HAL
open science

A feature and information theoretic framework for semantic similarity and relatedness

Giuseppe Pirrò, Jérôme Euzenat

To cite this version:

Giuseppe Pirrò, Jérôme Euzenat. A feature and information theoretic framework for semantic similarity and relatedness. Proc. 9th international semantic web conference (ISWC), Nov 2010, Shanghai, China. pp.615-630, 10.1007/978-3-642-17746-0_39 . hal-00793283

HAL Id: hal-00793283

https://inria.hal.science/hal-00793283v1

Submitted on 22 Feb 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness

Giuseppe Pirró*, Jérôme Euzenat

INRIA Rhône-Alpes, Montbonnot, France
{*Giuseppe.Pirro*, *Jerome.Euzenat*}@*inrialpes.fr*

Abstract. Semantic similarity and relatedness measures between ontology concepts are useful in many research areas. While similarity only considers subsumption relations to assess how two objects are alike, relatedness takes into account a broader range of relations (e.g., part-of). In this paper, we present a framework, which maps the feature-based model of similarity into the information theoretic domain. A new way of computing IC values directly from an ontology structure is also introduced. This new model, called Extended Information Content (*eIC*) takes into account the whole set of semantic relations defined in an ontology. The proposed framework enables to rewrite existing similarity measures that can be augmented to compute semantic relatedness. Upon this framework, a new measure called FaITH (Feature and Information THeoretic) has been devised. Extensive experimental evaluations confirmed the suitability of the framework.

Key words: Semantic Similarity, Feature Based Similarity, Ontologies

1 Introduction

Semantic similarity and relatedness investigates how alike two or more objects are, and plays an important role in many contexts. Generally speaking, similarity allows to infer knowledge and categorize objects into kinds. This is important when either it is not possible to exactly state what properties are salient for an object, or when it is not easy to separate an object into distinct properties [5, 26]. Semantic similarity has a long tradition in psychology and cognitive science where different models have been postulated. Among these, the *geometric* model enables to asses similarity between entities by considering them as points in a dimensionally organized metric space. The *feature-based* model, leverages features (i.e., characteristics) of the examined objects and assumes that similarity is a function of both common and distinctive features [24]. Recently, findings in information theory have been considered in computing similarity [19]. From a computer science perspective, similarity measures exploit some source of knowledge such as search engines [3] or ontologies such as WordNet [13]. More recently, similarity measures have been defined in Description Logics (DLs) [2, 4]. In [4]

* This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme.

several similarity measures are described, which take into account ontology instances for assessing the similarity between two concepts. While similarity only considers subsumption relations to assess how two objects are alike, relatedness takes into account a broader range of relations (e.g., part-of). The work presented in this paper focuses on computing similarity and relatedness by exploiting the terminological definition of ontology concepts. We leave the investigation about how this method can be applied to DLs as a future work.

Computing semantic similarity between ontology concepts is an important issue since having many applications in different contexts including: Information Retrieval, to improve the performance of current search engines [8], ontology matching, to discover correspondences between entities belonging to different ontologies [16], semantic query routing, to choose among the set of possible peers only those relevant, bioinformatics to assess the similarity between proteins [25] just to cite a few. This paper presents a semantic similarity framework, which is based on two main pillars. One is the projection of the feature-based model of similarity into the information theoretical domain. The reason to combine these two models is twofold. On one hand, the feature-based model has a solid theoretical underpinning supported by several psychological studies [24] and is more flexible than other theoretical models (e.g., geometric). On the other hand, the information-theoretic formulation of similarity allows to compare concept features not by simply counting object properties but taking into account the informativeness of the concepts being compared. The second pillar, is a new way to obtain IC values called Extended Information content (*eIC*). *eIC* considers the whole set of semantic relations defined in an ontology and assigns a score of informativeness to each concept without referring to external corpora as usually done by traditional IC-based approaches where time expensive and corpus-dependent occurrence count has to be performed.

The generality of this framework enables to rewrite several existing similarity measures that can be augmented to compute semantic relatedness. This aspect has been investigated and resulted in an improvement of existing similarity measures as will be discussed in Section 4. Finally, a new measure called FaITH (Feature and Information Theoretic) has been designed, which is a versatile tool to compute both similarity and relatedness. Extensive experimental evaluation of similarity and relatedness show the suitability of the proposed framework and FaITH in particular.

The remainder of this paper is organized as follows. Section 2 provides some background and surveys on popular measures. Section 3 presents the new similarity framework and the logical path toward its definition; here the FaITH measure and the *eIC* are discussed. Section 4 presents an extensive evaluation campaign. Section 5 concludes the paper.

2 Definitions and Background

We consider an ontology O as a graph, where nodes represent concepts and edges represent relations between concepts. If we consider the hierarchical structure of the ontology, each concept can have a set of sub-concepts (its descendants) in

the hierarchy. However, an ontology usually includes a broader set of semantic relation such as part-of. Figure 1 reports an excerpt of an ontology. Hereafter we will consider WordNet as reference ontology even if the same reasoning applies to any other ontology. In WordNet, the definition of a concept consists of its immediate superordinate(s) followed by a relative clause that describes how this concept differs from all others. For example *Fortified Wine* is distinguished from *Wine* because *"... alcohol (usually grape brandy)"* has been added just as the gloss accompanying its definition mentions.

