A new model to compute the Information Content of concepts from taxonomic knowledge

David Sánchez*
Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Spain

Montserrat Batet
Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Spain

ABSTRACT

The Information Content (IC) of a concept quantifies the amount of information it provides when appearing in a context. In the past, IC used to be computed as a function of concept appearance probabilities in corpora, but corpora-dependency and data sparseness hampered results. Recently, some authors tried to overcome previous approaches, estimating IC from the knowledge modeled in an ontology. In this paper, we develop this idea, by proposing a new model to compute the IC of a concept exploiting the taxonomic knowledge modeled in an ontology. In comparison with related works, our proposal aims to better capture semantic evidences found in the ontology. To test our approach, we have applied it to well-known semantic similarity measures, which were evaluated using standard benchmarks. Results show that the use of our model produces, in most cases, more accurate similarity estimations than related works.

Keywords: Computational linguistics, Knowledge management, Semantic similarity, Information Content, Ontologies.

INTRODUCTION

The growth of the so-called Information Society has produced an enormous amount of textual electronic information. Contrary to numerical data, which can be directly managed by means of mathematical operators, the coherent interpretation of textual data is challenging. Since textual resources consist of words or noun phrases, which refer to concepts that represent their meanings, tools and techniques are needed to enable a *semantic* interpretation of textual terms.

Within the area of semantics, Information Theory has played an important role. Many authors (Jiang & Conrath, 1997; Lin, 1998; Pirrò, 2009; Resnik, 1995; Sánchez & Batet, 2011; Sánchez, Batet, Valls, & Gibert, 2010) have relied on the notion of Information Content (IC) of term conceptualizations to quantify term semantics. Since the IC of a concept states the amount of information provided by the associated term, it provides a numerical measure that enables quantitative comparisons of terms, according to their underlying semantics. The assessment of *semantic similarity* is one of the most interesting applications: it quantifies the taxonomic resemblance between compared terms (e.g. *wrestling* and *boxing* are similar because both represent concepts that are specializations of *contact sports*). Since semantic similarity is a fundamental principle by which humans organize objects, it has been extensively used in the past in a variety of natural language applications such as automatic spelling error detection and correction (Budanitsky & Hirst, 2001), word-sense disambiguation (Patwardhan, Banerjee, & Pedersen, 2003), question answering (Tapeh & Rahgozar, 2008) or synonym detection (Lin, 1998). In the field of knowledge management, assessment of semantic similarity has aided in tasks such as information extraction (Sánchez & Isern, 2011; Sánchez, Isern, & Millán, 2011), thesauri generation (Curran, 2002; Hwang & Kim, 2009), knowledge integration (Shakya, Takeda, & Wuwongse, 2009), information retrieval (Fanizzi, d'Amato, & Esposito, 2009; Hliaoutakis, Varelas, Voutsakis, Petrakis, & Milios, 2006) or ontology learning (Ramezani, 2011; Sánchez, 2010). Practical applications, such as privacy-preservation of textual data (Martínez, Sánchez, & Valls, 2012; Martínez, Sánchez, Valls, & Batet, 2012), can also be found.

From the variety of similarity paradigms available in the literature, IC-based measures are some of the most accurate and commonly used ones (Formica, 2008; Jiang & Conrath, 1997; Li, Bandar, & McLean, 2003; Resnik, 1995). These calculate the degree of resemblance between terms, as a function of the amount shared and non-shared information (i.e. IC) of their conceptualizations. Hence, a proper estimation of IC of concepts corresponding to the compared terms is crucial to enable accurate similarity assessments.

In the past, authors used to base the IC calculus on concept appearance probabilities in textual corpora (Resnik, 1995). Corpora-dependency introduced a series of problems such as data sparseness, when limited corpora was available, or scalability issues, due to the need of manual pre-processing of text (Sánchez, et al., 2010). To overcome these issues, several authors have proposed to base the IC calculus on the knowledge modeled in an ontology (Sánchez, Batet, & Isern, 2011; Seco, Veale, & Hayes, 2004; Zhou, Wang, & Gu, 2008). In these works, the knowledge structure that ontologies provide have been exploited to estimate the degree of generality/concreteness of concept and, hence, to quantify their IC.

This paper reviews related works based on this last principle and proposes a new ontology-based model to estimate the IC of concepts. Compared to state of the art methods, our proposal aims to better capture the semantic features modeled in the ontology for evaluated concepts, enabling a more accurate quantification of their meaning. Our method has been applied to three classic IC-based similarity measures (Jiang & Conrath, 1997; Lin, 1998; Resnik, 1995), and it has been compared against other IC calculus models for different benchmarks. Similarity results have also been compared with other ontology-based paradigms of similarity computation (i.e. edge-counting and feature-based). Evaluation figures show that, on the one hand, our model improves similarity assessments of other IC calculus methods based either on corpora or on ontologies; on the other hand, our model enables similarity assessments that are more accurate than those achieved by other paradigms of similarity computation.

The rest of the paper is organized as follows. The second section reviews IC calculus models. The third section presents and formalizes our proposal. The fourth section details evaluation results, by describing the benchmarks, similarity measures and knowledge bases used to test and compare our method. A discussion of these results is also provided. The final section contains the conclusions and proposes some lines of future work.

INFORMATION CONTENT AND CONCEPT SEMANTICS

Resnik was the first to apply the definition of IC in a semantic context. He computed the IC of a concept *c* as the inverse of its probability of occurrence in a corpus (Resnik, 1995):

$$IC(c) = -\log p(c) \qquad (1)$$

According to the definition of taxonomic subsumption, the probability of concepts should monotonically increase as one goes up the taxonomy. As a consequence, the IC of concepts increases as these become more specific in the taxonomy. The rule $\forall c_1, c_2 \mid c_1$ *is hypernym of* $c_2 \Rightarrow IC(c_1) \leq IC(c_2)$ should be fulfilled.

To produce coherent IC calculus, the probability estimation should be coherent with the taxonomic structure of the ontology. To achieve this, Resnik counted the frequency of appearance of ontological concepts in a corpus so that each noun occurring in the corpus was counted as an occurrence of all taxonomic concepts subsuming it.

$$p(c) = \frac{\sum_{w \in W(c)} count(w)}{N}, \qquad (2)$$

where $W(c)$ is the set of terms in the corpus whose senses are subsumed by c, and N is the total number of corpus terms that are contained in the taxonomy.

This calculus is coherent to the *positiveness* and *additiveness* of the IC definition (Ross, 1976). This states that, if the occurrence of a variable (concept c) is composed by several independent occurrences of other variables (taxonomic specialization of c), the information content of the first is the sum of the information content of each individual variable.

