2) There are no more than five links in an allowable path, so we let $n = 8$. For example, if all link types appearing in the *path*(*root,c*) are *hh* links, then, the *relationship reasonable threshold* of the concept c is: $Th(c) = w_{hh}^8 = 0.8^8 = 0.1678$, since there is $w_{max} = w_{hh} = 0.8$.

For two concepts c_i and c_j in an ontology O, and c_j is a lower concept of c_i, we mark the semantic relatedness between the two concepts c_i and c_j as $rel(c_i,c_j)$. Next, we give the definitions of unclean concept and clean concept suggested in this study.

Definition 4 (*Unclean concept or node*). A concept c_j in an ontology O is viewed as an *unclean concept or node* if:

- It is not a term in WordNet;
- It is a term in WordNet, and the semantic relatedness between two concepts c_i and c_j, $rel(c_i,c_j)$, is less than the relationship reasonable threshold of the concepts c_i in the ontology O (marked as $Th(c_j)$), that is, $rel(c_i,c_j) < Th(c_j)$, then the concept c_j will be viewed as an unclean concept.

For these unclean nodes, they and all of their descending nodes are viewed as ill-defined taxonomy and therefore will not be considered during the next steps.

Definition 5 (*Clean concept or node*). A concept c_k in an ontology O is viewed as an *clean concept or node* if it is not an unclean concept or node.

Based on the two definitions of unclean concept and clean concept, it is easy to detect all unclean concepts and all clean concepts present in an ontology O. In the following sections, we use UC set to describe the set of all unclean concepts contained in the ontology O, and use uc to represent the number of unclean nodes contained in UC set. We use $C\backslash UC$ set to describe the set of all clean concepts contained in the ontology O, and use $ccnum$ to describe the number of non-leaf nodes contained in the $C\backslash UC$ set, and there is $ccnum < n\text{-}uc$.

Assume that there is a non-leaf node c_k in $C\backslash UC$ set, its all *direct child node* set is marked as $AX(c_k)$ and its all child nodes are marked as $AX(c_k) = (c_{k_1}, c_{k_2}, ...c_{k_i}, ..., c_{k_j}, ..., c_{k_m})$. Compute all semantic relatedness between any two nodes c_{k_i} and c_{k_j}, $rel(c_{k_i}, c_{k_j})$, $i < j$, and $i \neq j$, $i, j = 1,2,...,m$. We mark the *math-ematical average* of all semantic relatedness values between any two nodes c_{k_i} and c_{k_j} as $\overline{rel}$.

Definition 6 (*Taxonomic consistency of a concept*). The taxonomic consistency of a non-leaf clean concept c_k in $C \setminus UC$ set is denoted as $Con(c_k)$, which satisfies:

$$Con(c_k) = \frac{\sum_{i,j}^{m} (rel(c_{k_i}, c_{k_j}) - \overline{rel})^2}{m} \tag{2}$$

Variance values are used to measure the size of fluctuations of some discrete data in statistics. In this study, we use the variance value to reflect the centralization tendency of all semantic relatedness between two nodes contained in an ontology. If the variance value of the concept c_k is larger, it means that there is a set of disperse values of semantic relatedness between its any two child nodes. Thus we can state that the taxonomy of the concept c_k is bad. Contrarily, a lower $Con(c_k)$ value means that

semantic relatedness between all of its child nodes is within an acceptable range, and we conclude that the taxonomy of the node c_k is consistent.

Our basic observations of taxonomy consistency for an ontology are: 1) All concepts in ontologies can be classifed into unclean concepts and clean concepts. Both of them have various impacts on the consistencies of ontologies. 2) For an ontology *O*, on one hand, if there are more unclean concepts, the consistency of the ontology is lower. On the other hand, the taxonomy consistency of the ontology *O* is determined also by the taxonomy consistency of all non-leaf clean concepts in the ontology *O*.

Based on these observations, we design a novel consistency measurement (simply S-Measure) for ontologies using two factors, *conWeight* and *conVal.* The first factor, *conWeight,* is the measurement of consistency impact of all unclean concepts over an ontology. The less unclean concepts there are, the less *conWeight* is. The second factor, *conVal*, is the consistency measurement of all non-leaf clean concepts in the ontology *O*. The bigger the value is, the less taxonomy consistency in the ontology is.

Definition 7 (*Taxonomic consistency of an ontology*). The *taxonomy consistency of an ontology O*, *Stax,* which satisfies:

$$\begin{aligned} &Stax = conWeight(uc) + conVal(ccnum) \\ &conWeight(uc) = uc / n \\ &conVal(ccnum) = \sum Con(c_k)/ccnum \end{aligned} \tag{3}$$

where the parameter *n* is the number of all concepts, the parameter *uc* is the number of all unclean concepts, and the parameter *ccnum* is the number of all non-leaf clean concepts respectively.

As we show, the semantic relatedness of two concepts is one of the key issues in the suggested ontology consistency measurement and ontology matching. In our previous work, we have shown that the more accurate measure of semantic relatedness among concepts in an ontology, the more accurate is the consistency measure for an ontology [28]. In Section 4, we explain our semantic relatedness measure for pairs of words that we denote as p-Measure. In Section 6.1 we discuss the evaluation results of p-Measure based on the benchmark test of Miller and Charles [30].

