A Bayesian Approach for Weighted Ontologies and Semantic Search

Anna Formica¹, Michele Missikoff², Elaheh Pourabbas¹ and Francesco Taglino¹

¹National Research Council, Istituto di Analisi dei Sistemi ed Informatica “Antonio Ruberti”,
Via dei Taurini 19 - 00185, Rome, Italy
²National 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. 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
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:

$$\text{Ont} = \langle C, \text{ISA} \rangle$$

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:

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

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:

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

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).
A Bayesian Approach for Weighted Ontologies and Semantic Search

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:

\[
P(x_1, \ldots, x_n) = \prod P(x_i|pa_i)
\]

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 defined for a set of (non-independent) random variables to represent the marginal probability of a single variable w.r.t. the others.

\(^2\)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.

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

\[
\begin{align*}
rv &= \{r_1, \ldots, r_n\}, \\
av &= \{a_1, \ldots, a_m\},
\end{align*}
\]

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

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

\[
WRO = \langle Ont, w \rangle
\]

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 \([0 \ldots 1]\).

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 \( \text{Thing} \) 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 \( c' \):

\[
w_p(c) = \frac{w_p(c')}{\text{number of children of } c'}
\]

For instance, let us consider the concept \( \text{Salon} \). The associated \( w_p \) is 0.05 because \( w_p(\text{Attraction}) = 0.2 \) and \( \text{Attraction} \) 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 \( \text{SPEC} \) be the reflexive and transitive closure of \( \text{ISA} \). Then, the semantics of a feature \( a \in F \) is defined as follows:

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

where \( \text{feat}(\text{b, res}) \) means that the resource \( \text{res} \) is in UDR is annotated by the feature \( \text{b} \in F \). The semantics of an \( OFV \), say:

\[
\text{OFV} = \{a_1, \ldots, a_n\}
\]

is therefore defined according to the \( \Gamma \) function as follows:

\[
\Gamma(\text{OFV}) = \bigcap_i \Gamma(a_i)
\]

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.
Figure 1: The WRO of our running example with the w_p and w_b concept weights.
defined 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 \((w_p(\text{Thing}) = w_p(\text{G}) = 1)\), therefore the weight \( w_p \) coincides with the weight \( w_p \), i.e., \( w_p(G) = w_p(G) \), where \( w_p(G) = P(G = T | \text{Thing} = T) = P(G = T) = 0.2 \), 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 parent 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 of its parent Gastronomy as follows:

\[
P(L = T) = \sum_{v \in \{T,F\}} P(L = T, G = v)
\]

\[
= P(L = T, G = T) + P(L = T, G = F)
\]

\[
= P(L = T)P(G = T) + P(L = T)P(G = F)
\]

\[
= 0.07 \times 0.2 + 0 \times 0.8 = 0.014
\]

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).

### 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 \( \text{semsim} \) 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:

\[
\text{IC} = -\log(w(c))
\]

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

\[
\text{consim}(c_i, c_j) = \frac{2 \times \text{IC}(\text{lab}(c_i), c_j))}{(\text{IC}(c_i) + \text{IC}(c_j))}
\]

where the \( \text{lab} \) represents the least abstract concept of the ontology that subsumes both \( c_i \) and \( c_j \). Given an instance \( RV \) and an instance of \( AV \), say \( rv \) and \( av \) respectively, the \( \text{semsim} \) function computes the \( \text{consim} \) 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>
<thead>
<tr>
<th>VegetarianMeal</th>
<th>T</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00042</td>
<td>0.99958</td>
<td></td>
</tr>
</tbody>
</table>
and \( \text{av} \) that maximizes the sum of the \( \text{consim} \) similarity values (maximum weighted matching problem in bipartite graphs) (Dulmage and Mendelsohn, 1958). In particular, given:

\[
\text{rv} = \{r_1, \ldots, r_n\} \\
\text{av} = \{a_1, \ldots, a_m\}
\]

as defined in Section 2, let \( S \) be the Cartesian Product of \( \text{rv} \) and \( \text{av} \):

\[
S = \text{rv} \times \text{av}
\]

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

\[
\mathcal{P}(\text{rv, av}) = \{P \subseteq S \mid \forall (r_i, a_j), (r_h, a_k) \in P, r_i \neq r_h, a_j \neq a_k, |P| = \min\{n, m\}\}
\]

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

\[
\text{semsim}(\text{rv, av}) = \max_{P \in \mathcal{P}(\text{rv, av})} \frac{\sum_{(r_i, a_j) \in P} \text{consim}(r_i, a_j)}{\max(n, m)}
\]

In (Formica et al., 2013), we defined \( \text{semsim-p} \) 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 \( \text{semsim-b} \) where the weights, \( w_b \), are defined by using a Bayesian Network, as described in Section 3.