Corpora-based IC calculus was widely used in the past (Jiang & Conrath, 1997; Lin, 1998; Pedersen, Pakhomov, Patwardhan, & Chute, 2007; Sánchez, et al., 2010). However, accurate estimation of concept probabilities from textual corpora is challenging. First, words appearing in corpora should be associated to their corresponding concepts. This requires manual semantic disambiguation for ambiguous words, which hampers the scalability of the method (Sánchez, et al., 2010). Moreover, large and suitable corpora are needed to obtain robust probability assessments and to avoid data sparseness. Authors have used large and general purpose corpora such as the Brown corpus together with WordNet (Fellbaum, 1998) as ontology. In recent years, wider corpora like the Web have been employed to minimize data sparseness (Sánchez, et al., 2010). In domain specific contexts (biomedicine), more concrete domain corpora (such as clinical trials) are needed to obtain accurate results (Pedersen, et al., 2007). The availability of suitable corpora also compromises the IC calculus (Sánchez, Batet, et al., 2011).

To overcome these limitations, in recent years, some authors have proposed models to estimate the IC of a concept intrinsically, according to its degree of specialization in an ontology (Sánchez, Batet, et al., 2011; Seco, et al., 2004; Zhou, et al., 2008). In most basic works, concept probabilities are approximated as the ratio between the number of concept hyponyms (*hypo(c)*) in the ontology with respect to the total amount of ontological concepts (*max_nodes*) (Seco, et al., 2004).

$$IC_{Seco\ et\ al.}(c) = 1 - \frac{\log(hypo(c)+1)}{\log(max\_nodes)} \qquad (3)$$

This is based on the principle of cognitive saliency (Pirrò, 2009): new taxonomic specializations for a concept are introduced when it is necessary to differentiate them from already existing ones. Hence, concepts with many hyponyms would refer to general entities, which provides less information than concepts located at the leaves of the hierarchy.

With Seco et al. approach, concepts with the same number of hyponyms are considered equally informative, even though they may present different degrees of taxonomic generality. As stated in works framed in the semantic similarity field (Wu & Palmer, 1994), the relative depth of concepts in the taxonomy should be considered as a relevant semantic feature. In (Zhou, et al., 2008), the depth of concept *c* in the taxonomy, *depth(c)*, is incorporated in the IC assessment.

$$IC_{Zhou\ et\ al.}(a) = k\left(1 - \frac{\log(hypo(a)+1)}{\log(max\_nodes)}\right) + (1-k)\left(\frac{\log(depth(a))}{\log(max\_depth)}\right) \quad (4)$$

Inversely to *hypo(c)*, as *depth(c)* increases, *c* becomes more specialized and its IC increases as consequence. The constant *max_depth* corresponds to the maximum depth of the taxonomy, whereas *k* is a tuning parameter that weighs the contribution of the two evaluated features (number of hyponyms and depth). As a drawback, *k* should be tuned by the user according to the background ontology and the input data in order to optimize the results.

Several modifications of the above measures were proposed in (Sánchez, Batet, et al., 2011). Firstly, instead of relying on the amount of hyponyms of a concept, only the number of taxonomic leaves was evaluated. In this manner, the dependence on the inner taxonomic structure, whose characteristics (granularity and branching factor) may vary for different ontologies (Sánchez, Batet, et al., 2011), is avoided. The second modification is based on considering the degree of concreteness of the evaluated concepts during the IC assessment, as well. Instead of using the depth as a measure of concreteness, the whole set of taxonomic subsumers was evaluated, considering, if available, multiple inheritance relationships. In this manner, additional taxonomic evidences were captured. The proposed measure was defined as follows:

$$IC_{Sanchez\ et\ al.}(c) = -\log p(c) \cong -\log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|}+1}{max\_leaves+1}\right), \quad (5)$$

where *leaves(c)* is the set of concepts at the end of the taxonomic tree under concept *c*, and *subsumers(c)* is the set of taxonomic ancestors of *c* including itself. The ratio is normalized by the least informative concept, that is, the root node, for which the number of leaves is the total amount of leaves in the taxonomy (*max_leaves*), and the number of subsumers including itself is 1. To produce values in the range 0..1 and to avoid *log(0)* values, 1 is added to both expressions.

A NEW MODEL FOR ONTOLOGY-BASED IC ASSESSMENT

In this section, we present a new model to compute the IC of a concept from the knowledge modeled in an ontology. We use similar knowledge features (taxonomic subsumers and leaves) to those used in related works, but we evaluate them in a different manner. We build our proposal on a redefinition of the notion of IC as a function of ontological knowledge, similarly to what Resnik did for corpora. Our goal is to better quantify the differences between concepts.

To achieve this, we evaluate *i) inner taxonomic concepts* and *ii) taxonomic leaves* in different ways. We consider that semantics of the former (and hence their IC) are a function of the semantics provided by their subsumed leaves. Consequently, the crucial point is the accurate assessment of the IC of leaves.

As stated in the previous section, an acknowledged issue of Seco et al.'s approach is the fact that all leaves have the same IC value. Authors assume that the IC of an ontological concept is maximal when it is not further specialized. The lack of a proper differentiation between the IC of leaves hampered their results (Sánchez, Batet, et al., 2011). This contrasts to Resnik's approach, in which concepts are evaluated according to their appearance probabilities. Hence, leaves can be distinguished according to their degree of *commonness*. For example, given *soccer* and *Greco-Roman wrestling*, which are leaves subsumed by the concept *sport*, the former refers to a commonly used concept that will be more likely to appear in a discourse than the latter one. Hence, in the context of information theory, an occurrence of *soccer* will provide less information than an occurrence of *Greco-Roman wrestling*. Moreover, Resnik's approach also considered instances of concepts to compute their IC (see eq. 2), which enables a finer-grained differentiation between leaves.

In the ontological domain, the fact that some leaves have more concrete meanings than others may depend on several factors. For example, the criterion of knowledge engineer or the knowledge modeling bottleneck that characterizes manual approaches, may result in some taxonomies more specialized than others. Moreover, even widely used ontologies like WordNet rarely incorporate detailed sets of conceptual instances that could be considered during the evaluation of leaves.

To mitigate these problems, we propose a strategy to evaluate the degree of *commonness* of each leaf. We assume that leaves with many subsumers have more concrete meanings than others with smaller amounts. We sustain this hypothesis in the notion of taxonomic specialization. That is, the meaning of a leaf is a strict subset of the meaning of its taxonomic ancestors, which is progressively constrained as more specialization steps (and, hence, subsumers) are introduced. Moreover, in case of multiple inheritance, in which a leaf is subsumed by ancestors of different taxonomic trees, the meaning of a leaf will be even more constrained, as a result of the intersection of the semantic features shared by its subsumers. In both cases, the more the subsumers (either vertically arranged in successive specialization steps or horizontally organized in case of multiple inheritance), the more concrete the meaning of the leaf will potentially be. This principle is also exploited by feature-base similarity measures (Maedche & Staab, 2002; Sánchez, Batet, Isern, & Valls, 2012). They base the similarity assessment on the number of common subsumers between compared terms; that is, the more common subsumers, the more similar the concepts are because they share a larger amount of semantic features. In (D. Sánchez, et al., 2012), a practical evaluation showed that this criterion highly correlated with human judgments of similarity.

