A semantic approach for the requirement-driven discovery of web services in the Life Sciences

María Pérez, Rafael Berlanga, Ismael Sanz

Universitat Jaume I, Spain
{mcatalan, berlanga, isanz}@uji.es

Abstract. Research in the Life Sciences depends on the integration of large, distributed and heterogeneous data sources and web services. The discovery of which of these resources are the most appropriate to solve a given task is a complex research question, since there is a large amount of plausible candidates and there is little, mostly unstructured, metadata to be able to decide among them. We contribute a semi-automatic approach, based on semantic techniques, to assist researchers in the discovery of the most appropriate web services to fulfill a set of given requirements.

1 Introduction

Contemporary research in the Life Sciences depends on the sophisticated integration of large amounts of data obtained by in-house experiments with reference databases available on the web. This is followed by analysis workflows that rely on highly specific algorithms, often available as web services. The amount of data produced and consumed by this process is prodigious; however, the sheer amount of available resources is a source of severe difficulties.

Within this huge set of resources, one of the main problems to the user is finding the right web services for a given research task. The landscape of Life Sciences web services is large and complex: there are thousands of resources [1], most of them available in public repositories, i.e. BioCatalogue [3], but unfortunately only a few are described by adequate metadata, which is essentially textual in nature and this makes the discovery and the integration difficult. In addition, there are many versions of different services that apparently provide the same broad functionality, but not enough metainformation is available to decide which of these services is the most appropriate for a precise task.

Given this context, it is a pressing question how to help researchers to discover the best possible mapping between their requirements and the available tools. We present a semi-automatic approach to assist the researcher in web service discovery, looking for web services that are appropriate to fulfill the information requirements in the Life Sciences domain. The whole process is driven by well-captured requirements, in order to avoid the high costs associated with non-disciplined, non-reusable, ad-hoc development of integration applications. The matching between the requirements and web services is based on a semantic normalization of both the requirements and the web services metadata.

2 Approach overview

The overall approach we propose to assist the selection of web services based on users requirements consists of three main phases: (i) *Requirements elicitation and specification*, (ii) *Normalization*, and (iii) *Web service selection* phase. Here, we focus mainly on the *normalization* and the *web service selection* phases.

1. **Requirements elicitation and specification**. The user's information requirements are the information that drives the discovery of web services. Therefore, the requirements are gathered and formally described in the *requirements model* using the *i** formalism [6]. More details about this phase can be found on [4]. Figure 1 shows a fragment of the requirements model in which a subgoal and the tasks to achieve this subgoal are specified.
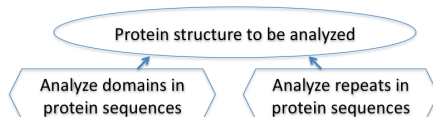

Fig. 1. A fragment of the requirements model.

2. **Normalization**. In the requirements model, task descriptions are expressed in natural language and, therefore, they must be normalized in order to be automatically processed. The normalization consists of a semantic annotation process in which the description of the tasks are processed and annotated with concepts from a reference ontology to allow the reconciliation of the user's requirements with the web services. It is carried out in two phases:

   2.1. **Domain specific annotations**. The purpose of this step is to identify the terms of the user-defined task related to the Life Sciences domain. We have selected a semantic annotator which is capable of using several ontologies to annotate biomedical texts [2]. In our case, it annotates the task description with concepts from the UMLS and ^{my}Grid ontologies. To determine the similarity between a text and a concept this annotator uses the following information-theoretic function:

$$sim(C,T) = max_{S \in lex(C)}(ratio(S,T))$$

$$ratio(S,T) = \frac{idf(cw(S,T)) - missing(S,T)}{idf(S)}$$

$$missing(S,T) = (idf(S) - idf(cw(S,T)))$$

The function $ratio(S,T)$ defines the ratio between the achieved information evidence for the text T, in our case the task description, and the information encoded in the lexicon form S. The function $missing(S,T)$

is the amount of information contained in S that have not been covered with T. $idf(S)$ measures the relevance of the terms in the string S, and $cw(S,T)$ is the set of terms in common between the concept string S and the text T. The function idf is defined as follows:

$$idf(S) = -\sum_{w \in S} log(P(w|UMLS))$$

The relevance of word is measured by means of its estimated probability within the whole UMLS lexicon (i.e. $P(w|UMLS)$). After applying the semantic annotator, each task is represented with a semantic vector with the $tf * idf$ values of each concept. For example, the task "Analyze domains in protein sequences" is represented with the semantic vector $\{'C1513868':8, 'D9000419':15\}$ where C1513868 is retrieved by "domains" and D9000419 is retrieved by "protein sequences".