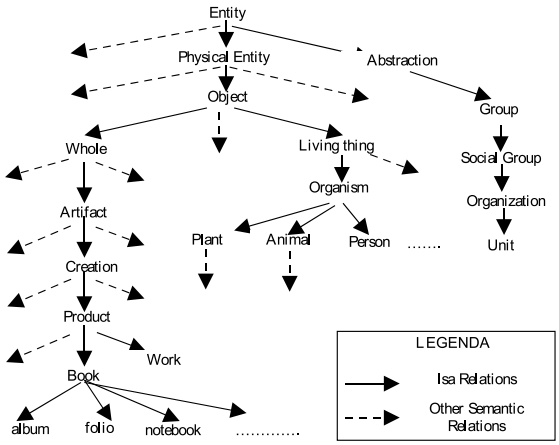

Fig. 1. An excerpt of ontology

An object *feature* (a concept in our case) can be seen as a property of the object. According to the definition above, concepts in the hierarchy inherit all the features of their superordinate even if they can have their own specific features. As an example, since *car* and *bicycle* both serve to transport people or objects, in other words they are both types of vehicles, they share all features pertaining to the concept *vehicle*. However, each concept has also its specific features as *steering wheel* for *car* and *pedal* for *bicycle*. Moreover, even if specialization relations constitute the majority in WordNet, there are other kinds of relations accompanying each definition that are useful to identify object features. For instance, *car* has a relation of type *part-of* with *engine* whereas *bicycle* has a *part-of* relation with *sprocket*. The use immediate concept features can be seen as a special case of semantic neighbourhood with radius equals to 1.

Similarity or relatedness measures, by looking at the ontology structure or by exploiting some additional information, address the problem of assessing (typically in terms of a numerical score) how alike two concepts are. As an example of similarity and relatedness, *car* and *bicycle* are similar whereas *car* and *wheel* are related. The choice to focus either on similarity or relatedness depends on the particular application context, even though many approaches to compute relatedness are extensions of similarity measures [6, 20]. The framework presented in this paper can be adopted to compute both similarity and relatedness.

2.1 State of the art

Similarity measures can be divided into different and not necessarily disjoint categories. In this work we consider information-theoretic approaches, ontology-based approaches and hybrid approaches.

Information Theoretic Approaches. Information theoretic approaches employ the notion of Information Content (IC), which quantifies the informativeness of concepts. Early IC approaches [19, 9, 12] obtained IC values by associating probabilities to each concept in an ontology on the basis of its occurrences in large text corpora. In the specific case of hierarchical ontologies, these probabilities are cumulative as we travel up from specific concepts to more abstract ones. This means that every occurrence of a concept in a given corpus is also counted as an occurrence of each concept containing it. IC values are obtained by computing the negative likelihood of encountering a concept in a given corpus. Note that this method ensures that IC is monotonically decreasing as we move from the leaves of the taxonomy to its roots.

Resnik [19] was the first to leverage IC for the purpose of semantic similarity. The basic intuition behind the use of the negative likelihood is that the more probable a concept is of appearing in a corpus the less information it conveys, in other words, infrequent words are more informative than frequent ones. Once IC values are available for each concept in the considered ontology, semantic similarity can be calculated. Resnik's formula to compute similarity states that similarity depends on the amount of information two concepts c_1 and c_2 share, which is given by the Most Specific Common Abstraction ($msca(c_1, c_2)$), that is, the concept that subsumes the two concepts being compared.

Starting from Resnik's work, Jiang and Conrath [9] and the Lin [12] proposed two measures, which calculate IC-values in the same manner as proposed by Resnik while correcting some problems with this similarity measure; if one were to calculate $sim_{res}(c_1, c_1)$ one would not obtain the maximal similarity value of 1, but instead the value given by IC(c_1). Besides, with Resnik's approach any two pairs of concepts having the same *msca* have exactly the same semantic similarity; for instance, in the WordNet ontology, $sim_{res}(Horse, Plant)$ $= sim_{res}(Animal, Plant)$ because in each case the $msca(Horse, Plant)$ and $msca(Animal, Plant)$ is *Living Thing*. However, in this case the semantic leap is not the same.

The Lin measure considers the ratio between the amount of information needed to state the commonality between two concepts and the information needed to describe them as discussed in [12].

Ontology based approaches. As for ontology based approaches, the work by Rada et al. [18] is similar to the Resnik measure since it also computes the $msca(c_1, c_2)$, but instead of considering the IC as the value of similarity, it considers the number of links that were needed to attain the $msca(c_1, c_2)$. Obviously, the less the number of links separating the concepts the more similar they are. The work by Hirst et al., which actually measures relatedness, is similar to the previous one but it uses a wider set of relations coupled with rules restricting the way concepts are transversed [6]. Nonetheless, the intuition also in this case

is that the number of links separating two concepts is inversely proportional to the degree of similarity.

Hybrid approaches. Hybrid approaches usually combine multiple information sources. Li et al. [11] proposed to combine structural semantic information in a nonlinear model. The authors empirically defined a similarity measure that uses shortest path length, depth and local density in a taxonomy and combine them.

In [22] the *OSS* distance function, combining *a-priori* scores of concepts with distance, is proposed. *OSS* performs the following steps to assess similarity between two concepts c_1 and c_2: (i) computing the score of the concept c_2 from the concept c_1; (ii) computing how much score has been transferred between the concepts; (iii) transforming the transfer of score into a distance measure.

Our previous work [17], defined a similarity measure combining features and information content that adopts Tversky's contrast model. This measure treats similarity between identical concepts as a special case and can give as output negative values, which make difficult the interpretation of results. The differences with the present work are: i) this paper describes a general framework, which can be used to rewrite even existing similarity measures; ii) here a new similarity measure is proposed, which adopts a different representation of the feature-based model; iii) in this paper a new way to compute IC values is proposed, which enables to compute both semantic similarity and relatedness; iv) an extensive evaluation of relatedness is proposed for FaITH and several other measures.

2.2 Comparison among measures

Each measure has its limitations. IC-based measures making use of corpora, though having a strong mathematical formalization, may sometimes fail to capture certain aspects of language. For instance, it is possible that corpus such as the British National Corpus, may not even mention certain words. Besides, values of IC are obtained through time intensive analysis of corpora and can heavily depend on the considered corpora (as discussed in Section 4). Ontology-based approaches require to work with consistent ontologies, that is, ontologies where distance between specific and more general concepts have the same interpretation. As an example it is obvious that the semantic leap between *Entity* and *Psychological Feature* is higher than that between *Canine* and *Dog* even if both couples are separated by one edge. Finally, hybrid approaches require the different information sources to be correctly "weighted". A common limitation of the considered approaches is that they can only compute either similarity or relatedness. The proposed framework, and in particular FaITH, are more flexible as the notion of *Extended Information Content* (*eIC*) can be exploited to compute both similarity and relatedness without depending on external corpora.