Given the above arguments, this paper considers leaves with many subsumers, that is, those with very concrete meanings, as *less common* than those with smaller amounts, that is, those with more general meanings. The evaluation of the complete amount of subsumers of a leaf is preferable to the evaluation of its relative depth like in (Zhou, et al., 2008). Since the latter corresponds to the length of the minimum path between the root node and the leaf, it may omit an amount of taxonomic knowledge (i.e., other subsumers which belong to different paths in cases of multiple inheritance), which produces less accurate results (Montserrat Batet, Sánchez, & Valls, 2011). With our approach, all subsumers modeled in the ontology are considered, which aids to differentiate leaves with the same depth but different number of ancestors (i.e. different degree of meaning concreteness).

Formally, we define the *commonness* of leaves as a function of their *subsumers*, as follows:

Definition 1. Let the concept subsumption (<) be a binary relation < : *C*×*C*, where *C* is the set of concepts in the ontology, and *l* <*s* means that *l* is a hierarchical specialization of *s*. We define the set of *subsumers* of a leaf *l* as:

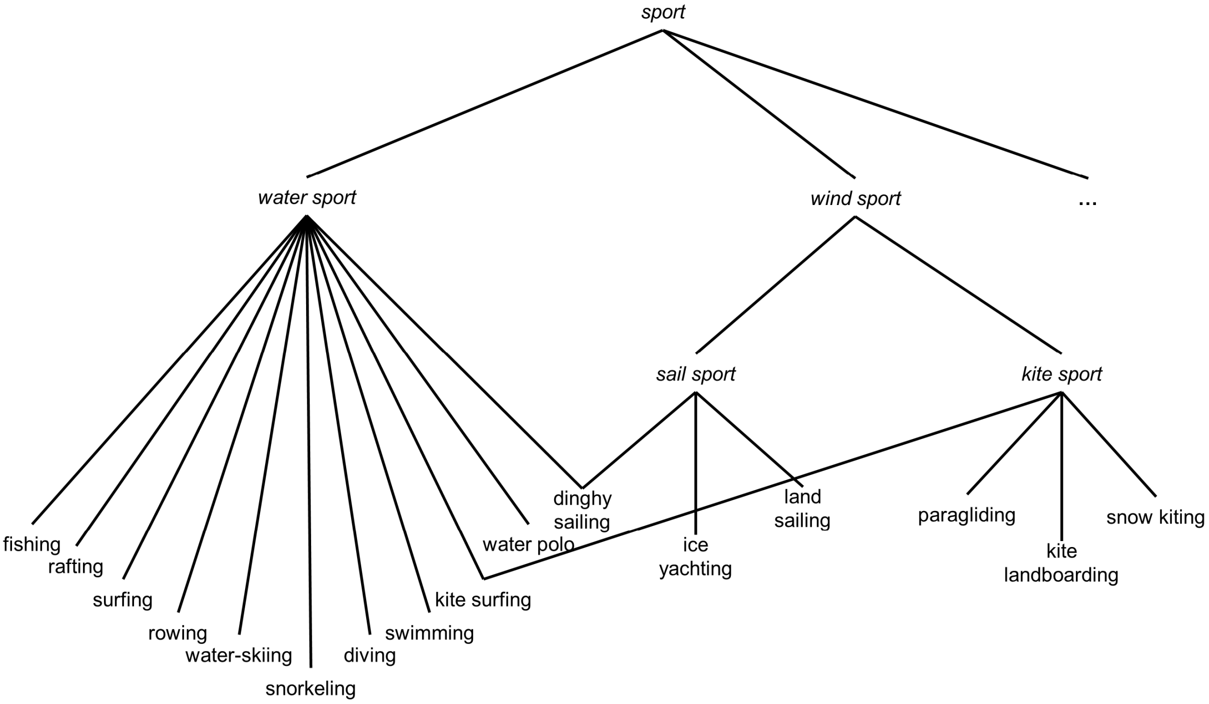
$$subsumers(l) = \{s \in C \mid l < s\} \cup \{l\} \qquad (6)$$

The fact of including the leaf itself in the set assumes the notion of dominance as a reflexive relation in the subsumption context (Partee, ter Meulen, & Wall, 1990) and avoids zero values in the subsequent calculations.

Definition 2. The *commonness* of a leaf *l* is inversely proportional to its number of subsumers:

$$Commonness(l) = \frac{1}{|subsumers(l)|} \qquad (7)$$

Figure 1. Ontology example.



Example 1. Given the sample ontology of Figure 1, the commonness of leaves of *water sport* (except *kite surfing* and *dinghy sailing*) is 0.33, because the cardinality of their sets of subsumers is three; for example, given that *subsumers(swimming)*={*sport, water sport, swimming*), we obtain *commonness(swimming)*=1/3=0.33. Likewise, the commonness of leaves of *wind sport* (except *kite surfing* and *dinghy sailing*) is 0.25, because the cardinality of their sets of subsumers is four; for example, given that *subsumers(paragliding)*={*sport, wind sport, kite sport,*

paragliding}, we obtain *commonness(paragliding)*=1/4=0.25. Finally, the commonness of *kite surfing* and *dinghy sailing* is 0.2, because that the cardinality of their subsumer sets (considering multiple inheritance) is five; for example, given that *subsumers(kite surfing)*={*sport, water sport, wind sport, kite sport, kite surfing*}, we obtain *commonness (kite surfing)*=1/5=0.2. One can see how the larger the number of subsumers is (either vertically or horizontally arranged), the lower the commonness values will be, which enables an accurate differentiation of leaves.

Once the commonness of leaves can be computed, as introduced above, we evaluate inner taxonomic concepts as a function of their leaves. We follow a parallel reasoning to Resnik's approach: where he computed the probability of a concept as the aggregation of appearances of all its specializations in corpora (eq. 2), we propose to compute the commonness of an inner taxonomic concept by adding the commonness of all its specializations. This is coherent both to the *additiveness* property of IC and to the notion of taxonomic generalization, in which the meaning of a taxonomic ancestor subsumes all meanings of its specializations. Hence, the commonness of an inner concept will be the sum of the commonness of its leaves, as follows.

