Semantic Web 0 (0) 1
IOS Press

How to improve Jaccard's feature-based similarity measure

Silvia Likavec, Ilaria Lombardi * and Federica Cena
Dipartimento di Informatica, Università di Torino, Corso Svizzera 185, 10149 Torino, Italy
E-mail: likavec@di.unito.it, lombardi@di.unito.it, cena@di.unito.it

Abstract. Similarity is one of the most straightforward ways to relate two objects and guide the human perception of the world. It has an important role in many areas, such as Information Retrieval, Natural Language Processing (NLP), Semantic Web and Recommender Systems. To help applications in these areas achieve satisfying results in finding similar concepts, it is important to simulate human perception of similarity and assess which similarity measure is the most adequate.

In this work we wanted to gain some insights into Tversky's and more specifically Jaccard's feature-based semantic similarity measure on instances in a specific ontology. We experimented with various variations of this measure trying to improve its performance. We propose Sigmoid similarity as an improvement of Jaccard's similarity measure. We also explored the performance of some hierarchy-based approaches and showed that feature-based approaches outperform them on two specific ontologies we tested. We also tried to incorporate hierarchy-based information into our measures and, even though they do bring some slight improvement, it seems that it is not worth complicating the measures with this information, since the measures only based on features show very comparable performance. We performed two separate evaluations with real evaluators. The first evaluation includes 137 subjects and 25 pairs of concepts in the recipes domain and the second one includes 147 subjects and 30 pairs of concepts in the drinks domain. To our knowledge these are some of the most extensive evaluations performed in the field.

Keywords: Similarity, properties, feature-based similarity, hierarchy, ontology, instances

1. Introduction

Similarity is one of the main guiding principles which humans use to categorise the objects surrounding them. Even though in psychology the focus is on how people organise and classify objects, in computer science similarity plays a fundamental role in information processing and finds its application in many areas from Artificial Intelligence to Cognitive Science, from Natural Language Processing to Recommender Systems. Semantic similarity can be employed in many areas, such as text mining, dialogue systems, Web page retrieval, image retrieval from the Web, machine translation, ontology mapping, word-sense disambiguation, item recommendation, to name just a few. Due to the widespread usage and relevance of semantic similarity, its accurate calculation which closely mirrors human judgement brings improvements in the above and many other areas.

Usually, semantic similarity measures have been tested on WordNet [11]. WordNet is a taxonomic hierarchy of natural language terms developed at Princeton University. The concepts are organised in synsets, which group the words with the similar meaning. More general terms are found at the top of the underlying hierarchy, whereas more specific terms are found at the bottom. Two important datasets used to test similarity measures on WordNet are the ones proposed by [28] and [20]. These datasets are manually composed and contain rated lists of domain-independent pairs of words.

However, with the diffusion of ontologies as knowledge representation structures in many areas, there is a need to find similar and related objects in specific domain ontologies used in various applications, rather than just testing the similarity of concepts in Wordnet

*Corresponding author. E-mail: lombardi@di.unito.it

. In these cases, the features of domain objects play an important role in their description, along with the underlying hierarchy which organises the concepts into more general and more specific concepts. The experiments with feature-based and hierarchy-based semantic similarity measures on specific domain ontologies are rare and without conclusive results [23,34]. Hence, we decided to carry out some experiments with Jaccard's feature-based similarity measure on two specific domain ontologies, trying to draw some conclusions on the best use practices and case specific characteristics. For our first experiment, we chose the domain of recipes using the slightly modified Wikitaaable dataset[1] used in [7]. For our second experiment, we chose the domain of drinks, developed previously by our research team for different purposes. Actually, it is difficult to find a publicly available ontology with defined properties, which is an important requirement for testing our approach. In addition, the domains which could be tested with non-expert users are very limited, since the users should be able to evaluate the similarity of all (or at least most of) the couples proposed in the test. This is the reason we could not have used any of the medical ontologies available, since in medical domain only expert evaluators are needed.

There are many different similarity measures around. It is not very clear which measure is the most suitable in which situation and comparative studies are rare [21,29]. Above all, the evaluation of the measures with users is limited often to very few participants. On the contrary, our experiments involve 137 and 147 real evaluators respectively, which is a significant number of participants compared to other studies.

Thus, we aimed to gain more insight into the behaviour of Jaccard's feature-based similarity measure for ontology instances calculated from property-value pairs for compared objects and compare its performance to hierarchy-based measures. We experimented with various variations of Jaccard's feature-based similarity measure and we report here our findings. The same variations of Jaccard's measure (and more generally Tversky's measure) could be applied to other feature-based similarity measures as well. Also, we tried to combine Jaccard's similarity measure with hierarchy based approaches.

Our aim was to avoid any dependency on the weighting parameters which mark the contribution of each feature (known also as relevance factors). These parameters can be tuned for each single domain, but we wanted to test the contribution of each feature with its equal share.

The main contributions of this work are the following:

1. a proposal for new feature-based similarity measures, which could take underlying hierarchy into account
2. new datasets which can be used for the evaluation of feature-based similarity measures.

The paper is organised as follows. In Section 2 we provide the background on the basic concepts we use in our work. In Section 3, we provide some details about the semantic similarity measures we will be dealing with: feature-based semantic similarity measures and hierarchy-based semantic similarity measures. For the sake of completeness we also give some background on Information Content similarity measures, although we will not be dealing with them in this paper. In Section 4 we report on the semantic similarity measures we dealt with and propose six possible improvements of the basic Jaccard's similarity measure. We also give details of the variations of each of these six similarity measures which might include or not the hierarchy-based similarity. The results of our extensive evaluation are reported in Section 5 followed by a Section 6 which summarises some additional works which exploit feature-based semantic similarity measures. We conclude in Section 7 drawing some conclusions and pointing some directions for future work.

2. Background on semantic knowledge representation

This section provides a background on the main notions used in this work.

2.1. Conceptual hierarchies

A *conceptual hierarchy* provides a taxonomy (a tree or a lattice) of concepts organised using the partial order IS-A relation, which specialises more general classes into more specific classes [3,32]. The IS-A relation is asymmetric and transitive and defines a hierarchical structure of the ontology, enabling the inheritance of characteristics from parent classes to descendant classes. W.r.t. to similarity calculation, it enables the employment of hierarchy-based approaches.

[1] http://wikitaaable.loria.fr/rdf/

2.2. Ontologies

Ontologies enable explicit specification of domain elements and their properties, hierarchical organisation of domain elements, exact description of any existing relationships between them and employment of rigorous reasoning mechanisms. An ontology can be seen as a "formal, explicit specification of a shared conceptualisation" [12]. They are expressed with standard formats and languages (e.g., OWL.[2]), which allow for extensibility and re-usability.

Two layers can be identified in an ontology: *ontology definition layer* which contains the classes which describe the concepts in the domain, and the *instance layer* which contains all the distinct individuals in the domain.

Relations between resources are defined by means of properties. Two types of properties exist:

(i) *object properties* linking individuals among themselves;

(ii) *data type properties* linking individuals and data type values.[3]

In this work, we only considered object properties, since the treatment of data type properties (such as literal values) requires further semantic analysis.

Instances in the ontology (also called individuals) describe concrete individuals. They are defined with individual axioms which provide their class memberships (property `rdf:type`), individual identities and values for their properties. All the properties of instances are inherited from the classes the instances belong to. A specific value is associated to each property and some properties can have more than one value. Properties of instances enable the employment of feature-based similarity measure to ontology instances.

3. Semantic similarity measures

In this section we give some details of the three main categories of semantic similarity measures, namely feature-based, hierarchy-based and information content-based. As a result of trying to improve and combine the above approaches, many hybrid similarity measures were born. In our experiments we only dealt with Tversky's feature-based measure, its improvements and its combination with hierarchy-based measures. We include information-content-based measures for the completeness sake only.