3 A Framework for Semantic Similarity and Relatedness

This section presents a new framework for computing semantic similarity and relatedness. After providing some preliminary definitions, the Tversky's formulation of similarity, which is based on a representation of concepts according to

their features, is introduced. This will serve as a basis to motivate the present framework. In more detail, the proposed framework adopts a ratio-based formulation of the Tversky's model of similarity and projects it into the information-theoretic domain. Section 3.4 describes the *Extended Information Content* (*eIC*), which can be used to compute relatedness between concepts. Note that the generality of this framework enables to rewrite several existing similarity measures, which can be augmented to compute relatedness.

3.1 Tversky's feature-based model of similarity

Amos Tversky, in his seminal work, proposed an alternative way to compute similarity by taking into account both common and distinguish "features "of the objects being compared. As an example of Tversky's formulation, *car* and *bicycle* both serve to transport people or objects (in other words they are both types of vehicles), then they share all features that pertain to the concept *vehicle*. However, each concept has also its specific features such as *steering wheel* for *car* or *pedal* for *bicycle*. Moreover, if we look beyond the hierarchical structure of their definitions we can find different kinds of relations with other concepts such as *engine part-of car* and *sprocket part-of bicycle*. The set of all relations can be exploited to further characterize concept features. Fig. 2 depicts an example of such reasoning. Early semantic similarity models, such as the geometric model,

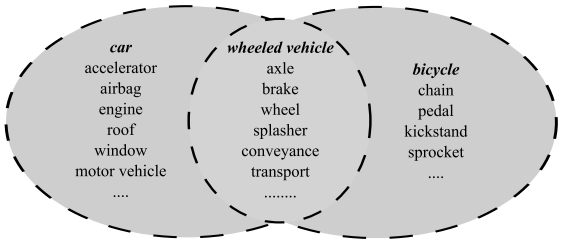

Fig. 2. An example of concept features

required to respect metric properties such as the triangle inequality or symmetry. Tversky's discussed several examples to support the idea that certain axioms, required by the geometric model, were not necessary in the process of similarity estimation. For instance, since Germany is judged to be more like Austria than Austria is to Germany [24] the symmetry property could not be respected in this case. According to the feature-based model, the similarity of a concept c_1 to a concept c_2 is a function of the features common to c_1 and c_2, those in c_1 but not in c_2 and those in c_2 but not in c_1. If we admit a function $\Psi(c)$ that yields the set of features relevant to c, Tversky's similarity model can be represented by the following equation, also known as *contrast model*:

$$sim_{tvr}(c_1, c_2) = \alpha F(\Psi(c_1) \cap \Psi(c_2)) - \beta F(\Psi(c_1) \setminus \Psi(c_2)) - \gamma F(\Psi(c_2) \setminus \Psi(c_1)) \ . \quad (1)$$

where F is some function that reflects the salience of a set of features, and α, β and γ are parameters that provide for differences in focus on the different

components. According to this model, features in common increase similarity whereas features that are unique to the two objects decrease similarity. However, note that the above formulation is not framed in information theoretic terms since it is based on sets of concept features.

3.2 A Ratio-based formulation of Tverky's similarity model

The difficulty with the contrast model described in equation (1) and discussed in our previous study [17] is that the more unique features a concept presents the lower the similarity. Moreover similarity values are not bounded between 0 and 1, which can make interpretation of results difficult. To overcome these issues, therefore, a *ratio model* is more appropriate since it is bounded between 0 and 1, irrespective of the size of the features being compared. Thus a more useful definition of feature-based similarity is:

$$sim_{tvr-ratio}(c_1, c_2) = \frac{F(\Psi(c_1) \cap \Psi(c_2))}{\beta F(\Psi(c_1) \setminus \Psi(c_2)) + \gamma F(\Psi(c_2) \setminus \Psi(c_1)) + F(\Psi(c_1) \cap \Psi(c_2))} \ . \tag{2}$$

Note that $\alpha = 1$ in the ratio model and then common features are maximally important in process of similarity estimation. At this point there are two main tasks we can perform:

1. Assess the degree to which concept c_1 and c_2 are similar to each other. In this case $\beta = \gamma$ since the similarity is not intended to be directional.
2. Assess the degree to which concept c_2 is similar to concept c_1. In this second task, similarity is directional and we are more interested in the features in c_1 than we are in the features unique to c_2. Here, β and γ do not need to be equal. This latter case is useful in many application contexts such as Information Retrieval (IR)or clustering where starting from a concept we are interested in finding what it is similar to.

Table 1 analyzes different scenarios obtained by manipulating the coefficients β and γ in equation (2). For the purpose of this paper, we consider $\beta = \gamma$ since we want to compute the similarity not directionally. Moreover, for the definition of the ratio based model described in equation (2) $\alpha = 1$, which maximizes the contribution of common features. We leave as future work the investigation of other values for these parameters in more targeted applications such as IR.

3.3 The FaITH similarity measure

This section describes the FaITH measure for semantic similarity and relatedness. The cornerstone of this measure is the $msca(c_1, c_2)$, which reflects the information shared by two concepts c_1 and c_2 in an ontology structure. In the information-theoretic domain, Resnik exploited the $msca(c_1, c_2)$ to assess the similarity between concepts. IC values are obtained by exploiting equation (3):

$$IC(c) = -log\ p(c) \ . \tag{3}$$

Table 1. Possible scenarios obtained by manipulating equation (2).