4 p-Measure

A straight way to measure semantic relatedness between two concepts is to use their path length in WordNet. As WordNet is a lexical hierarchical system, it is well known that concepts at upper levels of the hierarchy have less semantic relatedness between them than concepts at lower levels. This characteristic should be taken into account as a constraint in calculating the semantic relatedness of two concepts with depth factor. Therefore, the depth function should give a higher value when applied on nodes at lower levels. We considered the contribution of the depth to the relatedness as an exponential-growth functions as in [15, 17], a monotonically increasing function with respect to a depth *d:*

$$f(d) = \frac{e^{\alpha d} - e^{-\alpha d}}{e^{\alpha d} + e^{-\alpha d}} \tag{4}$$

where *d* is an integer, and α is a tuning parameter. In this study, we set $\alpha = 0.528$.

Next, the definition of the semantic relatedness between two concepts in this study is:

Definition 8 (*Semantic Relatedness between two concepts*).

Given two different concepts, c_1 and c_n, assume that the shortest path between them in WordNet is marked as $path(c_1,c_n)=\{c_1,e_{1,2},c_2,e_{2,3},...c_k,e_{k,k+1},c_{k+1},...,c_n\}$, $1\le k\le n-1$,then the *semantic relatedness* between the two concepts is denoted as $rel(c_1,c_n)$, which satisfies the following,

$$rel(c_1,c_n)=\begin{cases}\prod_{k=1}^{n-1} w_t(e_{k,k+1})\times f(depth(c_1))\times f(depth(c_n)), & dist(c_1,c_n)\le\theta,\\ 0, & dist(c_1,c_n)>\theta.\end{cases} \tag{5}$$

where

- $dist(c_1,c_n)$ is the distance between two concepts c_1 and c_n;
- $w_t(e_{k,k+1})$ is the *link type weight* of the link t between two concepts c_k and c_{k+1} which belongs to the $path(c_1,c_n)$.
- θ is a constant and refers to a *threshold*. If the distance $dist(c_1,c_n)$ is larger than θ, then the two concepts are viewed as unrelated. The threshold can be given by an expert or can be computed using the statistics data. Considering that there are maximum 16 words in a path of the noun hierarchy of WordNet, in this study, we set $\theta=20$.

If $c_1=c_n$, then $rel(c_1,c_n)$ equals to 1. In this study, we will measure the semantic relatedness of two concepts based on the definition of semantic relatedness, and we call the measure as p-Measure.

There are three reasons that the value range of a link weight is less than or equal to one: (1) The link strength and the type of link are two important facts for the edge-counting approaches. As discussed in Section 2.2, the link strength is computed by the information contents of two words, and the information content of a node is less than or equal to one. So the link strength will be less than or equal to one. (2) The relation type represents the maximum information content that this kind of link can carry. If the relation type is less than one, this type of link is considered being informative. (3) According to information content approaches, information contents contained in lowest nodes equal to one and that in higher nodes is less than one. Based on above three considerations, there is $0<w_t\le 1$.

Hirst and St-Onge [16] distinguished three kinds of relations: *extra-strong*, *strong*, and *medium-strong*. An extra-strong relation holds only between a word and its literal repetition; such relations have the highest weight of all relations. A strong relation has a lower weight than an extra-strong relation and a higher weight than a medium-strong relation. A word may have more synsets in WordNet, and each synset of them corresponds to a different sense of the word. Two words are strongly related if one of the following holds:

- They have a synset in common, for example, human and person.
- They are associated with two different synsets that are connected by the horizontal relation (e.g., antonymy, similarity) relation, such as precursor and successor.
- One of the words is a compound (or a phase) that includes the other and "there is any kind of link at all between a synset associated with each word" (e.g., school and private school).

Two words are said to be in a *medium-strong*, or *regular*, relation if there exists an allowable path connecting a synset associated with each word (e.g., carrot and apple). The intuition is that "the longer the path and the more changes of direction, the lower the weight". Unlike extra-strong and strong relations, medium-strong relations have different weights.

Due to above observations, in p-Measure, the value of link type weight corresponds to the "strength" of a given link type. Since different kinds of relations carry different strengths of information, the values of link type weights should reflect the types of links. In this study, we assign following values of the three types of links: For the link type $t = sa$, a higher value is assigned, for example, $w_{sa} = 0.9$. For the link type $t = hh$ *and* $t = hm$, a lower value is assgined, that is $w_{hh} = 0.8$ and $w_{hm} = 0.8$, respectively. Since all link type weights in the shortest path are smaller than one, there is $0 < rel(c_1, c_2) \leq 1$.

Similar to the hybird approach proposed in [17], by adding such structural information of the taxonomy, p-Measure can exploit all typical characteristics of a hierarchical structure when it is applied on such taxonomy. Moreover, such information can be tuned via parameters. The method is therefore flexible for many types of taxonomies; such as hierarchical structure or non-hierarchical structure.

5 Computing S-Measure

We propose a three step procedure for computing the S-Measure: (1) We apply the *Unclean Node Detection* (UND) algorithm to detect all unclean nodes for an ontology. (2) We compute the taxonomy consistency of each non-leaf clean node using the *Taxonomic Consistency of Node* (TCN) algorithm. (3) After collecting all *Con*s of all non-leaf clean concept nodes, we compute the *taxonomic consistency* value (Stax) for the ontology O.

