A Bayesian Approach for Weighted Ontologies and Semantic Search

Anna Formica[1], Michele Missikoff[2], Elaheh Pourabbas[1] and Francesco Taglino[1]

[1]*National Research Council, Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti",*
Via dei Taurini 19 - 00185, Rome, Italy
[2]*National Research Council, Istituto di Scienze e Tecnologie della Cognizione,*
Via S. Martino della Battaglia, 44 - 00185, Rome, Italy

Keywords: Semantic Search, Similarity Reasoning, Weighted Reference Ontology, Bayesian Network.

Abstract: Semantic similarity search is one of the most promising methods for improving the performance of retrieval systems. This paper presents a new probabilistic method for ontology weighting based on a Bayesian approach. In particular, this work addresses the semantic search method *SemSim* for evaluating the similarity among a user request and semantically annotated resources. Each resource is annotated with a vector of features (annotation vector), i.e., a set of concepts defined in a reference ontology. Analogously, a user request is represented by a collection of desired features. The paper shows, on the bases of a comparative study, that the adoption of the Bayesian weighting method improves the performance of the *SemSim* method.

1 INTRODUCTION

A weighted ontology is obtained by associating a weight with each concept. The adoption of such weights has proved to be beneficial in several ontology-based applications and services that range from ontology mapping to ontology-based decision making, to semantic search. Our main research objectives lay in the area of Semantic Similarity Reasoning for advanced search, where weighted ontologies represent a primary base for our search engine, called *SemSim* (Formica et al., 2013). There are various methods for weighting ontology concepts (see Related Work Section) and this great variety also depends on the meaning that such weights assume. In *SemSim* the concept weight is used to derive the information content (IC) of a concept in a hierarchy according to (Resnik, 1995), that represents the basis for computing the concept similarity.

According to the semantics adopted in *SemSim*, given a Universe of Digital Resources (UDR, i.e., the search space) the IC of a concept is directly related to its selectiveness and inversely related to the probability that, randomly selecting an instance in the UDR, such an instance belongs to the extension of the concept (i.e., the set of its instances). In a retrieval perspective, a concept with higher IC is expected to be more selective than a concept with lower IC, since the former has a lower probability than the latter. Furthermore, in an ontology, the concept weight decreases downward along the specialization hierarchy, proceeding from the root (the most general concept, e.g., *Thing*) towards the leaves, the most specific concepts. This is because the extension of a more specific concept is contained in the extension of a more general concept, and consequently, the likelihood of the former is lower than the likelihood of the latter. Symmetrically, the IC will progressively increase while we move downward along the hierarchy. For example, given a tourism domain, assuming that *Farmhouse* is a more specific concept of *Accommodation*, the latter has a lower information content than the former. Accordingly, at ontology level, the weight of concepts needs to be consistent with the set inclusion semantics of the ISA hierarchy, as described above.

The main focus of this paper concerns a method for assigning the weights to the concepts in the ontology, used to compute their IC. It may seem a marginal problem, but it is not, since a correct weighting strategy can significantly improve the performance of the semantic services based on weighted ontologies. In the literature there are several proposals, as reported in the Related Work Section.

In a previous work (Formica et al., 2013), we experimented two methods of concept weighting, both following the IC approach. The two methods, called *frequency-based* and *probabilistic-based*, differ for the way the likelihood of a concept c is computed. The frequency-based method is the most straightforward, it consists in counting the instances of each

Formica, A., Missikoff, M., Pourabbas, E. and Taglino, F.
A Bayesian Approach for Weighted Ontologies and Semantic Search.
DOI: 10.5220/0006073301710178
In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 2: KEOD*, pages 171-178
ISBN: 978-989-758-203-5

concept and then normalizing such a number over the total number of instances of concepts. The probabilistic-based method depends, in general, on the characteristics of the application domain and the structure of the taxonomy. Both approaches exhibit pros and cons. The frequency-based method depends on the number of UDR resources, and the more this number increases the more the concept weights are expected to be trustworthy. However, this method can be expensive, since it is necessary to maintain a count of the extensions of each concept in the ontology. This is only feasible in the *Closed World Assumption* and in the presence of a relatively stable UDR, otherwise for each update it is necessary to recompute a (more or less extensive) part of ontology weights. Conversely, the probabilistic-based approach is valid for both *Closed* and *Open World Assumptions* and, with *large* UDR, tends to be more stable. In fact, UDR updates usually evolve according to a given probability distribution (e.g., in a University, the ratio of students to professors tend to be stable, despite a significant number of new enrollments every year). Here, the problem is to be able to *guess* what are the more appropriate probabilities to be associated with each concept. In the perspective of addressing large application domains, in this paper we adopt the probabilistic approach.