Case	Coefficients	Description
Commonalities between c_1 and c_2	$\beta = \gamma = 0$	If there exists any commonality then $sim_{tvr'}(c_1, c_2) = 1$
Given c_1 assess to which degree c_2 is similar to it	$\beta = 1$, $\gamma = 0$	When the full set of features of c_1 are contained in c_2 then $sim_{tvr'}(c_1, c_2) = \frac{\alpha F(\Psi(c_1) \cap \Psi(c_2))}{\beta F(\Psi(c_1) \setminus \Psi(c_2)) + \alpha F(\Psi(c_1) \cap \Psi(c_2))}$
	$\beta = 0$, $\gamma = 1$	When the set of features of c_1 contains the features of c_2 then $sim_{tvr_r}(c_1, c_2) = \frac{\alpha F(\Psi(c_1) \cap \Psi(c_2))}{\gamma F(\Psi(c_2) \setminus \Psi(c_1)) + \alpha F(\Psi(c_1) \cap \Psi(c_2))}$
Given c_1 and c_2 assess to which degree they are similar to each other	$\beta = \gamma = 1$	Tversky's similarity is represented in terms of Tanimoto index.
	$\beta = \gamma = 0.5$	Tversky's similarity is represented in terms of Dice index.

where c is a concept and *p(c)* is the probability of encountering c in a given corpus. Note that this method ensures that IC is monotonically decreasing as we move from the leaves of the taxonomy to its roots.

In Fig. 2, the $msca(car, bicycle)$ is *wheeled vehicle* and these two concepts share all the features belonging to their *msca*. In a feature-based formulation of similarity, the $msca(c_1, c_2)$ can be seen as the intersection of features from c_1 and c_2. Therefore, one can speculate that the function F, that reflects the saliency of features, can be substituted by the function IC in the information theoretic domain (this new IC is referred to as $IC_{features}$). Starting from this assumption, by looking at Fig. 2, it is immediate to infer that the set of features specific to *car* (resp. *bicycle*) is given by $IC_{features}(car) - IC_{features}(wheeled_vehicle)$ (resp. $IC_{features}(bicycle) - IC_{features}(wheeled_vehicle)$). These three analogies, generalized in Table 2, are the building blocks of the proposed framework.

Table 2. Mapping between feature-based and information theoretic similarity models.

Description	Feature-based model	Information-theoretic model
Common features	$\Psi(c_1) \cap \Psi(c_2)$	$IC(msca(c_1, c_2))$
Features of c_1 alone	$\Psi(c_1) \setminus \Psi(c_2)$	$IC(c_1) - IC(msca(c_1, c_2))$
Features of c_2 alone	$\Psi(c_2) \setminus \Psi(c_1)$	$IC(c_2) - IC(msca(c_1, c_2))$

Moreover, as it will be discussed in Section 3.4, the way we compute the IC values for each concept (i.e., *eIC*) can take into account the different features of an object defined both in terms of the hierarchical structure and other kinds of semantic relations. By substituting the analogies from Table 2 in equation (2) the similarity measure called FaITH, reported in equation (4), is obtained.

$$sim_{FaITH}(c_1, c_2) = \frac{IC(msca(c_1, c_2))}{\beta(IC(c_1) - IC(msca(c_1, c_2))) + \gamma(IC(c_2) - IC(msca(c_1, c_2))) + IC(msca(c_1, c_2))} . \quad (4)$$

As we are concerned to compute how two concepts c_1 and c_2 are similar to each other we set the values of β and γ to 1 (see Table 1) thus obtaining:

$$sim_{FaITH}(c_1, c_2) = \frac{IC(msca(c_1, c_2))}{IC(c_1) + IC(c_2) - IC(msca(c_1, c_2))} \ . \tag{5}$$

Note that in the case of ontologies with multiple inheritance, the $msca(c_1, c_2)$ may be unique. In this case, FaITH considers the most informative *msca* (i.e., the *msca* with the highest information content).

3.4 Extended information Content (eIC)

The proposed framework combines the feature and the information theoretic models of similarity. One of the main difficulty with this model is that IC values have to be derived by analyzing large corpora, which may not even contain certain specific words. In order to overcome this issue, the intrinsic IC formulation proposed in [23] is adopted. The *intrinsic* IC (iIC) for a concept c is defined as:

$$iIC(c) = 1 - \frac{log(sub(c) + 1)}{log(max_{con})} \ . \tag{6}$$

where the function *sub* returns the number of subconcepts of a given concept c. Note that concepts representing leaves in the taxonomy will have an IC of one, since they do not have hyponyms. The value of one states that a concept is maximally expressed and is not further differentiated. Moreover max_{con} is a constant that indicates the total number of concepts in the considered taxonomy.

However, since an ontology usually contains relations beyond inheritance also useful to assess to what extent two concepts are alike, the Extended Information Content (eIC) is introduced. eIC by investigating each kind of ontological relation between concepts provides a better indicator about the features of concepts and then can be used to compute relatedness. For instance, by only focusing on *isa* relations, in the example in Fig. 2 we would lose some important information (e.g., that *car* has *part-of engine* or that *bicycle* has as *part-of sprocket*) that can help to further characterize commonalities and differences between two concepts. For each concept, the coefficient EIC is defined as follows:

$$EIC(c) = \sum_{j=1}^{m} \frac{\sum_{k=1}^{n} iIC(c_k \in C_{R_j}))}{|C_{R_j}|} \ . \tag{7}$$

This formula takes into account all the m kinds of relations that connect a given concept c with other concepts. Moreover, for all the concepts at the other end of a particular relation (i.e., each $c_k \in C_{R_j}$) the average iIC is computed. This enables to take into account the expressiveness of concepts to which a given concept is related in terms of their information content. The final value of *Extended Information Content* (eIC) is computed by weighting the contribution of the iIC and EIC coefficients thus leading to:

$$eIC(c) = \zeta iIC(c) + \eta EIC(c) \ . \tag{8}$$

The two parameters ζ and η can be settled in order to give more or less emphasis to the hierarchical IC of the two concepts. At this point, we can rewrite equation (5) thus obtaining:

$$sim_{FaITH}(c_1, c_2) = \frac{eIC(msca(c_1, c_2))}{eIC(c_1) + eIC(c_2) - eIC(msca(c_1, c_2))} \ . \tag{9}$$

This similarity measure corrects some drawbacks of existing approaches. First, it exploits features of concepts, expressed in terms of IC, and not only their position in the ontology structure. Second, it corrects the problem with Resnik's measure, in fact, $sim_{FaITH}(c_1, c_1) = 1$. Finally, by taking into account relations beyond inheritance, FaITH allows to compute semantic relatedness.