Definition 3. The *commonness* of an inner taxonomic concept *s* (i.e., *s* is not a leaf) is computed as the sum of the commonness values of all the leaves subsumed by *s*:

$$Commonness(s) = \sum_{\forall l | l \text{ is a leaf} \wedge l < s} commonness(l) \tag{8}$$

We hypothesize that the evaluation of the commonness of an inner concept as a function of its leaves will result in more accurate estimations than its evaluation as a function of its subsumers. This is because, as one moves up in the hierarchy, the number of subsumers decrease. Hence, very general ones, for example, direct specializations of the root node, will become indistinguishable (i.e. all of them have a unique ancestor). Even though one may consider that concepts modeled at the same level of abstraction are equally informative, it turns out that concepts (especially the most abstract ones) rarely present identical levels of informativeness. In WordNet, for example, there is a significant variation in the number of hyponyms subsumed by concepts located at the same level of abstraction, with a proportion up to 40:1 (Devitt & Vogel, 2004). As argued by several authors (Sánchez, Batet, et al., 2011; Seco, et al., 2004), one can assume that a conceptual abstraction with 40 times more hyponyms than another one is likely to appear more frequently (and hence, to provide less IC), since it can be referred by means of all its taxonomical specializations. In these cases, basing the IC calculus on the amount of concept subsumers will be too coarse. On the contrary, by considering that leaves of an ontology frame and define in a detailed manner the scope and boundaries of the modeled domain (Sánchez, Batet, et al., 2011), the evaluation of the set of leaves of a general concept provides a more detailed quantification of its semantics. This is because its meaning is the result of the subsumption of all of its specializations. As a result, general concepts with the same amount of subsumers can be more accurately differentiated.

Applying definition 3 to inner concepts of Example 1, we obtain *commonness(water sport)*=9×1/3+2×1/5=3.4, *commonness(wind sport)*=5×1/4+2×1/5=1.65, *commonness(kite sport)*=3×1/4+1/5=0.95 and *commonness(sail sport)*=2×1/4+1/5=0.7. We observe how the larger the number of leaves under an inner concept and the higher their commonness values are, the higher the commonness of the inner concept is. We also observe how concepts located at the same level of abstraction are differentiated according to their subsumed leaves.

Finally, given the above definitions, the IC of a concept *c* (leaf or not) is computed as the inverse of its commonness. We assume that the more general the meaning of a concept is (i.e. the higher the commonness), the less information it provides when appearing in a context. Formally:

Definition 4. The IC of a concept *c* is defined as:

$$IC(c) = -\log \frac{commonness(c)}{commonness(root)} \qquad (9)$$

Note that, to produce values in the 0..1 range, the absolute value of commonness is divided by the value of the most general, and hence, less informative concept: the root node. In this manner, IC is normalized according to the size of the ontology, which enables a fair comparison of IC values computed from different ontologies. This also permits coherent similarity calculus of concept pairs belonging to different ontologies, since IC-based similarity measures quantify the resemblance of concepts as a function of their IC and those of their common ancestors (Jiang & Conrath, 1997; Lin, 1998). For example, we can fairly compare IC-based similarity assessments obtained from ontologies of different sizes, or produce coherent similarity estimations when each of the compared concepts belongs to a different ontology. In this last case, a methodology to integrate ontological knowledge can be used (M. Batet, Sánchez, Valls, & Gibert, in press; Sánchez, Solé-Ribalta, Batet, & Serratosa, 2012).

By applying definition 4 to example 1 and computing the commonness of the root node by considering solely (for simplicity) the leaves shown in Fig.1 (i.e., (*commonness(root)*= 9×1/3+5×1/4+2×1/5=4.65)), the IC of inner concepts is: *IC(water sport)*=-log(3.4/4.65)=0.45, *IC(wind sport)*= -log(1.65/4.65)= 1.49, *IC(kite sport)*=-log(0.95/4.65)=2.29 and *IC(sail sport)*=-log(0.7/4.65)=2.73. Coherently to commonness values obtained above, general inner nodes which subsume a large number of leaves with higher commonness values (like *water sport*) result in low IC values. By computing IC values for leaves we obtain, for example, *IC(swimming)*= -log(0.33/4.65)=3.82, *IC(paragliding)*= -log(0.25/4.65)=4.22 or *IC (kite surfing)*= -log(0.2/4.65)=4.54. In this case, the larger the number of subsumers is, the higher the IC value is.

Compared with related works, our approach provides some theoretical advantages. First, similarly to Seco et al., a concept is evaluated according to its specializations. However, our approach gives less importance to the inner taxonomic detail as only the commonness of leaves is considered. As stated in (Sánchez, Batet, et al., 2011), this is desirable to lesser the influence of the inner taxonomic granularity, a dimension that closely depends on the criterion of the knowledge expert and that may vary from one ontology to another, and even for different taxonomic trees of the same ontology. In addition, we weight the contribution of each leaf in the IC of a subsumer concept according to its assessed commonness. This provides a more accurate quantification and differentiation of concept semantics. Contrary to (Zhou, et al., 2008), this is done without depending on manually tuned parameters, which may hamper the generality of the approach. As in (Sánchez, Batet, et al., 2011), and contrary to (Zhou, et al., 2008), the commonness of leaves relies on the whole set of subsumers rather than only on the relative depth (i.e. minimum path to the root node). As a result, our approach considers more taxonomic knowledge in the assessment. Finally, the differentiated calculus of leaves and inner concepts commonness enable an accurate quantification of their semantics, which, in each case, is assessed according to the knowledge structure that gives more detailed evidences; that is, the number of subsumers for leaves and the aggregation of commonness of leaves for subsumer concepts.

In order to provide coherent IC values according to the background ontology, as stated in the second section, those must monotonically increase as one moves down in the taxonomy. This property is proven for our proposal below.

Proposition 1. The function of eq. 8 monotonically increases according to concept specializations in the taxonomy.

Proof.

Let *C* be the set of concepts of the ontology, and *leaves(c)* be the set of hyponyms of *c* that are not further specialized, that is $leaves(c) = \{l \in C \mid l < c \wedge \neg\exists h \mid h < l\}$. Then, $\forall\, c_1, c_2 \in C \mid c_1 < c_2 \Rightarrow leaves(c_1) \subseteq leaves(c_2) \Rightarrow commonnes(c_1) \leq commonnes(c_2)$. That is, due to the fact that c_1 is a specialization of c_2, *leaves(*c_1*)* will necessarily be a subset of *leaves(*c_2*)*; hence, the commonness of the first (computed from *leaves(*c_1*)*) will necessarily be lower or equal to the commonness of the second (computed from *leaves(*c_2*)*). In consequence, as one moves down in the hierarchy, the final IC value, which is the inverted logarithm of the commonness values, monotonically increases.