3.1. Feature-based similarity measures

Calculation of similarity based on properties goes back to Tversky's work on Features of Similarity [31] where similarity between two objects O_1 and O_2 is a function of their common and distinctive features:

$$\text{SIM}_T(O_1, O_2) = \\ = \frac{\alpha(\psi(O_1) \cap \psi(O_2))}{\beta(\psi(O_1) \setminus \psi(O_2)) + \gamma(\psi(O_2) \setminus \psi(O_1)) + \alpha(\psi(O_1) \cap \psi(O_2))}. \quad (1)$$

In the formula above $\psi(O)$ describes all the relevant features of the object O, and $\alpha, \beta, \gamma \in \mathbb{R}$ are constants which allow for different treatment of the various components. For $\alpha = 1$ common features of the two objects have maximal importance. For $\beta = \gamma$ non-directional similarity measure is obtained. Depending on the values for α, β, γ, we obtain the following variations of Tversky's similarity:

- Jaccard's or Tanimoto similarity for $\alpha = \beta = \gamma = 1$;
- Dice's or Sørensen's similarity for $\alpha = 1$ and $\beta = \gamma = 0.5$.

We will be using the following notation:

- *common features of* O_1 *and* O_2: $\text{CF}(O_1, O_2) = \psi(O_1) \cap \psi(O_2)$,
- *distinctive features of* O_1: $\text{DF}(O_1) = \psi(O_1) \setminus \psi(O_2)$ and
- *distinctive features of* O_2: $\text{DF}(O_2) = \psi(O_2) \setminus \psi(O_1)$.

In order to calculate the above similarities for domain objects O_1 and O_2, we need to calculate for each property p they have in common, how much it contributes to common features of O_1 and O_2, distinctive features of O_1 and distinctive features of O_2, respectively. We denote these values by CF_p, DF_p^1 and DF_p^2.

Hence, we have to compare the property-value pairs between instances for each property they have in common. We will include in common features the cases when the two objects have the same value for the given property p. We will include in distinctive features of each object the cases when the two objects have different values for the given property p.

We consider equal the properties defined with owl:EquivalentProperty. Repeating the above process

[2] `http://www.w3.org/TR/owl-ref`

[3] In OWL, there is also the notion of annotation property (owl:AnnotationProperty) and ontology property (owl:OntologyProperty), used in OWL DL.

for each property O_1 and O_2 have in common, we obtain all common and distinctive features of O_1 and O_2:

$$\text{CF}(O_1, O_2) = \Sigma_{i=1}^{n} \text{CF}_{p_i}(O_1, O_2)$$

$$\text{DF}(O_1) = \Sigma_{i=1}^{n} \text{DF}_{p_i}^{1}(O_1) \qquad \text{DF}(O_2) = \Sigma_{i=1}^{n} \text{DF}_{p_i}^{2}(O_2)$$

where n is the number of common properties defined for O_1 and O_2. Then the above similarity measures become:

Jaccard's similarity:

$$\text{SIM}_J(O_1, O_2) = \frac{\text{CF}(O_1, O_2)}{\text{DF}(O_1) + \text{DF}(O_2) + \text{CF}(O_1, O_2)} \quad (2)$$

Dice's similarity:

$$\text{SIM}_D(O_1, O_2) = \frac{2\text{CF}(O_1, O_2)}{\text{DF}(O_1) + \text{DF}(O_2) + 2\text{CF}(O_1, O_2)} \quad (3)$$

3.1.1. Mathematical properties

Here we provide the reader with some basic mathematical properties of the Jaccard's similarity measure which we would deal with in the rest of the paper.

1. Boundaries
   $\forall O_1, O_2, 0 \leq \text{SIM}_J(O_1, O_2) \leq 1$.
2. Maximal similarity
   If $O_1 \equiv O_2$, then $\text{SIM}_J(O_1, O_2) = 1$.
3. Commutativity
   $\forall O_1, O_2, \text{SIM}_J(O_1, O_2) = \text{SIM}_J(O_2, O_1)$.
4. Monotonicity
   If $\text{CF}(O_1, O_2) \leq \text{CF}(O_1, O_3)$ and $\text{DF}(O_1) = \text{DF}(O_2)$, then $\text{SIM}_J(O_1, O_2) \leq \text{SIM}_J(O_1, O_3)$.
   If $\text{CF}(O_1, O_2) \leq \text{CF}(O_1, O_3)$ and $\text{DF}(O_2) = \text{DF}(O_3)$, then $\text{SIM}_J(O_1, O_3) \leq \text{SIM}_J(O_2, O_3)$.

3.2. Hierarchy-based similarity measures

Hierarchy-based or distance-based similarity measures use the underlying conceptual hierarchy directly and calculate the distance between concepts by calculating the number of edges or the number of nodes which have to be traversed in order to reach one concept from another. The hierarchy-based measure was first introduced in [24] as a simple shortest path connecting the compared concepts and was the basis for the development of many measures of semantic similarity. A discussion and comparison with information content-based approaches can be found in [5]. In this section we give more details about three hierarchy-based measures we used in our experiments: Leacock and Chodorow's measure [15], Wu and Palmer's measure [33] and Li's measure [16].

3.2.1. Leacock and Chodorow's similarity measure

The first measure we will look at is Leacock and Chodorow's similarity measure [15] which was originally used for word sense disambiguation in a local context classifier. The most similar nouns from the training set are substituted for the ambiguous ones in testing. In order to calculate the distances between words, the authors use the normalised path length in WordNet [11] between all the senses of the compared words and measure the path length in nodes:

$$\text{SIM}_{LC}(a, b) = -\log\left(\frac{Np}{2 \times \text{MAX}}\right). \quad (4)$$

Np is the number of nodes in the path p from a to b, whereas MAX is the maximum depth of the hierarchy. The distance between two words belonging to the same synset is 1.

If we want to express this measure as a function of distances between nodes we obtain the following formula:

$$\text{SIM}_{LC}(a, b) = -\log\left(\frac{\text{DIST}(a, b)}{2 \times \text{MAX}}\right). \quad (5)$$

$\text{DIST}(a, b)$ is the distance between a and b calculated as the shortest path length between these two nodes.

The disadvantage of this similarity measure is that many pairs of non-similar words are estimated as similar, due to the equal edge lengths in their hierarchy.

3.2.2. Wu and Palmer's similarity measure

Wu and Palmer's similarity measure [33] is based on the depths in the hierarchy of the two words being compared and on the depth of their common subsumer, which characterises their commonalities. If we denote by c the first subsuming node for the two compared nodes a and b, their similarity is calculated as follows:

$$\text{SIM}_{WP}(a, b) = \frac{2N_c}{N_a + N_b}. \quad (6)$$

N_n is the number of nodes along the path from the node n to the root.

This measure can also be expressed as a function of distances between nodes as follows:

$$\text{SIM}_{WP}(a, b) = \frac{2\text{DIST}(c, r)}{\text{DIST}(a, r) + \text{DIST}(b, r)}. \quad (7)$$

DIST(n, r) is the distance of n from the root, again calculated as the shortest path length between the two nodes. In [33] this measure was used in lexical selection problems in machine translation where the performance of inexact matches based on verb meanings is evaluated.

3.2.3. Li's similarity measure

Li et al. [16] proposed an approach for calculating the similarity between sentences which uses semantic information and word order information in the sentence. Similarity of singular words is calculated using the shortest path length between words DIST and the depth h of their common subsumer as follows:

$$\text{SIM}_L(a, b) = e^{-\alpha \text{DIST}(a,b)} \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}. \quad (8)$$

where $\alpha \in [0, 1], \beta \in (0, 1]$ are parameters which control the contribution of shortest path length and depth, respectively. As $\beta \to \infty$, the depth of a word in the semantic nets is not considered. Their optimal values depend on the knowledge base used and for WordNet they are $\alpha = 0.2$ and $\beta = 0.45$. For the proposed benchmark dataset, the optimal values are $\alpha = 0.2$ and $\beta = 0.6$ (obtained experimentally). If the words a and b belong to the same synset then DIST$(a, b) = 0$, if they do not belong to the same synset but the synsets for a and b contain one or more common words, then DIST$(a, b) = 1$ and for the remaining cases the actual path length between a and b is calculated. This measure was introduced for purely theoretical purposes, as an improvement of the existing similarity measures.