4 Evaluation

This section discusses the evaluation of the FaITH similarity measure and its comparison w.r.t. the state of the art. In the first experiment we evaluated FaITH as a semantic similarity measure while in the second experiment we evaluated FaITH as a semantic relatedness measure as using the *eIC* formulation. Finally, in order to have an insight of how FaITH works with more domain-related ontologies, we performed an evaluation using couples of concepts taken from the MeSH biomedical ontology. In each experiment, we evaluated the performance of the different methods in two settings. The first one (denoted as $F + eIC$) by exploiting the proposed framework along with the *eIC* while the second one using the classical approach to compute IC without mapping features in the IC domain. In particular, the *SemCor*(S) *Brown* (B) and *BNC* (*Bnc*) text corpora, of increasing size, have been used to obtain IC values. For the Li measure we adopted the same optimal parameter values as indicated by authors in [11].

In order to have an idea of the improvement using the $F + eIC$ formulation we computed for each measure and corpus the loss (L) in performance, which represents how much the performance of a given measure decrease when using the classical IC formulation. Besides, for each evaluation, as statistical test of significance, we computed the *p-value*. The analyzed similarity measures have been implemented in the Java WordNet Similarity Library available upon request, along with the datasets, at `http://grid.deid.unical.it/similarity`.

4.1 Experiment 1: evaluating FaITH on similarity

In the first experiment, we evaluate the FaITH measure on a dataset collected by an online similarity experiment described in our previous work [17]. The dataset contains similarity judgments for 65 word pairs [21] (referred to as $S_{R\&G}$) which are commonly used, along with a subset of 28 word pairs [14] (referred to as $S_{M\&C}$), to measure accuracy of similarity measures. The word pairs in the dataset have been originally chosen to range from very similar (e.g., *car-automobile*) to *semantically unrelated* (e.g., *chord-smile*) as discussed in [21]. Figure 3 reports the ratings of similarity provided by both human participants and computational methods. Values of correlation (ρ) for the different measures

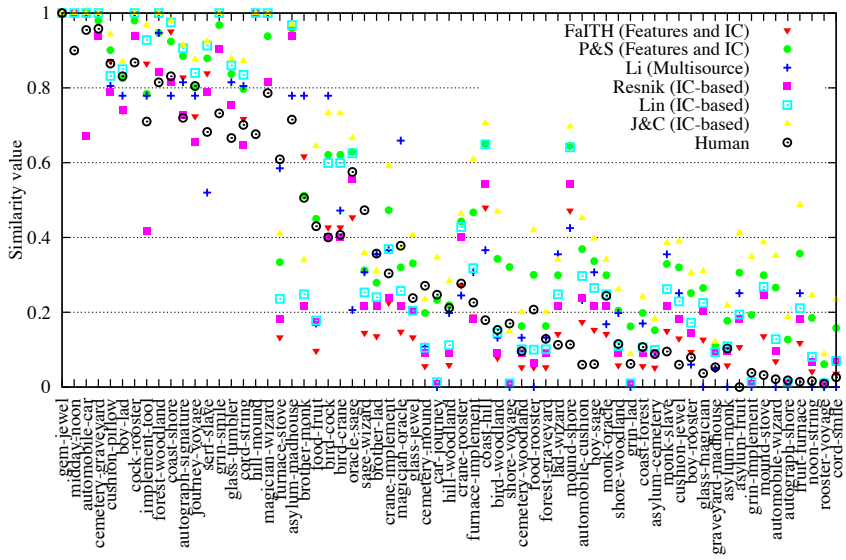

Fig. 3. Results for similarity measures and human ratings

are reported in Table 3. The first column indicates the correlation by using the IC formulation introduced in Section 3.4. Each other column considers the correlation according one corpus and the loss as compared to the result in the first column. The second column, for instance, indicates that by using the *SemCor(S)* corpus, the correlation of the Resnik measure is 0.71, with a loss of 16.4 % .

Table 3. Correlation values with $F + eIC$ (ρ) and different corpora

	Correlation on $S_{M\&C}$				**Correlation on $S_{R\&G}$**			
	ρ	$\rho_S/\boldsymbol{L}(\%)$	$\rho_B/\boldsymbol{L}(\%)$	$\rho_{Bnc}/\boldsymbol{L}(\%)$	ρ	$\rho_S/\boldsymbol{L}(\%)$	$\rho_B/\boldsymbol{L}(\%)$	$\rho_{Bnc}/\boldsymbol{L}(\%)$
Lenght	0.61	0.61	0.61	0.61	0.58	0.58	0.58	0.58
Depth	0.84	0.84	0.84	0.84	0.80	0.80	0.80	0.80
Li	0.91	0.91	0.91	0.91	0.90	0.90	0.90	0.90
Resnik	0.85	0.71/16.4	0.73/14.5	0.75/11.8	0.87	0.83/5.1	0.84/4.1	0.85/2.4
Lin	0.87	0.69/**20.2**	0.74/**15.0**	0.75/**14.2**	0.89	0.75/**15.0**	0.79/**10.9**	0.80/**9.8**
J&C	0.88	0.72/17.8	0.80/9.3	0.81/8.1	0.87	0.82/6.3	0.83/5.4	0.84/4.0
P&S	0.91	0.86/4.7	0.87/4.2	0.89/2.3	0.90	0.87/3.7	**0.88**/3.0	0.89/1.8
FaITH	**0.92**	**0.87**/5.5	**0.88**/4.6	**0.90**/2.4	**0.91**	**0.88**/3.3	0.88/2.9	**0.90**/1.0