The main steps of the Unclean Node Detection (UND) algorithm are the following: starting from the root node of an ontology *O*, the UND algorithm searches the ontology as far as possible in depth, and each traversal provides a path from the root node to one of its leaf nodes. We compute semantic relatedness of each concept pair (father-node, child-node) appearing in the path using p-Measure. The Algorithm 1 describes the detailed process.

After using the UND algorithm, all unclean nodes in the ontology *O* will be detected and the remaining, i.e., clean nodes, will be contained in the *C\UC* set. As discussed above, we use *ccnum* to describe the number of all non-leaf nodes contained in the *C\UC* set. If there is a non-leaf node c_k contained in the *C\UC* set, then the Taxonomic Consistency of Nodes (TCN) algorithm will be used to compute its taxonomic relationship consistency.

Analysis of S-Measure:

As described above, for the taxonomy of an ontology *O,* S-Measure not only determines whether the taxonomy is integrated and exhaustive, but also illustrates if the taxonomy is rational or consistent. If *Stax* is large, it indicates that there are some unclean concepts or classification conflicts within the taxonomy. Otherwise, if *Stax* is small, it indicates that the taxonomy is good.

Algorithm 1: The Unclean Node Detection Algorithm (UND)

Input: The node set C and relation set R of an ontology O
Output: The node set $C \setminus UC$, and the result relation set $\breve{R}$
1: Let $UC = \varphi$
2: for each $c_i \in C$
3: mark c_i as *unvisited*
4: compute $Th(c_i)$
5: end for
6: for each $c_i \in C$
7: if c_i is *unvisited* then
8: mark c_i as *visited*
9: for $r(c_i, c_j) \in R$
10: if c_j can not be found in WordNet then
11: $UC = UC \cup c_j$
12: move c_j and its all sub nodes and relations from C and R
13: else
14: calculate $rel(c_i, c_j)$
15: if $rel(c_i, c_j) < Th(c_j)$ then
16: $UC = UC \cup c_j$
17: move c_j and its all sub nodes and relations from C and R
18: end if
19: end for
20: end if
21: end for

Algorithm 2: The Taxonomic Consistency of Node Algorithm (TCN)

Input: The node set $C\backslash UC$ and the corresponding relations set $\breve{R}$ of an ontology O
Output: $Con(c_k)$
1: mark $c_k \in C \setminus UC$ as *unvisited*
2: for each c_k
3: if c_k is *unvisited* and is not a leaf-node
4: for any two sub-nodes c_{k_i}, c_{k_j} of c_k
5: calculate $rel(c_{k_i}, c_{k_i})$
6: end for
7: mark c_k *visited*
8: end if
9: end for
10: compute *rel*
11: compute $Con(c_k)$
12: end for

For a consistent ontology, the UND algorithm will not find any unclean nodes in it, there is $uc = 0$. Moreover, applying the TCN algorithm, for each non-leaf node c_k contained in the set $C\backslash UC$, the semantic relatedness values among its all direct child-nodes are equal. This means the taxonomic

consistency score of each node c_k is zero, that is $Con(c_k) = 0$, which cause $\sum Con(c_k) = 0$. Based on above analysis, for an ideal situation, there is $Stax = 0$.

6 S-Measure Evaluation

In this section we describe evaluations of p-Measure and S-Measure. First, in Section 6.1, we present an evaluation of p-Measure based on the Miller and Charles test [30]. Section 6.2 contains an analytical evaluation of the efficiency of S-Measure. Subsequently we present some empirical evaluations, where we will compare S-Measure with LF-measure [7], using the PROTON ontology.

6.1 Evaluating p-Measure

The Miller and Charles test is a well-know method for measuring semantic relatedness. In their benchmark dataset there are 30 pairs of words. For each pair of word, a significant number between 0 and 4 of "synonymy judgment" has been assigned by a person. Testing a semantic measure consists of computing the correlation factor (the Pearson-product moment correlation factor) between the Miller and Charles vector and the vector generated by p-Measure.

To make fair comparisons we decided to use an independent software package that will calculate similarity based on WordNet, developed by Pedersen and Michelizzi [31]. The package implements semantic similarity measures described, for example, by Leacock & Chodorow [23], Jiang & Conrath, Resnik, and Wu & Palmer [36]. The coefficient factors of Hirst & St-Onge, Hong-Minh & Smith, Mazuel & Sabouret, Yang & Powers are collected from [16], [17], [22] and [37], respectively.

