Linked Data Based Approach to Similarity Reasoning

Anna Formica, Michele Missikoff, Elaheh Pourabbas, Francesco Taglino

National Research Council, Istituto di Analisi dei Sistemi ed Informatica "A. Ruberti"
Viale Manzoni 30, I-00185 Rome, Italy

{name.surname}@iasi.cnr.it

Abstract: *SemSim* is a semantic similarity reasoning method that has been conceived to be used as a service for the Semantic Web. *SemSim* is based on a Weighted Reference Ontology, which is used to semantically annotate a collection of digital resources (e.g., documents) to be searched. In this paper we present a new approach to *SemSim* implementation based on Linked Data, that significantly increments its usability in the Semantic Web.

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

1. INTRODUCTION

Innovation generally starts from a creative idea, triggered by a specific problem requesting a non trivial solution or from an offered opportunity, e.g., by new technological solutions. But the innovative idea is just the starting point of a long undertaking, a seed that needs to be 'watered', 'fertilized', and cared to grow into a concrete value proposition for the enterprise. A key 'fertilizer' for innovations is represented by knowledge. Given a brilliant idea, we need to verify how promising it really is. First of all, checking if a similar idea has been explored in the past: is our idea really innovative or is just a step towards evolution? Are there previous similar experiences? Were they successful or not? If negative, what were the difficulties and obstacles encountered? These are some among the initial questions for which we would like to get an answer. Nowadays, there is an emerging movement, referred to as Open Innovation [1] that makes such questions easier to find an answer. But Open Innovation is not easy to practice, both for socio-economic and technological motivations. The work presented in this paper intends to address two problems that fall in the latter group: (i) the difficulty we encounter in finding (over the Internet, but also within a single company and its knowledge resources) the knowledge that appears relevant (to what extent? Is it possible to assess its relevance?) to the proposed innovative idea; (ii) the possibility to concretely extract and use the identified knowledge resources, despite of the different formats and applications used to generate them.

In this paper we address the two mentioned problems by proposing a representation of the notions underlying the *SemSim* semantic similarity method [2], according to the *Linked Data* approach [3]. *SemSim* is a method used to retrieve and assess the degree of similarity between a request and a knowledge resource, which is

based on a weighted ontology and semantically annotated resources. Linked Data is an approach aiming at encoding knowledge resources in a way that they can be easily accessed and reused in different contexts. Furthermore, Linked Data is the base for the inclusion of the knowledge resources in the vast open knowledge network belonging to the Semantic Web. With this approach we are able to publish *SemSim* as a service on the Semantic Web that can be freely invoked as long as the weighted ontology and resources' annotations are made accessible in a format compliant with Linked Data.

Regarding the related work, currently, there are several proposals following on the one hand the Linked Data principles for defining vocabularies and describing data (e.g., documents, people, etc.), and on the other hand semantic similarity approaches. Concerning the Linked Data initiatives, the most popular is DBpedia[1], which aims at extracting structured content from Wikipedia and representing it in a RDF format. With regard to semantic similarity approaches, see [2] for a detailed related work.

Concerning a joint approach the literature is still quite limited. It is worth mentioning the work in [6], although it adopts the Linked Data approach for representing the resources while we extend it to the semantic search engine itself.

The paper is organized as follows. In Section 2, we briefly recall the *SemSim* method. In Section 3, the knowledge space underlying our approach is presented, which is organized according to four levels. In Section 4, we propose an RDF representation of the *SemSim* notions in order to enabling Linked Data mechanisms. Finally, the Conclusion follows.

2. THE SEMSIM METHOD: AN OVERVIEW

The *Universe of Digital Resources* (UDR) represents the knowledge space where *SemSim* operates, it consists in a collection of digital resources that are semantically annotated using a reference ontology. In our work we address a simplified notion of ontology, *Ont*, consisting of a set of concepts organized according to a ISA hierarchy. In particular, *Ont* is a *taxonomy* defined by the pair $Ont = <C, H>$, where C is a set of concepts and H is a set of ordered pairs of concepts of C such that if $(c_i,c_j) \in H$, then c_j ISA c_i, i.e., c_i is a more general concept than c_j.