The *SemSim* similarity reasoning method is based on semantic annotations in the form of Ontology-based Feature Vectors (OFV), where each component is a concept in a reference ontology (referred to as feature when used for annotation). Each resource in the UDR is associated with an annotation vector (referred to as AV) while the user request is represented by a request vector (RV). The *SemSim* search engine then computes the semantic similarity by contrasting the RV against each AV, producing a ranked list of resources.

In this paper, we present a new probabilistic method for ontology weighting based on a Bayesian Network (BN). The importance of integrating BN into ontologies is discussed in (Grubisic et al., 2013), where they are considered a valid support to experts in the modeling of a specific problem domain. The new Bayesian weighting method defined here can be considered as an extension of the previous method defined in (Formica et al., 2013), i.e., it starts with the same probability assignments (w_p) given in the mentioned paper, used here as *a priori* probabilities. Then, such probabilities are refined taking into consideration the subsumption relations and the inherent dependencies among concepts. To this end, we build a BN isomorphic to the ISA hierarchy of the ontology referred to as Onto-Bayesian Network (OBN). It is constructed as follows: each concept in the ontology corresponds to a node in OBN and each subsumption relation corresponds to a probabilistic dependency. Considering the above example, we assume that the probabilistic weight of *Farmhouse* depends on the weight of its subsumer *Accommodation*. The method, referred to as *SemSim-b*, is illustrated in Section 3, while Section 4 shows its validation, by examining the correlation, precision, and recall and comparing them with the experiment given in (Formica et al., 2013). The results demonstrate that *SemSim-b* outperforms the previous *SemSim* method. Finally, Section 5 concludes the paper.

2 WEIGHTED ONTOLOGIES

In this section, we recall the probabilistic approach defined in (Formica et al., 2010), (Formica et al., 2013) in order to assign weights to a given ontology.

The UDR is the set of digital resources that are semantically annotated with a reference ontology (an ontology is a formal, explicit specification of a shared conceptualization (Gruber, 1993)). In our work we address a simplified notion of ontology, *Ont*, consisting of a set of concepts organized according to a specialization hierarchy. In particular, *Ont* is a taxonomy defined by the pair:

$$Ont = < C, ISA >$$

where $C = \{c_i\}$ is a set of concepts and *ISA* is the set of pairs of concepts in C that are in subsumption (subs) relation:

$$ISA = \{(c_i, c_j) \in C \times C | subs(c_i, c_j)\}$$

Given two concepts $c_i, c_j \in C$, their least upper bound, $lub(c_i, c_j)$, is always uniquely defined in C (we assume the hierarchy is a tree). It represents the least abstract concept of the ontology that subsumes both c_i and c_j.

Each resource in the UDR is annotated with an *OFV*[1], which is a vector that gathers a set of concepts of the ontology *Ont*, aimed at capturing its semantic content. The same also holds for a user request. It is represented as follows:

$$ofv = (c_1,...,c_n), \text{ where } c_i \in C,\ i = 1,...,n$$

Note that, when an *OFV* is used to represent the semantics of a user request, it is referred to as semantic *Request Vector* (*RV*) whereas, if it is used to represent the semantics of a resource, it is referred to as

[1] The proposed *OFV* approach is based on the *Term Vector* (or *Vector Space*) Model approach, where terms are substituted by concepts (Salton et al., 1975).

semantic *Annotation Vector* (*AV*). They are denoted as follows, respectively:

$$rv = \{r_1, \ldots, r_n\},$$
$$av = \{a_1, \ldots, a_m\},$$

where $\{r_1, \ldots, r_n\} \cup \{a_1, \ldots, a_m\} \subseteq C$.

Finally, a *Weighted Reference Ontology* (*WRO*) is defined as follows:

$$WRO =< Ont, w >$$

where w, the concept weighting function, is a probability distribution defined on C, such that given $c \in C$, $w(c)$ is a decimal number in the interval $[0\ldots1]$.

Figure 1 shows the WRO drawn upon the tourism domain that will be used in the running example. In this figure, the weights w_p have been assigned according to the probabilistic-based approach defined in (Formica et al., 2013). In particular, the weight of the root of the hierarchy, referred to as *Thing* is equal to 1, and the weights of the concepts of the hierarchy are assigned according to a top-down approach, as follows. Given a concept c, let c' be the parent of c, $w_p(c)$ is equal to the probability of c', divided by number of the children of c':

$$w_p(c) = \frac{w_p(c')}{|children(c')|}$$

For instance, let us consider the concept *Salon*. The associated w_p is 0.05 because $w_p(Attraction) = 0.2$ and *Attraction* has four children.

Below, the semantics of a feature vector OFV is presented, by extending its first formulation given in (Formica et al., 2008). When a concept is used in an OFV to annotate a resource or to represent a *RV*, it is referred to as a *feature*.

