Weighted Ontology for Semantic Search

Anna Formica, Michele Missikoff, Elaheh Pourabbas, and Francesco Taglino

Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti"
Consiglio Nazionale delle Ricerche, Viale Manzoni 30
00185 Rome, Italy
{anna.formica,michele.missikoff,elaheh.pourabbas,
francesco.taglino}@iasi.cnr.it

Abstract. This paper presents a method, *SemSim*, for the semantic search and retrieval of digital resources (DRs) that have been previously annotated. The annotation is performed by using a set of characterizing concepts, referred to as *features*, selected from a reference ontology. The proposed semantic search method requires that the features in the ontology are weighted. The weight represents the probability that a resource is annotated with the associated feature. The SemSim method operates in three stages. In the first stage, the similarity between concepts (*consim*) is computed by using their weights. In the second stage, the concept weights are used to derive the semantic similarity (*semsim*) between a user request and the DRs. In the last stage, the answer is returned in the form of a ranked list. An experiment aimed at assessing the proposed method and a comparison against a few among the most popular competing solutions is given.

Keywords: Similarity Reasoning, Reference Ontology, Information content, Digital Resources.

1 Introduction

Similarity reasoning is a very challenging research area. After some decades of research, there is an enormous corpus of scientific results available, but still there is not a single solution that is clearly emerging, capable of outperforming all the others. Probably, it will never be the case, due to the great variety of problems requiring similarity reasoning, and situations in which such problems arise. There is a nice example, reported in [12], concerning the perceived similarity between the elements of the triple: *(pig, pick-up, donkey).* At first, one would assert a higher similarity between the two terms that represent living entities: *pig* and *donkey*, since the term *pick-up* denotes a mechanical artefact. However, if we change perspective, considering a situation where we need to transport, say, potatoes, then *pick-up* and *donkey* are the most alike. Such divergent outcomes derive from a shift of perspective, and therefore a change in what are the relevant characteristics taken into consideration to determine the similarity. An effective similarity reasoner will be endowed with multiple methods and strategies, and the capacity of analyzing a situation to determine the most promising method.

R. Meersman and Z. Tari (Eds.): OTM 2008, Part II, LNCS 5332, pp. 1289–1303, 2008.
© Springer-Verlag Berlin Heidelberg 2008

1.1 Semantic Searching

One of the primary uses of similarity reasoning is in method processing a user request to retrieve a set of (digital) resources from a given repository (or, more in general, from the Web). Such resources can be the actual target of the retrieval process (textual or multimedia documents, images, Web services, business process diagrams, etc.) or digital surrogates of non-digital objects (people, organizations, cars, furniture, hotels, etc.). In our work, we are only dealing with digital surrogates, i.e., semantic annotations[1] that we will refer to as *feature vectors*[2]. Therefore, even if we intend to search objects of the first category, for instance digitalized documents, we are not going to consider them directly, in the search phase, e.g., by applying Natural Language or Information Retrieval techniques. Our search method is based on feature vectors, that we assume have been preliminary built (with some *feature extraction* techniques[3]) and made accessible.

Searching over features vectors instead of target resources has several advantages:

- feature vectors are homogeneous structures, independent of the nature of the resources they are associated with; therefore,
- the semantic search method can be unified for different kinds of resources (e.g., text, photo, video, etc.), once they have been properly annotated; therefore it is possible to search different repositories of a different nature by using the same method; then,
- the retrieved results can be reported in an homogeneous form, and ranked in a unique list, even if the concrete forms of the retrieved resources are very different.

Beside feature vectors, a second characterization of the proposed search method is represented by the use of a *weighted reference ontology (WRO)*, containing the relevant concepts in the given domain.

Here, we wish to recall the definition of an ontology, taken from the OMG Ontology Definition Metamodel [4]:

"An *ontology* defines the common *terms* and *concepts* (meaning) used to describe and represent an area of knowledge. An ontology can range in expressivity from a *Taxonomy* (knowledge with minimal hierarchy or a parent/child structure), to a *Thesaurus* (words and synonyms), to a *Conceptual Model* (with more complex knowledge), to a *Logical Theory* (with very rich, complex, consistent and meaningful knowledge)."

In our case, we restrict our view of the ontology to a taxonomy of concepts. This simplified view, as anticipated, is then enriched with a weight associated with each concept. Intuitively, the weight of a concept represents its *featuring power*, i.e., how

[1] Semantic annotation is a very active research area [25] whose description goes beyond the scope of the paper.

[2] A *feature vector* is an n-dimensional vector of (numerical) features that represent some object... [it is obtained as a] reduced representation of the key characteristics of the object (see http://en.wikipedia.org/wiki/Feature_vector).

[3] Feature extraction is an important topic that will not be addressed in this paper for lack of space [24].

selective is such a concept in characterizing the resources of our universe. A high weight corresponds to a low selectivity level, i.e., many resources are characterized by the concept. Conversely, a low weight corresponds to a high selectivity, and therefore its use in a request will return less instances. In accordance with the Information Theory [23], a concept weight will be used to determine its (relative) information content.