Consider an ontology $Ont = <C, H>$. A *request feature vector* (*request vector* for short) rv is defined by a set of ontology concepts (the order of the concepts is irrelevant), i.e., $rv = (c_1,...,c_n)$ where $c_i \in C$. Analogously, given a digital resource $dr_i \in$ UDR, an ontology feature vector ofv_i associated with dr_i is defined by a set of ontology concepts describing the resource as follows: $ofv_i = (c_{i,1},...,c_{i,m})$, where $c_{i,j} \in C$, $j = 1,...,m$.

A *Weighted Reference Ontology (WRO)* is a pair $WRO = <Ont, w>$, where w is a function defined on C, such that given a concept $c \in C$, $w(c)$ is a decimal number in the interval $[0,...,1]$.

In [2], we have experimented the method in the tourism domain, therefore the examples below are drawn upon it. We are soon going to adopt it in the BIVEE project, focusing on innovation. In this case, the rv will be, e.g., a problem and the

[1] http://dbpedia.org

UDR will be a knowledge space from which we wish to extract the documents relevant to this problem.

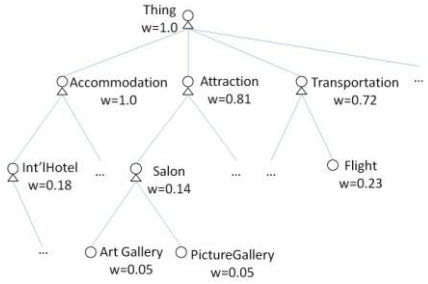

Fig. 1. A fragment of the Weighted Reference Ontology

In our experiment the digital resources are vacation packages for visiting a European capital, which are offered by a tourism agency. Each package is annotated with one ontology feature vector defined by using the concepts of the WRO. The UDR contains 22 vacation packages, which are indicated as *h1..h22*. A fragment of the WRO is shown in Figure 1. For instance, below some of the 22 *ofvs* are recalled:

ofv1 = (InternationalHotel, FrenchMeal, Cinema, Flight)
ofv2 = (Pension, VegetarianMeal, ArtGallery, ShoppingCenter)
ofv3 = (CountryResort, MediterraneaMeal, Bus)
...
ofv7 = (RegularAccommodation, RegularMeal, Salon, Flight)
...
ofv15 =(InternationalHotel, PictureGallery, Flight)
....

Suppose a tourist wants to visit a European capital and, in order to buy a vacation package, he/she expresses some preferences. For instance, she/he wants to travel by plane, sleep in a international hotel, have international food, and enjoy art galleries. According to *SemSim*, these preferences can be formulated by using the following request feature vector:

rv = (InternationalHotel, InternationalMeal, ArtGallery, Flight)

The *SemSim* method allows the user to choose among the 22 vacation packages offered by the tourism agency the one that better satisfies his/her needs. In particular, it evaluates the similarity between feature vectors, which is based on the notion of similarity between concepts (features), referred to as *consim*. Given a WRO, the *consim* notion relies on the information content approach defined by Lin [4], according to which the information content of a concept *c* is defined as *-log w(c)*, where *w* is the weight associated with the concept *c* in the WRO. Therefore, as the weight of a concept increases the informativeness decreases hence, the more abstract a concept the lower its information content.
On the basis of the *consim*, the *SemSim* method allows us to compute the semantic similarity between a request vector *rv* and an *ofv*, indicated as *semsim(rv,ofv)*. Such a computation essentially focuses on the pairs of concepts, one from the *rv* and the

other one from the *ofv*, that exhibit high affinity, computed according to the so called *stable marriage problem* [5]. Given a request vector *rv*, on the basis of the *semsim(rv,ofv)* values, a *Ranked Solution Vector* (*RSV*) is defined, which provides a ranked list of *ofvs* most similar to the *rv*. In the *RSV* each *ofv* is a associated with the related *semsim score*, from the highest to the lowest values, down to a threshold. For instance, in our experiment the threshold has been fixed to 0.5, and in the case of the request vector *rv*, the resulting RSV is the following:

RSV(*rv*) = <(*ofv15*, 0.66), (*ofv7*, 0.60), (*ofv1*, 0.52)>

Therefore, according to our approach, the *ofv15*, which refers to the *h15* resource, is the most similar vacation package among the 22 available to the user preferences. In fact, *ofv15* and *rv* have both the features *InternationalHotel* and *Flight* which match exactly. Furthermore, *ofv15* has the feature *PictureGallery* sharing the information content of *Salon* with the feature *ArtGallery* of *rv* (see the WRO in Figure 1). The similarity between *rv* and *ofv7* is lower because they have only one feature with an exact match (*Flight*) and all the remaining features sharing some information content in the taxonomy whose overall similarity does not exceed that with *ofv15*. Analogously, in the case of *ofv1*, although it has two features matching exactly with *rv* (*InternationaHotel* and *Flight*).

3. THE KNOWLEDGE SPACE ORGANIZATION

The Linked Data approach requires the adoption of standard vocabularies in representing the information structures to be exposed, shared, and connected to other pieces of data, information, and knowledge in the Semantic Web[2] [4]. Such vocabularies are described by metadata, classified and interlinked. In this paper we focus on shared vocabularies and utilize them to represent the whole knowledge space, called *reference knowledge space*, that include the UDR of a given domain (like Tourism or Business Innovation) but also the meta-knowledge used by *SemSim* service. The reference knowledge space, as shown in Figure 2, is composed of four levels as follows:

- The *Vocabulary* level defines the terminology used in the Linked Data implementation of *SemSim*. It is based on the well-known OWL[3], RDF[4], RDFS[5], XML Schema[6], and SKOS[7]. In particular, SKOS (Simple Knowledge Organization System) in our work has been used as a common data model for defining the *Domain Concepts* and the *SemSim* glossaries. The former is addressed to organize the knowledge of the given domain and the latter is conceived to model the data structures of the *SemSim* method introduced in the previous section.

---

2 http://linkeddata.org/
3 http://www.w3.org/2002/07/owl#
4 http://www.w3.org/1999/02/22-rdf-syntax-ns#
5 http://www.w3.org/2000/01/rdf-schema#
6 http://www.w3.org/2001/XMLSchema#
7 http://www.w3.org/2004/02/skos/core#

- The *Knowledge schema* level represents the schemes of the main components of the *weighted taxonomy* (e.g., broader, narrower) and the *SemSim* method, which are *ofv*, *rv*, and *RSV*. This level essentially provides further details about the structure and constraints implementing *SemSim*.
- The *Knowledge fact* level represents the extensions of the elements defined at the schema level. Essentially they are organized as a weighted hierarchy of concepts and a set of *ofv*s.
- The *Domain resource* level refers to the resources of the selected UDR on the basis of a specific application domain (e.g., Tourism or Business Innovation), each of which is annotated with one *ofv*s.

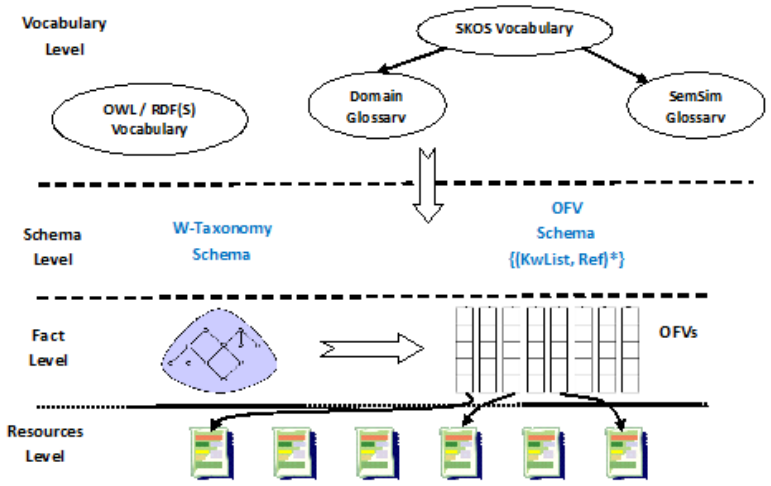

Fig. 2. The reference Knowledge Space

4. A LINKED DATA SOLUTION

In this section the *Knowledge schema* and *Knowledge fact* levels, presented in the previous section, are further detailed and modelled according to RDF. To this end we first provide in Figure 3 a graphical representation of both these levels. Successively, the diagrams are represented in RDF Turtle syntax[8].