Consider an ontology *Ont*, a set of features F, $F \subseteq C$, let *SPEC* be the reflexive and transitive closure of *ISA*. Then, the semantics of a feature $a \in F$ is defined as follows:

$$\Gamma(a) = \{res \in UDR \mid feat(b, res), (b, a) \in SPEC\}$$

where $feat(b, res)$ means that the resource $res \in$ UDR is annotated by the feature $b \in F$. The semantics of an *ofv*, say:

$$ofv = \{a_1,\ldots, a_n\}$$

is therefore defined according to the Γ function as follows:

$$\Gamma(ofv) = \bigcap_j \Gamma(a_j)$$

Overall, we emphasize the difference between the concept semantics and feature semantics. A concept denotes the set of the instances whose type is such a concept; while, if a concept is used as a feature, it denotes all the resources in the UDR annotated with such a feature.

3 BAYESIAN NETWORKS FOR WEIGHTED ONTOLOGIES

The core of the proposed solution is the adoption of a Bayesian approach for ontology weighting.

Bayesian Networks (BN), also known as belief networks, are probabilistic graphical models based on DAGs. They have been defined in the late 1970s within cognitive science and artificial intelligence as the method of choice for uncertain reasoning (Pearl and Russell, 2001). A BN is therefore a graphical structure that allows us to represent and reason about an uncertain domain. In particular, each node in the graph represents a random variable, that can take different values associated with probabilities, while the edges represent probabilistic dependencies of the corresponding random variables. In particular, an edge from a node X_i (parent node) to a node X_j (child node) indicates that a value taken by the variable X_j depends on the value taken by the variable X_i or, roughly speaking, the variable X_i "influences" X_j. Note that nodes without parents (roots) are not influenced by any node, nodes without children (leaves) do not influence any node, while any node affects its children. Therefore, nodes that are not directly connected in the BN represent variables that are conditionally independent of each other. According to the global semantics of BN, the full joint distribution is:

$$P(x_1,\ldots,x_n) = \prod_i P(x_i|pa_i)$$

where x_i is a value of the variable X_i, pa_i is a set of values for the parents of X_i, and $P(x_i|pa_i)$ denotes the conditional probability distribution of x_i given pa_i.

Therefore, in a BN, each node is associated with a probability function that takes, as input, a particular set of values for the node's parent variables, and gives (as output) the probability of the variable represented by the node. These probability functions are specified by means of (conditional) probability tables[2], one for each node of the graph.

In our approach, as mentioned in the Introduction, we build a BN isomorphic to the ISA hierarchy, referred to as *Onto-Bayesian Network* (*OBN*). In the *OBN*, according to the BN approach, the concepts are boolean variables and the Bayesian weight associated with a concept c, indicated as w_b, is the probability P that the concept c is True (T), i.e.:

$$w_b(c) = P(c{=}T)$$

As mentioned in the Introduction, in order to compute the weights w_b, conditional probability tables are de-

[2] A (conditional) probability table is defined for a set of (non-independent) random variables to represent the marginal probability of a single variable w.r.t. the others.