3.3. Information content-based similarity measures

The measures seen above work on single knowledge structures and they do not need external sources for similarity calculation. In this section we present the most important approach which uses an external corpus to compute the similarity. The foundational work on information content-based similarity is due to Resnik [25,26]. His approach is based on the fact that the more abstract classes provide less information content, whereas more concrete and detailed classes lower down in the hierarchy are more informative. The closest class that subsumes two compared classes, called a *most informative subsumer* is the class which provides the shared information for both, and measures their similarity. In order to calculate the information content of a concept in a IS-A taxonomy, Resnik turns to an external text corpus and calculates the probability of occurrence of the class in this corpus as its relative frequency (each word in the text corpus is counted as an occurrence of each class that contains it.). The information content of a class in a taxonomy is given by the negative logarithm of the probability of occurrence of the class in a text corpus as follows:

$$\text{SIM}_R(a, b) = max_{c \in S(a,b)}[-\log p(C)] \quad (9)$$

where $p(c)$ is the probability of encountering an instance of concept c, and $S(a, b)$ is the set of all concepts that subsume a and b. According to Resnik this approach performs better than hierarchy-based approaches, based on human similarity judgements as a benchmark. He used this similarity measure to resolve the problems of ambiguity in natural language.

3.4. Advantages and disadvantages of hierarchy-based and information content-based similarity measures - a discussion

Since they only depend on the underlying hierarchy of the domain ontology, the hierarchy-based similarity measures are fairly simple and require a low computational cost. The known problem with hierarchy-based similarity measures is that all the edges in the hierarchy are considered to be of the same length, so many similarity values are not correct. The accuracy of these measures have been surpassed by more complex approaches which exploit additional semantic information.

The problem with Resnik's similarity measure is the excessive information content value of many polysemous words (i.e. word senses not taken into account) and multi-worded synsets. Also, the information content values are not calculated for individual words but for synsets, hence the synsets containing commonly occurring ambiguous words would have excessive information content values. To deal with the problem of excessive information content value of many polysemous words, Resnik proposes *weighted word similarity* which takes into account all senses of the words being compared.

The main problem with information content-based similarity measures is their dependance on external corpora for the calculation of term frequencies. This requires disambiguation and annotation of terms in the corpus, very often done manually, hence affecting the applicability of this approach to large corpora. Also, with the change of corpora or the ontology, the term frequencies have to be recalculated. One step towards mitigating this problem was the introduction of

intrinsic computation of information content [30] as we will see in Section 6. The methods based on intrinsic information content outperform corpora-relying approaches.

In this work, we particularly focus on feature-based and hierarchy-based similarity measures since they do not require external knowledge sources for their application, rather they rely solely on the domain ontology. Feature-based similarity measures evaluate both common and distinctive features of compared objects, hence exploiting more semantic information than hierarchy-based approaches. But this information has to be available, which is not always the case. Otherwise the applicability and accuracy of these measures is hindered. Feature-based similarity measures can also be applied in cross-ontology settings.

4. Variations of Jaccard's feature-based similarity measure

In this section we present the variations of Tversky's, or more preciselly, Jaccard's similarity measure from Section 3.1 which we evaluated in our experiments. We experimented with 6 basic modifications of Jaccard's similarity (see Equation 2). For each of these 7 measures, we present 2 further variations, which aim to take the underlying hierarchical structure into account in different ways. We actually did the same calculations also with Dice's measure (see Equation 3) but the results were almost always worse, so we will not tackle Dice's measure any further.

4.1. Basic modifications of Jaccard's similarity measure

Our first assumption was that common features contribute to the similarity calculation in more substantial way than distinctive features. Hence, the first modification of Jaccard's measure, called **Common-squared Jaccard's similarity**, was to consider squares of only the values two objects have in common, giving more importance to common features. The following formula illustrates this measure:

$$\text{SIM}_{CSQ}(O_1, O_2) = \frac{\text{CF}^2(O_1, O_2)}{\text{DF}(O_1) + \text{DF}(O_2) + \text{CF}^2(O_1, O_2)} \tag{10}$$

For comparison purposes, the second modification of Jaccard's measure, called **Squared Jaccard's similarity**, was to consider squares of all the values as in the following formula:

$$\text{SIM}_{SQ}(O_1, O_2) = \frac{\text{CF}^2(O_1, O_2)}{\text{DF}^2(O_1) + \text{DF}^2(O_2) + \text{CF}^2(O_1, O_2)} \tag{11}$$

We also tried the following two modifications: **Normalised Jaccard's similarity** (Eq 12) and **Normalised common-squared Jaccard's similarity** (Eq 13), given below:

$$\text{SIM}_N(O_1, O_2) = \\ = \frac{\frac{\text{CF}(O_1,O_2)}{(\text{CF}(O_1,O_2)+\text{DF}(O_1))(\text{CF}(O_1,O_2)+\text{DF}(O_2))}}{\frac{\text{DF}(O_1)}{\text{CF}(O_1,O_2)+\text{DF}(O_1)} + \frac{\text{DF}(O_2)}{\text{CF}(O_1,O_2)+\text{DF}(O_2)} + \frac{\text{CF}(O_1,O_2)}{(\text{CF}(O_1,O_2)+\text{DF}(O_1))(\text{CF}(O_1,O_2)+\text{DF}(O_2))}} \tag{12}$$

$$\text{SIM}_{NCSQ}(O_1, O_2) = \\ = \frac{\frac{\text{CF}^2(O_1,O_2)}{(\text{CF}(O_1,O_2)+\text{DF}(O_1))(\text{CF}(O_1,O_2)+\text{DF}(O_2))}}{\frac{\text{DF}(O_1)}{\text{CF}(O_1,O_2)+\text{DF}(O_1)} + \frac{\text{DF}(O_2)}{\text{CF}(O_1,O_2)+\text{DF}(O_2)} + \frac{\text{CF}^2(O_1,O_2)}{(\text{CF}(O_1,O_2)+\text{DF}(O_1))(\text{CF}(O_1,O_2)+\text{DF}(O_2))}} \tag{13}$$

The Normalised common-squared Jaccard's similarity measure for ontological objects was first introduced in [6] and was further developed in [17] as a way to calculate similarity between shapes and in [18] for improving recommendation accuracy and diversity.

We also tried to converte Li's similarity formula [16] into feature based formula. Since the similarity measure increases with the increasing number of common features, common features can be taken as an argument of the sigmoid function. Furthermore, the similarity values should decrease with the increasing number of distinctive features, hence the distinctive features should be an argument of the negative sigmoid function translated by 1. So we obtained the following function:

$$\text{SIM}_S(O_1, O_2) = \frac{e^{\text{CF}(O_1,O_2)} - 1}{e^{\text{CF}(O_1,O_2)} + 1}\left(1 - \frac{e^{\text{DF}(O_1)+\text{DF}(O_2)} - 1}{e^{\text{DF}(O_1)+\text{DF}(O_2)} + 1}\right) \\ = \frac{2(e^{\text{CF}(O_1,O_2)} - 1)}{(e^{\text{CF}(O_1,O_2)} + 1)(e^{\text{DF}(O_1)+\text{DF}(O_2)} + 1)} \tag{14}$$

This result is similar to just taking the distinctive features as an argument to inverse exponential function since these two functions have similar graphs and behaviour for arguments greater than 0. But the exponential in the denominator increases very fast, so the similarity values were getting extremely small very quickly. Hence we decided to leave just the distinctive features in the denominator. Adding 1 prevents the case of the division with zero when there are no distinctive features among the compared objects. Also, we need to divide by 2 so that the final result is in the interval $[0, 1)$. So the final **Sigmoid similarity** has the following formula:

$$\text{SIM}_S(O_1, O_2) = \frac{e^{\text{CF}(O_1,O_2)} - 1}{(e^{\text{CF}(O_1,O_2)} + 1)(\text{DF}(O_1) + \text{DF}(O_2) + 1)} \tag{15}$$

Finally, the last measure which we considered was the sigmoid function of Jaccard's similarity as follows:

$$\text{SIM}_{JS}(O_1, O_2) = \frac{e^{\text{SIM}_J(O_1,O_2)} - 1}{e^{\text{SIM}_J(O_1,O_2)} + 1} \tag{16}$$

We call this measure **Sigmoid Jaccard's similarity.**

4.1.1. Mathematical properties

It is straightforward that the previously introduced mathematical properties (boundaries, maximal similarity, commutativity and monotonicity) are preserved for SIM_{CSQ}, SIM_{SQ}, SIM_N and SIM_{NCSQ}, since they are simple modifications of the original SIM_J measure.

