Metric of Intrinsic Information Content for Measuring Semantic Similarity in an Ontology

Md. Hanif Seddiqui **Masaki Aono**

Department of Electronics and Computer Engineering,
Toyohashi University of Technology,
1-1 Hibarigaoka, Tempaku, Toyohashi, Japan,
Email: hanif@kde.ics.tut.ac.jp, aono@ics.tut.ac.jp

Abstract

Measuring information content (IC) from the intrinsic information of an ontology is an important however a formidable task. IC is useful for further measurement of the semantic similarity. Although the state-of-art metrics measure IC, they deal with external knowledge base or intrinsic hyponymy relations only. A current complex form of ontology conceptualizes a class (also often called as a concept) explicitly with the help of the hyponymy classes and the asserted relations and restrictions. Therefore, we propose a modified metric for measuring IC intrinsically taking both the concept-to-concept and the concept-to-property relations. We evaluate our system theoretically and with experimental data. Our evaluation shows the effectiveness of our modified metric for extracting intrinsic information content to measure semantic similarity among concepts in an ontology.

Keywords: Concept, Ontology, Information Content, Semantic Similarity

1 Introduction

"An ontology is an explicit specification of a conceptualization" is a prominent definition by T.R. Gruber in 1995 (Gruber 1995). The definition was then extended by R. Studer et al., in 1998 as *"an ontology is an explicit, formal specification of a shared conceptualization of a domain of interest"* (Studer et al. 1998). Ontology is the backbone to fulfill the semantic web vision (Berners-Lee et al. 1999, Maedche & Staab 2001) and is a knowledge base to enable machines to communicate each other effectively. The knowledge captured in ontologies can be used to annotate data, to distinguish homonyms and polysemies, to drive intelligent user interfaces and even to retrieve new information.

An ontology contains core ontology, axioms or asserted rules, knowledge base and lexicon. Furthermore, core ontology is defined by a set of concepts, a set of properties, concept hierarchy, property hierarchy and functions to relate properties with concepts.

There are usually various size of ontologies, small-scale or large-scale. Large-scale ontologies often represent distributed knowledge area within a problem domain. Ontology segmentation or alignment of large-scale ontologies often requires method of ontology partitioning. In this regard, concept to concept semantic relatedness or similarity measure is necessary. The partition is often performed by the semantic relatedness or similarity measure among concepts of ontology.

The state-of-art metrics by Resnik (Resnik 1999), Lin (Lin 1998), Jiang and Conrath (Jiang & Conrath 1997), and Seco et al. (Seco et al. 2004) used extrinsic or intrinsic information content for semantic similarity measure. Resnik, Lin and Jiang and Cornath used the external source of information content. Although Seco et al. measured the information content within ontology, they used hyponyms of concepts only. They applied their metric to the trivial taxonomy of concepts like WordNet (Miller et al. 1990). However, ontologies, such as those developed by the Web Ontology Language (OWL) (McGuinness et al. 2004), are significantly more complex in data structures than the taxonomy of concepts only.

Our proposed metric of information content extends to take concept, properties and their relations of ontology into account. Therefore, it can be applied in both cases of a simple taxonomy and a complex ontology with concept-properties relations.

The scope of the this work has a well accepted field of ontology partitioning to achieve the scalability. As ontologies grow in size they become more and more difficult to create, use, understand, maintain, transform and classify. Therefore, Stuckenschimidt et. al. (Stuckenschmidt & Klein 2004) and Grau et. al. (Grau et al. 2005*a*,*b*, 2006) focus on partitioning OWL ontologies. Seidenberg and Rector (Seidenberg & Rector 2006) suggested segmentation of gigantic large ontologies to solve the scaling problems. Hu et al. (Hu, Cheng, Zheng, Zhong & Qu 2006, Hu, Zhao & Qu 2006, Hu et al. 2008) proposed partition based block matching for aligning large ontologies. Therefore, ontology partitioning with the help of semantic similarity measurement is necessary in segmentation, aligning large ontologies or obtaining scalability in large ontologies.

This work is to integrate with our scalable and efficient algorithm of ontology alignment called Anchor-Flood algorithm (Seddiqui & Aono 2008), which performs the best running time in the OAEI-2008 campaign.

The rest of the paper is organized as follows. **Section 2** introduces the state-of-art techniques of semantic similarity metrics, while **Section 3** focuses on the ontology structure of semantic web. **Section 4** describes the limitation of the state-of-art metrics. **Section 5** includes the detailed elaboration of our proposed metric. **Section 6** includes experiments and evaluation to show the effectiveness of our proposed metric. Concluded remarks and some future

This study was supported by Global COE Program "Frontiers of Intelligent Sensing" from Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT).

directions of our work is described in **Section 7**.

2 State-of-art Metrics

Semantic similarity metric is an important measure for partitioning a taxonomy of an ontology. A taxonomy is a hierarchical representation of a semantic network with a partial ordering, typically given by the concept inclusion relation (ISA). The concepts in a semantic network or taxonomy of an ontology have proximities among other concepts connected by edges.

