Computers in Industry 64 (2013) 191–202

Contents lists available at SciVerse ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Semantic search for matching user requests with profiled enterprises

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

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

ARTICLE INFO

Article history:
Received 19 September 2012
Accepted 20 September 2012
Available online 11 December 2012

Keywords:
Similarity reasoning
Weighted Reference Ontology
Information content
Digital resources

ABSTRACT

Semantic search is an important approach that promises significant improvements for customers to identify products of their interest. To perform semantic search, enterprises need to publish semantically enriched descriptions of their offered goods and services; then a customer expresses his/her request, in an easy Google like fashion, by providing a list of desired features. If enterprise offerings and customer requests are based on the same vocabulary (i.e., ontology), they can be semantically matched by using advanced semantic methods. In this paper, we propose an ontology-based method aimed at finding the best matches between a user request and the services offered by different enterprises. We assume that in a given business ecosystem (in the paper, as an example, the tourism sector) a group of SMEs agree on the adoption of a reference ontology, used to build the company profiles on the basis of the offered services. Accordingly, a user request, represented by a set of desired features, is expressed in terms of the reference ontology terminology (i.e., concepts). In this paper, we illustrate *SemSim*, a method used to collectively search the SME profiles to identify the services that match at best the user request. *SemSim* is based on the well-known *information content* approach used to evaluate the semantic similarity between concepts. The experimental results show that our proposal performs better than some of the most representative similarity search methods proposed in the literature.

1. Introduction

SMEs represent the backbone of the European industrial system. With the advent of the globalized markets and an increasing international competition, SMEs are pushed more and more to be connected in a business ecosystem and to operate in it as a unique virtual enterprise, depending on the market demands. When networking, SMEs can reach the critical mass required by the expanding markets but, at the same time, they face a problem of heterogeneity and fragmentation that may jeopardize their performance. To cope with these problems, a good solution is the development of a common semantic knowledge base, represented by a shared reference ontology. The first objective of such an ontology is to achieve an homogeneous and agreed understanding of the business domain for the different SMEs participating in the virtual enterprise. At the same time, a reference ontology is essential to build a more transparent market, where customers can easily make informed choices. Since an ontology is based on a formal paradigm (e.g., the OWL[1] representation), it is also a computational knowledge resource, on top of which it is possible to build a number of semantic services. Among these services, similarity reasoning, and semantic search and retrieval are used to identify, for instance, the SMEs offerings that match at best with a customer request.

In the literature, there are many proposals on semantic search and retrieval (see next section), but there is still no specific solution that clearly emerges. For this reason, we present a new proposal based on a notion that is currently gaining momentum in the field: the *information content* approach [14]. Along this line, we enrich the reference ontology by associating each concept with a weight such that, as the weight of a concept increases the informativeness decreases, hence the more abstract a concept the lower its information content.

In this paper, as an application domain, we focus on tourism sector which has been used to develop our running example. A very common concept in this domain is *accommodation*. It has a low information content, and when used to perform a search, a lot of resources (i.e., documents about various accommodation solutions) will be returned. To have a more focused answer, we need to use more specific concepts, i.e., with additional information content, in the request (e.g., *pension* or *campsite*). A search that uses a focused concept, with a higher information content, will get a smaller, more relevant answer set.

The key problem addressed in this paper concerns the optimal satisfaction of a customer (a tourist, in our example) searching for a number of facilities that are offered by different tourism SMEs, in a fragmented way (as opposed to the large tourism resorts capable of

* Corresponding author. Tel.: +39 06 7716461.
E-mail addresses: anna.formica@iasi.cnr.it (A. Formica), michele.missikoff@iasi.cnr.it (M. Missikoff), elaheh.pourabbas@iasi.cnr.it (E. Pourabbas), francesco.taglino@iasi.cnr.it (F. Taglino).

[1] Ontology Web Language (http://www.w3.org/TR/owl-features/).

0166-3615/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.compind.2012.09.007

offering 'one stop shops'). In our approach, the idea is that, let us say, a digital document (e.g., a Web page) in the tourism domain will aggregate accommodations, transportations, attractions, gastronomy, etc. offered by different SMEs in a unique digital vacation package available to the tourist.

We propose to annotate a digital vacation package with a vector of features, referred to as *ontology feature vector*, where each feature corresponds to an ontology concept. A tourist request, typically concerning a number of desirable facilities, will be represented with an ontology-based request feature vector. The addressed problem is to find the vacation package that matches at best the tourist desires. To this end, we propose the *SemSim* method that computes the similarity between the request vector and each different ontology feature vector annotating the vacation packages. In particular, two different approaches will be presented: the probabilistic (*semsim-p*) and frequency (*semsim-f*) approaches. Both approaches start by computing the similarity between pairs of concepts (*consim*), where, intuitively, the similarity of two concepts corresponds to the overlapping degree of the respective information contents. Then, the similarity between two ontology feature vectors is evaluated by using the Hungarian algorithm, a method that is used for solving the maximum weighted matching problem in bipartite graphs, which is isomorphic to our application problem (see details in Section 4).