### 4.1 Validation

In order to validate \( \text{semsim-b} \), we refer to the experiment proposed in (Formica et al., 2013). In that experiment, we considered four request vectors, namely \( \text{rv}_i, i = 1, \ldots, 4 \), which are recalled below:

\[
\begin{align*}
\text{rv}_1 &= \{\text{Campsite, EthnicMeal, RockConcert, Bus}\} \\
\text{rv}_2 &= \{\text{InternationalHotel, InternationalMeal, ArtGallery, Flight}\} \\
\text{rv}_3 &= \{\text{Pension, MediterraneanMeal, Cinema, ShoppingCenter}\} \\
\text{rv}_4 &= \{\text{CountryResort, LightMeal, ArcheologicalSite, Museum, Train}\}
\end{align*}
\]

and 22 annotated resources, represented by their annotation vectors, namely \( \text{av}_1, \text{av}_2, \ldots, \text{av}_{22} \). Below, only 10 of them are recalled for lack of space:

\[
\begin{align*}
\text{av}_1 &= \{\text{InternationalHotel, FrenchMeal, Cinema, Flight}\} \\
\text{av}_2 &= \{\text{Pension, VegetarianMeal, ArtGallery, ShoppingCenter}\} \\
\text{av}_3 &= \{\text{CountryResort, MediterraneanMeal, Bus}\} \\
\text{av}_4 &= \{\text{CozyAccommodation, VegetarianMeal, Museum, Train}\} \\
\text{av}_5 &= \{\text{InternationalHotel, ThaiMeal, IndianMeal, Concert, Bus}\}
\end{align*}
\]

\[
\begin{align*}
\text{av}_6 &= \{\text{SeasideCottage, LightMeal, ArcheologicalSite, Flight, ShoppingCenter}\} \\
\text{av}_7 &= \{\text{RegularAccommodation, RegularMeal, Salon, Flight}\} \\
\text{av}_8 &= \{\text{InternationalHotel, VegetarianMeal, Ship}\} \\
\text{av}_9 &= \{\text{FarmHouse, MediterraneanMeal, CarRental}\} \\
\text{av}_{10} &= \{\text{RegularAccommodation, EthnicMeal, Museum}\}
\end{align*}
\]

For each request vector, we computed the \( \text{semsim} \) 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 \( \text{SemSim} \) 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 \( \text{semsim-b} \) (SS-b) achieves a higher correlation with HJ with respect to \( \text{semsim-p} \) (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 \( \text{semsim-b} \) is equal to 1 for all the four \( \text{RV} \), and for three of them, namely \( \text{rv}_1, \text{rv}_3, \) and \( \text{rv}_4 \), it is higher with respect to the Precision obtained by applying \( \text{semsim-p} \). We also observe that both \( \text{semsim-b} \) and \( \text{semsim-p} \) achieve the same Recall (equal to 1) for three out of four \( \text{RV} \), while the Recall for the remaining \( \text{RV} \) (i.e. \( \text{rv}_1 \)) by \( \text{semsim-b} \) is lower than the one by \( \text{semsim-p} \).