There are a number of researches to measure the similarity among the concepts in a semantic network. Of the many approaches presented in the literature (Collins & Quillian 1969, Rada et al. 1989, Knappe et al. 2007, Hirst & St-Onge 1998, Sussna 1993, Wu & Palmer 1994), we divide them into two classes: edge-relative approaches and information theoretic approaches. Edge-related approaches focus on counting the edges or pre-assigned weight of edges, while information theoretic approaches analyze the essence of taxonomy of ontology. Each of them are described below for understanding the content of this paper easily.

2.1 Edge relative Approaches

Rada et al. (Rada et al. 1989) assumes that the similarity is proportional to the number of edges separating concepts. Sussna et al. (Sussna 1993) introduces a depth-relative scaling approach, based on the observation that siblings deep in the tree are more closely than siblings higher in the tree. Wu and Palmer (Wu & Palmer 1994) define their conceptual similarity based on the principle of depth-relative scaling as:

$$sim(c_i, c_j) = \frac{2 * depth(c_{ij})}{depth(c_i) + depth(c_j)}, \qquad (1)$$

where c_{ij} is the common super class of c_i and c_j, and $depth(c_k)$ gets the depth of c_k in the original class hierarchy.

In Eq. 1 each edge has the weight of unity regardless of their direction in the subsumption relation of a taxonomy. On the contrary, some proposed distance metrics use different weights of a particular edge for differentiating their direction in subsumption. A taxonomy of an ontology contains a network of a directed graph. This means that the edge between two concepts represents a relationship in the direction of the edge. In Fig. 1, a *dog* ISA *animal* and not the other way around. When we move in the direction of edge, we get generalization, whereas we obtain specialization while moving in an opposite direction of the orientation. Therefore, concept inclusion (ISA) intuitively implies strong similarity in the opposite direction from inclusion (specialization). In addition, the direction of inclusion (generalization) must contribute some degree of affinity. However, for the same reasons as in the case of specializations, transitive generalizations should contribute a decreased degree of similarity.

To make the edge influence the similarity, weight factor δ and γ is introduced for expressing similarity of immediate specialization and generalization respectively. The similarity function is then defined as:

$$sim_{wsp}(x, y) = \max_{j=1...m} \left\{ \sigma^{s(P_j)} \gamma^{g(P_j)} \right\}, \qquad (2)$$

where P_1 ... P_m are all paths connecting x and y. P_k is an edge, $s(P_j)$ is the number of edges toward specialization and $g(P_j)$ is the number of edges toward generalization.

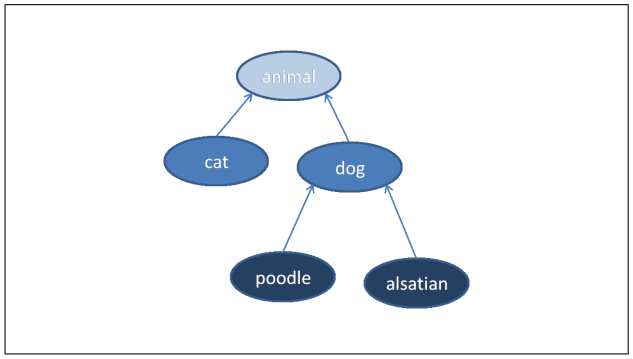

Figure 1: An example of a pet ontology

This similarity can be derived from an ontology by transforming the ontology into a directional weighted graph with σ as a downward and γ as a upward weights, and the similarity is the product of the weight on the path 2. An example ontology is displayed in Fig. 2.

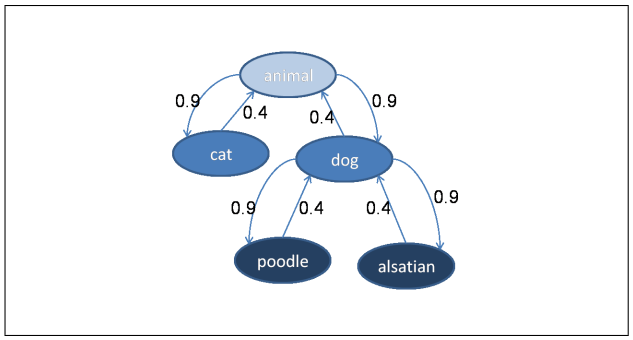

Figure 2: The ontology transformed into directed weighted graph, with the immediate specialization and generalization similarity values σ = 0.9 and γ = 0.4 respectively. Similarity is derived as the maximal (multiplicative) weighted path length, and thus $sim(poodle, alsatian)$ = 0.4*0.9 = 0.36.

The edge between concepts in a taxonomy represents some degree of affinity. However the edge weights are defined by different factors. The deeper concepts in a taxonomy has more specification than the shallower concepts in the taxonomy. A taxonomy is organized by a directed network. The direction in a network has a influence over the affinity. Defining the weight of an edge to measure the affinity of concept pair depends on the factors. Therefore, several researchers focus on the information theoretic approaches to define the weights to measure semantic affinities between concepts by introducing IC. Measuring IC removes the complexity of assigning dynamic weights to every edges and of considering direction of edges as well. Information theoretic approaches are discussed in more details below.