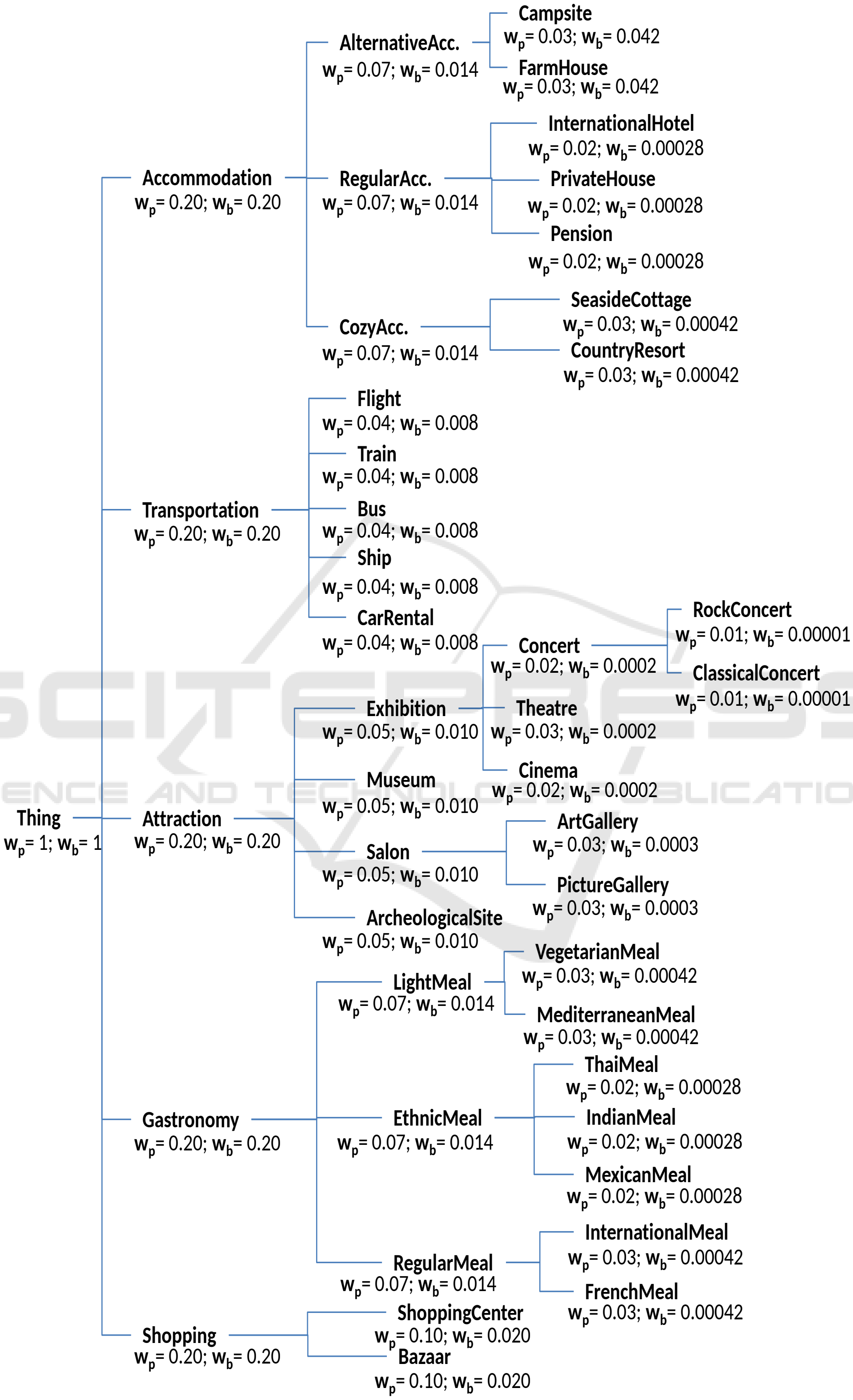
Thing
w_p= 1; w_b= 1
Accommodation
w_p= 0.20; w_b= 0.20
AlternativeAcc.
w_p= 0.07; w_b= 0.014
Campsite
w_p= 0.03; w_b= 0.042
FarmHouse
w_p= 0.03; w_b= 0.042
RegularAcc.
w_p= 0.07; w_b= 0.014
InternationalHotel
w_p= 0.02; w_b= 0.00028
PrivateHouse
w_p= 0.02; w_b= 0.00028
Pension
w_p= 0.02; w_b= 0.00028
CozyAcc.
w_p= 0.07; w_b= 0.014
SeasideCottage
w_p= 0.03; w_b= 0.00042
CountryResort
w_p= 0.03; w_b= 0.00042
Transportation
w_p= 0.20; w_b= 0.20
Flight
w_p= 0.04; w_b= 0.008
Train
w_p= 0.04; w_b= 0.008
Bus
w_p= 0.04; w_b= 0.008
Ship
w_p= 0.04; w_b= 0.008
CarRental
w_p= 0.04; w_b= 0.008
Attraction
w_p= 0.20; w_b= 0.20
Exhibition
w_p= 0.05; w_b= 0.010
Concert
w_p= 0.02; w_b= 0.0002
RockConcert
w_p= 0.01; w_b= 0.00001
ClassicalConcert
w_p= 0.01; w_b= 0.00001
Theatre
w_p= 0.03; w_b= 0.0002
Cinema
w_p= 0.02; w_b= 0.0002
Museum
w_p= 0.05; w_b= 0.010
Salon
w_p= 0.05; w_b= 0.010
ArtGallery
w_p= 0.03; w_b= 0.0003
PictureGallery
w_p= 0.03; w_b= 0.0003
ArcheologicalSite
w_p= 0.05; w_b= 0.010
Gastronomy
w_p= 0.20; w_b= 0.20
LightMeal
w_p= 0.07; w_b= 0.014
VegetarianMeal
w_p= 0.03; w_b= 0.00042
MediterraneanMeal
w_p= 0.03; w_b= 0.00042
EthnicMeal
w_p= 0.07; w_b= 0.014
ThaiMeal
w_p= 0.02; w_b= 0.00028
IndianMeal
w_p= 0.02; w_b= 0.00028
MexicanMeal
w_p= 0.02; w_b= 0.00028
RegularMeal
w_p= 0.07; w_b= 0.014
InternationalMeal
w_p= 0.03; w_b= 0.00042
FrenchMeal
w_p= 0.03; w_b= 0.00042
Shopping
w_p= 0.20; w_b= 0.20
ShoppingCenter
w_p= 0.10; w_b= 0.020
Bazaar
w_p= 0.10; w_b= 0.020

Figure 1: The WRO of our running example with the w_p and w_b concept weights.