The WRO is an important element of our proposal, since we restrict the elements of a feature vector to the terms that represent concepts in the ontology. For this reason, we will refer to the former as: *Ontology-based Feature Vector (OFV)*. The same is true for a user request that takes the form of a *Request Vector (RV)*.

The *SemSim* method proposed in this paper supports a user whishing to retrieve digital resources from a given UDR (Universe of Digital Resources). An UDR can be a document repository maintained by one enterprise, or can be a shared, distributed content repository hosted in different organizations belonging to a Virtual Enterprise, or even it can be the Web as a whole. When searching, the user indicates a set of desired features (in the form of a request vector *rv*) expecting to have in return a set of resources (partially) satisfying such features. Similarly to Google search, the output of SemSim is a ranked list of resources, sorted according to their similarity to the *rv*. The semantic search method is mainly based on the notion of similarity of ontology-based feature vectors, and therefore on the similarity of the concepts that compose the two structures. Similarity reasoning is a challenging job.

The following list reports the primary dimensions that can be considered in performing similarity reasoning:

- *Terminological*, if the concepts are characterized by a set of terms (e.g., WordNet synset, user generated keywords, etc.);
- *Linguistic*, determined contrasting the textual descriptions (if available) of the concepts;
- *Structural*, when we consider the information structures (e.g., attributes and associations) of the contrasted concepts;
- *Taxonomic*, when the similarity is determined by the hierarchal organization of concepts in the ontology;
- *Extensional*, when the similarity is derived considering the instances of the contrasted concepts;
- *Operational*, based on operations associated to the contrasted concepts.

An effective similarity reasoning system should be able to take into consideration more than one dimension from the above list. In this paper, we will present a similarity reasoning method, SemSim, that operates along the taxonomic and extensional dimensions. In fact, a central issue is the weighting of concepts in the ontology that is based on both the position in taxonomy and extension of each concept, seen as the set of annotated DRs. To maintain a precise count of the annotated resources is a difficult job, especially in a dynamic domain. Therefore, we propose a probabilistic approach, where the weight of a concept represents the probability that a resource in the domain is characterized by that concept.

1.2 Promising Application Domains for Semantic Search

This work has started in the context of a large industrial conglomerate (Finmeccanica) that develops large engineering systems: from air traffic control (ATC) to integrated

civil protection networks. Each project consists of thousands of parts, devices, apparatuses, subsystems, and interconnected systems. In previously accomplished projects, an incredible wealth of knowledge has been accumulated in the form of blueprints, design drafts, data sheets, CAD/CAM files, test cases and measures, technical notes, installation manuals, troubleshooting plans, etc. All these documents are stored in digital forms, but are placed in different sites, different formats, created with different software tools; therefore they are not easy, for an interested user, to be identified and retrieved. When a new project starts, it is extremely useful (and cost saving) to have the possibility to effectively access the wealth of knowledge produced by previous projects. To this end, the availability of a global search engine, based on semantic technologies, is particularly promising. This is the application context in which the SemSim method has been initially developed.

Another application domain is represented by the tourism industry. Here instead of having a single large industrial conglomerate (with a closed UDR), we have a network of SMEs (e.g., providing transportation, accommodation, food, cultural services, natural parks access) able to provide an integrated offer for a tourist[4]. In the tourism domain we have again a large variety of digital resources, available on different web sites, in different locations. Here the variety of formats is less important (essentially we have web documents), but the fragmentation and the possibility of retrieving documents on the basis of their semantic content is equally important. Moreover, this domain is more open, dynamic, and less regulated than the previous one. Since tourism is more intuitive, we drew from it the running example used in this paper.

The rest of the paper is organized as follows. Section 2 is dedicated to the related work, while Section 3 presents the basic notions and the structures used in SemSim. In Section 4, the definition of a Weighted Reference Ontology is given and a running example is introduced. In Section 5, the proposed SemSim method for evaluating the semantic similarity of Ontology-based Feature Vectors is presented. In Section 6, we present an assessment of the SemSim method, contrasting it against a few other methods. Finally, Section 7 concludes the paper with indications of future work.

2 Related Work

In the vast literature available (see for instance [1,5,17,18]), we will focus on the proposals tightly related to our approach. We wish to recap that the focus of our work is on the method to compute the similarity between concept vectors. To this end, we need to build a two stages method: firstly computing the pair-wise concept similarity, and then deriving the similarity of two vectors of concepts. Thus, we adopted a technique based on the information content [22], which has been successively refined by Lin in [15]. The Lin's approach shows a higher correlation with human judgement than other methods, such as the *edge-counting* approach [21] and Wu-Palmer [26].

We need to emphasise that we deal with specialised domains (e.g., systems engineering, tourism, etc.), requiring specialised domain ontologies. The large majority of existing proposals make use of WordNet. This is a lexical ontology that is generic

[4] We exclude the big tour operators and hotel chains, to address the constellation of SMEs and small to micro tourism services providers.