2.2 Information Theoretic Approaches

Information theoretic approaches are well defined in a couple of research works by (Jiang & Conrath 1997, Lin 1998, Resnik 1995, Seco et al. 2004). They obtain their needed IC values by statistically analyzing corpora. They associate probabilities to each concept in the taxonomy based on word occurrences in a given corpus. These probabilities are cumulative as we go up the taxonomy from specific concepts to more abstract concepts. The IC value is then obtained by considering the negative log likelihood (Resnik 1995, 1999):

$$ic_{res}(c) = -log\ p(c) \quad (3)$$

where c is any concept in WordNet and $p(c)$ is the probability of encountering c in a given corpus. It should be noted that this method ensures that IC is monotonically decreasing as we move from the leaves of the taxonomy to its roots. (Resnik 1995) was the first to consider the use of this formula, that stems from the work of Shannon (Shannon & Weaver 1948), for the purpose of semantic similarity judgments. The basic intuition behind the use of the negative likelihood is that the more probable a concept is of appearing then the less information it conveys, in other words, infrequent words are more informative than frequent ones. Resnik describes an implementation in using WordNet's (Miller et al. 1990) taxonomy of noun concepts. According to Resnik, semantic similarity depends on the amount of information two concepts have in common, this shared information is given by the most specific common abstraction i.e. the super-concept that subsumes both concepts. In order to find a quantitative value of shared information we must first discover the super-concept, if one does not exist then the two concepts are maximally dissimilar, otherwise the shared information is equal to the IC value of the super-concept. Formally, semantic similarity is defined as:

$$sim_{res}(c_1, c_2) = \max_{c \in S(c_1,c_2)} \{ic_{res}(c)\}, \quad (4)$$

where $S(c_1, c_2)$ are the set of concepts that subsume c_1 and c_2.

Lin (Lin 1998) proposes a modification of measuring the semantic similarity using Resnik's equation defined in Eq. 3 and 4. He states that the similarity between concepts c_1 and c_2 is measured by the ratio between the amount of information needed to state the commonality of c_1 and c_2 and the information needed to fully describe what c_1 and c_2 are. The definition might be expressed as:

$$sim_{lin}(c_1, c_2) = \frac{2 * sim_{res}(c_1, c_2)}{ic_{res}(c_1) + ic_{res}(c_2)} \quad (5)$$

Jiang and Conrath (Jiang & Conrath 1997) proposes a combined model taking the shortest path, edge-counting methods, Resnik's information content into account and by adding decision factors. They calculate the weights between two concepts of parent and child relation.

$$wt(c_c, c_p) = \left(\beta + (1-\beta)\frac{\overline{E}}{E(c_p)}\right)\left(\frac{d(c_p)+1}{d(c_p)}\right)^{\alpha} [ic_{res}(c_c) - ic_{res}(c_p)]\, T(c_c, c_p) \quad (6)$$

where $d(c_p)$ is the depth of the node (concept) c_p, $E(c_p)$ is the number of children of c_p, the local density $(\overline{E})$ is the average density in the entire taxonomy, and $T(c_c, c_p)$ is the link relation/type factor. The parameters α ($\alpha \geq 0$) and β ($0 \leq \beta \leq 1$) control the influence of node depth and density, respectively.

If we consider the case where node depth (as we will consider node depth indirectly by considering the number of children a particular node contains) is ignored and link type and local density both have a weight of 1. In this special case, the dissimilarity metric is:

$$dist_{jcn}(c_1, c_2) = (ic_{res}(c_1) + ic_{res}(c_2)) - 2 * sim_{res}(c_1, c_2) \quad (7)$$

2.3 Intrinsic Information Content Metric

The classical way of measuring IC of concepts combines knowledge of their hierarchical structure from an ontology with the statistics on their actual usage in text as derived from a large corpus. However, Seco et al. (Seco et al. 2004) derived a wholly intrinsic measure of IC that relies on hierarchical structure alone and applied their derivation to large taxonomy of WordNet. They report a competitive correlation value between human and machine similarity judgment on the dataset of Miller and Charles (Miller & Charles 1991) against WordNet.

Seco et al. argue that the more hyponyms a concept has the less information it expresses, otherwise there would be no need to further differentiate it. Likewise, concepts, that are leaf nodes, are the most specified in the taxonomy so the information they express is maximal. Formally they define:

$$ic_{seco}(c) = 1 - \frac{log(hypo(c) + 1)}{log(max_{wn})} \quad (8)$$

where the function *hypo* returns the number of hyponyms of a given concept and max_{wn} is a constant that is set to the maximum number of concepts that exists in the taxonomy.

Seco's IC decreases monotonically as we transverse from leaf to root. The information content of the imaginary top node of WordNet would yield an IC value of 0. This metric gives the same score to all leaf nodes in the taxonomy regardless of their overall depth. All leaves have same maximum value 1.