Table 1: Probability table of *Gastronomy*.

	Gastronomy	
	T	*F*
	0.20	0.80

Table 2: Conditional probability table of *LightMeal*.

	LightMeal	
Gastronomy	*T*	*F*
T	0.07	0.93
F	0	1

fined by using w_p proposed according to the probabilistic approach recalled in the previous section as *a priori* weights. In particular, given two concepts c_1, c_2, we assume that:

$$P(c_2{=}T|c_1{=}T){=}w_p(c_2)$$

where c_1 is the parent of c_2 according to the *ISA* relation.

For instance, consider the concepts *Gastronomy*, *LightMeal*, and *VegetarianMeal* of the ontology given in Figure 1. *Gastronomy* (G for short) has as parent $Thing$, which is always True ($w_p(Thing) = w_b(Thing) = 1$), therefore the weight w_b coincides with the weight w_p, i.e., $w_b(G) = w_p(G)$, where $w_p(G)$ = $P(G{=}T|Thing{=}T)$ = $P(G{=}T) = 0.2$, as shown in Table 1. Consider now *LightMeal* (L). In this case the conditional probability, which is given in Table 2, depends on the True/False (T/F) values of its father *Gastronomy*, and in particular $w_p(L) = P(L{=}T|G{=}T) = 0.07$ (therefore $P(L{=}F|G = T) = 0.93$). Then, the weight w_b associated with L, $w_b(L)$ = $P(L{=}T)$, is computed starting from the probability of its parent *Gastronomy* as follows:

$$\begin{aligned} P(L{=}T) &= \textstyle\sum_{v\in\{T,F\}} P(L{=}T, G = v) \\ &= P(L{=}T, G{=}T) + P(L{=}T, G{=}F) \\ &= P(L{=}T|G{=}T)P(G{=}T) \\ &\quad + P(L{=}T|G{=}F)P(G{=}F) \\ &= 0.07 \times 0.2 + 0 \times 0.8 = 0.014 \end{aligned}$$

taking into account the Kolmogorov definition for two given variables A and B:

$$P(A,B) = P(A|B)P(B).$$

Therefore, the Bayesian weight $w_b(L) = P(L{=}T) = 0.014$, as shown in Table 4. Analogously, in Table 5 the probability of *VegetarianMeal* (V) is given. The weight $w_b(V)$ has been computed by using the conditional probabilities given in Table 3, and by taking into account the weight $w_b(L)$ computed above (see Table 4).

Table 3: Conditional probability table of *VegetarianMeal*.

	VegetarianMeal	
LightMeal	*T*	*F*
T	0.03	0.97
F	0	1

Table 4: Probability table of *LightMeal*.

	LightMeal	
	T	*F*
	0.014	0.986

4 *SemSim* AND VALIDATION

The *SemSim* method has been conceived to search for the resources in the UDR that best match the *RV*, by contrasting it with the various *AV*, associated with the searchable digital resources. This is achieved by applying the *semsim* function, which has been defined to compute the semantic similarity between *OFV*.

In *SemSim*, the weights are used to derive the IC of the concepts that, according to (Lin, 1998), represents the basis for computing the concept similarity. In particular, according to the information theory, the IC of a concept c, is defined as:

$$IC = -log(w(c))$$

The *semsim* function is based on the notion of similarity between concepts (features), referred to as *consim*. Given two concepts c_i, c_j, it is defined as follows:

$$consim(c_i, c_j) = \frac{2 \times IC(lub(c_i, c_j))}{IC(c_i) + IC(c_j)}$$

where the *lub* represents the least abstract concept of the ontology that subsumes both c_i and c_j. Given an instance of *RV* and an instance of *AV*, say *rv* and *av* respectively, the *semsim* function computes the *consim* for each pair of concepts belonging to the set formed by the Cartesian product of the *rv*, and *av*.

However, we focus on the pairs that exhibit high affinity. In particular, we adopt the exclusive match philosophy, where the elements of each pair of concepts do not participate in any other pair. The method aims to identify the set of pairs of concepts of the *rv*

Table 5: Probability table of *VegetarianMeal*.