EVALUATION

In order to test the accuracy of the proposed IC calculus in comparison with related works, we have evaluated them by means of classical IC-based similarity measures and several benchmarks, which are detailed below.

Similarity measures based on Information Content

As stated in the introduction, the notion of IC has been used to develop semantic similarity measures. Resnik's seminal work (Resnik, 1995) proposes a similarity measure which relies relies solely on the amount of information shared by the compared terms (10). Such common information was estimated as the IC of the most specific common ancestor subsuming both terms in an ontology, that is, the Least Common Subsumer (LCS).

$$sim_{res}(c_1,c_2) = IC(LCS(c_1,c_2)) \tag{10}$$

The fact of relying only on a common abstraction makes concept pairs which shares the same LCS indistinguishable from point of view of the semantic similarity. This problem was solved by (Lin, 1998) and (Jiang & Conrath, 1997) incorporated the IC of the compared concepts in the assessment by using different similarity coefficients.

$$sim_{lin}(c_1,c_2) = \frac{2 \times IC(LCS(c_1,c_2))}{(IC(c_1)+IC(c_2))} \tag{11}$$

$$sim_{j\&c}(c_1,c_2) = IC(LCS(c_1,c_2)) - (IC(c_1)+IC(c_2)) \tag{12}$$

It is important to note that similarity measures are usually applied to words extracted from text. Since the IC calculus is performed at a conceptual level, conceptualizations modeled in the ontology, which correspond to the compared words, should be identified. Since some of the evaluated words may be polysemous, several concepts (i.e, one per word sense) may be retrieved. A proper disambiguation of input terms can solve ambiguity, by assigning input words to unique ontological concepts. However, as stated in the introduction, semantic similarity is usually applied to non-disambiguated inputs (Budanitsky & Hirst, 2006), which is the case of the benchmarks used for evaluation purposes.

In order to enable a fair comparison of results, we reproduced the behavior of previous works which deal with polysemic words by *i)* retrieving all ontological concepts corresponding to all senses of each word, *ii)* computing individual similarities for each combination of concepts, and *iii)* selecting the maximum similarity value as the final result. The rationale for this criterion is that, given two non-disambiguated words, human subjects pay more attention to their similarities, that is, their most related senses (Tversky, 1977). Formally, given two potentially polysemous words (w_1, w_2), their similarity is computed as:

$$sim_{XXX}(w_1, w_2) = \max_{(i,j)}(sim_{XXX}(c_{1i}, c_{2j})), \quad (13)$$

where c_{1i} and c_{2j} are each of the concepts retrieved from the ontology (i.e. one per word sense).

In the following tests, WordNet version 2 has been used as ontology, which is the usual version used in related works. The JWNL Java API (http://sourceforge.net/projects/jwordnet/) has been used to access WordNet concepts and to explore their taxonomical trees. Since these can be completely pre-loaded in memory in a Hash table, and their taxonomic links can be explored efficiently by using memory pointers, the runtime for intrinsic IC calculus models (ours and those of related works) is in the order of milliseconds for each concept on a 2.6Ghz Intel Core CPU.

Results

In order to evaluate the accuracy of the similarity measures for different IC computation models mentioned above, we followed the usual procedure of comparing to the extent to which the computerized similarity ratings correlate with human judgments. Correlation values for a given measure will provide a quantification of the accuracy of the IC calculus, which enables an objective comparison against other approaches.

In order to enable reproducible and comparable evaluations, several benchmarks of word pairs with human-based similarity assessments have been proposed (Agirre et al., 2009; Miller & Charles, 1991; Rubenstein & Goodenough, 1965). From these, the Miller and Charles benchmark is the most commonly used (Jiang & Conrath, 1997; Lin, 1998; Patwardhan & Pedersen, 2006; Petrakis, Varelas, Hliaoutakis, & Raftopoulou, 2006; Resnik, 1995; Seco, et al., 2004). It consists of 30 English noun pairs with averaged similarity values provided by 38 students in a scale from 0 (unrelated) to 4 (synonyms) (Miller & Charles, 1991). Evaluation results for this benchmark, the above similarity measures, and IC computation models proposed by related works are already reported in the literature, which enables a direct and objective comparison with our own model (see Table 1). First, we have taken correlation results reported in Patwardhan and Pedersen (Patwardhan & Pedersen, 2006), in which Resnik's, Lin's and Jiang and Conrath's measures were evaluated by using a corpora-based IC calculus. Results are summarized in the first three rows of Table 1. Regarding intrinsic IC computation models (Seco et al., Zhou et al. and Sanchez et al.), correlation values have been taken from (Seco, et al., 2004) and (Sánchez, Batet, et al., 2011); see rows 4 to 6, 7 to 9 and 10 to 12, respectively. Results obtained by the IC calculus model proposed in this paper, which were obtained under the same conditions, are reported in rows 13 to 15 of Table 1.

Table 1. Correlation values for IC-based similarity measures for the Miller and Charles benchmark. From left to right: type of IC calculus, similarity measure, reference in which the measure/method was proposed, correlation against human judgments and reference in which correlations were reported. ***Bold*** *lines represent our method.*

IC Calculus	Measure	Proposed in	M&C	Evaluated in
	Resnik	(Resnik, 1995)	0.72	(Patwardhan & Pedersen, 2006)
Corpora-based	Lin	(Lin, 1998)	0.7	(Patwardhan & Pedersen, 2006)
(Resnik, eq.1)	J&C	(Jiang & Conrath, 1997)	0.73	(Patwardhan & Pedersen, 2006)
	Resnik	(Seco, et al., 2004)	0.77	(Seco, et al., 2004)
Intrinsic	Lin	(Seco, et al., 2004)	0.81	(Seco, et al., 2004)
(Seco et al., eq. 3)	J&C	(Seco, et al., 2004)	0.84	(Seco, et al., 2004)
	Resnik	(Zhou, et al., 2008)	0.82	(Sánchez, et al., 2010)
Intrinsic	Lin	(Zhou, et al., 2008)	0.82	(Sánchez, et al., 2010)
(Zhou et al, eq. 4)	J&C	(Zhou, et al., 2008)	0.82	(Sánchez, et al., 2010)
	Resnik	(Sánchez, et al., 2010)	0.84	(Sánchez, et al., 2010)
Intrinsic	Lin	(Sánchez, et al., 2010)	0.85	(Sánchez, et al., 2010)
(Sánchez et al., eq. 5)	J&C	(Sánchez, et al., 2010)	0.87	(Sánchez, et al., 2010)
	Resnik	**This work**	**0.83**	**This work**
Intrinsic	**Lin**	**This work**	**0.86**	**This work**
(this work, eq. 9)	**J&C**	**This work**	**0.89**	**This work**

First of all, it is noticed that corpora-based IC calculus provides lower correlations than intrinsic versions (0.7-0.73 vs. 0.77-0.89). As stated in the second section, the accuracy of corpora-based methods is hampered by the adequacy and availability of data and on the appropriate word tagging so that concept usage can be properly estimated. Instead, intrinsic IC calculus relies on smaller but better structured knowledge sources: ontologies. As a result, they are scalable and easily applicable and, as shown in the results, they better mimic human judgments of similarity.