Like Resnik's and Lin's measure, Seko's metric of semantic similarity yields result in $[0, .., 1]$. Seco et. al. formulated their metric as:

$$sim_{seco}(c_1, c_2) = 1 - \frac{ic_{seco}(c_1) + ic_{seco}(c_2) - 2 * sim'_{res}(c_1, c_2)}{2}, \quad (9)$$

where sim'_{res} corresponds to Resnik's similarity function but accommodating Seco's IC values.

3 Ontology in Semantic Web

Apart from the general discussion about measuring IC against WordNet which contains a complete list of concepts, the definition of a domain ontology would reveal the fact about the current complex form of ontology of semantic technology. According to (Ehrig 2007), ontology contains core ontology, logical mapping, knowledge base, and lexicon. Furthermore, a core ontology, S, is defined by a tuple of five entities as follows:

$$S = (C, \leq_C, R, \sigma, \leq_R),$$

where C and R are two disjoint sets called "concepts" and "relations" respectively. A relation is known as a property of a concept, or a restriction on a property about a concept. Throughout the paper, we use the term "relation" to represent property of a concept, or a restriction on a property about a concept. A function represented by $\sigma(r) =< dom(r), ran(r) >$ where $r \epsilon R$, domain is $dom(r)$ and range is $ran(r)$. A partial order $\leq_R$ represents on R, called relation hierarchy, where $r_1 \leq_R r_2$ iff $dom(r_1) \leq_C dom(r_2)$ and $ran(r_1) \leq_C ran(r_2)$. The notation $\leq_C$ represents a partial order on C, called concept hierarchy or "taxonomy". In a taxonomy, if $c_1 <_C c_2$ for $c_1, c_2 \epsilon C$,

then c_1 is a sub-concept of c_2, and c_2 is a super-concept of c_1. If $c_1 <_C c_2$ and there is no $c_3 \epsilon C$ with $c_1 <_C c_3 <_C c_2$, then c_1 is a direct sub-concept of c_2, and c_2 is a direct super-concept of c_1 denoted by $c_1 \prec c_2$ (Ehrig 2007).

Ontology relationship among its concepts is defined by their taxonomy, where concepts are maintained in a hierarchical or super or sub-concept organization along with the properties and restrictions. Therefore properties along with the restrictions are sometimes referred to as relations. They usually play important roles to define the relationships among concepts. Relations of an ontology are differentiating our metric from the other.

4 Limitation of the State-of-art Metrics

The metrics we describe so far are used to measure the semantic similarity among concepts where concepts are organized in a hierarchy or a taxonomy. They are only considering the concept to concept relation, i.e. the metrics consider only super-concept and sub-concept organization. However description logic (DL) based domain ontology of semantic technology usually contains properties, restrictions and other complex relations in addition to the trivial taxonomy or concept hierarchy.

WordNet does not heavily depend on the properties, rather it has a complete list of concepts to define another concept. As it is a thesaurus, we can have a large text corpora having connection with WordNet. Unlike WordNet, description logic based domain ontology only focuses on a particular domain of interest. It is seldom complete by its concepts alone as it may contain a limited number of concepts of one's interest. Different ontologies of a particular domain might widely be different and influenced by its targeted users and the knowledge of the its developers. Moreover, it has seldom large text corpora to define its concepts. On the contrary, DL ontology has an explicit specification to define a concept not only by the concept alone, but also with the help of the other concepts, its properties and restrictions and the other logical assertions available inside the ontology.

The classical metrics of measuring semantic similarity often use the available concepts, a large text corpora or a large complete hierarchy for using the hyponymy relations to measure the IC of a concept. These metrics cannot be used against domain ontologies at their current states. However, semantic relatedness measure plays an important role in ontology for partitioning or segmenting large ontologies for resolving scalability issues.

5 Proposed Modification in IC Metric

To overcome the limitation of the state-of-art metrics of computing semantic similarity among concepts within a domain ontology and to cope with the new ontologies with the introduced complex description logics, we propose a modified metric of computing intrinsic information content. The metric can be applied to a simple taxonomy and to a recent complex OWL ontology as well.

The primary source of IC in ontology is obviously concepts and concept hierarchy. However, OWL ontology also contains properties, restrictions and other logical assertions, often called as relations. Properties are used to define functionality of a concept explicitly to specify a meaning. They are related to concept by means of domain, range and restrictions.