Let us first consider SIM_{JS} measure since it is simpler than SIM_S measure. The values of SIM_{JS} belong to $[0, \frac{e-1}{e+1})$ since the values of the sigmoid function belong to $[0, 1)$, for positive arguments. Hence, its maximum value is $\frac{e-1}{e+1}$ when $\text{SIM}_J = 1$. Sigmoid function is a monotone function so the monotonicity is preserved. Commutativity is straightforward.

As far as SIM_S measure is concerned, its values belong to $[0, 1)$ and the maximal value is obtained when the compared objects do not have any distinctive features. Commutativity and monotonicity are again straightforward.

Next, we will see how we tried to include the hierarchical information into these measures.

4.2. Feature-based similarity combined with hierarchical information

The basic question we wanted to answer with these experiments was how much the actual hierarchical information contributes to the similarity of two objects and if this information can be simulated by properties. Basically, the hierarchical information could be either integrated into the measures by considering the `rdf:type` property (basic measures) or it could be integrated into the measures by excluding the `rdf:type` property and including the hierarchical information in some different way. We also tried to see how including both would affect the performance. We decided to distinguish the following two variations of each of the above described basic measures:

V1: the measures *without considering* the property `rdf:type` but including the hierarchical information in the feature-based similarity formula. In this case, since the greater distance between two objects means that they are less similar, we decided that the distance between objects counts as their distinctive feature. In the following formulas $\text{DIST}(O_1, O_2)$ is the number of edges along the path connecting O_1 and O_2 and MAX is the maximum depth of the class hierarchy.

$$\text{SIM}_{Jnth}(O_1, O_2) = \\ = \frac{\text{CF}(O_1, O_2)}{\text{DF}(O_1) + \text{DF}(O_2) + \text{CF}(O_1, O_2) + \frac{\text{DIST}(O_1,O_2)}{2\text{MAX}}} \tag{17}$$

$$\text{SIM}_{SQnth}(O_1, O_2) = \\ = \frac{\text{CF}^2(O_1, O_2)}{\text{DF}^2(O_1) + \text{DF}^2(O_2) + \text{CF}^2(O_1, O_2) + \frac{\text{DIST}^2(O_1,O_2)}{(2\text{MAX})^2}} \tag{18}$$

The equations for SIM_{CSQnth}, SIM_{Nnth}, $\text{SIM}_{NCSQnth}$ are analogous to the equations above. The ones that are worth writing out explicitly are SIM_{Snth} and SIM_{JSnth}.

$$\text{SIM}_{Snth}(O_1, O_2) = \\ = \frac{2(e^{\text{CF}(O_1,O_2)} - 1)}{(e^{\text{CF}(O_1,O_2)} + 1)(\text{DF}(O_1) + \text{DF}(O_2) + 1 + \frac{\text{DIST}(O_1,O_2)}{2\text{MAX}})} \tag{19}$$

$$\text{SIM}_{JS}(O_1, O_2) = \frac{e^{\text{SIM}_{Jnth}(O_1,O_2)} - 1}{e^{\text{SIM}_{Jnth}(O_1,O_2)} + 1} \tag{20}$$

V2: the original measures *considering* the property `rdf:type` *and* including the hierarchical information in the feature-based similarity formula. This approach counts the hierarchical information twice in a way but in two subtle ways. Including the property `rdf:type` takes into account all the objects which are of the same recipe type, whereas including the hierarchical information accounts for the distance between the compared objects. These measures (SIM_{Jh}, SIM_{SQh}, SIM_{CSQh}, SIM_{Nh}, SIM_{NCSQh}, SIM_{Sh} and SIM_{JSh}) showed slightly better performance results with respect to the original measures, as we will see in Section 5. But, in our opinion, the gained improvement does not justify the increased complexity and execution times of the proposed variants.

Let us illustrate our ideas with some simple examples. In the recipe domain, all the recipes could be instances of the class `Recipe` or there could exist an underlying hierarchy of recipe types and each recipe could be an instance of its corresponding recipe type. In our case, instead of having all the recipes instances of `DishType`, we made the recipes instances of various dish types, such as `BreadDish`, `CakeDish`, `PastaDish` etc. This choice has the following consequences:

- instead of having the same values for the property `rdf:type` for all the dishes, we can distinguish them according to various values for the property `rdf:type`;
- in case we want to take the hierarchical information into account, we can calculate the distance between various dishes, rather than assume that they all have the same similarity, since they all have the same parent. Since we deal with the instances in the ontology, we decided to consider them "descendants" of their classes, otherwise the instances of one class would all be equal.

5. Evaluation

The most commonly used datasets to test similarity measures are the ones proposed by [28] and [20]. Rubenstein and Goodenough's experiment dates back to 1965. They asked 51 native English speakers to assess the similarity of 65 English noun pairs on a scale from 0 (semantically unrelated) to 4 (highly related). Miller and Charles' experiment in 1991 considered a subset of 30 noun pairs and their similarity was reassessed by 38 undergraduate students. The correlation w.r.t Rubenstein and Goodenough results was 0.97. [26] repeated the same experiment in 1995 with 10 subjects. The correlation w.r.t. Miller and Charles results was 0.96. Finally, [22] repeated the experiments in 2009 with 101 participants, both English and non-English native speakers. His average correlation w.r.t. Rubenstein and Goodenough was 0.97, and 0.95 w.r.t. Miller and Charles. We can see that similarity judgements by various groups of people over a long period of time remain stable.

All the experiments cited above [28,20,26,22] were dealing with common English nouns and the correlation with these results was mostly used to test similarity measures on WordNet [11]. But our focus is different. We wanted to experiment with the similarity measures on specific domain ontologies, which represents more complex entities. We needed data representation where features of domain objects are explicitly provided, which is not the case with WordNet.

Our experiment was conducted with the goal of evaluating the feature-based similarity of instances in its various forms and its comparison with hierarchy-based approaches. In our first experiment we evaluated our approach in the domain of recipes using the slightly modified Wikitaaable dataset[4] used in [7]. In our second experiment we evaluated our approach in the drinks domain using an ontology developed previously by our research team. We assumed that both datasets represent information known by a wide range of people, without the need be an expert to assess the similarity of the proposed domain items. The datasets used in our experiments are available here: http://www.di.unito.it/ lombardi/SimilarityTest/

5.1. Hypotheses

We wanted to verify that:

H1: Jaccard's feature-based similarity measure performs better than hierarchy-based approaches;

H2: it is possible to improve Jaccard's feature-based similarity;

H3: hierarchical information is encoded better with features than with underlying hierarchy;

H4: combining the hierarchy and feature-based approach beyond linear combination further improves the Jaccard's similarity measure and its variations.

[4] `http://wikitaaable.loria.fr/rdf/`

5.2. Subjects

A total of 137 people (60 female and 77 male, average age 30) took part in the first test and 147 (65 female and 82 male, average age 32) took part in the second one.

They were selected among authors' colleagues and among the second and third year undergraduate students at the Faculty of Philosophy at the University of Torino, Italy. The participants were recruited according to an availability sampling strategy.[5] All the participants were native Italian speakers.

5.3. Materials

We designed two experiments to test our hypotheses.

In the first experiment, we designed a web questionnaire with 25 pairs of recipes, chosen from Wikitaaable dataset, covering a range of recipes expected to be judged very similar to not similar at all. The original recipes from the dataset were translated to Italian. The original dataset contains 1666 recipes. The properties defined for the recipes are the following: `rdf:type`, `can_be_eaten_as`, `has_ingredient`, `suitable_for_diet`, `not_suitable_for_diet` and `has_origin`. This dataset provided us with a good setting to test our approach. We only made the following slight modifications: in the original ontology all the recipes were instances of `Recipe`, whereas we needed to use the hierarchical structure of the ontology to test hierarchy-based similarity measures, as well as incorporate the hierarchical information into our measures, hence we made the recipes instances of various `Dish_Type`'s. In the original ontology `rdf:type` is a property, whereas in our case, we once used it as a property and once as an underlying hierarchical information. Figure 1 shows the basic Wikitaaable taxonomy of recipes, but only including the top categories and the ones from which we used the instances to test our approach. The properties are omitted for clarity.