Regarding the comparison of intrinsic IC calculus models, results show what was expected from the theoretical reasoning discussed in the third section. The approach by Seco et al. provides high baseline results which are progressively improved as more knowledge is considered during the IC assessment. Particularly, Zhou et al., who incorporated the notion of depth as a weighted feature into the calculus, better differentiated concepts with the same amount of hyponyms. Note that the weight value recommended by authors (k=0.5) was used in this test. The approach by Sanchez et al. provides even more accurate results, thanks to the exploitation of additional taxonomic knowledge (subsumers instead of relative depth).

The model presented in this paper provides the highest correlation in most cases. We believe that this is motivated by the more accurate assessment of concept generality/concreteness, as a function of their subsumers (in case of leaf nodes) or of their hyponyms (in case of inner taxonomic nodes). Precisely, the differentiation of both types of concepts better distinguishes them because the most detailed taxonomic feature (leaves or subsumers) is evaluated in each case.

To further evaluate the influence of this design decision, we compared how each of these dimensions (i.e., number of hyponyms and number of subsumers) correlated *individually* against human assessments. On the one hand, computing IC as the *inverse function of the number of hyponyms of a concept* is precisely what Seco et al. proposes (eq. 3). Results are shown in the first three rows of Table 2, which are obviously identical to those reported in Table 1. On the other hand, we computed IC as a *direct function of the number of subsumers of a concept*, normalized by the maximum number of possible subsumers, as follows:

$$IC_{subsumers}(c) = \log\left(\frac{|\,subsumers(c)\,|}{max\_subsumers}\right) \tag{14}$$

Results are reported in rows 4th to 6th in Table 2.

Table 2. Correlation values for similarity measures varying the IC calculus criterion. ***Bold*** *lines represent our method.*

IC Calculus	Measure	M&C
	Resnik	0.77
inv_f(hyponyms)	Lin	0.81
(Seco et al., eq. 3)	J&C	0.84
	Resnik	0.72
f(subsumers)	Lin	0.71
(eq. 14)	J&C	0.69
	Resnik	**0.83**
This work	**Lin**	**0.86**
(eq. 9)	**J&C**	**0.89**

Results show that the exclusive evaluation of the set of hyponyms provides more accurate assessments than the analysis of the subsumer set. This is because the former provides finer-grained and more detailed semantic evidences than the latter. However, both approaches present disadvantages: the fact that leaves become indistinguishable, in the first case, and that general concepts with the same number of subsumers are considered equally informative, in the latter case. The integration of both dimensions as proposed in this paper manages to improve individual results, as shown in the last three rows of Table 2. With our model, leaves are distinguished according to their subsumers, and inner concepts are differentiated according to the degree of commonness of their leaves. The proposed model is also coherent with the theoretical principles of IC and adapts the idea proposed by Resnik for corpora to the ontological domain: to compute the IC of a concept as a function of the IC of its specializations.

Finally, it is important to note that correlations achieved with our model (up to 0.89) are very close to the degree of agreement between human judgments for this benchmark. Concretely, a correlation value of 0.9015 among ratings of different human experts was reported by Resnik when reproducing the Miller and Charles' experiment, a value that represents an upper-bound for computerized approaches.

Extending the evaluation framework

In order to provide more robust evaluation evidences, we also used an additional recent benchmark named *Wordism similarity goldstandard* (Agirre, et al., 2009), which is a part of the well-known WordSim353 test collection (Finkelstein et al., 2002), but focused on the evaluation of semantic similarity. It consists of a set of 203 word pairs with associated ratings that also includes those of Miller and Charles. Due to its recentness, very few authors have reported evaluation results for this benchmark. Hence, we implemented and tested related works on intrinsic IC calculus to enable an objective comparison with our proposal. Correlation values are reported in Table 3.

Table 3. Correlation values for IC-based similarity measures for Wordism similarity goldstandard. From left to right: type of IC calculus, similarity measure, reference in which the measure/method was proposed and correlation against human judgments. ***Bold*** *lines represent our method.*

IC Calculus	Measure	Proposed in	Wordsim
	Resnik	(Seco, et al., 2004)	0.66
Intrinsic	Lin	(Seco, et al., 2004)	0.64
(Seco et al., eq. 3)	J&C	(Seco, et al., 2004)	0.62
	Resnik	(Zhou, et al., 2008)	0.62
Intrinsic	Lin	(Zhou, et al., 2008)	0.63
(Zhou et al, eq. 4)	J&C	(Zhou, et al., 2008)	0.63
	Resnik	(Sánchez, Batet, et al., 2011)	0.68
Intrinsic	Lin	(Sánchez, Batet, et al., 2011)	0.66
(Sánchez et al., eq. 5)	J&C	(Sánchez, Batet, et al., 2011)	0.64
	Resnik	**This work**	**0.68**
Intrinsic	**Lin**	**This work**	**0.68**
(this work, eq. 9)	**J&C**	**This work**	**0.68**

Firstly, one notices that all correlation values are significantly lower than those reported for Miller and Charles' benchmark (Table 1). This is caused by its larger size and the higher heterogeneity of terms, which makes Wordsim a more general and also a more challenging benchmark (Pirrò & Euzenat, 2010). In any case, similar conclusions as those enounced in the previous section can be extracted by analyzing the differences of correlation values: Seco et al.'s method is improved by Sanchez et al.'s and this one by the proposed model. A difference regards Zhou et al.'s method, which provides the lowest accuracy. As discussed in the second section, this method relies on a weighing parameter to aggregate the contribution of hyponym cardinality and depth. The parameter was tuned empirically for classical similarity benchmarks and lacks a theoretical base. The different behavior observed for the two benchmarks suggests that a different parameter tuning could be performed to improve the results in this last case.

Another difference regards the fact that, in this test, Resnik's measure tends to provide equal or even better results than Jiang & Conrath's and Lin's methods, whereas, in Miller and Charles' one, the opposite behavior was observed. As discussed in (Pirrò, 2009), this suggests that the LCS of the compared concepts, in which Resnik's method is solely based, better represents the commonalties between terms in this last benchmark.