## 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 $rv_2$.

<table>
<thead>
<tr>
<th>AV</th>
<th>$rv_1$</th>
<th>$rv_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HJ</td>
<td>SS-p</td>
</tr>
<tr>
<td>av1</td>
<td>0.10</td>
<td>0.54</td>
</tr>
<tr>
<td>av2</td>
<td>0.10</td>
<td>0.34</td>
</tr>
<tr>
<td>av3</td>
<td>0.25</td>
<td>0.50</td>
</tr>
<tr>
<td>av4</td>
<td>0.18</td>
<td>0.49</td>
</tr>
<tr>
<td>av5</td>
<td>0.51</td>
<td>0.64</td>
</tr>
<tr>
<td>av6</td>
<td>0.14</td>
<td>0.40</td>
</tr>
<tr>
<td>av7</td>
<td>0.16</td>
<td>0.51</td>
</tr>
<tr>
<td>av8</td>
<td>0.10</td>
<td>0.37</td>
</tr>
<tr>
<td>av9</td>
<td>0.10</td>
<td>0.46</td>
</tr>
<tr>
<td>av10</td>
<td>0.21</td>
<td>0.49</td>
</tr>
<tr>
<td>av11</td>
<td>0.15</td>
<td>0.45</td>
</tr>
<tr>
<td>av12</td>
<td>0.10</td>
<td>0.25</td>
</tr>
<tr>
<td>av13</td>
<td>0.89</td>
<td>0.71</td>
</tr>
<tr>
<td>av14</td>
<td>0.10</td>
<td>0.38</td>
</tr>
<tr>
<td>av15</td>
<td>0.10</td>
<td>0.33</td>
</tr>
<tr>
<td>av16</td>
<td>0.10</td>
<td>0.39</td>
</tr>
<tr>
<td>av17</td>
<td>0.93</td>
<td>0.87</td>
</tr>
<tr>
<td>av18</td>
<td>0.26</td>
<td>0.46</td>
</tr>
<tr>
<td>av19</td>
<td>0.50</td>
<td>0.73</td>
</tr>
<tr>
<td>av20</td>
<td>0.34</td>
<td>0.51</td>
</tr>
<tr>
<td>av21</td>
<td>0.77</td>
<td>0.85</td>
</tr>
<tr>
<td>av22</td>
<td>0.46</td>
<td>0.72</td>
</tr>
<tr>
<td>Corr</td>
<td>1.00</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 7: Correlation about $rv_3$ and $rv_4$.

<table>
<thead>
<tr>
<th>AV</th>
<th>$rv_3$</th>
<th>$rv_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HJ</td>
<td>SS-p</td>
</tr>
<tr>
<td>av1</td>
<td>0.10</td>
<td>0.55</td>
</tr>
<tr>
<td>av2</td>
<td>0.62</td>
<td>0.80</td>
</tr>
<tr>
<td>av3</td>
<td>0.29</td>
<td>0.36</td>
</tr>
<tr>
<td>av4</td>
<td>0.10</td>
<td>0.44</td>
</tr>
<tr>
<td>av5</td>
<td>0.10</td>
<td>0.38</td>
</tr>
<tr>
<td>av6</td>
<td>0.31</td>
<td>0.56</td>
</tr>
<tr>
<td>av7</td>
<td>0.10</td>
<td>0.45</td>
</tr>
<tr>
<td>av8</td>
<td>0.10</td>
<td>0.38</td>
</tr>
<tr>
<td>av9</td>
<td>0.12</td>
<td>0.36</td>
</tr>
<tr>
<td>av10</td>
<td>0.18</td>
<td>0.45</td>
</tr>
<tr>
<td>av11</td>
<td>0.78</td>
<td>0.85</td>
</tr>
<tr>
<td>av12</td>
<td>0.38</td>
<td>0.52</td>
</tr>
<tr>
<td>av13</td>
<td>0.10</td>
<td>0.39</td>
</tr>
<tr>
<td>av14</td>
<td>0.42</td>
<td>0.63</td>
</tr>
<tr>
<td>av15</td>
<td>0.10</td>
<td>0.28</td>
</tr>
<tr>
<td>av16</td>
<td>0.31</td>
<td>0.47</td>
</tr>
<tr>
<td>av17</td>
<td>0.10</td>
<td>0.51</td>
</tr>
<tr>
<td>av18</td>
<td>0.18</td>
<td>0.43</td>
</tr>
<tr>
<td>av19</td>
<td>0.10</td>
<td>0.49</td>
</tr>
<tr>
<td>av20</td>
<td>0.22</td>
<td>0.40</td>
</tr>
<tr>
<td>av21</td>
<td>0.10</td>
<td>0.53</td>
</tr>
<tr>
<td>av22</td>
<td>0.10</td>
<td>0.50</td>
</tr>
<tr>
<td>Corr</td>
<td>1.00</td>
<td>0.81</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS-b</td>
<td>SS-p</td>
</tr>
<tr>
<td>$rv_1$</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td>$rv_2$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$rv_3$</td>
<td>1.00</td>
<td>0.67</td>
</tr>
<tr>
<td>$rv_4$</td>
<td>1.00</td>
<td>0.50</td>
</tr>
</tbody>
</table>

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

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.

In (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
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


