Semantic-based Taxonomic Categorization of Web Services

Miguel Ángel Corella and Pablo Castells

Universidad Autónoma de Madrid, Escuela Politécnica Superior
Campus de Cantoblanco, 28049 Madrid, Spain
{miguel.corella, pablo.castells}@uam.es

Abstract. With the envisioned proliferation of Web services available on the WWW and private repositories, new and better support techniques are needed for service discovery and organization to stay manageable. Service classification under hierarchic taxonomies is commonly a key feature for properly organizing service repositories in a rational way, as well as a good foundation for sophisticated retrieval techniques. In this paper, a heuristic approach for the semi-automatic classification of (semantic) Web services is proposed, based on matching new unclassified services to previously classified ones in a given corpus. This hypothesis is validated by an experimental test and the comparison with results achieved by other approaches.

1 Introduction

A major envisioned area where the emerging ontology-based technologies [3] are expected to bring key benefits is that of using semantic-based descriptions to endow Web services with a qualitatively higher potential for automation [19]. The basis of this trend is to add semantic information beyond current WSDL-based [6] descriptions, as a means to enable the manipulation of services by software programs easing the automation of such tasks as service selection, invocation, composition, and discovery [11]. Further service management tasks, besides the latter, can also take advantage of the additional semantics to provide with new automation facilities. In this paper we address the categorization (i.e. specification of their domain or business focus) of Web services in a repository with respect to a predefined taxonomy.

Nowadays, the most widely accepted and used protocol for publishing and searching Web services is UDDI [15]. This protocol enables the creation of service repositories, in which services are organized based on standard or proprietary taxonomies (e.g. UNSPSC – United Nations Standard Products and Service Codes, NAICS – North American Industry Classification System, etc.). Each service is classed under one (or more) categories, in order to ease different repository tasks performed by service publishers and consumers, as well as repository administrators.

Since the categorization of services and maintenance of repositories has to be done manually by human entities, the classification task becomes considerably difficult, heavy and error-prone in practice, due to several issues (e.g. huge size of taxonomies in

real-world applications, multiple people involved in maintaining or sharing services in a common repository, several distributed repositories being shared, etc.).

According to this, it is our goal to provide automatic mechanisms to assist service publishers in the categorization task, in order to reduce the effort required, and promote globally consistent classification decisions, even when several users are involved. To do so, we propose a heuristic-based classification system that computes a ranked list of candidate classes from a given taxonomy, in which a new service better fits, by comparing the new service with the services that are already classified and published in a given repository. Besides this classification aid, the categorization information thus obtained can be used for the enhancement of further tasks, such as the ones related to semi-automatic service annotation and discovery [12].

The paper is organized as follows: Section 2 introduces some related work already presented in the domain of Web service classification. The definition of the problem of Web service classification in a formal way, and the motivation of why service semantics are needed in order to successfully solve it are given in Section 3. Our classification heuristic is presented in Section 4. Finally, Section 5 provides conclusions and outlines future work directions.

2 Related work

The problem of the automatic categorization of Web services has been addressed in prior work from two main approaches, that we may class as heuristic (e.g. [16]) and non-heuristic (e.g. [5] and [10]). In the proposal presented in [10] the authors offer two different strategies, the first one consisting of using the information contained in WSDL-based descriptions to select a category in which the service fits best. The second one aims at dynamically creating the classification taxonomy by using WSDL descriptions as input. Their procedure consists of the following steps: a) extraction of meaningful words from service descriptions, using Natural Language Processing techniques, b) construction of term vectors with those words, and c) application of different classification techniques for vector categorization, namely machine learning in their first strategy, and clustering techniques in the second. So, service classification problem is solved by a text classification problem approach.

The approach to classification in [5] proposes similar steps as the ones described in [10]. The main difference is the usage of Support Vector Machines as term vector classification mechanism. Again, service classification is reduced to a text classification problem. In addition, this proposal provides service publishers with extra information after the classification. More precisely, a concept lattice, extracted using Formal Concept Analysis over the term vectors, is presented to the publishers. This lattice allows developers to know how the words used in service descriptions (the ones extracted during classificaiton) contribute to the selection of a specific category. With this information, service developers could, for example, modify some description words which may cause ambiguity in the classification process.

The authors of [16] propose a framework for the automation of the semantic annotation of services. Using domain ontologies (i.e. the ones used for the annotation) as service categories, service classification is used to select the most suitable ontology in

the taxonomy of domain ontologies. To do so, an algorithm to match Web service data types (in XML Schema) and concepts is defined, based on schema matching.