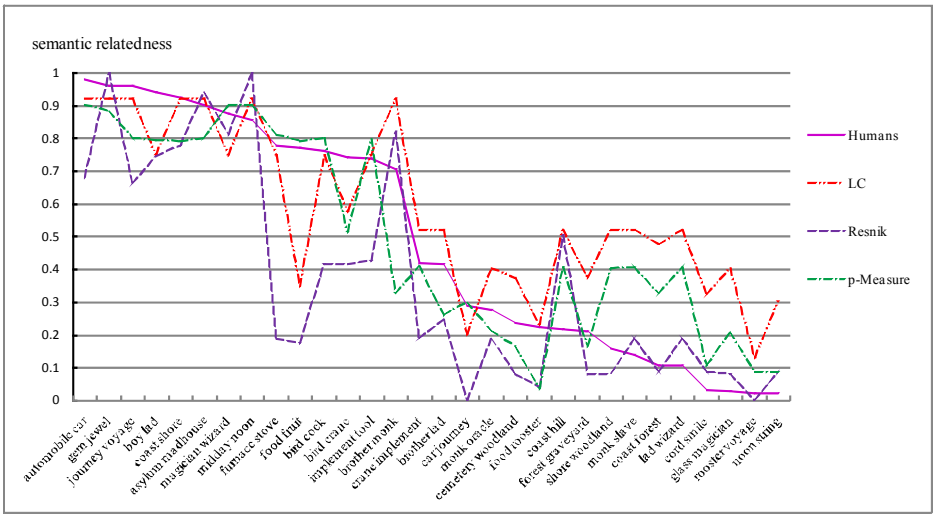

Figure 2 The relatedness values for the 30 pairs dataset obtained by Humans, Leacock and Chodorow, Resnik and p-Measure in the range of [0, 1].

In order to make the visual comparison easier with p-Measure, we scaled the output values obtained by Humans in the range [0, 4], Leacock & Chodorow in the range [0, 4], and Resnik in the range [0, 12] are scaled to [0, 1]. Figure 1 shows a chart with the similarity values obtained by some existing algorithms and the values obtained by p-Measure.

In this experiment, we set $w_{sa} = 0.9$, $w_{hh} = w_{hm} = 0.8$ and $\theta = 20$, respectively. With these parameters, we obtained a p-Measure of 0.882. Compared with other existing algorithms (presented in Table 1), it is clear that the results obtained by p-Measure for pairs of words outperform previous results, with the exception of two measures, Yang & Powers and Mazuel & Sabouret. Note that Yang & Powers propose an algorithm with a number of seven parameters to be fine tuned, while our proposal requires only one parameter. Compared with Mazuel & Sabouret's measure, p-Measure has the virtues of being simple and achieving relatively high accuracy.

Measures	Correlation Coefficient (ρ)
Resnik	0.808
Leacock & Chodorow	0.840
Yang & Powers	0.900
Hirst & St-Onge	0.847
Jiang & Conrath	0.807
Mazuel & Sabouret, $TC_X = 0.4$	0.902
Hong-Minh & Smith	0.88
p-Measure	0.882

Table 1. Pearson product-moment correlation factor for eight different approaches

The correlation factor of Hong-Minh & Smith is close to the correlation factor of p-Measure. Hong-Minh & Smith pays attention to the strength of link type using information theoretic measures while p-Measure distinguishes three types of links and assigns three different values to weight their strengths while carrying information. Both p-Measure and Hong-Minh & Smith use the tuning parameter α, while α=0.3057 is used based on their experiments, we use α=0.528 based on our experiments.

6.2 Analytical Evaluation of S-Measure

The example ontology in Figure 3 will be used to show the effectiveness of S-Measure. The example ontology is based on a cleaned ontology produced by CleOn system and is described in Fig. 7 in [33]. In the example ontology, there are fifteen cleaned concepts that all can be founded in WordNet. In order to show the effectivenss of S-Measure, we add four new concepts with different color to the example ontology. They are Legal agent, Bus, Train and State.

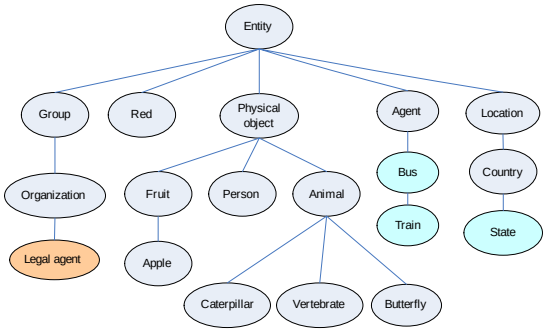

Figure 3 The example ontology Entity.

The analytical evaluation steps are: First, compute the semantic relatedness between each concept pair (father-node, its child-nodes) using the UND algorithm. After this step, we can detect all unclean concepts and break these corresponding links, as a consequence to reduce the successive amount of work. Second, the semantic relatedness of each two child nodes belonging to same father concept in the remaining taxonomy will be computed and the taxonomy consistency values of all non-leaf nodes will be given by the TCN algorithm. Finally, the taxonomy consistency value of the example ontology, *Stax*, will be computed and analyzed. The detailed analytical steps are:

Step 1: Appling the UND algorithm.

In each UND traversal path, we get all pairs of concept nodes, find their shortest path and the types of relationships to measure their semantic relatedness according to Definition 8. For example, the shortest path of the word pair (animal, butterfly) in WordNet is:

{animal, *hyponym* , invertebrate, *hyponym*, arthropod, *hyponym*, insect, *hyponym*, lepidopterous insect, *hyponym*, butterfly}

As discussed in Section 4, we set $w_{sa} = 0.9$, $w_{hh} = w_{hm} = 0.8$, $\alpha = 0.528$ respectively, there is, $rel(animal, butterfly) = 0.3265$. The detailed results are shown in Table 2.