	VegetarianMeal	
	T	*F*
	0.00042	0.99958

and *av* that maximizes the sum of the *consim* similarity values (*maximum weighted matching problem in bipartite graphs* (Dulmage and Mendelsohn, 1958)). In particular, given:

$$rv = \{r_1,..., r_n\}$$
$$av = \{a_1,..., a_m\}$$

as defined in Section 2, let *S* be the Cartesian Product of *rv* and *av*:

$$S = rv \times av$$

then, $\mathcal{P}(rv, av)$ is defined as follows:

$$\mathcal{P}(rv,av) = \{P \subset \mathrm{S} \mid \forall\ (r_i, a_j), (r_h, a_k) \in P,$$
$$r_i \neq r_h,\ a_j \neq a_k,\ |P| = min\{n,m\}\}.$$

Therefore, on the basis of the *maximum weighted matching problem in bipartite graphs*, *semsim*(*rv*,*av*) is given below:

$$semsim(rv,av) = \frac{\max\limits_{P\in\mathcal{P}(rv,av)} \left\{ \sum\limits_{(r_i,a_j)\in P} consim(r_i,a_j) \right\}}{max\{n,m\}}$$

In (Formica et al., 2013), we defined *semsim-p* in which the weights w_p of the concepts in the WRO are computed by using the *probabilistic approach*. Analogously, in this paper we introduce *semsim-b* where the weights, w_b, are defined by using a Bayesian Network, as described in Section 3.

4.1 Validation

In order to validate *semsim-b*, we refer to the experiment proposed in (Formica et al., 2013). In that experiment, we considered four request vectors, namely rv_i, $i = 1,...4$, which are recalled below:

$rv_1 = \{Campsite, EthnicMeal, RockConcert, Bus\}$
$rv_2 = \{InternationalHotel, InternationalMeal, ArtGallery, Flight\}$
$rv_3 = \{Pension, MediterraneanMeal, Cinema, ShoppingCenter\}$
$rv_4 = \{CountryResort, LightMeal, ArcheologicalSite, Museum, Train\}$

and 22 annotated resources, represented by their annotation vectors, namely av_1, av_2, ..., av_{22}. Below, only 10 of them are recalled for lack of space:

$av_1 = \{InternationalHotel, FrenchMeal, Cinema, Flight\}$
$av_2 = \{Pension, VegetarianMeal, ArtGallery, ShoppingCenter\}$
$av_3 = \{CountryResort, MediterraneanMeal, Bus\}$
$av_4 = \{CozyAccommodation, VegetarianMeal, Museum, Train\}$
$av_5 = \{InternationalHotel, ThaiMeal, IndianMeal, Concert, Bus\}$
$av_6 = \{SeasideCottage, LightMeal, ArcheologicalSite, Flight, ShoppingCenter\}$
$av_7 = \{RegularAccommodation, RegularMeal, Salon, Flight\}$
$av_8 = \{InternationalHotel, VegetarianMeal, Ship\}$
$av_9 = \{FarmHouse, MediterraneanMeal, CarRental\}$
$av_{10} = \{RegularAccommodation, EthnicMeal, Museum\}$
...

For each request vector, we computed the *semsim* value against the 22 annotation vectors, and then, we calculated the Pearson correlation index (Corr) against human judgment (*HJ*) scores. While the original experiment demonstrated that *SemSim* outperforms some of the most representative similarity methods defined in the literature (i.e., Dice, Jaccard, Cosine, and Weighted Sum (Formica et al., 2013)), in this work, we show in Tables 6 and 7 that *semsim-b* (*SS-b*) achieves a higher correlation with *HJ* with respect to *semsim-p* (*SS-p*).

Furthermore, Table 8 reports the values of the Precision and Recall measures obtained by a threshold fixed to 0.60. As we observe, the Precision achieved by applying *semsim-b* is equal to 1 for all the four *RV*, and for three of them, namely rv_1, rv_3, and rv_4, it is higher with respect to the Precision obtained by applying *semsim-p*. We also observe that both *semsim-b* and *semsim-p* achieve the same Recall (equal to 1) for three out of four *RV*, while the Recall for the remaining *RV* (i.e. rv_1) by *semsim-b* is lower than the one by *semsim-p*.

5 RELATED WORK

In this section, we first recall some of the existing proposals concerning the weighting of the concepts of an ontology. Successively, a second line of research is recalled regarding the integration of BN and Ontologies.

5.1 Ontology Weighting Methods

In (Gao et al., 2015) an approach based on edge-counting and information content theory for measuring semantic similarities has been presented. In particular, different ways of weighting the shortest path length are proposed, although they are essentially based on WordNet frequencies. As also mentioned in (Formica et al., 2008), we do not adopt the WordNet frequencies for several reasons. Firstly, because

Table 6: Correlation about rv_1 and rv_2.