According to Resnik, semantic similarity depends on the shared information. As Resnik introduces the IC which represents the expressiveness of a particular concept. Classical metric of IC are based on the available concepts in a taxonomy or in a large text corpora. However, as time passes on, the definition and the content of ontology becomes more and more complex. The expressiveness of a concept is not only rely on the concept taxonomy but also on the other relations like properties and property-restrictions. Fig. 3 shows an example ontology of concepts *Biblo*, *Institution*, *Reference*, *InProceedings*, *Article*, *School* and *Publisher* with the support of 37 different properties which is playing an important role to define concepts and distinguish a concept from the others. Let us consider the concepts *Biblo* and *Institution*, where they are sharing no property and although *Institution* has only three properties to specify its meaning and is distinguish from the root. However, it does not contain other properties to specify its children more concisely. They have the only difference in their hyponym or subsumption relation. On the other hand, concepts *Biblo* and *Reference* has no common property as well. However *Reference* is expressed with 21 properties. Concept *Reference* is defined concisely and specifically with 21 properties. The fact is that *Reference* and *Institution* are subsumed by a concept *Biblo* though, *Reference* has more expressiveness than *Institution*. Therefore, information content of *Reference* is larger than that of concept *Institution*. In the figure, *Reference*, *Article* and *InProceedings* has close semantic relatedness. Likewise, *Institution*, *Publisher* and *School* has semantic relatedness. However, *Institution* and *Reference* has weak semantic relatedness because of their less sharing information. Therefore, we can consider that the information content or expressiveness of a concept is directly proportional to the number of properties it is related to by means of property functions or property restrictions.

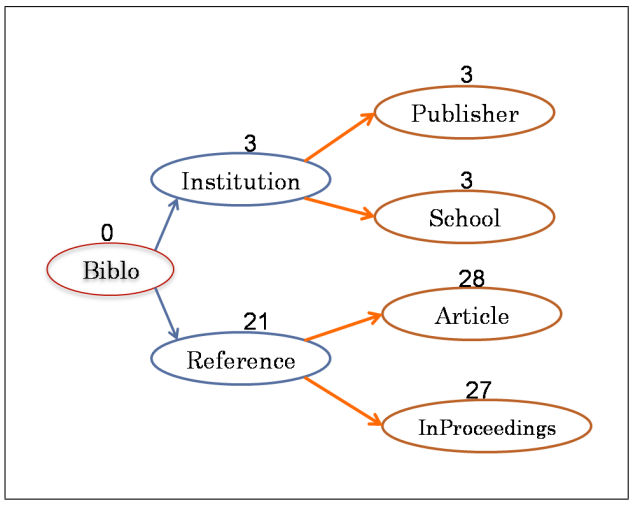

Figure 3: Illustration of the impact of properties on semantic similarity.

5.1 Proposed IC Metric

We already have discussed about the probable sources of IC or the expressiveness of semantic similarity among the concepts of ontology. We find that the IC of a concept is negatively related to the probability of a concept in an external large text corpora (Resnik 1995). We also find that the IC of a concept is inversely related to the number of hyponyms or the concepts it subsumes (Seco et al. 2004). Moreover, we observe that description logic (DL) based ontol-

ogy of semantic technology is formal and explicit in its conceptualization with the help of relations. Every concept is defined with sufficient semantic embedding with the organization, property functions, property restrictions and other logical assertions. Current ontology of semantic technology is defined as "an explicit specification of a conceptualization" (Gruber 1995). Although the most domain ontologies are not as complete as WordNet in terms of concepts and concept organization, they have well support from logical assertions to define a concept concisely. Therefore, we can obtain sufficient IC of a concept without depending on the external large text corpora heavily, required that we use intrinsic information of the concept. One of the good source of intrinsic information of a concept is its relations by means of property functions and property restrictions. Our relation based IC is defined as:

$$ic_{rel}(c) = \frac{log(rel(c)+1)}{log(total\_rel+1)}, \tag{10}$$

where *"rel"* stands for the relation of properties, property function and restrictions, $rel(c)$ denotes the number of relations of a concept c and $total\_rel$ represents the total number of relations available in the ontology.

As long as the information content of a concept depends both on the hyponyms or subsumption relations of a concept and the related properties of the concept, we need to integrate the $ic_{re}(c)$ with the Seco's metric of intrinsic information content defined in Eq. 8. This integration introduces a coefficient factor ρ and the equation becomes as:

$$ic(c) = \rho \cdot ic_{rel}(c) + (1-\rho) \cdot ic_{seco}(c), \tag{11}$$

where the coefficient factor ρ is defined by the nature of an ontology. While a small size of ontology is often incomplete by its concepts alone, the coefficient factor tends to increase to focus on relations. On the contrary, when relations are inadequate to define a concept and there are a large number of concepts in the taxonomy, ρ tends to decrease its value. However, we definitely need a trade-off to select the coefficient factor and we define it as:

$$\rho = \frac{log(total\_rel+1}{log(total\_rel) + log(total\_concept)},$$

where $total\_rel$ is the maximum number of relations while $total\_concepts$ is the maximum number of concepts available in an ontology.

Moreover, the semantic similarity is computed as described in Eq. 9 by replacing ic_{seco} with ic defined in Eq. 11 and the sim'_{res} is measured by our modified information content metric. Then the semantic similarity $sim_{proposed}$ is defined as below:

$$sim_{proposed}(c_1, c_2) = 1 - \frac{ic(c_1) + ic(c_2) - 2 * sim'_{res}(c_1, c_2)}{2}, \tag{12}$$

6 Experiments and Evaluation

For experiment with our modified metric of IC, we obtained a reference ontology of a benchmarks from the Ontology Alignment Evaluation Initiative (OAEI), 2009[1] displayed in Fig. 4. We can download the ontology from its homepage[2].