2.2. **Application specific annotations**. The next step is to determine the functionality the task is describing. For that purpose, we use the taxonomy of categories defined by BioCatalogue to classify the user-defined tasks. The matching between the tasks and the categories is made by using the ISub metric [5] that performs a string matching algorithm. For example, the task "Analyze domains in protein sequences" is annotated with the category "Protein Sequence Analysis" with a score of 0.5586.

3. **Web services selection**. The discovery of the suitable web services for the user's requirements is based on the matching between the annotations of the tasks and the metadata of the web services. However, most web service registries suffer from the lack of metadata. To address this problem, we have automatically annotated the description (or the documentation in case there is no description available), the categories and the tags of the 1729 web services registered in BioCatalogue. The annotations of each service are stored as a vector that contains the $tf * idf$ of each concept.

The discovery process is made by two independent searches: (i) search the web services that are annotated with the same category as the task, and (ii) search the web services that have concepts in common with the user-defined task. The results of both searches are combined and scored based on the following linear combination:

$$score = C\text{-}score * w_1 + S\text{-}score * w_2$$

where w_1 and w_2 are weights that depend on the relevance of each search and fulfill the condition $w_1 + w_2 = 1$, $C\text{-}score$ is the score calculated by the ISub metric, that is, the score of the string matching between the categories of the taxonomy and the user-defined task, and $S\text{-}score$ is the cosine similarity between the vector of the task and the vector of the service.

Table 1 shows the highest ranked services for the task "Analyze domains in protein sequences".

At the end, the user obtains a set of ranked lists of web services that are supposed to provide the functionality required by the tasks. In case the results are not those expected by the user, she can refine the process at the three phases of the guide.

Service	Shared annotations	C-score	S-score	Score
GlobPlot	C1513868, D9000419	0	0.6934	0.5547
Uniprot	D9000419	0.5586	0.5427	0.5459
GenesilicoProteinSilicoSOAP	C1513868, D9000419	0.5586	0.4725	0.4897
Emboss tmap	D9000419	0.5586	0.4443	0.4671
ELMdb	D9000419	0.5586	0.4379	0.4621

Table 1. Ranked list of services for the task *Analyze domains in protein sequences.*

3 Conclusions and Future work

We have presented a semi-automatic approach that guides researchers in the Life Sciences to the discovery of web services that respond to their informational requirements. Due to the importance of the semantic normalization, we have annotated the available information of the services registered in BioCatalogue by using a biomedical semantic annotator. In turn, the user requirements are also annotated with the same ontologies, thus allowing the application of a semantic search technique to find mappings between requirements and services. The result of the process is that the user is provided with a set of ranked lists of web services that are appropriate for her stated information needs.

Some direct follow-ups of this work are the refinement of some details of the semantic techniques, the exploitation of other sources of metadata (bibliographical information, referenced web pages), and the creation of a GUI to facilitate its application.

Acknowledgements

References

1. *Nucleic Acids Research*, volume 38 suppl 2 (Web Server Issue), July 2010.
2. Rafael Berlanga, Victoria Nebot, and Ernesto Jimenez. Semantic annotation of biomedical texts through concept retrieval. In *BioSEPLN 2010*, 2010.
3. J. Bhagat, F. Tanoh, E. Nzuobontane, T. Laurent, J. Orlowski, M. Roos, K. Wolstencroft, S. Aleksejevs, R. Stevens, S. Pettifer, R. Lopez, and C.A. Globe. BioCatalogue: a universal catalogue of web services for the life sciences. *NAR*, 2010.
4. María Pérez, Sven Casteleyn, Ismael Sanz, and María José Aramburu. Requirements gathering in a model-based approach for the design of multi-similarity systems. In *MoSE+DQS '09*. ACM, 2009.
5. Giorgos Stoilos, Giorgos B. Stamou, and Stefanos D. Kollias. A string metric for ontology alignment. In *The Semantic Web - ISWC 2005*, pages 624–637, 2005.
6. Eric Yu. *Modelling Strategic Relationships for Process Reenginering*. PhD thesis, University of Toronto, Canada, 1995.