(i.e., not focused on a specific domain) and, furthermore, contains only simple terms, no multi-word terms are reported (e.g., terms such as "power supply" or "farm house" are not available in WordNet). Therefore, our approach is different from all other proposals that use any generic ontology.

SemSim is based on an ontology with weighted concepts. In [6] there is an interesting proposal that makes use of an ontology enriched with a typical Natural Language Processing method, based on *term frequency* and *inverse document frequency* (*tf-idf*). With respect to this work, our proposal abstracts from the linguistic domain during the annotation phase, allowing therefore for a pure semantic approach. Furthermore, in weighting the terms connected to the elements of the ontology, [6] relies on a rigid approach, i.e., it proposes 5 relevance levels that correspond to 5 constant values: *direct(1.0), strong(0.7), normal(0.4), weak(0.2), irrelevant(0.0)*. In our method, the weights, and the relationships among concepts, are not discrete and take any value between 0 and 1.

The work presented in [13] shares some similarity with our approach. It proposes a bottom up method that, starting from the weight associated with concept nodes, determines the concept similarity by building vectors of weights. Therefore, the objective is the similarity of concepts that depends on the topology of the ontology and the position of concepts therein. However, our scope is wider: similarity of concepts (*consim*) is just a step of a more comprehensive method aimed at determining the similarity of two concept vectors (*semsim*). We could have selected the method proposed in [13] for the first phase of our work (concept similarity), but it was not completely convincing, since its assessment is based on the well known Miller and Charles experiment [19] that, being based on WordNet, is not conceived for specialized domains.

In [14], a similarity measure between words is defined, where each word is associated with a concept in a given ISA hierarchy. The proposed measure essentially combines path length between words, depth of word subsumer in the hierarchy, and local semantic density of the words. However, similar to [13], the authors evaluate their method using Miller and Charles experiment that was conceived for general domains and is not appropriate for specialized applications.

Other research results concern the similarity between two sets (or vectors) of concepts. In the literature the *Dice* [9,16] and *Jaccard* [11] methods are often adopted in order to compare vectors of concepts. However, in both Dice and Jaccard concept similarity is computed by using exact match, with 0 or 1 response. Therefore, the matchmaking of two concept vectors is based on their intersection, without considering the positioning of the concepts in the ontology. More recent works (see [2]) introduce the ontology, hence proposing a more elaborated concept matching. Our proposal is based on a more refined semantic matchmaking, since the match of two concepts is based on their shared information content, and the vectors similarity is based on the optimal concepts coupling.

[3] introduces two new algorithms for computing the semantic distance/similarity between sets of concepts belonging to the same ontology. They are based on an extension of the Dijkstra algorithm[5] to search for the shortest path in a graph. With respect to our approach, here the similarity is based on the distance between concepts rather than the

[5] http://en.wikipedia.org/wiki/Dijkstra's_algorithm

information content carried by each concept. Furthermore, the similarity between two sets of concepts is computed by considering the similarity between each concept from a set and all the concepts from the other. Finally, the similarity between adjacent concepts is supposed to be decided at design-time by the ontology developer and consequently introduces a certain degree of rigidity and bias on the results.

This brief overview has mainly the goal of positioning our work with respect to the most relevant results in the literature. As shown, in the various cases, our approach is either more focused (i.e., for specialised domains) or more elaborated (e.g., considering the information content of concepts with respect to the UDR and their positioning in the ontology). But we know that more elaborated solutions may not perform better than simpler ones. For this reason, we decided to conduct an experiment to assess the results of the SemSim method against a few among the most promising competitors. These results are reported in Section 6.

3 Basic Definitions

In this section we introduce the basic notions and the structures used in the SemSim method. Summarising, SemSim is based on the following structures:

- a Universe of Digital Resources (UDR) over which the search is performed;
- a Weighted Reference Ontology (WRO);
- a Semantic Annotation Repository (SAR) containing the ontology-based feature vectors (OFVs), one for each digital resource in UDR;
- a Request Vector (RV);
- a Ranked Solution Vector (RSV), subset of the UDR resources, whose OFVs are similar to the RV, filtered by a given threshold.

In this section, we provide a formal account of the structures that are used in the SemSim method.

Definition 1. *Universe of Digital Resources (UDR). The UDR* is the totality of the digital resources that are semantically annotated.

Definition 2. *Ontology.* An *Ontology* is a formal, explicit specification of a shared conceptualization [10]. In our work we address a simplified notion of ontology, *Ont*, that focuses on a set of concepts organized according to a specialization hierarchy. In particular, *Ont* is a *tree* structure defined by the pair:

$$Ont = \langle C, H \rangle$$

where C is a set of concepts and H is the set of pairs of concepts of C that are in subsumption (*subs*) relation:

$$H = \{(c_i, c_j) \in C \times C \mid subs(c_i, c_j)\}$$

Since we assume that *Ont* is a tree, given two concepts c_i, $c_j \in C$, the *least upper bound* of c_i, c_j, $lub(c_i,c_j)$, is always defined in C. It represents the less abstract concept of the ontology that subsumes both c_i and c_j.