In Figure 3(a) the notion of *ofv* has been generalized in order to model both the *annotation vector* (*AnnotV*), and the *request vector* (*RV*) as specializations of a generic set of concepts referred to as *OFV*. In particular, the former is always associated with a resource whereas the latter is used to express the user preferences. Furthermore, the *RSV element* (*RSVElem*) allows us to model the elements of the *ranked solution vector* (*RSV*). Each *RSVElem* has two properties, namely *hasAnnotV* and *hasScore* whose ranges are *AnnotV* and *float*, respectively. The *weighted taxonomy* can be organized as a set of concepts related by the *narrower*/*broader* relationship. Each concept in the taxonomy has one property, namely *hasWeight*

[8] http://www.w3.org/TR/turtle/

whose range is *float*. In Figure 3(b) a fragment of the weighted taxonomy in the tourism domain and a possible instance of *AnnotV* (*ofv15*) are illustrated.

In Table 1, see below, we present a RDF representation of the *semsim* namespace. The *SemSim* notions are defined in terms of existing RDF-based vocabularies, which are adopted by referring to their namespaces. In particular, we use *rdf*, *rdfs*, *owl*, *skos* and *xsd* as prefixes for RDF, RDFS, OWL, SKOS and XML Schema namespaces, respectively.

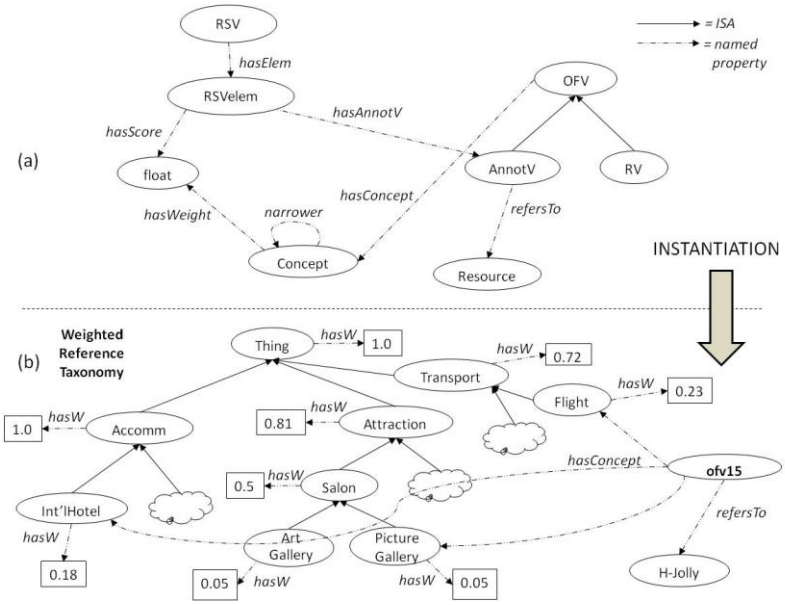

Fig. 3. A graph-based representation of the *SemSim* vocabulary (a) and its application (b)

In accordance with Figure 3(a), we present in Table 1 the *SemSim* data structures and properties, where their initial letter is denoted in upper case (e.g., *AnnotV*) and lower case (e.g., *hasConcept*), respectively.

To build the WRO, we refer to SKOS, a W3C recommendation designed for the representation of thesauri, classification schemes, taxonomies, or any other type of structured controlled vocabulary. The main reason for adopting SKOS is that it is widely used in the context of the Semantic Web, and therefore it allows us the reuse of existing taxonomies and, in turn, publish the *SemSim* vocabulary for a wide use on the Web. Since SKOS does not support the notion of the weight of a concept, as required in our approach, in the *semsim* namespace the *hasWeight* property has been introduced.

In accordance with Figure 3(b), we report in Table 2 an example of *weighted taxonomy* and *annotation vector*, by using the *SemSim* vocabulary defined above, and the Turtle syntax.