Concept pair		SemanticRelatedness (*rel*)	Relation reasonable threshold (*Th*)
Entity	Group	0.2847	0.1678
Group	Organization	0.5823	0.1678
Entity	Red	0.1580	0.1678
Entity	Physical object	0.5018	0.1678
Physical object	Fruit	0.3011	0.1678
Fruit	Apple	0.6399	0.1678
Physical object	Person	0.3656	0.1678
Physical object	Animal	0.3761	0.1678
Animal	Caterpillar	0.6392	0.1678
Animal	Vertebrate	0.6392	0.1678
Animal	Butterfly	0.3275	0.1678
Entity	Agent	0.2406	0.1678
Agent	Bus	0.1303	0.1678
Bus	Train	0.6377	0.1678
Entity	Location	0.2406	0.1678
Location	Country	0.3976	0.1678
Country	State	0.8736	0.1678

Table 2 Results gained after applying the UND algorithm

After applying the UND algorithm, we can find that three concept nodes are marked as unclean concepts. One is the concept node "Legal agent*"*, which can not be found in WordNet, other two nodes are "Bus*"* and "Red". From the last two columns of Table 2, two semantic relatedness (*rel*s) of two concept pairs, (Entity, Red), (Agent, Bus), are smaller than their corresponding relation reasonable thresholds. According to Definition 4, the two nodes "Red" and "Bus" are unclean concepts, so we mark them and all of their sub-nodes (e.g., Train) as UC nodes. The nodes will not be considered in future measure steps, even though the semantic relatedness value of the concept pair (Bus, Train) is large enough. The remained concept nodes will be contained in clean concepts set (*C\UC)*.

Step 2: Appling the TCN algorithm.

To the taxonomy collection of the node "Entity" AX_{c_1} (i.e., Group, Physical object, Agent, Location), the node "Physical object" AX_{c_2} (they are Fruit, Person, Animal), and the node "Animal" AX_{c_3} (they are Caterpillar, Vertebrate, Butterfly), we first compute their value of semantic relatedness, and then compute their variances of semantic relatedness.

	Group	Physical object	Agent	Location
Group	-	0.3461	0.2925	0.2925
Physical object	-	-	0.4570	0.7141
Agent	-	-	-	0.3863
Location	-	-	-	-

Table 3 Measuring the semantic relatedness between any two nodes in the taxonomy collection of the concept "Entity".

	Fruit	Person	Animal
Fruit	-	0.2036	0.2094
Person	-	-	0.6208
Animal	-	-	-

Table 4 Measuring the semantic relatedness between any two nodes in the taxonomy collection of the concept "Physical object".

	Caterpillar	Vertebrate	Butterfly
Caterpillar	-	0.4095	0.2097
Vertebrate	-	-	0.2097
Butterfly	-	-	-

Table 5 Measuring the semantic relatedness between any two nodes in the taxonomy collection of the concept "Animal".

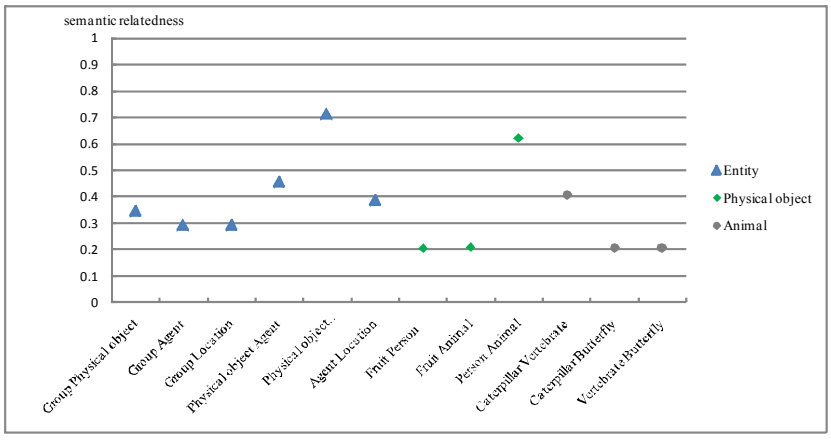

Figure 4 The Semantic relatedness among all child nodes of three concepts "Entity", "Physical object" and "Animal".

As shown in Table 6, *Con*(*Entity*) and *Con*(*Animal*) are smaller than *Con*(*Physical object*), which means the taxonomies of two concepts "Entity" and "Animal" are better than that of the concept "Physical object". Figure 4 shows that the points' distributions of two nodes "Entity" and "Animal" are more centered in a smaller range than that of the node "Physical object". This means that the two

taxonomies of two concepts "Entity" and "Animal" have more taxonomy consistency than that of the concept "Physical object".

Concept (c_k)	Con (c_k)
Entity	0.0254
Physical object	0.0572
Animal	0.0133

Table 6 Measuring the taxonomic consistencies of three concepts ("entity", "physical object" and "animal").

Step 3: Compute the Stax of the example ontology.