Definition 3. *Weighted Reference Ontology (WRO).* Given an ontology $Ont = \langle C, H \rangle$, a *WRO* is a pair:

$$WRO = <Ont, w>$$

where w is a function defined over C, such that given a concept $c \in$ C, $w(c)$ is a rational number in the interval [0..1] standing for a weight associated with the concept c in the ontology *Ont*.

Definition 4. *Ontology Feature Vector (OFV)*. Given an ontology *Ont* = <C, H> and a digital resource $dr_i \in$ UDR, an OFV associated with *dr*, ofv_i, is a set of ontology concepts defined as follows:

$$ofv_i = (c_{i,1},...,c_{i,m}) \text{ where } c_{i,j} \in \text{C}, j = 1...m$$

To actually link a concept to the resources that it characterises, it is necessary to introduce the notion of a *featured extension*.

Definition 5. *Featured Extension*. Given an ontology *Ont* = <C, H>, and a concept (feature) $c \in$ C, the *featured extension* of c is defined according to the extension function F_{Ont} as follows:

$$F_{Ont}(c) = \{dr \in \text{UDR} \mid c \in ofv_{dr}\}$$

Therefore, given a feature c, $F_{Ont}(c)$ provides all the digital resources in UDR whose OFVs contain c, i.e., all the digital resources that are annotated by the feature c.

Definition 6. *Similarly Featured Extension.* Given an ontology *Ont* = <C, H>, and a concept (feature) $c \in$ C, the *similarly featured extension* of c is defined according to the extension function S_{ONT} as follows:

$$S_{Ont}(c) = \{dr \in \text{UDR} \mid \exists\, c' \in ofv_{dr}\ consim(c,c') > k\ \}$$

where *consim* is the concept similarity that will be formally introduced in Section 5, and k is a threshold suitably defined according to the cases. In this paper we assumed k=0.3. Therefore, given a feature c, $S_{Ont}(c)$ provides all the digital resources in UDR whose OFVs contain a feature c' whose similarity with c is higher than a fixed threshold. In other words, it provides all the digital resources that are annotated by a feature similar to that required, up to a threshold.

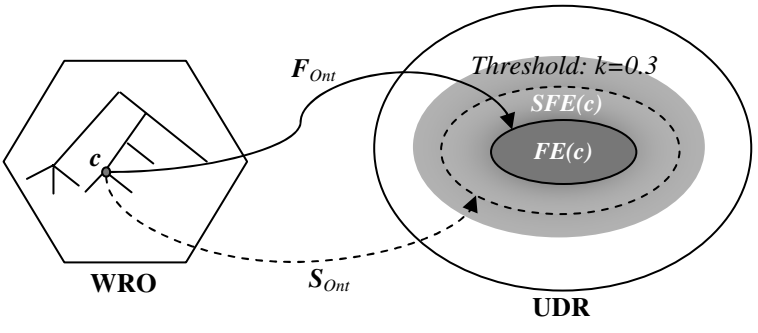

Fig. 1. Relationship between a concept c and its extensions

Figure 1 visually depicts the relationship among a concept c and its extensions: Featured Extension (*FE(c)* represented by the inner set) and Similarly Featured Extension (*SFE(c)* represented by the dash-bordered set).

Definition 7. *Semantics of an OFV.* Given a repository UDR annotated with OFVs, the semantics of an OFV, *ofv*, is defined according to the extension function E_{ofv} as follows:

$$E_{ofv}(ofv) = \cap_{i=1, \ldots, m} F_{Ont}(c_i)$$

where $F_{Ont}(c_i)$ is the *featured extension* of the concept c_i, for i = 1...m. Therefore, $E_{ofv}(ofv)$ provides all the digital resources in UDR characterized by the features in the OFV.

Definition 8. *Semantics of a Request Vector.* Given an ontology *Ont* = <C, H> and a Request Vector *rv*:

$$rv = (c_1,\ldots,c_n)$$

where $c_i \in$ C for i = 1...n, the semantics of *rv* is defined according to the extension function E_{RV} as follows:

$$E_{RV}(rv) = \cup_{i=1, \ldots, n} (F_{Ont}(c_i) \cup S_{Ont}(c_i))$$

where $F_{Ont}(c_i)$ and $S_{Ont}(c_i)$ are respectively the *featured extension* and the *similarly featured extension* of the concept c_i for i = 1...n. Therefore, $E_{RV}(rv)$ provides all the digital resources in UDR whose OFVs contain at least one feature of *rv*, or one feature that is similar to at least one feature of *rv*.

Definition 9. *Ranked Solution Vector.* Given an ontology *Ont* = <C, H> and a Request Vector *rv*, the Ranked Solution Vector associated with *rv*, RSV(*rv*), is defined as follows:

$$RSV(rv) = \{(dr, semsim) \mid dr \in E_{RV}(rv) \text{ AND } semsim(dr,rv) > h\}$$

where *semsim(dr,rv)* is the semantic similarity between *dr* and *rv* that will be introduced in Section 5, and *h* is a threshold suitably defined according to the cases. Therefore, the Ranked Solution Vector of a Request Vector provides all the digital resources of UDR whose similarity with the Request Vector is higher than the given threshold.