Another research area related to the work presented here is that of service matchmaking (see e.g. [13] and [17]). It is related to Web service classification in that our approach computes similarity degrees between services in order to assign them to a common category, and service matchmaking aims to find services that match a concrete capability description. The main difference between both research areas is that while service classification admits some degree of fuzziness in service matching, i.e. continuous similarity measures, service matchmaking typically does not; using discrete matching levels (e.g. "no match", "complete match", "partial match", etc.).

3 Semantics for service classification

As mentioned earlier, service classification is a common necessity to make service administration and retrieval manageable for human users. Moreover, it can serve as a complementary aid for automatic service discovery and selection techniques. Nevertheless, there are usability problems involved in service categorization which cause difficulties for this categorization in real-world environments. Such problems include:

- Classification taxonomies can be extremely large, comprising thousands of categories (e.g. UNSPSC ~ 20,000 classes, NAICS ~ 2,300 classes).
- The number of services in a repository can grow quite large, making it impossible for administrators to validate the information published along with a service.
- The placement of a service under a proper category requires a considerable amount of knowledge of the complete taxonomy in order to make appropriate decisions.

The work presented here aims at alleviating the administrator's work, and reducing the categorization effort for service providers. Our proposal approaches the classification problem as follows. Given a set of services already classified under a given taxonomy, and a new service description to be published, the unclassified service is compared with the classified ones, whereby a measure of the likelihood that the service should be assigned a certain category is computed.

WSDL descriptions provided by current technologies are not suitable for this purpose, as they only focus on the syntactic view of the services, which is not sufficient to support valid service classification criteria in practice. As stated in the specification of WSDL standard, descriptions in this language provide details about the operations a service supports, and the input / output information involved in their invocation. The fact is that, although this information enables comparisons between services, a much accurate matching is possible if semantic information is added to the descriptions. Consider this example: take two Web services, the first one defining currency conversion capabilities, and the second one, a service to compute distances between cities. The currency converter service could have one operation, involving:

- An input message with two currency codes, of type string.
- An output message with one part containing the conversion rate (a double).

On the other hand, the distance calculator would have also one operation, having:

- An input message with one part containing two city names of string type.
- An output message with one part containing the distance between them (a double).

From a conceptual point of view, these services should yield a low similarity measure value when compared. However, since their interfaces are syntactically equivalent (same number of operations and messages, same data types), comparing their WSDL descriptions would produce a very high result value. So, extra information beyond mere syntactic WSDL definitions is needed. There are at least two possibilities:

- Using identifiers of the different WSDL elements. The hypothesis here is that the names given to operations, parameters, messages, etc., are often meaningful, which, from our point of view, is too idealistic and often fails.
- Using semantic Web service descriptions. In the proposed example, it is clear that the relevant information about service parameters is their semantic meaning (e.g. currency codes, city names, etc.). WSDL descriptions do not foresee this kind of semantics, but ontology-based service descriptions do. Thus, supporting discrimination between syntactically equivalent parameters having different semantics.

The need for service semantics is thus clear. The syntactic information in WSDL-based descriptions is not sufficient, and would often lead to inconsistent similarity values, and therefore, to service misclassification. Semantic Web service descriptions can solve this problem by providing means to describe service inputs and outputs from a conceptual point of view. The approach here presented is compatible with every semantic description languages such as WSMO [18], OWL-S [14], WSDL-S [1] or SWSO [2].

4 Heuristic classification

The heuristic is divided into three granularity levels, corresponding to the comparison between different service elements involved in the categorization procedure.

Service category level. Since services have to be assigned a category as a result of the classification procedure, this level is needed in order to find evidence that a service should belong to a specific category (used to sort the ranked category list).