According to Definition 7, we can compute the *Stax* of the example ontology. After Step 1, three concepts, *Legal agent, Bus* and *Red*, are marked as unclean nodes. Since there are three non-leaf collection nodes contained in the *C\UC* set, then there is *ccnum*=3. So the Stax of the example ontology is:

$$conWeight(uc) = 3/19 = 0.1579$$
$$conVal(ccnum) = (0.0254 + 0.0572 + 0.0133)/3$$
$$Stax = conWeight(uc) + conVal(ccnum) = 0.1899$$

6.3 Empirical Evaluation of S-Measure

In this section, we compare S-Measure with LF-measure in a real ontology evaluation. The F-measure is often used in conjunction with Precision (P) and Recall (R), as a weighted average of the two. Dellschaft and Staab [7] describe a method for calculating the precision and recall of an ontology O, when compared to a reference ontology O_R. Lexical Precision $LP(O,O_R)$ is defined as:

$$LP(O,O_R) = \frac{|O \cap O_R|}{|O|} = \frac{Number\ of\ concepts\ common\ to\ O\ and\ O_R}{Number\ of\ concepts\ in\ O} \tag{6}$$

whereas the Lexical Recall $LR(O,O_R)$ is defined as:

$$LR(O,O_R) = \frac{|O \cap O_R|}{|O_R|} = \frac{Number\ of\ concepts\ common\ to\ O\ and\ O_R}{Number\ of\ concepts\ in\ O_R} \tag{7}$$

The Lexical F-measure (LF) $LF(O,O_R)$ is used for giving a summarizing overview and for balancing the precision and recall values. The LF-measure is the harmonic mean of *LP* and *LR*.

$$LF(O,O_R) = \frac{2 * LP(O,O_R) * LR(O,O_R)}{LP(O,O_R) + LR(O,O_R)} \tag{8}$$

The corpus used in our experiments is an ontology called PROTON, an acronym for Proto Ontology, which has been created and used in the KIM platform for semantic annotation, indexing, and retrieval [18]. The PROTON ontology consists of around 300 classes and 100 properties. It contains three unique beginners (top level concepts) together with a number of their child concepts, viewed as sub-ontologies. We will use three sub-ontologies in the experiments, namely the *Agent* sub-ontology, the *Object* sub-ontology, and the *Happening* sub-ontology.

The first experiment process is designed as follows: for each experimental dataset, we first compute $LF(O,O_R)$ value by the LF-measure with the reference ontology WordNet, and then compare

its *Stax* value computed using S-Measure. The result of the first experiment is shown in the first three lines of Table 8.

Ontology name	Number of concepts	Maximum depth	Number of unclean concepts
O: Agent sub-ontology	41	7	8
O: Happening sub-ontology	41	7	5
O: Object sub-ontology	38	4	9
$O_{M:}$: Agent sub-ontology	41	7	8
$O_{M:}$: Happening sub-ontology	41	7	5
$O_{M:}$: Object sub-ontology	38	4	9

Table 7 Details of three experimental ontologies.

It is important for an ontology measure approach to evaluate the rationality of each child concept classification in ontology as well as the taxonomy consistency of a whole ontology. For example, if the concept "Apple" appears in the taxonomy of the concept "Car", generally we will view the former concept as an uncleanly defined. Similar, if the concept "Car door" exists in the taxonomy set {Taxi, Ambulance} of "Car", it is also viewed as inappropriate. S-Measure has advantage compared with other approaches in this aspect.

In order to demonstrate that S-Measure considers not only the three kinds of relationships between a parent and all its child concepts but also the rationality of each child node to be classified, we made some modifications to the three experimental ontologies shown in Table 9. The second experiment process is similar to the former experiment. The second experimental results are shown in the last three lines of Table 8.

	$LP(O,O_R)$	$LR(O,O_R)$ *1000	$LF(O,O_R)$ *1000	*conWeight*	*conVAL*	*Stax*(O)
O: Agent sub-ontology	0.4146	0.1446	0.2965	0.1951	0.0172	0.2123
O: Happening sub-ontology	0.6341	0.2211	0.4534	0.1220	0.093	0.215
O: Object sub-ontology	0.5789	0.1871	0.3837	0.2368	0.0175	0.2543
$O_{M:}$: Agent sub-ontology	0.4146	0.1446	0.2965	0.1951	0.0011	0.1962
$O_{M:}$: Happening sub-ontology	0.6341	0.2211	0.4534	0.1220	0.0196	0.1416
$O_{M:}$: Object sub-ontology	0.5789	0.1871	0.3837	0.2368	0.0101	0.2469

Table 8 Evaluation of three experimental ontologies with LF-measure and S-Measure.

Ontologies	Modifications
O_M: Agent sub-ontology	Replace five concepts "bank", "airline", "university", "woman" and "team" with "school", "station", "school", "mouse" and "key" respectively.
O_M: Happening sub-ontology	Exchange the position of the concept "role" and "situation"
O_M: Object sub-ontology	Exchange the position of "account" and "group", replace two concepts "service" and "order" with "weight" and "tree" respectively.

Table 9 The modifications of the three experimental ontologies in the second experiment.

Analysis: Table 8 presents the experimental results when comparing the values of LF-measure and S-Measure for the three experimental ontologies and the three modified experimental ontologies. Compared with the first experiment, it is easy to see that the three *Stax* values for the three experimental ontologies in the second experiment lowered after the modifications. However, the three FL-measure values remained the same. This shows that S-Measure can reflect the differences of taxonomies for sub-concepts of ontologies better than that of LF-measure.

According to three criteria for a good evaluation measure suggested by Dellschaft and Staab [7], our ontology measure has three specific characteristics: (1) S-Measure evaluates ontologies from multiple dimensions, e.g., from the number of unclean concepts and the consistency of sub-trees in ontologies. (2) We also weight different kinds of errors existing in ontologies. This enables better analysis of strengths and weaknesses of an ontology. (3) In S-Measure, the effect of an error is proportional to the depth of a concept in an ontology. For example, an error near the root of a concept hierarchy has a stronger effect on the evaluation compared to an error that occurs nearer to the leaf nodes.