The main dish types are: `RollDish`, `CrepeDish`, `BakedGoodDish`, `BeverageDish`, `SoupAndStuffDish`, `PancakeDish`, `MousseDish`, `SweetDish`, `SaladDish`, `SaltedDish`, `SauceDish`, `WaffleDish`, `PreserveDish`, `SpecialDietDish`. The recipes we used in our evaluation did not belong to all recipe groups, since the complexity of the test would have been too high and we would have obtained random answers from the users.

In the second experiment, we again used a web questionnaire, this time containing 30 pairs of drinks chosen from the ontology describing drinks. The original ontology has 148 classes, with the main drink classes being: `Water`, `AlcoholicBeverage`, `DrinkInACup`, `PlantOriginDrink` and `SoftDrink`. The properties defined for the drinks are the following: `has_alcoholic_content`, `has_caloric_content`, `has_ingredient`, `is_sparkling`, `is_suitable_for` and `rdf:type`. Figure 2 shows the basic taxonomy of drinks, again only including the main categories and omitting the properties.

In each experiment the participants were asked to asses the similarity of these 25 pairs (respectively 30 pairs) by anonymously assigning them a similarity value on the scale from 1 to 10 (1 meaning not similar at all, 10 meaning very similar). The ordering of pairs was random. The users' similarity values were then turned into the values from the [0, 1] interval to make them match the similarity values produced by various algorithms.

In each experiment, the participant group P was divided into two groups: P1 was used as a reference group and P2 was used as a control group to measure the correlation of judgments among human subjects, as in [25]. We experimented with different partitions of P1 and P2 and obtained similar results.

In this paper we considered the following feature-based similarity measures:

1. Tversky's similarity where $\alpha = \beta = \gamma = 1$, also known as Jaccard's or Tanimoto similarity SIM_J;
2. Common-squared Jaccard's similarity SIM_{CSQ};
3. Squared Jaccard's similarity SIM_{SQ};
4. Normalized Jaccard's similarity SIM_N;
5. Normalized common-squared Jaccard's similarity SIM_{NCSQ};
6. Sigmoid similarity SIM_S;
7. Sigmoid Jaccard's similarity SIM_{JS};

For the comparison, we considered the following edge-based similarity measures:

1. Leacock and Chodorow's similarity SIM_{LC};
2. Wu and Palmer's similarity SIM_{WP};
3. Li's similarity SIM_L.

[5] Although a more suitable way to obtain a representative sample is random sampling, this strategy is time consuming and not financially justifiable. Hence, much research is based on samples obtained through non-random selection, such as the availability sampling, i.e. a sampling of convenience, based on subjects available to the researcher, often used when the population source is not completely defined.

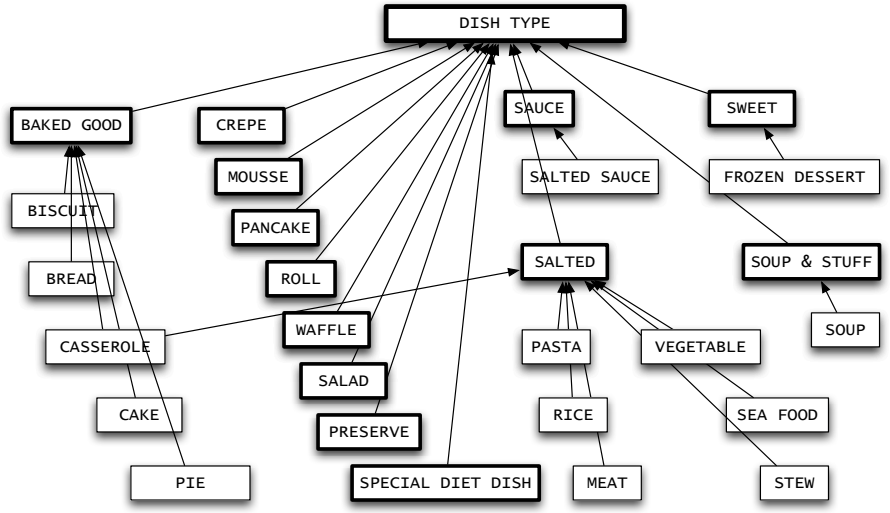

Fig. 1. Recipe taxonomy

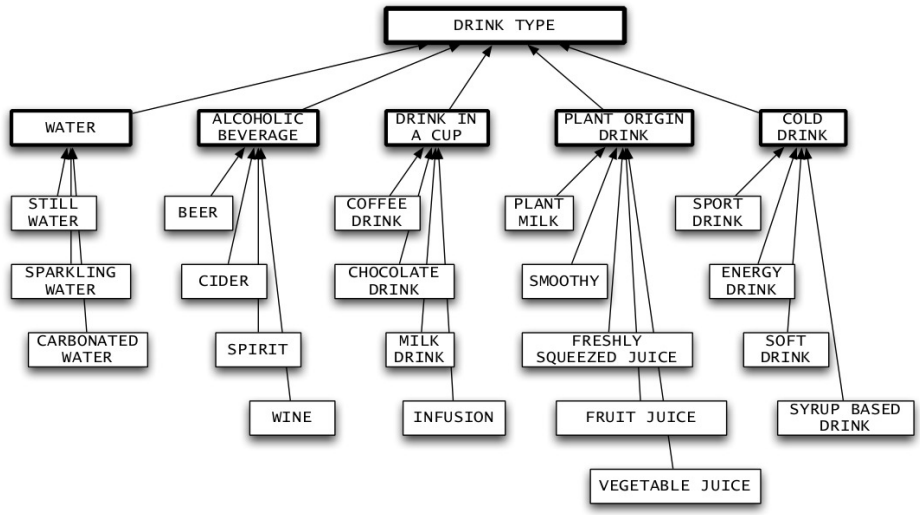

Fig. 2. Drinks taxonomy

For each of the feature-based measures, we considered the following variants:

V1. the measure *without* the `rdf:type` property but combined with underlying hierarchy (measures SIM_{Jnth}, SIM_{SQnth}, SIM_{CSQnth}, SIM_{Nnth}, $\text{SIM}_{NCSQnth}$, SIM_{Snth}, SIM_{JSnth});

V2. the measure *with* the `rdf:type` property combined with underlying hierarchy (measures SIM_{Jh}, SIM_{SQh}, SIM_{CSQh}, SIM_{Nh}, SIM_{NCSQh}, SIM_{Sh}, SIM_{JSh}).

For all the feature-based measures we also tested the Dice's similarity, but in almost all the cases it showed worse performance than Jaccard's similarity, so we will not report the results here.

5.4. Measures

We used the Spearman rank correlation coefficient ρ to measure the accuracy of similarity judgement.

The Spearman rank correlation coefficient measures statistical dependence between two ranked variables. It actually describes the relationship between two variables using a monotonic function. It is equal to the Pearson correlation between the ranked variables. For a sample of size n, the two sets of values

$X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$ are converted to ranks $\{r(x_1), \ldots, r(x_n)\}$ and $\{r(y_i), \ldots, r(y_n)\}$ and then the Pearson coefficient is calculated as follows:

$$\rho = \frac{\Sigma_{i=1}^{n}(r(x_i) - \overline{r(x)})(r(y_i) - \overline{r(y)})}{\sqrt{\Sigma_{i=1}^{n}(r(x_i) - \overline{r(x)}^2(r(y_i) - \overline{r(y)})^2}}$$

where $\overline{r(x)}$ and $\overline{r(y)}$ are the sample means of the two sets of ranked values. The value of Spearman coefficient ranges from +1 (indicating a strong similar rank) to −1 (indicating a strong dissimilar rank). Value 0 means there is no correlation.

5.5. Implementation and performance

The test software to compare the different measures has been developed in Java, using the Apache Jena library to extract data from the ontology. The test has been performed on a MacBook Pro with a 2.66 GHz Intel Core i7 processor and 8 GB 1067 MHz DDR3 RAM. Tables 1 and 2 show the execution time for the different measures obtained in the Recipes and the Drinks experiments, respectively. In both cases, feature-based approaches significantly outperform hierarchy-based ones. A possible explanation for this could be that hierarchy-based approaches need to traverse the hierarchy to calculate the distance between two nodes, while feature-based ones simply need to extract properties and their values for the two nodes.