The proposed measure works as follows. Let $\mathcal{S}$ be the set of Web services in a repository, and let $\mathcal{C}$ be the classification taxonomy. If we allow a service to be classified under several categories of the taxonomy, we may define the classification by $\mathcal{C}$ as a mapping $\tau : \mathcal{S} \rightarrow 2^{\mathcal{C}}$. Given a new service s to be added to $\mathcal{S}$, we want to find the categories in $\mathcal{C}$ that best suit s. Given $c \in \mathcal{C}$, let $\mathrm{P}(s:c)$ be the probability that c is an appropriate classification for s, estimated here by comparison of s with the services classified under c. If we take $\mathrm{P}(s:c) \sim 0$ if $\{x \in \mathcal{S} \mid c \in \tau(x)\} = \emptyset$, we can write:

$$\mathrm{P}(s:c) \sim \mathrm{P}\left(s:c \wedge \left(\bigvee_{x \in \mathcal{S}} c \in \tau(x)\right)\right)$$

By rewriting the right hand-side using the inclusion-exclusion principle applied to probability [20], it can be seen that:

$$\mathrm{P}(s:c) \sim \sum_{A \subset \mathcal{S}} (-1)^{|A|+1} \prod_{x \in A} \mathrm{P}\left(c \in \tau(x)\right) \cdot \mathrm{P}\left(s:c \mid c \in \tau(x)\right)$$

provided that $s{:}c \wedge c \in \tau(x)$ are pairwise independent for all $x \in \mathcal{S}$. Since $c \in \tau(x)$ is true iff $x \in \{x \in \mathcal{S} \mid c \in \tau(x)\}$, and assuming a crisp service classification (i.e. $c \in \tau(x)$ is either true or false, as opposed to fuzzy classification where $P(c \in \tau(x)) \in [0,1]$), we have:

$$P(s:c) \sim \sum_{A \subset \tau^{-1}(c)} (-1)^{|A|+1} \prod_{x \in A} P(s:c \mid c \in \tau(x))$$

Now we shall estimate $P(s{:}c \mid c \in \tau(x))$ by a measure of similarity sim (s, x), that is:

$$P(s:c) \sim \sum_{A \subset \tau^{-1}(c)} (-1)^{|A|+1} \prod_{x \in A} \mathrm{sim}(s,x)$$

whereby the appropriateness of a category for a service is computed in terms of the similarity between the service and the services classified under that category.

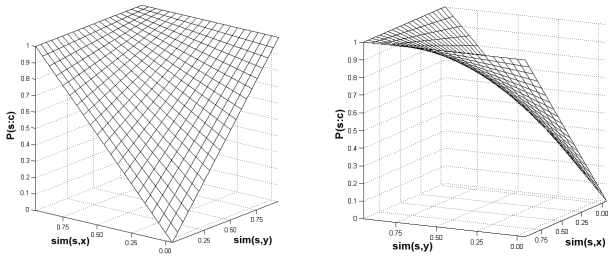

Fig 1. Graph (displayed from two angles) showing the behavior of the measure with respect to sim(s,x) and sim(s,y) (i.e. similarity between a service s and a category c having two services).

Note that $P(s{:}c) \in [0,1]$, provided that sim $(s, x) \in [0,1]$, and increases monotonically with respect to sim (s, x). Figure 1 shows how $P(s{:}c)$ behaves with respect to the sim (s, x) value.

Service description level. The comparison between services is based on the assumption that services of the same category deal with similar semantic concepts. Therefore, operation structures (i.e. conceptual roles and grouping of the service parameters) are relevant for service-level comparisons. In fact, the similarity between services is measured in terms of the similarity between operation sets, as follows:

$$\mathrm{sim}(s,s') = \frac{\sum_{i=1}^{\min(|OP|,|OP'|)} \mathrm{sim}(\mathrm{top}(OP_i, OP'_i))}{\max(|OP|,|OP'|)} \left\{ \mathrm{sim}(op,op') = \mathrm{sim}(I_{op}, I_{op'}) \cdot \mathrm{sim}(O_{op}, O_{op'}) \right\} \left\{ \mathrm{sim}(P,P') = \frac{\sum_{i=1}^{\min(|P|,|P'|)} \mathrm{sim}(\mathrm{top}(P_i, P'_i))}{\max(|P|,|P'|)} \right.$$

where OP and OP' are the operation sets of services s and s'; I_{op}, $I_{op'}$, O_{op}, $O_{op'}$, are the sets of inputs and outputs of two operations op and op'; and P and P' are the set of parameters used as inputs/outputs in the services.

Note that the service comparison defined here returns values in the range [0,1], provided that the similarity between ontology concepts is also within that range. Figure 2 shows how the sim(s, x) behaves with respect to sim(I_{op}, $I_{op'}$) and sim(O_{op}, $O_{op'}$).