7 Conclusion and Future Work

In this paper, we proposed a new taxonomy consistency measure for ontologies, S-Measure, which computes the semantic relatedness between two concepts existing in the ontology. Computing S-Measure involves three steps: (1) Search and mark all unclean concepts present in an ontology using the UND algorithm, based on p-Measure for two words; (2) Measuring the taxonomy rationality of each non-leaf clean concept existing in the ontology using the TCN algorithm; and (3) Computing *Stax* for the whole ontology, based on former two steps.

Regarding to S-Measure, one of the key issues is the semantic relatedness measure for pairs of words. For this, we propose a new WordNet-based semantic measure, p-Measure. We demonstrated the effectiveness of p-measure, by comparing it with seven other popular measures on a same dataset. Analytical evaluation has shown the detailed procedures of S-Measure for the example ontology, and the experimental evaluations have shown the effectiveness of S-Measure for the PROTON ontology, its three sub-ontologies and three modified experimental ontologies.

We realise that there is still room for improvements to both the suggested similarity measure and the suggested ontology measure approach. In the previous section, we tested with ontologies in English, neigher p-Measure nor S-Measure would work with other languages, because the two suggested measures rely on WordNet. For some words used in ontologies but they are not found in WordNet, they will be viewed as unclean concepts by S-Measure even they are meaningful. This also poses some limitations on our methods: if the ontologies use significantly different set of words as WordNet, the quality of metrics could be poor. This limitation could be alleviated by the use of mappings, but this needs to be further investigated. In our experiments, the sizes of the example ontology and all experimental ontologies were small or medium, we did not test on web-scale.

In our future work we plan to address the following three issues: (1) In this study, we considered concepts that are not terms in WordNet as unclean concepts. However in some commercial or industrial ontologies, some terms are not in WordNet, although they are very popular, such as the term “delivery note” and the term “argan”. For these terms, we will build a search tool which relates some terms to concepts from WordNet by means of mappings; (2) We plan to incorporate more external

knowledge into the p-Measure computation. For example, WordNet glosses are used to measure semantic relatedness between two words, since a gloss contains the functional description of a concept (by means of using other concepts) [1]. Term frequencies in Internet (e.g., Wiki) based approaches are also suggested to measure semantic relatedness, thus some often used words which are not terms in WordNet can be processed also; (3) There are still some other consistency measures for taxonomies that should be considered in S-Measure. As described by Gómez Pérez [10], the evaluation of an ontology includes inspecting the taxonomy of the ontology from three aspects: inconsistency, incompleteness, and redundancy. In this study, we consider only the inconsistency aspect. Therefore, in our future work, we will investigate how to evaluate two further aspects of the taxonomy of an ontology, namely incompleteness and redundancy.

Acknowledgements

This work was supported by a grant from the National High Technology Research and Development Program of China (863 Program) (No. 2009AA01A348), a key grant from the Shanghai Science and Technology Foundation (No. 09dz1500800) and the Opening Project of Shanghai Key Laboratory of Integrate Administration Technologies for Information Security (AGK2010004). Zoltán Miklós was partially supported by the FP7 EU Project NisB (contract no. ICT-256955). The authors wish to thank Mika Timonen and anonymous reviewers for their valuable comments.

References

1. Alvarez, M.A. and Lim, S.J., A Graph Modeling of Semantic Similarity between Words. in: ICSC2007: First IEEE International Conference on Semantic Computing, (Irvine, California, 2007), 355-362.
2. Avesani, P., Giunchiglia, F. and Yatskevich, M., A Large Scale Taxonomy Mapping Evaluation. in ISWC2005: International Semantic Web Conference, Gil, Y. e. a., editor, LNCS 3729, 67-81. Berlin: Springer-Verlag, 2005.
3. Brank, J., Grobelnik, M. and Mladenić, D., A survey of ontology evaluation techniques. in SIKDD 2005, (Ljubljana, Slovenia, 2005), 166-169.
4. Brewester, C., *et al.*, Data driven ontology evaluation. in LREC2004: Proceedings of Int. Conf. on Language Resources and Evaluation, (Lisbon, Portugal, 2004), 26-28.
5. Budanitsky, A. and Hirst, G., Semantic distance in Wordnet: An experimental, application-oriented evaluation of five measures. in Proceedings of the North American chapter of the association for computational linguistics, (Pittsburgh, 2001).
6. Ding, L., *et al.*, Swoogle: A search and metadata engine for the semantic web. in Proc. CIKM2004: ACM Thirteenth Conference on Information and Knowledge Management, (Washington, USA, 2004), 652-659.
7. Dellschaft, K. and Staab, S., On how to perform a gold standard based evaluation of ontology learning. in ISWC2006: Proceedings of the 5th International Semantic Web Conference, (Athens, USA, 2006), 228-241
8. Gangemi, A., Catenacci, C., Ciaramita, M. and Lehmann, J., Modelling Ontology Evaluation and Validation. in ESWC2006: The 3rd Annual European Semantic Web Conference, (Budva, Montenegro, 2006), 140-154.
9. Gómez-Pérez, A. Evaluation of ontologies. International Journal of Intelligent Systems, 2001, 16(3). 391-409.
10. Gómez-Pérez, A., Fernández-López, M. and Corcho, O. Ontological Engineering. Springer, 2004.