Table 1. Semsim data structures and properties

OFV	rdf:type	owl:Class.
AnnotV	rdf:type	owl:Class;
	rdfs:subClassOf	OFV.
RV	rdf:type	owl:Class;
	rdfs:subClassOf	OFV.
RSV	rdf:type	owl:Class.
hasWeight	rdf:type	owl:DatatypeProperty;
	rdfs:domain	skos:Concept;
	rdfs:range	xsd:float.
hasConcept	rdf:type	owl:ObjectProperty;
	rdfs:domain	OFV;
	rdfs:range	skos:Concept.
refersTo	rdf:type	owl:ObjectProperty;
	rdfs:domain	AnnotV.
hasScore	rdf:type	owl:DatatypeProperty;
	rdfs:domain	RSVElem;
	rdfs:range	xsd:float
hasAnnotV	rdf:type	owl:ObjectProperty;
	rdfs:domain	RSVElem;
	rdfs:range	AnnotV.
hasRSVElem	rdf:type	owl:ObjectProperty;
	rdfs:domain	RSV;
	rdfs:range	RSVElem.

5. CONCLUSION

In this paper, we have proposed a RDF representation of the notions underlying the *SemSim* semantic similarity method. *SemSim* works on a weighted ontology, on the basis of which resources of interest are annotated by using Ontology Feature Vector (OFV) structures. Leveraging on the weighted ontology and the OFVs, the method is able to assess the semantic similarity between a given request (*request vector*, RV) and available resources, by returning a ranked list of best matches (*ranked solution vector*, RSV). OFV, RV and RSV structures have been modelled by re-using very popular vocabularies in the Semantic Web such as OWL, RDF(S), XML Schema, while the reference taxonomy has been represented in SKOS. Such a RDF-based representation allows us to define *SemSim* structures in accordance with the Linked Data approach. Accordingly, we are able to publish a Linked Data compliant *SemSim* service. In order to invoke such a web service, a weighted SKOS taxonomy, and the annotations of the resources of interest need also to be available on the Web, by adopting the *semsim* namespace specifications, as described in Section 4.

Table 2. Weighted taxonomy and annotation vector

Thing	rdf:type	skos:Concept;
	semsim:hasWeight	'1.0'.
Accommodation	rdf:type	skos:Concept;
	semsim:hasWeight	'1.0'.
Int'lHotel	rdf:type	skos:Concept;
	skos:narrower	Accommodation;
	semsim:hasWeight	'0.18'.
Attraction	rdf:type	skos:Concept;
	semsim:hasWeight	'0.81'.
Salon	rdf:type	skos:Concept;
	skos:narrower	Attraction;
	semsim:hasWeight	'0.5'.
ArtGallery	rdf:type	skos:Concept;
	skos:narrower	Salon;
	semsim:hasWeight	'0.05'.
PictureGallery	rdf:type	skos:Concept;
	skos:narrower	Salon;
	semsim:hasWeight	'0.05'.
Transportation	rdf:type	skos:Concept;
	semsim:hasWeight	'0.72'.
Flight	rdf:type	skos:Concept;
	skos:narrower	Transportation;
	semsim:hasWeight	'0.23'.
ofv15	rdf:type	semsim:AnnotV;
	semsim:refersTo	H-Jolly;
	semsim:hasConcept	Int'lHotel,
		PictureGallery,
		Flight.

REFERENCES

1. Chesbrough, H., Open Innovation: The New Imperative for Creating and Profiting from Technology, Boston Harvard Business School Press, 2003, ISBN: 1-57851-837-7.
2. Formica A., Missikoff M., Pourabbas E., Taglino F. (2013). Semantic search for matching user requests with profiled enterprises. Computers in Industry, 64: 191-202.
3. Heath T., Bizer C. (2011). Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology February 2011, 136 pages, (doi:10.2200/S00334ED1V01Y201102WBE001) .
4. Lin, D. (1998). An Information-Theoretic Definition of Similarity. In Proceedings of 15th the International Conference on Machine Learning. Madison, Wisconsin, USA, Morgan Kaufmann. Shavlik J. W. (ed.), 296-304.
5. Mairson H. (1992). The Stable Marriage Problem. The Brandeis Review, 12(1).
6. Sheng H., Chen H., Yu T., Feng Y. (2010). Linked data based semantic similarity and data mining. In Proc of 11th IEEE International Conference on Information Reuse and Integration (IRI 2010). Las Vega, USA, 104-108.