AV	rv_1			rv_2		
	HJ	*SS-p*	*SS-b*	*HJ*	*SS-p*	*SS-b*
av_1	0.10	0.54	0.20	0.72	0.80	0.68
av_2	0.10	0.34	0.12	0.21	0.55	0.43
av_3	0.25	0.50	0.37	0.16	0.35	0.18
av_4	0.18	0.49	0.22	0.10	0.49	0.26
av_5	0.51	0.64	0.37	0.10	0.47	0.34
av_6	0.14	0.40	0.18	0.20	0.49	0.34
av_7	0.16	0.51	0.24	0.71	0.90	0.77
av_8	0.10	0.37	0.20	0.10	0.49	0.38
av_9	0.10	0.46	0.29	0.10	0.35	0.18
av_{10}	0.21	0.49	0.32	0.40	0.46	0.30
av_{11}	0.15	0.45	0.16	0.10	0.44	0.28
av_{12}	0.10	0.25	0.12	0.10	0.23	0.10
av_{13}	0.89	0.71	0.67	0.10	0.34	0.16
av_{14}	0.10	0.38	0.16	0.44	0.55	0.41
av_{15}	0.10	0.33	0.13	0.86	0.70	0.64
av_{16}	0.10	0.39	0.18	0.25	0.54	0.40
av_{17}	0.93	0.87	0.77	0.10	0.48	0.21
av_{18}	0.26	0.46	0.17	0.10	0.39	0.21
av_{19}	0.50	0.73	0.37	0.10	0.46	0.23
av_{20}	0.34	0.51	0.32	0.10	0.41	0.21
av_{21}	0.77	0.85	0.52	0.10	0.48	0.26
av_{22}	0.46	0.72	0.56	0.10	0.46	0.20
Corr	1.00	0.90	0.93	1.00	0.83	0.88

Table 7: Correlation about rv_3 and rv_4.

AV	rv_3			rv_4		
	HJ	*SS-p*	*SS-b*	*HJ*	*SS-p*	*SS-b*
av_1	0.10	0.55	0.43	0.10	0.39	0.21
av_2	0.62	0.80	0.68	0.10	0.36	0.23
av_3	0.29	0.36	0.30	0.45	0.48	0.41
av_4	0.10	0.44	0.26	0.88	0.75	0.68
av_5	0.10	0.38	0.25	0.10	0.38	0.21
av_6	0.31	0.56	0.43	0.50	0.66	0.58
av_7	0.10	0.45	0.29	0.10	0.43	0.27
av_8	0.10	0.38	0.26	0.10	0.37	0.25
av_9	0.12	0.36	0.30	0.10	0.37	0.25
av_{10}	0.18	0.45	0.29	0.14	0.42	0.33
av_{11}	0.78	0.85	0.70	0.14	0.40	0.30
av_{12}	0.38	0.52	0.33	0.16	0.34	0.25
av_{13}	0.10	0.39	0.16	0.18	0.48	0.29
av_{14}	0.42	0.63	0.40	0.20	0.42	0.33
av_{15}	0.10	0.28	0.17	0.10	0.29	0.16
av_{16}	0.31	0.47	0.39	0.31	0.59	0.51
av_{17}	0.10	0.51	0.24	0.10	0.49	0.32
av_{18}	0.18	0.43	0.30	0.84	0.86	0.75
av_{19}	0.10	0.49	0.32	0.32	0.57	0.46
av_{20}	0.22	0.40	0.30	0.36	0.71	0.55
av_{21}	0.10	0.53	0.37	0.21	0.50	0.37
av_{22}	0.10	0.50	0.23	0.29	0.58	0.44
Corr	1.00	0.81	0.86	1.00	0.88	0.93

we deal with specialized domains (e.g., tourism), requiring specialized domain ontologies and WordNet is a generic lexical ontology. Secondly, there are concepts in WordNet for which the frequency is not given (e.g., accommodation), or is irrelevant, as in the case of meal (the frequency is 20).

In (Rusu et al., 2014), the importance of measuring semantic similarity between concepts of an ontology is emphasized. In particular, the authors show that their proposal improves the basic distance metric, although the information content approach is not addressed in the experiment.

The method proposed in (Seco et al., 2004) is based on the assumption that the more descendants a concept has the less information it expresses. Concepts that are leaf nodes are the most specific in the taxonomy and their information content is maximal. Analogously to our proposal, in this method the information content is computed by using only the structure of the specialization hierarchy. However, it forces all the leaves to have the same IC, independently of their depth in the hierarchy.

5.2 Ontology and Bayesian Networks

In (Rajput and Haider, 2011) a framework, called BNOSA, has been proposed that uses an ontology to conceptualize a problem domain. In this framework, a BN has been adopted to predict missing values and/or to resolve conflicts of multiple values. In contrast, in our approach the BN has been used to assign weights to concepts of the reference ontology.

(Clark and Radivojac, 2013) applies BN for computing the information content of concepts in a taxonomy and, more in general, of a sub-graph, by summing the information content of each concept in the sub-graph. However, with respect to our work, it does not focus on any specific way for building the conditional probability tables of the BN, and such tables are assumed to be given.