5.6. Results and discussion

For the first experiment (Recipes) Table 3 reports the Spearman rank correlation between the reference group P1 and control group P2, as well as the Spearman rank correlation between the participant group P and the hierarchy-based approaches. Tables 4, 5, 6 report the Spearman rank correlation between each of the above measures and the participant group P for the base case and for the 2 variations introduced in Section 4. The best performing measure in each group is reported in bold.

For the second experiment (Drinks) Table 7 reports the Spearman rank correlation between the reference group P1 and control group P2, as well as the Spearman rank correlation between the participant group P and the hierarchy-based approaches. Tables 8, 9, 10 report the Spearman rank correlation between each of the above measures and the participant group P for the base case and for the 2 variations introduced in Section 4.

5.6.1. Participants group and the hierarchy-based approaches

The correlation for both experiments with the human subjects, i.e. the comparison of the two groups of participants (row P1-P2), is 0.945 (respectively 0.977) and similar to the one reported in [25]. From this good correlation we can conclude that the participants' responses were coherent among themselves in both experiments and that we can trust the human ratings.

Furthermore, w.r.t. the correlation of participants group with hierarchy-based approaches we can see that in the first experiment (Recipes) with hierarchy-based approaches we obtained similar results for all three measures (the best one being Wu and Palmer's measure) and they all have a relatively weak positive correlation with the human judgement. In the second experiment (Drinks) the performance is better (the best one again being Wu and Palmer's measure).

This can be explained with the fact that the underlying hierarchy in the ontology of drinks is designed mirroring better the human categorisation than the ontology of recipes (which was indeed flat at the beginning and we performed only minimal changes to obtain the main categories of recipes). But it also shows how dependant the measure is on the ontology design.

5.6.2. Base case group of feature-based similarities and the comparison with hierarchy-based approaches

The first group of feature-based similarity measures contains the original Jaccard's similarity measure SIM_J and the six basic modifications proposed in Section 4 (SIM_{CSQ}, SIM_{SQ} SIM_N, SIM_{NCSQ}, SIM_S and SIM_{JS}).

Tables 4 and 8 show the results for the base case for all 7 similarity measures in both experiments, where the property `rdf:type` was included in the calculations and no further hierarchical information was taken into account.

Looking at the Tables 4 and 8 we can see that in each experiment there are a few measures which improve the basic Jaccard's similarity SIM_J . But the one that is consistently better is Sigmoid similarity SIM_S.

Hence, we confirmed our hypothesis H2 that it is possible to improve the original Jaccard's formulation of similarity measure by using the Sigmoid similarity measure SIM_S.

Furthermore, in the first experiment (Recipes) all the feature-based measures in Table 4 perform better than the hierarchy-based measures from Table 3. In

MEASURE	execution time (in seconds)
Leacock and Chodorow	0.413
Wu and Palmer	0.225
Li et al.	0.267
Jaccard	0.028
Sq. Jaccard	0.010
Comm-sq. Jaccard	0.011
Norm. Jaccard	0.008
Norm. comm-sq. Jacc.	0.010
Sigmoid	0.007
Sigmoid Jaccard	0.005

Table 1
Execution time for the Recipes ontology

MEASURE	execution time (in seconds)
Leacock and Chodorow	0.061
Wu and Palmer	0.047
Li et al.	0.064
Jaccard	0.018
Sq. Jaccard	0.008
Comm-sq. Jaccard	0.006
Norm. Jaccard	0.005
Norm. comm-sq. Jacc.	0.004
Sigmoid	0.003
Sigmoid Jaccard	0.003

Table 2
Execution time for the Drinks ontology

MEASURE	ρ
P1 - P2	**0.945**
Leacock and Chodorow	0.583
Wu and Palmer	0.585
Li et al.	0.576

Table 3
Recipes - P1 - P2 and edge-based

MEASURE	ρ
Jaccard	0.633
Sq. Jaccard	0.644
Comm-sq. Jaccard	0.639
Norm. Jaccard	0.636
Norm. comm-sq. Jacc.	0.629
Sigmoid	**0.664**
Sigmoid Jaccard	0.633

Table 4
Recipes - base case

MEASURE	ρ
Jaccard no t. + h.	**0.614**
Sq. Jaccard no t. + h.	0.608
Comm-sq. Jaccard no t. + h.	0.606
Norm. Jaccard no t. + h.	0.578
Norm. comm-sq. Jacc. no t. + h.	0.605
Sigmoid no t. + h.	0.583
Sigmoid Jaccard no t. + h.	0.613

Table 5
Recipes - base case without `rdf:type` with hierarchy

MEASURE	ρ
Jaccard + h.	0.643
Sq. Jaccard + h.	0.639
Comm-sq. Jaccard + h.	0.644
Norm. Jaccard + h.	0.636
Norm. comm-sq. Jacc. + h.	0.638
Sigmoid + h.	**0.664**
Sigmoid Jaccard + h.	0.643

Table 6
Recipes - base case with hierarchy

MEASURE	ρ
P1 - P2	
Leacock and Chodorow	0.680
Wu and Palmer	0.891
Li et al.	0.781

Table 7
Drinks - P1 - P2 and edge-based

MEASURE	ρ
Jaccard	0.876
Sq. Jaccard	0.870
Comm-sq. Jaccard	0.851
Norm. Jaccard	0.899
Norm. comm-sq. Jacc.	0.877
Sigmoid	**0.900**
Sigmoid Jaccard	0.876

Table 8
Drinks - base case

MEASURE	ρ
Jaccard no t. + h.	0.888
Sq. Jaccard no t. + h.	0.882
Comm-sq. Jaccard no t. + h.	0.868
Norm. Jaccard no t. + h.	0.889
Norm. comm-sq. Jacc. no t. + h.	0.898
Sigmoid no t. + h.	**0.900**
Sigmoid Jaccard no t. + h.	0.888

Table 9
Drinks - base case without `rdf:type` with hierarchy

MEASURE	ρ
Jaccard + h.	0.861
Sq. Jaccard + h.	0.861
Comm-sq. Jaccard + h.	0.839
Norm. Jaccard + h.	0.889
Norm. comm-sq. Jacc. + h.	0.877
Sigmoid + h.	**0.900**
Sigmoid Jaccard + h.	0.861

Table 10
Drinks - base case with hierarchy

the second experiment (Drinks) all the feature-based measures in Table 8 perform better than Leacock and Chodorow's measure and Li et al.'s measure and SIM_N and SIM_S perform better even than Wu and Palmer's measure.

This shows that we can obtain better similarity results by considering properties for instances in an ontology, rather than hierarchy underlying the ontology. This confirms our hypothesis H1 that the feature-based similarity shows better performance than hierarchy-based approaches. Often, other proposed methods in the literature do not surpass Li's measure, even when they surpass other hierarchy-based methods. On both of our datasets, the best performing hierarchy-based similarity measure is Wu and Palmer's measure. But all feature-based measures in the first experiment and some of the measures in the second experiment surpass Wu and Palmer's measure and all surpass the other two hierarchy-based measures. This means that properties play more important role than the underlying hierarchy when describing ontological instances and their mutual similarity.

5.6.3. *Substituting `rdf:type` with hierarchical information*

Tables 5 and 9 show the results for the 7 similarity measures in both experiments, where the property `rdf:type` was excluded from the calculations and where we tried to simulate this information with hierarchical information. We tried to incorporate the hierarchy-based similarity by including the normalised distance between concepts (calculated as $\frac{\text{DIST}}{2\,\max}$) in the feature-based formulae as a part of distinctive features.

In the first experiment (Recipes), better results are obtained by the simple base measure, hence by using `rdf:type` instead of hierarchical information (Table 4), whereas in the second experiment (Drinks) results are mostly better with the hierarchical information but without `rdf:type` (Table 9. But even in this case, the Sigmoid measure $\text{SIM}_{S\,nth}$ performs better in base case, showing consistent improvement.

But we cannot confirm our hypothesis H3 that hierarchical information is encoded better with features than with underlying hierarchy, even though in the case of best performing Sigmoid measure $\text{SIM}_{S\,nth}$ it is.

5.6.4. *Including both, `rdf:type` and hierarchical information*

Finally, we wanted to combine hierarchical information with features to see if the similarity values could be improved. Just a linear combination of feature-based similarity and hierarchy-based similarity would not yield better results, since hierarchy-based similarity would just decrease the correlation. Hence, we included the hierarchy information as in the above. The results are reported in Table 6 and Table 10.

