AI Communications 29 (2016) 231–232
DOI 10.3233/AIC-140649
IOS Press

Thesis

Semantic-based approach for the discovery of Life Sciences web resources driven by rich user's requirements

María Pérez Catalán
Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I, Castellón de la Plana, Spain
E-mail: mcatalan@uji.es

Abstract. In recent years, the number of resources available on the Web has increased quickly. In Life Sciences, researchers are publishing their research results on the Web as web resources with the aim of collaborating with other researchers. However, the discovery of web resources relevant to a specific requirement is a challenging task. The thesis aims to assist the users in the discovery of the most suitable resources for their specific requirements.

Keywords: Web resource discovery, semantics, information retrieval, knowledge extraction

1. Introduction

Web resources have been gaining popularity as providers of relevant data, whether those stored in datasets or those resulting from the execution of complex functions. Although the discovery of web resources has been largely studied, it is still a challenging research task due to the high dependency current search engines have on the characteristics of the available metadata. In some domains like Life Sciences, this dependency becomes even worse due to the heterogeneity of data, and the lack of adequate metadata [1].

Current web resource registries allow users to search for resources that fulfill their information needs. The discovery in these registries is mainly based on the use of well-defined metadata, which is usually limited and very specific, and on the string matching of the user's query keywords, which is hampered by the heterogeneity of data.

The main objective of this thesis is to assist the users in the discovery of the most appropriate resources for their information needs, specifically in the Life Sciences domain.

2. Contributions of the thesis

This thesis [2] has reviewed the discovery of web resources in the Life Science domain. Currently, there are web resource registries on the Web that allow users to discover web resources that are supposed to be relevant to their requirements. However, most of them present some limitations that hinder the web resource discovery: (i) poor representation of user's requirements, (ii) high discovery dependency on the characteristics of the resources metadata, and (iii) low assistance to the user during the whole discovery process, specifically in the selection of the most appropriate resource.

The goal of this thesis is to assist the user in the discovery of the resources that are the most appropriate for her requirements, by addressing the main limitations of current registries. The main characteristic of the proposed approach is that the whole discovery process is driven by the user, who is assumed to provide a rich specification of her requirements, and who can modify the discovery parameters in order to customize the process.

In the proposed approach, the specification of the user's requirements is a rich description of what the user needs, which includes not only the functionality, but also relevant features of the required resource, such as the input/output parameters or the species involved by the resource. With the aim of being widely adopted by users, our approach is not restricted to a specific requirement specification technique. Currently, the implemented prototype, BioUSeR [3], supports textual descriptions and i^* models as requirements specification techniques.

One of the most important limitations of current registries is their high dependency on the characteristics of metadata, both structural and lexical. With respect to the structural dependency, many registries define specific fields to describe specific features of the resources. However, evidence shows that most of the resources features are implicitly described in the textual descriptions and, consequently, they are not identified as feature values by the search engines. On the other hand, regarding the lexical dependency, the lack of widely accepted standards increases the heterogeneity of data describing the resources and, therefore, users have to know which vocabulary has been used in the metadata in order to specify their requirements with the same vocabulary. The discovery process proposed in this thesis alleviates this dependency by using normalization techniques. First, to address the heterogeneity and ambiguity of data, all data involved in the discovery process are automatically semantically annotated with domain knowledge resources. Afterwards, knowledge extraction techniques are used to automatically identify relevant information about the resources features, which improve their characterization. Then, the discovery is based on the semantic mapping between the normalized requirements specification and the normalized resources metadata. The semantic mapping allows to retrieve resources that are described with different formats and vocabularies. Therefore, we can conclude that the dependency on the characteristics of the metadata is considerably reduced by the use of normalized data.

With the aim of assisting the user until the end of the discovery process, the discovered resources are ranked according to their relevance to the user's requirement. The relevance of a resource is estimated considering how well the resource fulfils not only the functionality, but also the features required by the user. At the end, the user gets a ranked list in which the most appropriate resources for her requirements are in the top positions. Finally, if the resources are not those expected by the user, she can modify the discovery process by modifying the requirements specification, the information about the facets automatically identified, and other discovery parameters.

Therefore, we can conclude that the main limitations of current registries in Life Sciences have been alleviated in our approach by: (i) allowing the user to provide a rich specification of her information needs and to modify discovery parameters and information that have been automatically identified (e.g., facets values), (ii) using normalization techniques in order to alleviate the dependency on the data characteristics, and (iii) providing relevant information to the user, such as the automatically extracted facets values, the semantic annotation of her requirements specification, and the ranking of resources.

The discovery approach has been validated by evaluating each one of its phases. Moreover, we have further validated it by comparing it with other IR techniques, and with one of the most popular web resource registries in Life Sciences, BioCatalogue. This later experiment [4] has demonstrated that our approach obtains more precise results with less iterations and fewer effort than current registries.

Finally, the proposed discovery process has been implemented as part of a prototype called BioUSeR [3]. BioUSeR visualizes each phase of the discovery process, and allows the user to modify some parameters during the whole process. Its simplicity and the visualization of relevant information make the discovery of relevant web resources easier and less error-prone.

Acknowledgements

The author is grateful to Dr. Rafael Berlanga and Dr. Ismael Sanz for supervising her doctoral dissertation.

References

[1] C. Goble, R. Stevens, D. Hull, K. Wolstencroft and R. Lopez, Data curation + process curation = data integration + science, *Briefings in Bioinformatics* **9**(6) (2008), 506–517.

[2] M. Pérez, Semantic-based approach for the discovery of Life Sciences web resources driven by rich user's requirements, PhD thesis at Universitat Jaume I, Spain, directed by Dr. Rafael Berlanga and Dr. Ismael Sanz.

[3] M. Pérez, R. Berlanga, I. Sanz and M.J. Aramburu, A semantic approach for the requirement-driven discovery of web resources in the Life Sciences, *Knowledge and Information Systems* **34** (2013), 671–690.

[4] M. Pérez, R. Berlanga, I. Sanz and M.J. Aramburu, BioUSeR: A semantic-based tool for retrieving Life Sciences resources driven by text-rich user requirements, *Journal of Biomedical Semantics* **4** (2013), 12.