11. Gómez-Pérez, A., Some Ideas and Examples to Evaluate Ontologies. in AAAI1995: Proceeding of the 11th Conference on Artificial Intelligence, (Los Angeles, CA, 1995), 299-305.
12. Gruber, T. Ontology. The Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2008.
13. Guarino, N. and Welty, C.A. An overview of OntoClean. In S. Staab and R. Studer, editors, Handbook on Ontologies. Springer Verlag, 2003.
14. Guarino, N. Towards a Formal Evaluation of Ontology Quality. IEEE Intelligent Systems, 2004, 19(4). 78-79.
15. Haase, P. and Stojanovic, L., Consistent evolution of OWL ontologies. in Proceedings of the Second European Semantic Web Conference, (Heraklion, Greece, 2005), vol. 3532 of Lecture Notes in Computer Science, 182-197.
16. Hirst, G. and St-Onge, D. Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum 1998, 305-332.
17. Hong-Minh, T. and Smith, D., Word similarity In WordNet. in Third International Conference on High Performance Scientific Computing, (Hanoi, Vietnam), Bock, H.G., *et al.*, editors. 293-302. Berlin: Springer-Verlag, 2008.
18. http://proton.semanticweb.org/, 2009.
19. http://wordnet.princeton.edu/, 2009.
20. Jiang, J.J. and Conrath, D.W., Semantic similarity based on corpus statistics and lexical taxonomy. in Proc. on International Conference on Research in Computational Linguistics, 1997, 19-33.
21. Qin, P., Lu, Z., Yan, Y. and Wu, F., A New Measure of Word Semantic Similarity based on WordNet Hierarchy and DAG Theory. in WISM2009: 2009 International Conference on Web Information Systems and Mining, (Shanghai, China, 2009), 181-185.
22. Mazuel, L. and Sabouret, N., Semantic relatedness measure using object properties in an ontology. in ISWC2008: 7th International Semantic Web Conference, (Karlsruhe, Germany, 2008), Springer-Verlag, 681-694.
23. Leacock, C. and Chodorow, M. Combining local context and WordNet similarity for word sense identification. MIT Press, London, 1998, 265-283.
24. Lewen, H., Supekar, K., Noy, N.F. and Musen, M.A., Topic-Specific Trust and Open Rating Systems: An Approach for Ontology Evaluation. in EON2006: Proc. of the 4th International Workshop on Evaluation of Ontologies for the Web (EON2006) at the 15th International World Wide Web Conference, (Edinburgh, UK, 2006).
25. Lin, D., An information-theoretic definition of similarity. in Proceedings of the 15th International conference on Machine Learning, (Wisconsin, USA, 1998), 296-304.
26. Lozano-Tello, A. and Gómez-Pérez, A. Ontometric: A method to choose the appropriate ontology. Journal of Database Management, 2004, 15(2). 1-18.
27. Lu, Z., Miklos, Z. Cai, S. and Gu, J., Measuring Taxonomic Consistency of Ontologies Using Lexical Semantic. in ICDIM2001: The Fifth International Conference on Digital Information Management, (Thunder Bay, Canada, 2010), 242-247.
28. Lu, Z., ontoMATCH: A Probabilistic Architecture for Ontology Matching. in ICIS2010: The 3rd International Conference on Information Sciences and Interaction Sciences, (Chengdu, China, 2010), 174-180.
29. Maedche, A. and Staab, S. Measuring similarity between ontologies. in Proceedings of CIKM2002, LNAI vol.2473, 251-263.
30. Miller, G.A. and Charles, W.G. Contextual correlates of semantic similarity. Language and Cognitive Processes, 1991, 6(1).1-28.
31. Pedersen, T. and Michelizzi, J. http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi, 2009.

32. Resnik, P., Using information content to evaluate semantic similarity in a taxonomy. in IJCAI2005: 14th International Joint Conference on Artificial Intelligence, (Edinburgh, Scotland, UK, 2005), 448-453.
33. Reul, Q., Sleeman, D. and Fowler, D. CleOn: Resolution of lexically incoherent concepts in an engineering ontology. Technical report, University of Aberdeen, 2008.
34. Sure, Y., *et al.* Why Evaluate Ontology Technologies? Because It Works! IEEE Intelligent Systems, 2004, 19(4). 74-81.
35. Velardi, P., Navigli, R., Cucchiarelli, A. and Neri, F. Evaluation of OntoLearn: a methodology for automatic learning of domain ontologies. Ontology Learning from Text: Methods, Evaluation and Applications, Buitelaar, P., editors. IOS Press, 2005.
36. Wu, Z. and Palmer, M., Verb Semantics and Lexical Selection. in 32nd Annual Meeting of the Association for Computational Linguistics, (Las Cruses, New Mexico, 1994), 133-138.
37. Yang, D. and Powers, D.M.W., Measuring Semantic Similarity in the Taxonomy of WordNet. in ACSC2005: The Twenty-Eighth Australasian Computer Science Conference, (Newcastle, Australia, 2005), 315-322.