In the first experiment (Recipes) we obtained small improvements for some basic measures, whereas in the second experiment (Drinks) only the Sigmoid measure $\text{SIM}_{S\,h}$ brings very small improvement.

Hence, w.r.t. the hypothesis H4 we can conclude that combining the hierarchy and feature-based approach beyond linear combination provides good correlation with human judgement but sometimes better results are obtained without incorporating the underlying hierarchy.

5.6.5. *Concluding considerations*

We can see that in both domains, the consistently best performing measure is the Sigmoid similarity SIM_S. In the domain of Recipes we obtained the improvement of 4.9% for the base case, whereas in the domain of Drinks we obtained the improvement of 2.74% for the base case.

In both cases very small additional improvement is obtained by considering the underlying hierarchy together with `rdf:type`. Hence, there is little benefit in adding this additional information to the basic similarity measures.

Also, in both cases, there is an improvement w.r.t. the best hierarchy-based measure (Wu and Palmer): 13.5% in the case of Recipes and 1.01% in the case of Drinks. This again shows how dependant the hierarchy-based similarity measures are on the design of the underlying conceptual hierarchy.

We can see that even with relatively small number of properties defined for the concepts in the ontology, the feature-based similarity outperforms hierarchy-based approaches. Of course, the number of defined properties plays an important role in semantic similarity calculation.

The dataset used also plays an important role in the similarity calculations. To the best of our knowledge, the datasets used in this work include significantly higher number of participants than many works in the field. Usually, the similarity measures are tested on Wordnet [11], but we are providing the community with yet another rich dataset to experiment with.

We summarise here our main findings as the responses to our hypotheses.

H1: Jaccard's feature-based similarity measure performs better than hierarchy-based approaches;

H2: it is possible to improve Jaccard's feature-based similarity. We propose Sigmoid similarity measure as the improvement of the original Jaccard's measure which brings the improvement of 4.9% in the domain of recipes and of 2.74% in the domain of drinks;

H3: we cannot say if the hierarchical information is encoded better with features than with underlying hierarchy;

H4: combining the hierarchy and feature-based approach beyond linear combination further improves the Jaccard's similarity measure and its variations but to a very small degree, hence it is questionable if these modifications are worth implementing.

5.7. Comparison with the performance on WordNet

Since most of the similarity measures in the literature have been tested on WordNet, we include here the comparison of our results with the corresponding results provided in [29]. We include only the results for Miller and Charles' dataset [20], since the ones for [28] are not always available.

We can see that the hierarchy-based measures (Leacock and Chodorow, Wu and Palmer and Li et al.) perform better on WordNet than on Wikitaaable dataset but the performance is similar to the Drinks dataset. This is due to the hierarchical structure of Wikitaaable and Drinks datasets. The hierarchy in Wikitaaable is rather shallow, hence the information obtained from the underlying conceptual hierarchy is not so rich. On the other hand, the Drinks ontology has a deeper underlying hierarchy and provides more precise information. We can see that Tversky's similarity measure performs better on Wikitaaable and Drinks datasets, since there are more properties defined for the concepts. We include also the results for Tversky + hierarchy, Sigmoid Tversky + hierarchy, Norm. com. sq. Tversky, Norm. com. sq. Tversky + hierarchy (with type) and Norm. com. sq. Tversky + hierarchy (no type) which perform even better than simple Tversky's measure on Wikitaaable and Drinks datasets.

6. Related work

In this section we give a brief summary of other works which deal with feature-based similarity. These approaches calculate feature-based similarity in different ways, starting from Tversky's similarity measure but taking into account different aspects w.r.t. us (antescendant classes, descendant classes etc., whereas we compare property-value pairs). We include these works here to have a more complete picture of feature-based similarity measures. We did not test these measures since the scope of our work was to evaluate the performance of Tversky's similarity measure and its variations calculating them from property-value pairs for compared objects. The same variations of Tversky's measure could be applied to other feature-based similarity measures as well. Also, some of these measures are not applicable in our context. For example, we cannot calculate the number of descendant classes since we deal with instances in the ontology.

An interesting approach to feature-based similarity calculation is given in [22] and [23]. Both works translate the feature-based model into information content (IC) model, with a slightly different formulation of Tversky's formula where Tversky's function describing the saliency of the features is substituted by the information content of the concepts. In [22] *Intrinsic Information Content* iIC, introduced in [30], is used taking into account the number of subconcepts of a concept and a total number of concepts in a domain. In [23] *Extended Information Content* EIC is used instead of iIC where iIC is combined with EIC as a average iIC for all the concepts connected to a certain concept with different relations. Both approaches use

MEASURE	WordNet	Wikitaaable	Drinks
Leacock and Chodorow	0.74	0.583	0.680
Wu and Palmer	0.74	0.585	0.891
Li et al.	0.82	0.576	0.781
Tversky/Jaccard	0.73	0.633	0.875
Tversky + hierarchy	N/A	0.643	0.861
Sigmoid	N/A	0.664	0.900
Sigmoid + hierarchy (no type)	N/A	0.583	0.900
Sigmoid + hierarchy (with type)	N/A	0.664	0,900

Table 11
Comparison with WordNet

the underlying ontology structure directly, where all the defined semantic relations are used, rather than relying on an external corpus. Their new similarity measure called *FaITH* is based on this novel framework. Also, this new IC calculation can be used to rewrite the existing similarity measures in order to calculate relatedness, in addition to similarity.

An important work on feature-based similarity regarding ontological concepts is described by [34]. They start from Tversky's assumption that similarity is determined by common and distinctive features of the compared objects and consider the relations between concepts as their features. They linearly combine two similarities. The first similarity is obtained from direct connections between two objects, as well as common features shared between both concepts (in this case the similarity between relations is calculated using Wu and Palmer's measure [33]). The second similarity is based on distinctive features of each object. Their approach can be used to calculate similarity at the class level, as well as the similarity of instances. Furthermore, it is possible to take into account only specific relations which leads to context-aware similarity. The problem is that the proposed method is assessed only on 4 pairs of concepts.

[29] also introduce a new feature-based approach for calculating ontology-based semantic similarity based on taxonomical features. They evaluate the semantic distance between concepts by considering as features the set of concepts that subsume each of them. Practically, the degree of disjunction between their feature sets (non-common subsumers) model the distance between concepts, whereas the degree of overlap (common subsumers) models their similarity. The problem with this approach is that If the input ontology is not deep enough or built with enough taxonomical details or it does not consider multiple inheritance, the knowledge needed for similarity calculation might be scarse. The authors also provide a detailed survey of most of the ontology-based approaches and compare their performance on WordNet 2.0. From this analysis they draw important conclusions about the advantages and limitations of these approaches and give directions on their possible usage. A slightly different version of this measure was used by [2] on SNOMED CT ontology to evaluate the similarity of medical terms.

[27] and [21] propose similar measures for calculating semantic similarity based on matching of their synonym sets, semantic neighbourhoods (semantic relations among classes) and features which are classified into parts, functions and attributes. This enables separate treatment of these particular class descriptions and introduction of specific weights which would reflect their importance in different contexts. In [27] these similarities are calculated using Tversky's formula, where parameters in the denominator are calculated taking into account the depth of the hierarchies of different ontologies. Synonym sets and semantic neighbourhoods are useful when detecting equivalent or most similar classes across ontologies. Features are useful when detecting similar but not equivalent classes. [21] eliminate the need for the parameters in the denominator in Tversky's formula, hence they do not rely on the depth of the corresponding ontologies. This leads to matching based only on common words for synset similarity calculation. Also, set similarities are calculated per relationship type. Finaly, their similarity does not have weights for different similarity components. The novelty of their work is the application of this and various other similarity measures to MeSH ontology (Medical Subject Headings). Both methods can be used for cross-ontology similarity calculation.

A semantic similarity measure for OWL objects introduced by [13] is based on Lin's information theoretic similarity measure [19]. They compare semantic

description of two services and define their semantic similarity as a ratio between the shared and total information of the two objects. The semantic description of an object is defined using its description sets which contain all the statements (triples) describing the given object and their information content is based on their "inferencibility", i.e. the number of new RDF statements that can be generated by applying a certain set of inference rules to the predicate. They use their measure to determine the similarity of semantic services annotated with OWL ontologies.