For the *Length* measure, lower values correspond to higher similarity values. For instance, the two word pairs (i.e., *gem-jewel* and *automobile-car*) have a length equal to zero since belonging to the same WordNet synset respectively and then are maximally similar according to the WordNet's design principle. On the other hand, examples of unrelated words are the couples *rooster-voyage* and *chord-smile* having a path length of 30. The *Depth* measure obtained a value of correlation of about 30% better than the *Path* measure. This measure assesses similarity by considering the depth of the $msca(c_1, c_2)$. Edge counting

approaches reach the lowest correlation w.r.t. human ratings in both datasets. That is because these approaches work well only when the values computed have a "consistent interpretation", that is, when the length of the path (resp. depth of the $msca(c_1, c_2)$) between two general concepts and that between two specific ones express the same semantic leap, which is not the case of WordNet.

As for IC-based approaches, Resnik's measures obtained the lowest value of correlation. However, the usage of the *IC(*$msca(c_1, c_2)$*)* brings better results in terms of correlation as compared to path-based measures. The other two IC-based measures (i.e., Lin and J&C) obtained better results since considering the IC of the two concepts as well. As for hybrid approaches, the Li measure, which combines the depth of the $msca(c_1, c_2)$ and the length of the path between two concepts, obtained a higher value of correlation. However, note that this measure to correctly weights the contributions of the different information sources requires the tuning of two coefficients as described in [11]. The P&S measure, described in [17], obtained a remarkable value of correlation in both $S_{R\&G}$ and $S_{M\&C}$. However, the P&S formulation treats the computation of similarity between identical concepts as a special case as discussed in [17]. Moreover, in some cases, $sim_{P\&S(c_1,c_2)} < 0$, which makes the interpretation of results difficult.

Moreover, this measure is not as flexible as FaITH, which can be adopted to different contexts as discussed in Section 3.2. The FaiTH measure obtained the best value of correlation in both $S_{R\&G}$ and $S_{M\&C}$. Note that in all cases the $F + eIC$ formulation brings better results. The loss L can reach the 20% and 15% with the Lin measure in $S_{R\&G}$ and $S_{M\&C}$ respectively. Moreover, using classical approaches the performance heavily depend on the adopted corpus even if it can be noted that larger corpora bring better results. The *p-values* in both evaluations are $p - value < 0.001$, which indicate that the results are significant. Finally, one note about the couple *car-journey*. The two words, even if generally related since a *car* can be the means to do a *journey*, are not *similar*. This is because similarity, which is a special case of relatedness, only considers the relations of hypernymy/hyponymy (i.e., isa). The FaITH measure assigned a similarity score of 0.007 to this couple while the J&C, Resnik, Lin and P&S assigned 0.346, 0.009, 0.013 and 0.233 respectively. In this case, the FaITH measure since giving the lowest value of similarity seems to better comply with the definition of similarity. In summary, our intuition to exploit a ratio-based representation of Tversky's similarity model and project it into the information theoretic domain is consistent.

4.2 Experiment 2: evaluating FaITH on relatedness

In this experiment, FaITH has been evaluated as a semantic relatedness measure by using the *eIC* formulation. For the evaluation, the WordSim353 dataset, which is a test collection for measuring word relatedness often used in the literature has been adopted. Further detail on the dataset are available in [1]. Even in this case, for each measure, the Pearson correlation coefficient w.r.t. human ratings of similarity has been computed. In this evaluation we compare FaITH with more relatedness measures. In particular, we also considered the Leacock & Chodorow (referred to as *Lch*) [10] and the Wu & Palmer (referred to as

Wup) [27] measures. We also used a measure of relatedness between two words (referred to as *Ovp*), which assesses the overlap score between two concepts by augmenting glosses with glosses of related concepts [15]. The optimal values for the parameters ζ and η, experimental determined, are 0.4 and 0.6 respectively.

Table 4. Evaluation on relatedness

Measure	ρ	$\rho_S/\boldsymbol{L\%}$	$\rho_B/\boldsymbol{L\%}$	$\rho_{Bnc}/\boldsymbol{L\%}$
Lch	0.36	0.36	0.36	0.36
Wup	0.32	0.32	0.32	0.32
Ovp	0.21	0.21	0.21	0.21
Resnik	0.40	0.36/**11.1**	0.36/**9.9**	0.38/5.4
Lin	0.404	0.37/7.9	0.378/6.4	0.38/5.7
J&C	0.40	0.38/4.0	0.38/2.8	0.39/1.8
P&S	0.41	0.38/5.4	0.38/5.1	0.39/4.7
FaITH	**0.43**	**0.40**/7.0	**0.40**/6.3	**0.40/5.8**

While similarity measures perform extremely well on small similarity datasets such as the M&C and R&G discussed in Section 4.1, their performance drastically decrease when applied to a larger dataset such as WordSim353. The values of correlation reported in Table 4 are related to the word pairs contained in WordNet. Note that for the *Lch*, *Wup* and *Ovp* measures the results are the same as they are not based on IC.

As can be observed, FaITH performs clearly better than the other measures, which substantiate our intuition of adopting the $F + eIC$ strategy. Besides, all the similarity measures perform worse when not using $F + eIC$. The loss (L) in performance is reported in Table 4. In particular, all the IC-based measures take advantage of this formulation, with the Resnik measure improving of about 11%. In the case of not adopting the $F + eIC$, correlation values heavily depend on the considered corpus. Overall, FaITH and the *eIC* formulation represent a promising technique to compute similarity and relatedness between words and help to augment and improve existing similarity measures.