4 Weight Assignment in the WRO

Prior to addressing the method to associate weights with the concepts in the ontology, we introduce our example drawn from the tourism domain. In the example we assume to have a dozen of hotels that accepted to annotate their leaflets by using a common reference ontology. Each annotation is therefore an OFV, as reported in Table 1.

Table 1. OFV-based annotation of Digital Resources

ofv_1 = (InternationalHotel, Golf, InternationalMeal, Theatre)
ofv_2 = (Pension, FrenchMeal, Biking, Reading)
ofv_3 = (CountryResort, MediterraneanMeal, Tennis)
ofv_4 = (CozyAccommodation, ClassicalMusic, InternationalMeal)
ofv_5 = (InternationalHotel, ThaiMeal, IndianMeal, ClassicalMusic)
ofv_6 = (CountryResort, LightMeal, ClassicalMusic)
ofv_7 = (SeasideCottage, EthnicMeal, CulturalActivity)
ofv_8 = (CountryResort, VegetarianMeal, CulturalActivity)
ofv_9 = (SeasideCottage, MediterraneanMeal, Golf, Biking)
ofv_{10} = (RegularAccommodation, RegularMeal, Biking)
ofv_{11} = (SeasideCottage, VegetarianMeal, Tennis)
ofv_{12} = (SeasideCottage, VegetarianMeal)

The above Semantic Annotation Repository (SAR) has been built starting with the WRO reported in Figure 2.

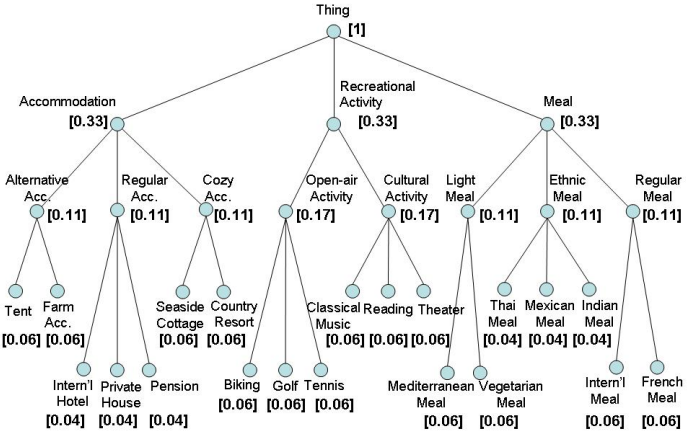

Fig. 2. Concept weights as uniform probabilistic distribution

Our approach in weighting is based on the probability distribution along the hierarchy of concepts starting from the root, namely *Thing*, that stands for the most abstract concept, whose weight $w_p(Thing)$ is equal to *1*. Here we adopt a uniform probabilistic distribution, therefore, given a number n of children (c_i, $i=1…n$) of this top concept, the probability of each child is $w_p(c_i)=1/n$. Accordingly, for any other concept c, $w_p(c)$ is equal to the probability of the father of c, divided by the number of the children of the father of c (i.e., the fan-out). In Figure 2, an example of the probabilistic distribution over concepts of our ISA hierarchy is illustrated. For instance, let us consider the concept *LightMeal*, where $w_p(LightMeal) = 1/9$, since $w_p(Meal)=1/3$ and *Meal* has *3* sub-concepts.

In the next section, we will show how the set of OFVs will be used to perform a semantic search of the hotels, starting from a vector of desired features indicated by the user.

5 The SemSim Method for Semantic Search and Retrieval

Consider the following user request:

"I would like to stay in a seaside hotel, where I can have vegetarian food, play tennis, and attend sessions of classical music in the evening".

It can be formulated according to the request feature vector notation as follows:

rv = (SeasideCottage, VegetarianMeal, Tennis, ClassicalMusic)

Once the *rv* has been specified, the SemSim method is able to evaluate the semantic similarity (*semsim*) among the *rv* and each available OFV. As already mentioned, in order to compute the *semsim* between two feature vectors, it is necessary first to compute the similarity (*consim*) between pairs of concepts.

5.1 Computing Concept Similarity: *consim*

Given a WRO, the notion of *consim* relies on the probabilistic approach defined by Lin [15], which is based on the notion of information content. According to the standard argumentation of information theory, the *information content* of a concept *c* is defined as $-log\ w_p(c)$, therefore, as the weight of a concept increases the informativeness decreases, hence, the more abstract a concept the lower its information content.