In (Yazid et al., 2014), a similarity measure for the retrieval of medical cases has been proposed. This approach is based on a BN, where the a priori probabilities are given by experts on the basis of cause-effect conditional dependencies. As already mentioned, in our approach the a priori probabilities are not given by experts and rely on the probabilistic-based approach.

In (Jung et al., 2010), an ontology mapping-based search methodology (OntSE) is proposed in order to

Table 8: Precision and Recall about the four request vectors.

	Precision		*Recall*	
	SS-b	*SS-p*	*SS-b*	*SS-p*
rv_1	1.00	0.50	0.67	1.00
rv_2	1.00	1.00	1.00	1.00
rv_3	1.00	0.67	1.00	1.00
rv_4	1.00	0.50	1.00	1.00

evaluate the semantic similarity between user keywords and terms (concepts) stored in the ontology, using a BN. Furthermore, in (Grubisic et al., 2013), the authors emphasize the need of having a non-empirical mathematical method for computing conditional probabilities in order to integrate a BN in an ontology. In particular, in the proposed approach the conditional probabilities depend only on the structure of the domain ontology. However, in the last two mentioned papers, the conditional probability tables for non-root nodes are computed starting from a fixed value, namely 0.9.

In line with (Grubisic et al., 2013), we also provide a non-empirical mathematical method for computing conditional probabilities, but our approach does not depend on a fixed value as initial assumption. In fact, in *SemSim-b* the conditional probabilities are computed on the basis of the weight w_p, which depends only on the structure of the domain ontology, i.e., the probability of the parent node divided by the number of sibling nodes.

6 CONCLUSION

In this paper we presented a new approach to semantic similarity reasoning based on the integration of Bayesian Networks and Weighted Ontologies. Such a solution improves the performance of the *SemSim* method proposed in (Formica et al., 2013). In essence, the proposed approach is based on the construction of a Bayesian Network, isomorphic to a given ontology, referred to as OBN (Onto Bayesian Network). Then, the OBN is used to compute the information content of each concept in the ontology. We have shown that the *SemSim* method achieves better performances by using the weights obtained from the OBN rather than the ones achieved according to the probabilistic-based approach. The *SemSim* method has been conceived assuming that the ontology is organized as a tree-shaped taxonomy. In a future work, we will focus on ontologies organized as DAG, therefore we will extend these results to ISA hierarchies with multiple inheritance.

REFERENCES

Clark, W. T. and Radivojac, P. (2013). Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics*, 29(13):i53–i61.

Dulmage, A. and Mendelsohn, N. (1958). Coverings of bipartite graphs. *Canadian Journal of Mathematics*, 10:517 – 534.

Formica, A., Missikoff, M., Pourabbas, E., and Taglino, F. (2008). *Weighted Ontology for Semantic Search*, pages 1289–1303. Springer Berlin Heidelberg, Berlin, Heidelberg.

Formica, A., Missikoff, M., Pourabbas, E., and Taglino, F. (2010). Semantic search for enterprises competencies management. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development (IC3K 2010)*, pages 183–192.

Formica, A., Missikoff, M., Pourabbas, E., and Taglino, F. (2013). Semantic search for matching user requests with profiled enterprises. *Computers in Industry*, 64(3):191 – 202.

Gao, J.-B., Zhang, B.-W., and Chen, X.-H. (2015). A wordnet-based semantic similarity measurement combining edge-counting and information content theory. *Engineering Applications of Artificial Intelligence*, 39:80 – 88.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220.

Grubisic, A., Stankov, S., and Perai, I. (2013). Ontology based approach to bayesian student model design. *Expert Systems with Applications*, 40(13):5363–5371.

Jung, M., Jun, H.-B., Kim, K.-W., and Suh, H.-W. (2010). Ontology mapping-based search with multidimensional similarity and bayesian network. *The International Journal of Advanced Manufacturing Technology*, 48(1):367–382.

Lin, D. (1998). An information-theoretic definition of similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann.

Pearl, J. and Russell, S. (2001). Bayesian networks. In Arbib, M. A., editor, *Handbook of Brain Theory and Neural Networks*, pages 157–160. MIT Press.

Rajput, Q. and Haider, S. (2011). Bnosa: A bayesian network and ontology based semantic annotation framework. *J. Web Sem.*, 9(2):99–112.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th Int. Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Rusu, D., Fortuna, B., and Mladenic, D. (2014). Measuring concept similarity in ontologies using weighted concept paths. *Applied Ontology*, 9(1):65–95.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Seco, N., Veale, T., and Hayes, J. (2004). An intrinsic information content metric for semantic similarity in WordNet. *Proc. of ECAI*, 4:1089–1090.

Yazid, H., Kalti, K., and Amara, N. E. B. (2014). A new similarity measure based on bayesian network signature correspondence for braint2 tumors cases retrieval. *Int. J. Computational Intelligence Systems*, 7(6):1123–1136.