Table 1: contains IC values measured by Seco's metric and our modified metric.

Concepts	Number of Relations	Number of Hyponyms	ic_{seco}	ic_{rel}	$ic_{modified}$
Date	3	0	1.000	0.332	0.641
PageRange	2	0	1.000	0.263	0.603
Organization	0	3	0.613	0.000	0.283
Institution	3	2	0.693	0.332	0.499
Publisher	3	0	1.000	0.332	0.641
School	3	0	1.000	0.332	0.641
List	0	1	0.807	0.000	0.373
PersonList	4	0	1.000	0.386	0.670
Journal	7	0	1.000	0.498	0.730
Address	3	0	1.000	0.332	0.641
Person	0	0	1.000	0.000	0.462
Conference	6	0	1.000	0.466	0.713
Reference	21	23	0.113	0.740	0.450
Academic	24	2	0.693	0.771	0.735
PhdThesis	26	0	1.000	0.790	0.887
MastersThesis	26	0	1.000	0.790	0.887
Misc	22	0	1.000	0.751	0.866
MotionPicture	22	0	1.000	0.751	0.866
Part	23	5	0.500	0.761	0.640
InCollection	26	0	1.000	0.790	0.887
InProceedings	27	0	1.000	0.798	0.891
Article	28	0	1.000	0.807	0.896
Chapter	26	0	1.000	0.790	0.887
InBook	27	0	1.000	0.798	0.891
Report	25	2	0.693	0.781	0.740
TechReport	25	0	1.000	0.781	0.882
Deliverable	25	0	1.000	0.781	0.882
Informal	21	4	0.551	0.740	0.653
Manual	23	0	1.000	0.761	0.871
Unpublished	23	0	1.000	0.761	0.871
Booklet	23	0	1.000	0.761	0.871
LectureNotes	22	0	1.000	0.751	0.866
Book	27	3	0.613	0.798	0.713
Collection	31	0	1.000	0.830	0.909
Monograph	30	0	1.000	0.823	0.905
Proceedings	34	0	1.000	0.852	0.920

As the Fig. 4 displays the concepts associated with its number of relations, we obtained the results displayed in Table 1 by using Eq. 8, 10 and 11. From the Table 1 and observing the Fig. 4, we see that hyponyms of "Reference" concepts has a close value of information content. Therefore, they are semantically close related and the other concepts are not as close as a group. However, the values of ic_{seco} in the Table 1 do not reveal a group of semantically related concepts because of their scatteredness. Moreover, Seco's IC values of all the leaves are unity regardless of their depth and position in ontology, which is certainly misleading to a group of semantically related concepts in an ontology.

From the experiments, we also observe that the deeper concepts have more expressiveness or larger IC values. Therefore, it guarantees that our modified IC metric takes the depth of a concept implicitly and the children of a concept explicitly. However, we do not take the link type and local concept density into account unlike expressed in (Jiang & Conrath 1997). As we consider the hyponyms by incorporating the Seco's IC metric, it consider the edges between subsumption concepts implicitly.

Using Table 1 we produce the semantic similarity between $Reference$ to each of its leaves considering similarity equation proposed by Seco et al. in Eq. 9 and proposed by us in Eq. 12. The semantic similarity is displayed in Table 2. We observe that our semantic similarity is close to the real proximity. $Article$, $Chapter$ and so on has close semantic relationship with $Reference$ in the domain of bibliography.

Furthermore, we also compute semantic similarity for every possible pair of concepts of the ontology depicted in Fig. 4 and derive only the semantically related groups or blocks of concepts using both equation. Our proposed method produces $Reference$ with its 23 children as a unique block with another small block of containing $Institution$, $Publisher$ and $School$. Therefore, our proposed metric produce two blocks containing 24 and 3 elements respectively. On

[1] http://oaei.ontologymatching.org/2009/

[2] http://oaei.ontologymatching.org/2009/benchmarks/101/onto.rdf

Figure 4: It shows a taxonomy of an ontology where each concept is associated with its number of relations.

the other hand, according to Eq. 9, we get several number of small blocks even among the children of *Reference* concept. It produces six small blocks having 4, 5, 3, 6, 3, and 4 elements. We also produce two real blocks manually by the domain experts with 24 and 4 elements. We evaluate the blocks by well-known Purity, Inverse-purity and F-measure metrics which rely on precision, recall and f-measure ideas of information retrieval.

Let B be a set of computed blocks ($|B| = n$) and R be a set of manual blocks produced by experts ($|R| = m$). b_i denotes a block in B, while r_j denotes a block in R. $|b_i|$ returns the number of entities in b_i , and $|r_j|$ is defined analogously. $b_i \cap r_j$ calculates the common entities in both b_i and r_j. N be the total number of blocked items. Then, Purity is computed by taking the weighted average of maximal precision values:

$$Purity(B) = \Sigma_{i=1}^{n} \frac{|b_i|}{N} max_j \ Precision(b_i, r_j), \quad (13)$$

where the Precision of a block b_i against a real block r_j is defined as:

$$prec(b_i, r_j) = \frac{|b_i \cap r_j|}{|b_i|}$$

Inverse-purity focuses on the block with maximum recall against a real block and is defined as:

$$Inverse{-}purity(B) = \Sigma_{j=1}^{m} \frac{|r_j|}{N} max_i \ Precision(r_j, b_i), \quad (14)$$

where, $Precision(r_j, b_i)$ is also called as $Recall(b_i, r_j)$.