The rest of the paper is organized as follows. After the next section presenting the related work, Section 3 introduces some basic notions about the probabilistic and frequency based weight assignments. The *SemSim* method is presented in Section 4, and evaluated in Section 5, where our experiment is illustrated. Finally, Section 6 concludes.

2. Related work

In the vast literature available (see for instance [1,5,9,16,17]), we will restrict our focus on the proposals tightly related to our approach. We emphasize that the focus of our work is both on the assignment of weights to the concepts of a reference ontology, and the method to compute the similarity between concept vectors.

In the large majority of papers proposed in the literature [5,17], weight assignment to the concepts of a reference ontology (or a taxonomy) is performed by using WordNet [22], see for instance [13], and also [14,20] which inspired our method. WordNet (a lexical ontology for the English language) provides, for a given concept (noun), the natural language definition, hypernyms, hyponyms, synonyms, etc, and also a measure of the *frequency* of the concept. The latter is obtained by using noun frequencies from the Brown Corpus of American English. Then, the *SemCor* project [7] made a step forward by linking subsections of Brown Corpus to senses in the WordNet lexicon (with a total of 88,312 observed nouns). We did not adopt the WordNet frequencies for two reasons. Firstly, we deal with specialized domains (e.g., systems engineering, tourism, etc.), requiring specialized domain ontologies. WordNet is a generic lexical ontology (i.e., not focused on a specific domain) that contains only simple terms. In fact, multi-word terms are not reported (e.g., terms such as "seaside cottage" or "farm house" are not defined in WordNet). Secondly, there are concepts in WordNet for which the frequency is not given (e.g., "accommodation").

Concerning weight assignment, in [6] a joint use of an ontology and a typical Natural Language Processing method, based on *term frequency* and *inverse document frequency* (*tf–idf*), is presented. In weighting the similarity between terms and elements of the ontology, the authors propose an approach based on five fixed relevance levels corresponding to five constants: *direct(1.0)*, *strong(0.7)*, *normal(0.4)*, *weak(0.2)*, *irrelevant(0.0)*. In our semantics-based approach, the weights and the similarity between concepts may take any value between 0 and 1.

The work presented in [13] shares some analogies with our approach with regard to the need of computing weights without relying on large text corpora. Therefore, they propose a method, referred to as CP/CV, such that each node in the taxonomy is associated with a concept vector, built on the basis of the topology of the ontology and the position of concepts therein. Then, the similarity of concepts is evaluated according to the *cosine* similarity of the related concept vectors. Conversely, in our work the similarity of concepts is conceived to determine the similarity of two concept vectors.

Regarding the methods to compute the similarity between concept vectors, our work proposes a two stages method, firstly computing the pair-wise concept similarity (*consim*), and then deriving the similarity between vectors of concepts (*semsim*, which stands for *semsim-p* or *semsim-f*). As anticipated, pair-wise concept similarity is performed according to the information content approach, originally proposed by Resnik [20] and successively refined by Lin [14]. The Lin's approach shows a higher correlation with human judgment than other methods, such as the *edge-counting* approach [19] and Wu–Palmer [23]. With regard to the second stage, we adopted a solution inspired by the *maximum weighted matching problem* in bipartite graphs. In the literature the *Dice*, *Jaccard* and *Cosine* [15] methods are often adopted in order to compare vectors of concepts. However, in these methods the matchmaking of two concept vectors is based on their intersection, without considering the position of the concepts in the ontology. According to the *Weighted Sum* approach [2] a fixed value (i.e., 0.5) is assigned to each pair of hierarchically related concepts. Our proposal is based on a more refined semantic matchmaking, since the match of two concepts is performed according to their shared information content, and the vector similarity is based on the optimal concept coupling.

In [3] two algorithms for computing the semantic distance/similarity between sets of concepts belonging to the same ontology are introduced. They are based on an extension of the Dijkstra algorithm [4] to search for the shortest path in a graph. With respect to our approach, in the mentioned paper the similarity is based on the distance between concepts rather than the information content of concepts.

3. Probabilistic and frequency based weight assignments

In this section, we first recall some of the basic definitions introduced in [10,11] and, successively, we illustrate the *probabilistic* and the *frequency* based weight assignment approaches.

The *Universe of Digital Resources* (UDR) is the totality of digital resources that are semantically annotated with a reference ontology (an *ontology* is a formal, explicit specification of a shared conceptualization [12]). 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:

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

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

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

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 uniquely defined in C (we assume the hierarchy is a lattice). It represents the least abstract concept of the ontology that subsumes both c_i and c_j.