Given two concepts c_i and c_j, their similarity, $consim(c_i,c_j)$, is defined as the maximum information content shared by the concepts divided by the information contents of the two concepts [11]. Note that, since we assumed that the ontology is a tree, the least upper bound of c_i and c_j, $lub(c_i,c_j)$, is always defined and provides the maximum information content shared by the concepts in the taxonomy. Formally, we have:

$$consim(c_i,c_j)=2\ log\ w_p(lub(c_i,c_j)) / (log\ w_p(c_i)+log\ w_p(c_j))$$

For instance, considering the pair of concepts *Biking* and *Tennis* of the WRO shown in Figure 2, the *consim* is defined as follows:

$$consim(Biking, Tennis) = 2\ log\ w_p\ (Open\text{-}airActivity) / (log\ w_p\ (Biking)+ log\ w_p\ (Tennis))=0.63$$

since, according to Figure 2, *Open-airActivity* is the *lub* of *Biking* and *Tennis* and therefore provides the maximum information content shared by the comparing concepts.

5.2 Computing Semantic Similarity Degree: *semsim*

In this section we show how we derive the semantic similarity of two vectors, *rv* and *ofv*, by using the *consim* function. In principle, we need to start from the cartesian product of the vectors:

$$rv \otimes ofv = \{ (c_i,c_j) \}$$

where: $i = 1..n$, $j = 1..m$, $n = |rv|$, $m = |ofv|$, $c_i \in rv$, and $c_j \in ofv$.

For each pair we can derive the concept similarity *consim*, as seen in the previous section. However, we do not need to consider all possible pairs, since in many cases the check is meaningless (e.g., contrasting a vegetarian meal with a classical music concert). Hence, we aim at restricting our analysis considering only the pairs that exhibit a higher affinity. Furthermore, we adopted the exclusive match philosophy (sometimes named *wedding* approach) where once a pair of concepts has been successfully matched, concepts do not participate in any other pair. In other words, assuming *rv* and *ofv* represent a set of boys and a set of girls respectively, we analyze all possible sets of marriages, when polygamy is not allowed. Our solution, for the computation of the semantic similarity, *semsim*(*rv*,*ofv*), makes use of the method based on the *maximum weighted bipartite matching* problem in bipartite graphs [7,8].

Essentially, the method aims at the identification of the sets of pairs of concepts of the two vectors that maximizes the sum of *consim*.

$$semsim(rv,ofv) = max(\Sigma\, consim(c_i,c_j)) / min(n,m)$$

In particular, the method that we adopted to solve this problem is based on the well-known Hungarian Algorithm [20].

For instance, in the case of *rv* and ofv_1 of our running example, the following set of pairs of concepts has the maximum *consim* sum:

{(SeasideCottage, InternationalHotel),
(VegetarianMeal, InternationalMeal)
(ClassicalMusic, Theater),
(Tennis, Golf)}

since:

consim(SeasideCottage, InternationalHotel)= 0.36
consim(VegetarianMeal, InternationalMeal)= 0.38
consim(ClassicalMusic, Theater)=0.62
consim(Tennis, Golf)=0.62

and any other pairing will lead to a smaller sum. Therefore:

$$semsim(rv, ofv_1) = (0.36 + 0.38 + 0.62 + 0.62) / 4 = 0.49$$

where the sum of *consim* has been normalized according to the minimal cardinality of the contrasted vectors (in this case 4 for both).

6 SemSim Assessment

In this section we present some preliminary results on the assessment of the proposed SemSim method. The assessment is based on the correlation of se*msim* with human judgment (HJ). Essentially, we contrasted the results of our method with those obtained by a selected group of 20 people. We asked them to express their judgement (on a scale of 0 to 3) on the similarity among the *rv*, and the set of resources at hand, i.e., the hotels H_i, i = 1...12, annotated with the ofv_i shown in Table 1. In Table 2, the human judgment (whose values have been normalized) and *semsim* scores are illustrated (see second and third column).

Table 2. Results of the comparison among human judgment, SemSim and some representative similarity methods

Feature Vectors	*HJ*	*SemSim*	*Dice*	*Jaccard*	*Salton's Cosine*	*Weighted Sum*
ofv_1	0.60	0.49	0.00	0.00	0.00	0.00
ofv_2	0.60	0.49	0.00	0.00	0.00	0.00
ofv_3	0.67	0.63	0.29	0.17	0.08	0.29
ofv_4	0.60	0.56	0.29	0.17	0.08	0.43
ofv_5	0.59	0.43	0.25	0.14	0.06	0.25
ofv_6	0.80	0.66	0.29	0.17	0.08	0.43
ofv_7	0.60	0.55	0.29	0.17	0.08	0.43
ofv_8	0.67	0.63	0.29	0.17	0.08	0.43
ofv_9	0.67	0.69	0.25	0.14	0.06	0.25
ofv_{10}	0.36	0.37	0.00	0.00	0.00	0.00
ofv_{11}	0.82	0.75	0.86	0.75	0.25	0.86
ofv_{12}	0.71	0.50	0.67	0.50	0.25	0.67
Correlation with HJ	**1.00**	**0.82**	**0.70**	**0.67**	**0.66**	**0.72**

Furthermore, we compared SemSim with some representative similarity methods proposed in the literature: Dice, Jaccard, Salton's Cosine[16] and the Weighted Sum defined in [2]. For the sake of simplicity, we recall their formulas below, where *X* and *Y* represent the *rv* and an *ofv*, respectively.

$$2\frac{|X \cap Y|}{|X|+|Y|}$$ Dice'coefficient

$$\frac{|X \cap Y|}{|X \cup Y|}$$ Jaccard's coefficient

$$\frac{|X \cap Y|}{|X| \times |Y|}$$ Salton's Cosine coefficient

$$2\frac{\sum Aff(X_i, Y_j)}{|X|+|Y|}$$ Weighted Sum function,
where $Aff(X_i, Y_j)$, the affinity b/w X_i and Y_j, is
1 if $X_i = Y_j$
0.5 if X_i is a broader or narrower concept of Y_j
0 otherwise

The experiment has shown that SemSim yields a higher correlation with human judgement (0.82) with respect to other representative methods.