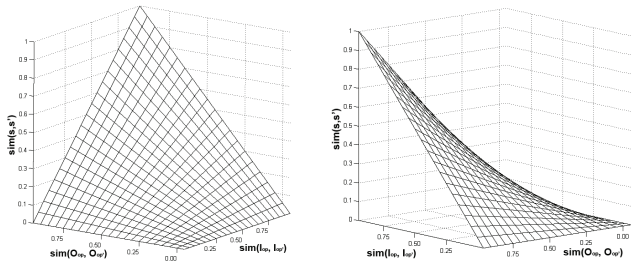

Fig. 2. Service to service similarity measure graph (displayed from two angles) showing the behavior of the similarity measure with respect to sim(I_{op},$I_{op'}$) and sim(O_{op},$O_{op'}$) (i.e. similarity between two services each one with only one operation).

Service parameter level. This last level aims to provide with a similarity value between annotated service parameters, this is, the ontology concepts used to annotate them, i.e. a measure enabling the comparison of domain ontology concepts. A lot of previous work has been developed in this area (e.g. [4], [8], [9]). Nevertheless, a new measure has been developed fitting our specific objectives.

This new measure works as follows. Let $\mathcal{T}$ denote the set of all concepts in the domain ontology. The similarity between two concepts is measured in terms of their distance in the ontology class hierarchy. Given two concepts $t,t' \in \mathcal{T}$, let t_0 be the lower common ancestor to t and t' in $\mathcal{T}$, and let $d = \text{dist}(t,t_0) + 1$, $d' = \text{dist}(t',t_0) + 1$ be the number of levels (plus 1) between t, t' and t_0 in the concept hierarchy. We define the similarity between t and t' as:

$$\text{sim}(t,t') = \left(1 - \frac{\alpha}{\text{h}(\mathcal{T})} \cdot \frac{|d-d'|}{d+d'}\right) \cdot \frac{1}{\min(d,d')} \cdot \left(1 - \frac{\max(d,d')-1}{\text{h}(\mathcal{T})}\right)$$

where:

- h($\mathcal{T}$) is the total height of the concept hierarchy, which is introduced to measure the distance between concepts as a proportion of the total depth of the ontology.
- The term $\frac{|d-d'|}{d+d'}$ increases (that is, the similarity decreases) with the difference in the depth level between t and t'.
- $\alpha \in [0,1]$ (in our test α has been empirically tuned to 0.8) is a parameter that ensures a minimum non-zero similarity value in a way that similarity ranges in some interval [*min*,1] above 0, relaxing the influence of this level in the whole heuristic.
- The factor $1 - \frac{\max(d,d')-1}{\text{h}(\mathcal{T})}$ reinforces the decrease of the similarity when one concept is super concept of the other (i.e. one of the concepts is the first common parent), to achieve a monotonic decrease of the similarity from 1 to 0.

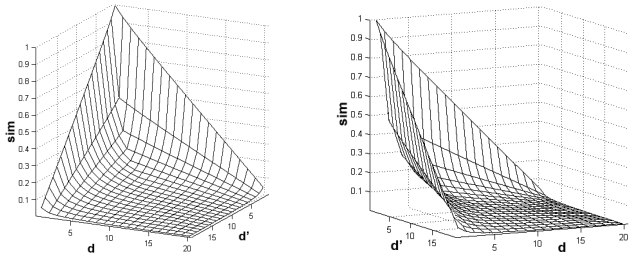

Fig. 3. Concept to concept similarity measure graph (displayed from two angles) showing the behaviour of the similarity measure with respect to *d* and *d*' in a twenty depth levels ontology.

Figure 3 shows how the similarity function sim(*t*,*t*') depends on the distance *d* and *d*' of the *t* and *t*' to their lowest common ancestor.

5 Conclusions and future work

We have presented a new classification approach based on the usage of conceptual service descriptions that can be used to assist service publishers, consumers and repository administrators, in the manual service categorization and retrieval tasks. Moreover, by proposing the classification of new services based on previous decisions made for similar services, consistency is implicitly enhanced against a potential drift over time or across multiple users. All the approach has been based on the hypothesis that the augmentation of service descriptions with semantic information enables a more precise comparison of service descriptions, and therefore an improved accuracy of the automatic classification. As part of the ongoing work we are working on a complete classification framework (implementing the heuristic presented) in order to have a platform in which formally and easily test our approach.