Comparison with other similarity paradigms

In order to put correlation values reported above into context, in this section we compare the results achieved by our model with other ontology-based similarity paradigms. Again, Miller and

Charles' benchmark and WordNet have been used thanks to the availability of evaluation results for other similarity paradigms.

Ontology-based methods proposed in the literature can be classified in several families other than IC-based measures discussed in this paper. On the one hand, *edge-counting measures* are based on counting the minimum number of taxonomic links separating two concepts in a given ontology (Leacock & Chodorow, 1998; Li, et al., 2003; Rada, Mili, Bichnell, & Blettner, 1989; Wu & Palmer, 1994). On the other hand, *feature-based approaches* estimate similarity according to the amount of common and non-common semantic features between concept pairs (Petrakis, et al., 2006; Rodríguez & Egenhofer, 2003; Tversky, 1977). By features, authors analyze taxonomic knowledge modeled in an ontology as well as synonym sets or concept descriptions (glosses) retrieved from dictionaries. The contribution of each feature is aggregated by means of weighing parameters. Correlation values reported in the literature for the above measures are compiled in Table 4.

Table 4. Correlation values for different similarity measures for the Miller and Charles benchmark. From left to right: type of similarity computation paradigm, measure name, reference in which it is proposed, correlation against human subjects and reference in which correlations were reported. ***Bold*** *lines represent our method.*

Type	Measure	Proposed in	M&C	Evaluated in
Edge counting	Rada et al.	(Rada, et al., 1989)	0.59	(Petrakis, et al., 2006)
	W&P	(Wu & Palmer, 1994)	0.74	(Petrakis, et al., 2006)
	L&C	(Leacock & Chodorow, 1998)	0.74	(Patwardhan & Pedersen, 2006)
	Li et al.	(Li, et al., 2003)	0.82	(Petrakis, et al., 2006)
Feature based	R&E	(Rodríguez & Egenhofer, 2003)	0.71	(Petrakis, et al., 2006)
	Tversky	(Tversky, 1977)	0.73	(Petrakis, et al., 2006)
	Petrakis et al.	(Petrakis, et al., 2006)	0.73	(Petrakis, et al., 2006)
Intrinsic IC (eq. 9)	**Resnik**	**This work**	**0.83**	**This work**
	Lin	**This work**	**0.86**	**This work**
	J&C	**This work**	**0.89**	**This work**

The most basic edge-counting measure (Rada, et al., 1989), which is based solely on the evaluation of the minimum path separating the two concepts, provides the lowest accuracy (0.59). Other methods, which incorporate additional semantic evidences such as the depth ((Wu & Palmer, 1994) and (Leacock & Chodorow, 1998)), are able to improve this value (0.74). Parameterized methods (Li, et al., 2003) or those integrating several knowledge features such as (Rodríguez & Egenhofer, 2003), (Tversky, 1977) or (Petrakis, et al., 2006), also improve the baseline correlation, at the cost of introducing weighing parameters that should be tuned for a specific problem to maximize the accuracy. This hampers their generality.

Results achieved by IC-based measures when using the proposed model are the highest of the bunch (0.83-0.89). This suggests that, even though all paradigms use the same knowledge to assess similarity, that is, the input ontology, our model better captures semantic evidences that enable a more accurate differentiation of concepts and, hence, a more precise similarity assessment.

CONCLUSION

The IC of a concept is a fundamental dimension that sustains many semantic analyses of textual resources (Resnik, 1995). By accurately quantifying the informativeness of concepts referred in a discourse, one can identify its main topics (Sánchez, Castellà-Roca, & Viejo, in press), build user profiles (Viejo, Sánchez, & Castellà-Roca, in press) or even detect potentially sensitive information (Sánchez, Batet, & Viejo, in press). Moreover, as discussed throughout the paper, IC has been exploited as the fundamental principle to estimate the similarity between terms (Formica, 2008; Jiang & Conrath, 1997; Li, et al., 2003; Resnik, 1995), being applied in many areas to improve the comprehension of textual resources (see some examples in the introduction). Since the above tasks rely on the quantification of IC of individual concepts, an improvement of the IC calculus, as it is proposed in this paper, produces an immediate positive impact on their results.

The proposed model solely exploits the taxonomic knowledge available in an ontology, which is the most commonly available one (Ding et al., 2004). Instead of a refinement of intrinsic calculus methods proposed in related works, we present a redefinition of the way in which IC is assessed. Its design aims to better capture taxonomic semantic evidences available in the ontology and, hence, to enable a better differentiation of the meaning of concepts. The fact that our proposal avoids the use of tuning parameters, together with its lack of corpora dependency, enables a general, efficient and easily applicable approach.

Evaluation results sustained the theoretical hypotheses. Our method provided, in most cases, higher correlations with human judgments than other IC calculus models when applied in the context of semantic similarity assessment. On the one hand, our approach improved other intrinsic and corpora-based methods, obtaining correlation values that were very close to human agreement. On the other hand, in comparison with other ontology-based similarity paradigms, our model enabled IC-based measures to provide the most accurate results, despite exploiting the same knowledge source.

As future work, we plan to study the behavior of our method in specific domains such as biomedicine, in which large ontologies are available. For that purpose, domain-specific similarity benchmarks which cover medical terms (Pedersen, et al., 2007) and medical ontologies such as SNOMED-CT (Spackman, 2004) can be considered. Finally, in order to test the generality of our method, we plan to apply it to similarity assessment scenarios in which concepts are spread through several ontologies. By adapting IC-based measures to support multiple input ontologies (similarly to methods proposed in (M. Batet, et al., in press) or (David Sánchez, et al., 2012)), the recall of the similarity assessment can be improved by solving cases in which concepts are missing in one ontology but found in another.

ACKNOWLEDGEMENTS

This work was partly funded by the Spanish Government through the project CONSOLIDER INGENIO 2010 CSD2007-0004 “ARES”, and by the Government of Catalonia under grant 2009 SGR 1135.

REFERENCES

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., & Soroa, A. (2009). A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proceedings of the Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL* (pp. 19-27). Boulder, Colorado.

Batet, M., Sánchez, D., & Valls, A. (2011). An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics, 44*(1), 118-125.

Batet, M., Sánchez, D., Valls, A., & Gibert, K. (in press). Semantic similarity estimation from multiple ontologies. *Applied Intelligence*.

Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 10-15). Pittsburgh, USA.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of semantic distance. *Computational Linguistics, 32*(1), 13-47.

Curran, J. R. (2002). Ensemble Methods for Automatic Thesaurus Extraction. In *Proceedings of the Empirical Methods in Natural Language Processing, EMNLP 2002* (pp. 222–229). Philadelphia, PA, USA.