In order to improve readability, the results given in Table 2 are also illustrated in Figure 3.

6.1 The Ranked Solution Vector

In Table 3, the lists of the DRs, obtained by human judgement and SemSim, are shown. They are ordered according to decreasing semantic similarity degrees with *rv*.

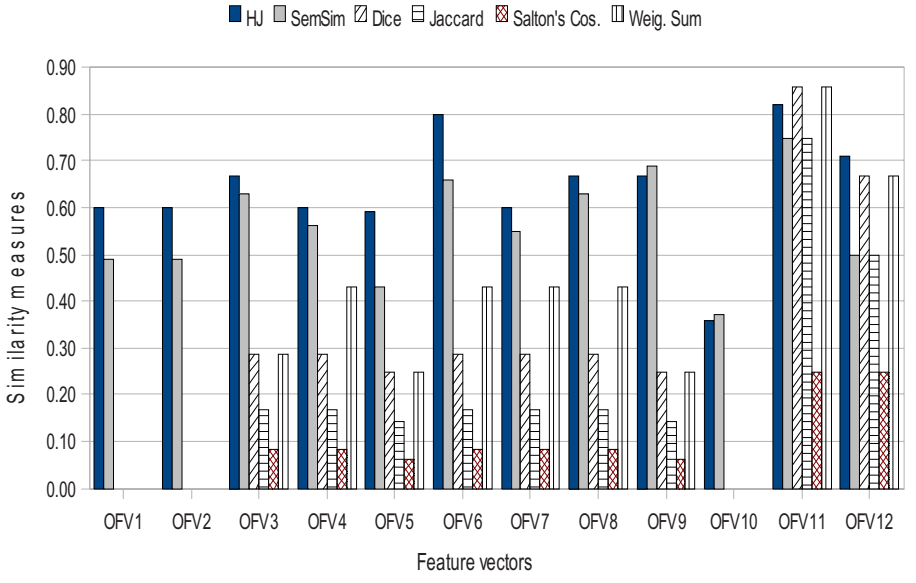

Fig. 3. Illustration of the comparison shown in Table 2

Table 3. Ranked Solution Vectors for HJ and SemSim

Human Judgment (HJ)		***SemSim***	
Ranked Resources	**Values**	**Ranked Resources**	**Values**
H6	0.83	H11	0.75
H11	0.78	H9	0.69
H3, H8, H9	0.75	H6	0.66
H12	0.70	H3, H8	0.63
H1, H2, H4, H7	0.67	H4	0.56
H5	0.63	H7	0.55
H10	0.37	H12	0.50
		H1, H2	0.49
		H5	0.43
		H10	0.37

In this table, the horizontal line separates DRs according to a threshold, here fixed to 0.55. Thus, the Ranked Solution Vector (RSV) of our running example is defined by the DRs above the horizontal line.

Analyzing Table 3, we are able to show the effectiveness of our method according to the precision and recall values. As usual, *precision* is obtained dividing the number of discovered valid resources (the intersection between the first and the third columns of Table 3) by the total number of discovered resources (third column of Table 3), while *recall* is computed dividing the number of discovered valid resources by the total number of valid resources (first column of Table 3), up to a threshold. Finally, the F-measure, which is two times the product of precision and recall divided by their sum, is also given.

In our case, assuming the threshold equal to 0.55, we have:

Precision	Recall	F-measure
1	0.64	0.78

Note that, in general, selecting higher thresholds results in higher precision values, while selecting lower thresholds leads to higher recall values.

7 Conclusions and Future Work

The problem of achieving new generations of search engines capable of exploiting the emerging semantic technologies is attracting much attention today. It is a widely shared opinion that we need to perform a "quantum leap" and achieve new generation search engines that exploit the semantic content of a resource when performing a search and retrieval task. In this paper, we presented the SemSim method that goes in this direction. Our method is innovative since it is based on the possibility of annotating each DR with a vector of characterizing features (OFV), selected from the concepts of an ontology. Our method is based on three key elements: (i) a Weighted Reference Ontology, where each concept in the ISA hierarchy is weighted using a probabilistic distribution approach; (ii) the use of the Lin method to determine the similarity between concepts (i.e., *consim*) in the Ontology; (iii) the use of the Hungarian Algorithm to compute the similarity degree between a *rv* and an *ofv*.

The SemSim method has been implemented and a number of tests have been carried out that show its high correlation with human judgment.