Similarity can find many applications is Recommender Systems. [9] developed a content-based movie recommender system based on Linked Open Data (LOD) in which they adapt a vector space model (VSM) approach to compute similarities between RDF resources. Their assumption is that two movies are similar if they have features in common. The whole RDF graph is represented as a 3-dimensional matrix where each slice refers to one property in the ontology and for each property the similarity between two movies is calculated using their cosine similarity.

Similarity among concepts can be also automatically learned. *Similarity learning* is an area of supervised machine learning, closely related to regression and classification, where the goal is to learn from examples a similarity function that measures how similar or related two objects are. Similarity learning is used in information retrieval to rank items, in face identification and in recommender systems. Moreover, many machine learning approaches rely on some similarity metric. This includes unsupervised learning such as clustering, which groups together close or similar objects, or supervised approaches like K-nearest neighbour algorithm. Metric learning has been proposed as a preprocessing step for many of these approaches.

Automatic learning of similarity among concepts in an ontology is used especially for ontology mapping (also known as ontology alignment, or ontology matching) [10], the process of determining correspondence between ontology concepts. This is necessary for integrating heterogeneous databases, developed independently with their own data vocabulary or different domain ontologies.

There are several works which have exploited ML techniques towards ontology alignment. [14] organised the ontology mapping problem into a standard machine learning framework, exploiting multiple concept similarity measures (i.e, synset-based, Wu and Palmer, description-based, Lin). In [8] a multi-strategy learning was used to obtain similar instances of hierarchies to extract similar concepts using Naïve Bayes (NB) technique. In [1], following a parameter optimisation process on SVM, DT and neural networks (NN) classifiers, an initial alignment was carried out. Then the user's feedback was used to improve the overall performance. All these works considered concepts belonging to different ontologies while we concentrated on concepts in a same ontology.

7. Conclusions and future work

In this work we present some modifications of Jaccard's feature-based similarity measure based on properties defined in an ontology and their comparison with hierarchy-based approaches. We also propose an improvement of Jaccard's similarity measure, namely Sigmoid similarity measure SIM_S.

We further proposed two variations of all the feature-based measures, to see how much the underlying hierarchical information contributes to accurate similarity measurement. We came to the conclusion that the underlying hierarchical information does provide some additional information when calculating similarity. However, the improved performance is very small, so it might not be worth adding the complexity to the similarity calculation.

In addition to Jaccard's similarity measure, we tested 6 modifications of this measure and for each of these measures additional 2 variations on slightly modified Wikitaaable dataset in the domain of recipes and on Drinks ontology designed by our researchers previously. Our first evaluation included 137 subjects and 25 pairs of concepts and our second evaluation included 147 subjects and 30 pairs of concepts. This is a significant number of real evaluators compared with other evaluations in the literature.

Our main finding is that the original Jaccard's similarity measure could be improved by using Sigmoid similarity.

As a future work, it would be interesting to see how the similarity measures would perform in the presence of more properties or on a different dataset. MeSH and SNOWMED are some possible candidate datasets, although in these cases expert opinion would be needed. Also, it would be interesting to add hierarchical structure among property values. For example, in the present approach we consider `Fusilli` and `Spaghetti` two different domain items (hence, two different property values). But `Fusilli` and `Spaghetti` are both descendants of `Pasta` so they

could be considered as "almost" the same values for properties.

Moreover, in this work we only consider object type properties. Taking data type properties, such as literals, into account might be an interesting area for future investigation. In this case, it would be necessary to determine when two literal values can be considered equal.

In addition, similar experiments can be applied to linked open data [4] or any other data structure where the objects are described by means of their properties.

References

[1] B. Bagheri Hariri, H. Abolhassani, and H. Sayyadi. A neural-networks-based approach for ontology alignment. In *SCIS & ISIS*, volume 2006, pages 1248–1252, 2006.

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

[3] G. Beydoun, A. A. Lopez-Lorca, F. García-Sánchez, and R. Martínez-Béjar. How do we measure and improve the quality of a hierarchical ontology? *Journal of System Software*, 84(12):2363–2373, 2011.

[4] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[5] A. Budanitsky and G. Hirst. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[6] F. Cena, S. Likavec, and F. Osborne. Property-based interest propagation in ontology-based user model. In *20th Conference on User Modeling, Adaptation, and Personalization, UMAP 2012*, volume 7379 of *LNCS*, pages 38–50. Springer, 2012.

[7] A. Cordier, V. Dufour-Lussier, J. Lieber, E. Nauer, F. Badra, J. Cojan, E. Gaillard, L. Infante-Blanco, P. Molli, A. Napoli, and H. Skaf-Molli. Taaable: A case-based system for personalized cooking. In S. Montani and L. C. Jain, editors, *Successful Case-based Reasoning Applications-2*, volume 494 of *Studies in Computational Intelligence*, pages 121–162. Springer Berlin Heidelberg, 2014.

[8] J. David. Association rule ontology matching approach. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 3(2):27–49, 2007.

[9] T. Di Noia, R. Mirizzi, V. C. Ostuni, D. Romito, and M. Zanker. Linked open data to support content-based recommender systems. In *8th International Conference on Semantic Systems, I-SEMANTICS '12*, pages 1–8. ACM, 2012.

[10] J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[11] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[12] T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition Journal*, 5(2):199–220, June 1993.

[13] J. Hau, W. Lee, and J. Darlington. A semantic similarity measure for semantic web services. In *Web Service Semantics Workshop at WWW (2005*, 2005.

[14] R. Ichise. Machine learning approach for ontology mapping using multiple concept similarity measures. In *Computer and Information Science, 2008. ICIS 08. Seventh IEEE/ACIS International Conference on*, pages 340–346. IEEE, 2008.

[15] C. Leacock and M. Chodorow. *Combining local context and WordNet similarity for word sense identification*, pages 305–332. In C. Fellbaum (Ed.), MIT Press, 1998.

[16] Y. Li, D. McLean, Z. Bandar, J. O'Shea, and K. A. Crockett. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150, 2006.

[17] S. Likavec. Shapes as property restrictions and property-based similarity. In O. Kutz, M. Bhatt, S. Borgo, and P. Santos, editors, *2nd Interdisciplinary Workshop The Shape of Things*, volume 1007 of *CEUR Workshop Proceedings*, pages 95–105. CEUR-WS.org, 2013.

[18] S. Likavec, F. Osborne, and F. Cena. Property-based semantic similarity and relatedness for improving recommendation accuracy and diversity. *International Journal on Semantic Web and Information Systems*, 11(4):1–40, 2015.

[19] D. Lin. An information-theoretic definition of similarity. In *15th International Conference on Machine Learning ICML '98*, pages 296–304. Morgan Kaufmann Publishers Inc., 1998.

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

[21] E. G. M. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou. X-similarity: Computing semantic similarity between concepts from different ontologies. *Journal of Digital Information Management*, 4(4):233–237, 2006.

[22] G. Pirró. A semantic similarity metric combining features and intrinsic information content. *Data and Knowledge Engineering*, 68:1289–1308, 2009.

[23] G. Pirrò and J. Euzenat. A feature and information theoretic framework for semantic similarity and relatedness. In *9th International Semantic Web Conference, ISWC '10*, volume 6496 of *LNCS*, pages 615–630. Springer, 2010.

[24] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Trans. on Systems Management and Cybernetics*, 19(1):17–30, 1989.

[25] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

[26] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[27] M. A. Rodriguez and M. J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 2000.

[28] H. Rubenstein and J. B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

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

[30] N. Seco, T. Veale, and J. Hayes. An intrinsic information content metric for semantic similarity in wordnet. In R. L. de Mántaras and L. Saitta, editors, *16th Eureopean Conference on Artificial Intelligence, ECAI '04, including Prestigious Applicants of Intelligent Systems, PAIS '04*, pages 1089–1090. IOS

Press, 2004.

[31] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

[32] C. A. Welty and N. Guarino. Supporting ontological analysis of taxonomic relationships. *Data Knowledge Engeneering*, 39(1):51–74, 2001.

[33] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *32nd Annual Meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[34] P. D. H. Zadeh and M. Reformat. Assessment of semantic similarity of concepts defined in ontology. *Information Sciences*, 250:21–39, 2013.