However, a more robust evaluation metric can be obtained by combining the ideas of Purity and Inverse-purity, called as F-measure and defined as:

$$F - measure(B) = \Sigma_{j=1}^{m} \frac{|r_j|}{N} max_i \left\{F(r_j, b_i)\right\}, \quad (15)$$

where

$$F(r_j, b_i) = \frac{2 \times Recall(r_j, b_i) \times Precision(r_j, b_i)}{Recall(r_j, b_i) + Precision(r_j, b_i)}$$

Now, we apply the Precision, Recall and F value against the blocks derived by both of the metrics and the results are displayed in Table 3 and Table 4 to represent the model of Seco et al. and our proposed model respectively.

Table 2: contains semantic similarity between *Reference* to each of its leaves considering Seco's metric and our proposed metric.

e_1	e_2	sim_{seco}	$sim_{proposed}$
Reference	PhDThesis	0.113	0.782
Reference	MastersThesis	0.113	0.782
Reference	InCollection	0.113	0.782
Reference	InProceedings	0.113	0.780
Reference	Article	0.113	0.777
Reference	Chapter	0.113	0.782
Reference	InBook	0.113	0.780
Reference	TechReport	0.113	0.784
Reference	Deliverable	0.113	0.784
Reference	Manual	0.113	0.790
Reference	Unpublished	0.113	0.790
Reference	Booklet	0.113	0.790
Reference	LectureNotes	0.113	0.792
Reference	Collection	0.113	0.771
Reference	Monograph	0.113	0.773
Reference	Proceedings	0.113	0.765

Table 3: Precision, Recall and F value for the suggested blocks by Eq. 9

$\|Block\|$	$\|Ref.Block\|$	Precision	Recall	F
4	24	1.000	0.167	0.286
5	24	1.000	0.208	0.344
3	24	1.000	0.125	0.222
6	24	1.000	0.250	0.400
3	24	1.000	0.125	0.222
4	4	1.000	1.000	1.000

We summarize Purity, Inverse-purity and F-measure for both of the metrics in Table 5. The Table 5 depicts that our proposed metric outperforms in the given example. Finally, we can state that our proposed modified metric for intrinsic information content works efficiently provided that the ontology concepts are specified by their properties. Once the ontology does not contain any properties, it behaves just as a Seco's model.

7 Conclusions and Future Works

In this paper, we describe the modified metric of information content (IC) that would be applicable to both of the domain ontologies of semantic technology and the simple however complete taxonomy like WordNet as well. Our proposed IC metric can be used to measure the semantic similarity among the concepts of an ontology regardless of its complex structure. In our modified metric we consider both the concept hyponyms or subsumptions and the relations as well with a coefficient factor of logarithmic ratio to combine them. We implemented this metric of IC to measure the semantic similarity among concepts. Ontology processing algorithms like ontology alignment and ontology segmentation can get benefits by detecting semantically related blocks of an ontology with the semantic similarity measure.

Table 4: Precision, Recall and F value for the suggested blocks by our proposed equation

$\|Block\|$	$\|Ref.Block\|$	Precision	Recall	F
24	24	1.000	1.000	1.000
3	4	1.000	0.750	0.857

Table 5: Purity, Inverse-purity and F-measure for the blocks suggested by Eq. 9 and by our proposed metric

Metric	Purity	Inverse-purity	F-measure
Seco	1.000	0.318	0.424
Proposed	1.000	0.970	0.984

We experimented with a small however representative ontology to observe the trend of our metric and we obtained more accurate semantic similarity among concepts and eventually we detected semantically related blocks of concepts. We measured the Purity, Inverse-purity and F-measure for our proposed metric and we observed that our proposed metric outperformed against Seco's model of intrinsic information content on ontology, where concepts are specified with properties. The semantic groups of closely related concepts can be further used in ontology segmentation and in large scale ontology alignment. The experiment shows that our modified IC metric works better.

Our future plan is to experiment more with large DL ontologies. Furthermore, we have a plan to integrate our modified metric with our fastest and scalable Anchor-Flood algorithm (Seddiqui & Aono 2008) to align large scale ontologies and to retrieve segmented alignment.

References

Berners-Lee, T., Fischetti, M. & Dertouzos, M. (1999), *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*, Harper San Francisco.

Collins, A. & Quillian, M. (1969), 'Retrieval time from semantic memory', *Journal of Verbal Learning and Verbal Behavior* **8**, 240–247.

Ehrig, M. (2007), *Ontology Alignment: Bridging the Semantic Gap*, Springer, New York.