4.3 Experiment 3: evaluation on the MeSH ontology

The MeSH Medical Subject Headings (MeSH) ontology is mainly a hierarchy of medical and biological terms. It consists of a controlled vocabulary and a *Tree*. The controlled vocabulary contains several different types of terms such as *Descriptors*, *Qualifiers*, *Publication Types*, *Geographics* and *Entry* terms. Entry terms are the synonyms or the related terms to descriptors. MeSH descriptors are organized in a tree, which defines the MeSH Concept Hierarchy. In the MeSH tree there are 15 categories each of which is further divided into subcategories. For each subcategory, its descriptors are arranged in a hierarchy from most general to most specific. This evaluation investigates how FaITH performs with domain related ontologies. Similarly to the first evaluation, a dataset of human similarity judgments has been exploited (refer to [7] for further details). Results obtained by computational methods are compared with those provided by humans in Table 6 whereas, Table 5 reports values of correlations.

Table 5. Correlation with $F + iIC$

Measure	ρ
Resnik	0.72
Lin	0.71
J&C	0.71
Li	0.70
P&S	0.72
FaITH	**0.74**

Table 6. Evaluation on MeSH

Word 1	**Word 2**	**Human**	Resnik [19]	Lin [12]	J&C [9]	Li [11]	P&S [17]	**FaITH**
Antibiotics	Antibacterial Agents	0.93	1.00	1.00	1.00	0.99	1.00	1.00
Measles	Rubeola	0.91	0.92	1.01	1.00	0.99	1.03	1.00
Chicken Pox	Varicella	0.97	1.00	1.00	1.00	0.99	1.00	1.00
Down Syndrome	Trisomy 21	0.87	1.00	1.00	1.00	0.99	1.00	1.00
Seizures	Convulsions	0.84	0.88	1.04	0.90	0.81	1.10	0.99
Pain	Ache	0.87	0.86	1.00	1.00	0.99	1.00	0.95
Malnutrition	Nutritional Deficiency	0.87	0.62	1.00	1.00	0.98	1.00	0.87
Myocardial Ischemia	Myocardial Infarction	0.75	0.59	0.92	0.89	0.80	0.85	0.83
Hepatitis B	Hepatitis C	0.56	0.65	0.82	0.86	0.66	0.70	0.79
Pulmonary Valve Stenosis	Aortic Valve Stenosis	0.53	0.65	0.78	0.81	0.66	0.64	0.76
Psychology	Cognitive Science	0.59	0.68	0.77	0.81	0.80	0.63	0.75
Asthma	Pneumonia	0.37	0.51	0.79	0.87	0.52	0.66	0.75
Diabetic Nephropathy	Diabetes Mellitus	0.50	0.61	0.76	0.79	0.77	0.61	0.74
Hypothyroidism -	Hyperthyroidism	0.41	0.62	0.73	0.75	0.63	0.57	0.72
Sickle Cell Anemia	Iron Deficiency Anemia	0.44	0.60	0.72	0.79	0.36	0.56	0.71
Carcinoma	Neoplasm	0.75	0.25	0.68	0.85	0.45	0.46	0.65
Urinary Tract Infection	Pyelonephritis	0.65	0.47	0.58	0.67	0.42	0.42	0.60
Hyperlipidemia	Hyperkalemia	0.15	0.33	0.48	0.47	0.51	0.32	0.56
Lactose Intolerance	Irritable Bowel Syndrome	0.47	0.47	0.47	0.40	0.30	0.30	0.47
Adenovirus	Rotavirus	0.44	0.27	0.33	0.45	0.35	0.20	0.40
Vaccines	Immunity	0.59	0.00	0.00	0.52	0.00	0.00	0.34
Migraine	Headache	0.72	0.23	0.24	0.37	0.17	0.14	0.80
Bacterial Pneumonia	Malaria	0.15	0.00	0.00	0.20	0.13	0.00	0.22
AIDS	Congenital Heart Defects	0.06	0.00	0.00	0.27	0.10	0.00	0.18
Sarcoidosis	Tuberculosis	0.40	0.00	0.00	0.25	0.07	0.00	0.17
Anemia	Appendicitis	0.03	0.00	0.00	0.19	0.13	0.00	0.13
Meningitis	Tricuspid Atresia	0.03	0.00	0.00	0.19	0.13	0.00	0.13
Failure to Thrive	Malnutrition	0.62	0.00	0.00	0.18	0.13	0.00	0.12
Sinusitis	Mental Retardation	0.03	0.00	0.00	0.36	0.13	0.00	0.11
Hypertension	Kidney Failure	0.50	0.00	0.00	0.21	0.13	0.00	0.11
Breast Feeding	Lactation	0.84	0.00	0.00	0.04	0.08	0.00	0.03
Dementia	Atopic Dermatitis	0.06	0.00	0.00	0.16	0.10	0.00	0.00
Osteoporosis	Patent Ductus Arteriosus	0.15	0.00	0.00	0.03	0.10	0.00	0.00
Amino Acid Sequence -	AntiBacterial Agents	0.15	0.00	0.00	0.15	0.00	0.00	0.00
Otitis Media	Infantile Colic	0.15	0.00	0.00	0.07	0.08	0.00	0.00
Neonatal Jaundice	Sepsis	0.19	0.00	0.00	0.19	0.16	0.00	0.00

The P&S measure, which on WordNet similarity was the closest to FaITH, obtained even in this case a lower value of correlation. Note that the Li measure, which on WordNet obtained a remarkable value of correlation, obtained the lowest correlation on MeSH. We hypothesize that this can be due to two reasons. First, the Li measure depends on two parameters to correctly balance the contribution of the path between c_1 and c_2 to be compared and the depth of their $msca(c_1, c_2)$. Hence, it is possible that parameter values that achieved a good correlation in WordNet do not obtain the same (comparable) performance in MeSH. The second reason is related to the structure of the considered ontology. MeSH is a more domain-specific ontology than WordNet and therefore, in MeSH the combination of path and depth in a non linear function as suggested by the Li measure could not have the same consistent interpretation as in WordNet. The three information content measures obtained better correlation,

with Resnik's measure showing a slightly higher level of correlation. This trend is in contrast with the results obtained by the same measure on WordNet where it obtained the lowest correlation both on the $M\&C$ and $R\&G$ datasets. This fact can be justified assuming the in MeSH the $msca(c_1, c_2)$ better expresses the amount of information shared by two terms. Finally, even on this dataset the FaITH measure obtained the highest correlation. In this case the value of correlation is lower than that obtained on WordNet. Results are significant due to the very low value of *p-value* (i.e., $p - value < 0.001$).