Devitt, A., & Vogel, C. (2004). The topology of WordNet: Some Metrics. In *Proceedings of the 2nd Global Wordnet Conference, GWC 2004* (pp. 106-111). Brno, Czech Republic.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V., & Sachs, J. (2004). Swoogle: A Search and Metadata Engine for the Semantic Web. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management, CIKM 2004* (pp. 652-659). Washington, D.C., USA.

Fanizzi, N., d'Amato, C., & Esposito, F. (2009). Inductive Classification of Semantically Annotated Resources through Reduced Coulomb Energy Networks. *International Journal On Semantic Web and Information Systems, 5*(4).

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Massachusetts: MIT Press.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems, 20*(1), 116-131.

Formica, A. (2008). Concept similarity in Formal Concept Analysis: An information content approach. *Knowledge-Based Systems 21*(1), 80-87.

Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E. G. M., & Milios, E. E. (2006). Information Retrieval by Semantic Similarity. *International Journal on Semantic Web and Information Systems, 2*(3), 55-73.

Hwang, M., & Kim, P. (2009). A New Similarity Measure for Automatic Construction of the Unknown Word Lexical Dictionary. *International Journal On Semantic Web and Information Systems, 5*(1), 48–64.

Jiang, J. J., & Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics, ROCLING X* (pp. 19-33). Taipei, Taiwan.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification *WordNet: An electronic lexical database* (pp. 265-283): MIT Press.

Li, Y., Bandar, Z., & McLean, D. (2003). An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering, 15*(4), 871-882.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning, ICML 1998* (pp. 296-304). Madison, Wisconsin, USA.

Maedche, A., & Staab, S. (2002). Measuring similarity between ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management* (pp. 251-263). London, UK.

Martínez, S., Sánchez, D., & Valls, A. (2012). Semantic Adaptive Microaggregation of Categorical Microdata. *Computers & Security, 31*(5), 653-672.

Martínez, S., Sánchez, D., Valls, A., & Batet, M. (2012). Privacy protection of textual attributes through a semantic-based masking method. *Information Fusion, 13*(4), 304-314.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6*(1), 1-28.

Partee, B., ter Meulen, A., & Wall, R. (1990). *Mathematical Methods in Linguistics*: Kluwer Academic Publishers.

Patwardhan, S., Banerjee, S., & Pedersen, T. (2003). Using Measures of Semantic Relatedness for Word Sense Disambiguation. In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing and Computational Linguistics, CICLing 2003* (pp. 241-257). Mexico City, Mexico.

Patwardhan, S., & Pedersen, T. (2006). Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together* (pp. 1-8). Trento, Italy.

Pedersen, T., Pakhomov, S., Patwardhan, S., & Chute, C. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics, 40*(3), 288-299.

Petrakis, E. G. M., Varelas, G., Hliaoutakis, A., & Raftopoulou, P. (2006). X-Similarity:Computing Semantic Similarity between Concepts from Different Ontologies. *Journal of Digital Information Management, 4*, 233-237.

Pirrò, G. (2009). A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering, 68*(11), 1289-1308.

Pirrò, G., & Euzenat, J. (2010). A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In *Proceedings of the 9th International Semantic Web Conference* (pp. 615-630). Shangai, China.

Rada, R., Mili, H., Bichnell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics, 9*(1), 17-30.

Ramezani, M. (2011). Using Similarity-Based Approaches for Continuous Ontology Development. *International Journal On Semantic Web and Information Systems, 7*(2), 45-64.

Resnik, P. (1995). Using Information Content to Evalutate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995* (pp. 448-453). Montreal, Quebec, Canada.

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

Ross, S. (1976). *A First Course in Probability*: Macmillan.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8*(10), 627-633.

Sánchez, D. (2010). A methodology to learn ontological attributes from the Web. *Data & Knowledge Engineering 69*(6), 573-597.

Sánchez, D., & Batet, M. (2011). Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective *Journal of Biomedical Informatics, 44*(5), 749-759.

Sánchez, D., Batet, M., & Isern, D. (2011). Ontology-based Information Content computation. *Knowledge-based Systems, 24*(2), 297-303.

Sánchez, D., Batet, M., Isern, D., & Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications, 39*(9), 7718-7728.

Sánchez, D., Batet, M., Valls, A., & Gibert, K. (2010). Ontology-driven web-based semantic similarity. *Journal of Intelligent Information Systems, 35*(3), 383-413.

Sánchez, D., Batet, M., & Viejo, A. (in press). Detecting sensitive information from textual documents: an information-theoretic approach. In *Proceedings of the 9th International Conference on Modeling Decisions for Artificial Intelligence* Girona, Spain.

Sánchez, D., Castellà-Roca, J., & Viejo, A. (in press). Knowledge-Based Scheme to Create Privacy-Preserving but Semantically-Related Queries for Web Search Engines. *Information Sciences*.

Sánchez, D., & Isern, D. (2011). Automatic extraction of acronym definitions from the Web. *Applied Intelligence, 34*(2), 311-327.

Sánchez, D., Isern, D., & Millán, M. (2011). Content Annotation for the Semantic Web: an Automatic Web-based Approach. *Knowledge and Information Systems, 27*(3), 393-418.

Sánchez, D., Solé-Ribalta, A., Batet, M., & Serratosa, F. (2012). Enabling semantic similarity estimation across multiple ontologies: An evaluation in the biomedical domain. *Journal of Biomedical Informatics, 45*(1), 141-155

Seco, N., Veale, T., & Hayes, J. (2004). An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In *Proceedings of the 16th European Conference on Artificial Intelligence, ECAI 2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004* (pp. 1089-1090). Valencia, Spain.

Shakya, A., Takeda, H., & Wuwongse, V. (2009). Community-Driven Linked Data Authoring and Production of Consolidated Linked Data. *International Journal On Semantic Web and Information Systems, 5*(3), 23-48.

Spackman, K. (2004). SNOMED CT milestones: endorsements are added to already-impressive standards credentials. *Healthcare Informatics, 21*(9), 54-56.

Tapeh, A. G., & Rahgozar, M. (2008). A knowledge-based question answering system for B2C eCommerce. *Knowledge-Based Systems, 21*(8), 946-950.

Tversky, A. (1977). Features of Similarity. *Psycological Review, 84*(4), 327-352.

Viejo, A., Sánchez, D., & Castellà-Roca, J. (in press). Preventing Automatic User Profiling in Web 2.0 Applications. *Knowledge-Based Systems*.

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

Zhou, Z., Wang, Y., & Gu, J. (2008). A New Model of Information Content for Semantic Similarity in WordNet. In *Proceedings of the Second International Conference on Future Generation Communication and Networking Symposia, FGCNS 2008* (pp. 85-89). Sanya, Hainan Island, China.