Grau, B., Parsia, B. & Sirin, E. (2006), 'Combining OWL ontologies using E-connections', *Web Semantics: Science, Services and Agents on the World Wide Web* **4**(1), 40–59.

Grau, B., Parsia, B., Sirin, E. & Kalyanpur, A. (2005*a*), 'Automatic Partitioning of OWL Ontologies Using ε-Connections', *Proceedings of the International Workshop on Description Logics (DL'05), Edinburgh, Scotland* .

Grau, B., Parsia, B., Sirin, E. & Kalyanpur, A. (2005*b*), 'Modularizing OWL ontologies', *Proc. KCAP-2005 Workshop on Ontology Management, Banff, Canada* .

Gruber, T. (1995), 'Toward Principles for the Design of Ontologies Used for Knowledge Sharing', *International Journal of Human-Computer Studies* **43**(5/6), 907–928.

Hirst, G. & St-Onge, D. (1998), 'Lexical chains as representations of context for the detection and correction of malapropisms', *WordNet: An electronic lexical database* pp. 305–332.

Hu, W., Cheng, G., Zheng, D., Zhong, X. & Qu, Y. (2006), 'The Results of Falcon-AO in the OAEI 2006 Campaign', *Proceedings of Ontology Matching (OM-2006), Athens, Georgia, USA* pp. 124–133.

Hu, W., Qu, Y. & Cheng, G. (2008), 'Matching Large Ontologies: A Divide-and-Conquer Approach', *Journal of Data & Knowledge Engineering* **67**(1), 140–160.

Hu, W., Zhao, Y. & Qu, Y. (2006), 'Partition-based Block Matching of Large Class Hierarchies', *Proceedings of the 1st Asian Semantic Web Conference (ASWC2006), Beijing, China* pp. 72–83.

Jiang, J. & Conrath, D. (1997), 'Semantic similarity based on corpus statistics and lexical taxonomy', *Proceedings on International Conference on Research in Computational Linguistics, Taiwan* pp. 19–33.

Knappe, R., Bulskov, H. & Andreasen, T. (2007), 'Perspectives on ontology-based querying', *International Journal of Intelligent Systems* **22**(7).

Lin, D. (1998), 'An information-theoretic definition of similarity', pp. 296–304.

Maedche, A. & Staab, S. (2001), 'Ontology Learning for Semantic Web', *IEEE Intelligent Systems* **16**(2), 72–79.

McGuinness, D., Van Harmelen, F. et al. (2004), 'Owl web ontology language overview', *W3C recommendation, http://www. w3. org/TR/owl-features* **10**.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. (1990), 'WordNet: An on-line lexical database', *International journal of lexicography* **3**(4), 235–312.

Miller, G. & Charles, W. (1991), 'Contextual Correlates of Semantic Similarity.', *Language and cognitive processes* **6**(1), 1–28.

Rada, R., Mili, H., Bicknell, E. & Blettner, M. (1989), 'Development and application of a metric on semantic nets', *IEEE transactions on systems, man and cybernetics* **19**(1), 17–30.

Resnik, P. (1995), 'Using information content to evaluate semantic similarity in a taxonomy', *Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada* pp. 448–453.

Resnik, P. (1999), 'Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language', *Journal of artificial intelligence* pp. 95–130.

Seco, N., Veale, T. & Hayes, J. (2004), 'An intrinsic information content metric for semantic similarity in WordNet', *Proceedings of ECAIf2004, the 16th European Conference on Artificial Intelligence* **16**, 1089–1090.

Seddiqui, M. H. & Aono, M. (2008), 'Alignment Results of Anchor-Flood Algorithm for OAEI-2008', *Proceedings of Ontology Matching Workshop of the 7th International Semantic Web Conference, Karlsruhe, Germany* pp. 120–127.

Seidenberg, J. & Rector, A. (2006), 'Web Ontology Segmentation: Analysis, Classification and Use', *Proceedings of the 15th International Conference on World Wide Web (WWW2006), Edinburgh, Scotland* pp. 13–22.

Shannon, C. & Weaver, W. (1948), 'A mathematical theory of communication', *Bell System Technical Journal* **27**, 379–423.

Stuckenschmidt, H. & Klein, M. (2004), 'Structure-based Partitioning of Large Concept Hierarchies', *Proceedings of the 3rd International Semantic Web Conference (ISWC2004), Hiroshima, Japan* pp. 289–303.

Studer, R., Benjamins, V. & Fensel, D. (1998), 'Knowledge Engineering: Principles and Methods', *Journal of Data & Knowledge Engineering* **25**(1-2), 161–197.

Sussna, M. (1993), Word sense disambiguation for free-text indexing using a massive semantic network, *in* 'Proceedings of the second international conference on Information and knowledge management', ACM New York, NY, USA, pp. 67–74.

Wu, Z. & Palmer, M. (1994), Verb semantics and lexical selection, *in* 'Proceedings of the 32nd annual meeting of the Association for Computational Linguistics', Las Cruces, New Mexico, pp. 133–138.