In the future we intend first of all to carry out extensive experiments to acquire a better understanding of the characteristics of our method. A further direction is represented by the possibility of associating a weight with the elements of the request vector, allowing the user to specify a scale of importance on the desired features. This is the first requirement that emerged from the participants when they performed human test: not all the required features are equivalent, users would like to indicate what are the important (or even mandatory) features with respect to other features for which a compromise is acceptable.

Another line of activities will concern the WRO and the method to assign weights to concepts. Currently, the weights are defined according to a uniform distribution of probability. We wish to explore the behaviour of the SemSim method in presence of a skewed probability distribution that may be useful in many cases.

References

1. Alani, H., Brewster, C.: Ontology ranking based on the Analysis of Concept Structures. In: K-CAP 2005, Banff, Alberta, Canada (2005)
2. Castano, S., De Antonellis, V., Fugini, M.G., Pernici, B.: Conceptual Schema Analysis: Techniques and Applications. ACM Transactions on Databases Systems 23(3), 286–333 (1998)
3. Cordì, V., Lombardi, P., Martelli, M., Mascardi, V.: An Ontology-Based Similarity between Sets of Concepts. In: Proceeding of WOA 2005, pp. 16–21 (2005)
4. DSTC, IBM, Sandpiper Software; Ontology Definition Metamodel, Revised submission to OMG (2005), http://www.omg.org/docs/ad/05-01-01.pdf

5. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
6. Fang, W.-D., Zhang, L., Wang, Y.-X., Dong, S.-B.: Towards a Semantic Search Engine Based on Ontologies. In: Proc. of 4th Int'l Conference on Machine Learning, Guangzhou (2005)
7. Formica, A.: Concept similarity by evaluating Information Contents and Feature Vectors: a combined approach. Communications of the ACM (CACM) (to appear, 2008)
8. Formica, A., Missikoff, M.: Concept Similarity in SymOntos: an Enterprise Ontology Management Tool. Computer Journal 45(6), 583–594 (2002)
9. Frakes, W.B., Baeza-Yates, R.: Information Retrieval, Data Structure and Algorithms. Prentice Hall, Englewood Cliffs (1992)
10. Gruber, T.R.: A translation approach to portable ontologies. Knowledge Acquisition 5(2), 199–220 (1993)
11. Jaccard, P.: Bulletin del la Société Vaudoise des Sciences. Naturelles 37, 241–272 (1901)
12. Kasahara, K., Matsuzawa, K., Ishikawa, T., Kawaoka, T.: Viewpoint-Based Measurement of Semantic Similarity between Words. In: Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics, pp. 292–302. Fort Lauderdale (1995)
13. Kim, J.W., Candan, K.S.: CP/CV: Concept Similarity Mining without Frequency Information from Domain Describing Taxonomies. In: Proc. of CIKM 2006 Conference (2006)
14. Li, Y., Bandar, Z.A., McLean, D.: An Approach fro Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Transactions on Knowledge and Data Engineering 15(4), 871–882 (2003)
15. Lin, D.: An Information-Theoretic Definition of Similarity. In: Shavlik, J.W. (ed.) Proc. of 15th the International Conference on Machine Learning, Madison, Wisconsin, USA, pp. 296–304. Morgan Kaufmann, San Francisco (1998)
16. Maarek, Y.S., Berry, D.M., Kaiser, G.E.: An Information Retrieval Approach For Automatically Constructing Software Libraries. IEEE Transactions on Software Engineering 17(8), 800–813 (1991)
17. Madhavan, J., Halevy, A.Y.: Composing Mappings among Data Sources. In: VLDB 2003, pp. 572–583 (2003)
18. Maguitman, A.G., Menczer, F., Roinestad, H., Vespignani, A.: Algorithmic Detection of Semantic Similarity. In: Proc. of WWW 2005 Conference, Chiba, Japan (May 2005)
19. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Language and Cognitive Processes 6(1), 1–28 (1991)
20. Munkres, J.: Algorithms for the Assignment and Transportation Problems. Journal of the Society of Industrial and Applied Mathematics 5(1), 32–38 (1957)
21. Rada, L., Mili, V., Bicknell, E., Bletter, M.: Development and application of a metric on semantic nets. IEEE Transaction on systems. Man, and Cybernetics 19(1), 17–30 (1989)
22. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proc. of IJCAI (1995)
23. Shannon, C.E.: A Mathematical Theory of Communication. The Bell System Technical Journal 27, 379–423, 623–656 (1948)
24. Velardi, P., et al.: TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. In: 9th Conf. on Terminology and Artificial Intelligence TIA 2007, Sophia Antinopolis (2007)
25. Velardi, P., Navigli, R., Cuchiarelli, A., Neri, F.: Evaluation of ontolearn, a methodology for automatic population of domain ontologies. In: Buitelaar, P., Cimiano, P., Magnini, B. (eds.) Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press, Amsterdam (2005)
26. Wu, Z., Palmer, M.: Verb semantics and lexicon selection. In: The 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New, Mexico, pp. 133–138 (1994)