As a continuation of the work presented here we are investigating the potential of our classification approach to enhance service retrieval mechanisms. This line of research is already achieving promising results as reported in [12]. In addition, we envisage the extension of our algorithm to deal with complex descriptions of concepts using axiomatic descriptions of preconditions and post conditions (as supported by WSML) generalizing our matching functions in order to benefit from reasoning tools.

6 Acknowledgements

This research was supported by the Spanish Ministry of Industry, Tourism and Commerce (CDTI05-0436) and the Ministry of Science and Education (TIN2005-0685). Thanks are due to Rubén Lara for all his help and feedback on the research presented.

References

1. Akkiraju, R., Farrel, J., Miller, J., Nagarajan, M., Schmidt, M., Sheth, A., Verma, K: Web Service Semantics – WSDL-S, Technical Note, Version 1.0, 2005.
2. Battle, S., Bernstein, A., Booley, H., Grosof, B., Gruninger, M., Hull, R., Kifer, M., Martin, D., McIlraith, S., McGuiness, D., Su, J., Tabet, S.: Semantic Web Service Ontology (SWSO), Version 1.0, 2005.
3. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American, 2001.
4. Bernstein, A., Kaufmann, E., Bürki, C., Klein, M.: How similar is it? Towards personalized similarity measures in ontologies. In the 7th Internationale Tagung Wirtschaftsinformaitk. Bamberg, Germany, 2005, pp. 1347-1366.
5. Bruno, M., Canfora, G., Di Penta, M., Scognamiglio, R.: An approach to support web service classification and annotation. In Proceedings of the IEEE International Conference on e-Technology, e-Commmerce and e-Services (EEE 2005), Hong Kong 2005.
6. Christiensen, E. et al: Web Service Description Language (WSDL), v1.1.
7. Corella, M. A., Castells, P.: A Heuristic Approach to Semantic Web Services Classification. In Proceedings of the 10th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES 2006), Bournemouth, UK, 2006.
8. Culmore, R., Rossi, G., Merelli, E.: An ontology similarity algorithm for BioAgent. In NETTAB 02 Agents in Bioinformatics. Bologna, Italy, 2002.
9. Ehrig, M., Haase, P., Stojanovic, N.: Similarity for ontologies – a comprehensive framework. Workshop on Enterprise Modelling and Ontology at PAKM 2004. Austria, 2004.
10. Heβ, A., Kushmerick, N.: Automatically attaching semantic metadata to Web Services. In Workshop on Information Integration on the Web (IIWeb2003), Acapulco, Mexico, 2003.
11. Keller, U., Lara, R., Lausen, H., Polleres, A., Fensel, D.: Automatic Location of Services. In 2nd European Semantic Web Conference (ESWC 2005). LNCS Vol. 3532 pp. 1-16.
12. Lara, R., Corella, M.A., Castells.: A flexible model for the discovery of Web services. 1st International Workshop on Semantic Matchmaking and Resource Retrieval: Issues and Perspectives (SMR 2006), co-located with VLDB 2006. Seoul, Korea, 2006.
13. Li, L., Horrocks, I.: A software framework for matchmaking based on semantic web technology. In the International Journal of Electronic Commerce, 8(4):39 – 60. 2004.
14. Maritn, D., Burnstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., Narayanan, S., Paolucci, M., Parsia, B. et al: OWL-S: Semantic markup for web services, v1.1, 2004.
15. OASIS: UDDI: The UDDI technical white paper, 2004.
16. Oldham, N., Thomas, C., Sheth, A., Verma, K.: METEOR-S Web Service Annotation Framework with Machine Learning Classification. In Proc. of the 1st Int. Workshop on Semantic Web Services and Web Process Composition (SWSWPC'04), California, July 2004.
17. Paolucci, M., Kawamura, T., Payne, T., Sycara, K.: Semantic Matching of Web Service Capabilities. In Proceedings of the First International Semantic Web Conference, 2002.
18. Roman, D., Lausen, H., Keller, U., de Brujin, J., Bussler, C., Domingue, J., Fensel, D., Hepp, M., Kifer, M., König-Ries, B., Kopecky, J., Lara, R., Oren, E., Polleres, A., Scicluna, J., Stollberg, M.: Web Service Modeling Ontology (WSMO), 2005.
19. Terziyan, V. Y., Kononenko, O.: Semantic web enabled web services: State-of-the-art and industrial challenges. In Proc. International Conference on Web Services (ICWS), 2003.
20. Whitworth, W. A.: Choice and Chance, with one thousand exercises. Hafner Pub. Co. New York, 1965.