5 Concluding Remarks and Future Work

This paper described a new model of similarity combining features [24] and information-content [19]. In particular, by exploiting a ratio-based formulation of the feature model a family of similarity measures as reported in Table 2 has been defined. One on these measures, called FaITH, to quantify how two ontology concepts are similar to each other, has been presented. Another contribution of this paper is the definition of Extended Information Content (eIC) that enables to compute relatedness between concepts by taking into account relations beyond subsumption. The proposed framework enabled to rewrite existing IC-based measures with significant improvement in their performance.

There are at least two interesting strands for future research. One is how to extend the framework to Description Logics (DLs). The main aspect that should be addressed is how to express Extended Information Content values for concepts defined in DLs. Moreover, investigating how similarity depends on the expressiveness of the considered DL is another interesting concern.

The second aspect we want to address is how this strategy, and in particular FaITH, works in more targeted applications such as document clustering, information retrieval and query answering across ontologies.

References

1. Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., Soroa, A.: A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches In *Proc. of NAACL-HLT*, 2009.
2. Borgida, A., Walsh, T., Hirsh, T.: Towards Measuring Similarity in Description Logics In *Proc. of Description Logics*, 2005.
3. Danushka, B., Yutaka, M., Mitsuru, I.: Measuring Semantic Similarity Between Words using Web Search Engines. In *Proc. of WWW2007*, pp. 757-766, 2007.
4. D ' Amato, C.: Similarity-based Learning Methods for the Semantic Web. *PhD Thesis*, University of Bari, 2007.
5. Son, J. Y, Goldstone, R. L.: The Transfer of Scientific Principles using Concrete and Idealized Simulation. *The Journal of the Learning Sciences*, (14), pp. 69-110, 2005.
6. Hirst, G., St-Onge, D.: Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms In *C. Fellbaum (Ed.), WordNet. An Electronic Lexical Database*, Chp. 13, pp. 305-332.

7. Hliaoutakis, A.: Semantic Similarity Measures in MeSH Ontology and their Application to Information Retrieval on Medline,. Technical report, Technical Univ. of Crete, Dept. of Electronic and Computer Engineering, 2005.
8. Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E. G. M., Milios, E. E.: Information Retrieval by Semantic Similarity. *Int. J. SWIS*, 2(3), pp. 55-73, 2006.
9. Jiang, J.J., Conrath, D.W.: Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In *Proc. of ROCLING X*, 1997.
10. Leacock, C., Chodorow, M.: Combining Local Context and WordNet Similarity for Word Sense Identification. In *C. Fellbaum (Ed.), WordNet. An Electronic Lexical Database*, Chp. 11, pp. 265-283.
11. Li, Y., Bandar, A., McLean, D.: An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE TKDE*, 15(4), pp. 871-882.
12. Lin, D.: An Information-theoretic Definition of Similarity. In *Proc. of Conf. on Machine Learning*, pp. 296-304, 1998.
13. Miller, G.A.: WordNet an on-line Lexical Database. *International Journal of Lexicography*, 3(4), pp.235-312, 1990.
14. Miller, G.A., Charles W.G.: Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, (6), pp. 1-28, 1991.
15. Banerjee, S., Pedersen, T.,: Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proc. of IJCAI*, pp. 805-810, 2003.
16. Pirró, G. , Ruffolo, M., Talia, D.: SECCO: On Building Semantic Links in Peer to Peer Networks. *Journal on Data Semantics*, XII, pp. 1-36, 2009.
17. Pirró, G.: A Semantic Similarity Metric Combining Features and Intrinsic Information Content. *Data Knowl. Eng*, 68(11), pp. 1289-1308, 2009.
18. Rada, R., Mili, H., Bicknell, M., Blettner, E.: Development and Application of a measure on Semantic Nets. *IEEE TSMC*, (19), pp. 17-30, 1989.
19. Resnik, P.: Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proc. of IJCAI*, pp. 448-453, 1995.
20. Rodriguez, M.A., Egenhofer, M.J.: Determining Semantic Similarity among Entity Classes from Different Ontologies. *IEEE TKDE*, 15(2), pp. 442-456, 2003.
21. Rubenstein, H., Goodenough, J. B.: Contextual Correlates of Synonymy. *CACM*, 8 (10), pp. 627-633, 1965.
22. Schickel-Zuber, V., Faltings, B.: OSS: A Semantic Similarity Function based on Hierarchical Ontologies. In *IJCAI*, pp. 551-556, 2007.
23. Seco, N., Veale, T., Hayes, J.: An Intrinsic Information Content measure for Semantic Similarity in WordNet. In *Proc. of ECAI 2004*, pp. 1089-1090, 2004.
24. Tversky, A.:Features of Similarity. *Psychological Review*, 84 (2), pp. 327-352, 1977.
25. Wang, J., Du, Z., Payattakool, R., Yu, P., Chen, C.: A New Method to Measure the Semantic Similarity of GO Terms. *Bioinformatics*, 23(10), pp. 1274-1281, 2007.
26. Watanable, S.: *Knowing and Guessing: A Quantitative Study of Inference and Information*, Wiley, 1969.
27. Wu, Z., Palmer, M.: Verb semantics and Lexical Selection. In *Proc. of FQAS ACL-94*, pp. 133-138, 1994.