Under consideration for publication in Knowledge and Information Systems

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

María Pérez-Catalán[1], Rafael Berlanga[2], Ismael Sanz[1] and María José Aramburu[1]

[1]Department of Computer Science and Engineering
[2]Department of Computer Languages and Systems
Universitat Jaume I, Castelló de la Plana, Spain
{mcatalan,berlanga,isanz,aramburu}@uji.es

Abstract. Research in the Life Sciences depends on the integration of large, distributed and heterogeneous web resources (e.g. 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 are many candidate resources and there is little, mostly unstructured, metadata to be able to decide among them.

In this paper we contribute with a semi-automatic approach, based on semantic techniques, to assist researchers in the discovery of the most appropriate web resources to fulfill a set of requirements. The main feature of our approach is that it exploits broad knowledge resources in order to annotate the unstructured texts that are available in the emerging web-based repositories of web resource metadata.

The results show that the web resource discovery process benefits from a semantic-based approach in several important aspects. One of the advantages is that the user can express her requirements in natural language avoiding the use of specific vocabularies or query languages. Moreover, the discovery exploits not only the categories or tags of web resources, but also their description and documentation.

Keywords: Web resources discovery, requirements-driven methods, Life Sciences, knowledge resources

1. Introduction

Contemporary research in the Life Sciences depends on the sophisticated integration of large amounts of data obtained by in-house experiments, for instance

Received Jul 29, 2011
Revised Dec 16, 2011
Accepted Jan 29, 2012

DNA sequencing, with reference databases available on the web (Cochrane and Galperin, 2010). This is followed by complex analysis workflows that rely on highly specific algorithms, often available as web services (Burgun and Bodenreider, 2008). The amount of data produced and consumed by this process is prodigious, and the sheer amount of available resources to manage this data is a source of severe difficulties.

First of all, the existing data sources are very heterogeneous, in some cases as a consequence of a lack of standards for structure and content (Mesiti et al, 2009) and in other cases because the community does not use the available standards. The integration of biomedical data among resources distributed over the web is a major challenge. For example, (Loureno et al, 2010) remarks the challenges, due to the heterogeneity, of data integration in a specific domain from well-known resources which are supposed to contain related data. In addition, many data sources are wrapped as web services, which provide procedural APIs which are more specific, but far less flexible, than declarative query languages found in standalone databases.

Another problem is finding the right web resources for a given research task. The landscape of Life Sciences-oriented web resources is large and complex: there are thousands of available resources, but unfortunately only a few are described by adequate metadata for pursuing large-scale integration efforts.

In addition, there may be several resources that apparently provide the same broad functionality (a particular insidious example is the variety of resources providing variants of alignments of genes and proteins), but not enough metainformation is available to decide which of them is actually the most appropriate for a precise task (Smedley et al, 2010).

Here we claim that semantic technologies can provide a solution to the discovery of web resources in the context of the Life Sciences. common interface for registering, browsing and annotating Life Sciences web services. To enhance its accessibility and usability, BioCatalogue can be indexed by search engines such as Google, provides a programmable API and can be queried from a web browser. The catalogue does not host the services but provides a mechanism to discover and annotate them. BioCatalogue annotations explain what the services do and how to use them. These annotations are manually provided by the service providers and the user community plus some monitoring and usage analysis data obtained automatically by BioCatalogue servers. However, at the moment, most of these annotations are very far from being standard metadata specifications, as they are just free text elements that do not conform to the ^{my}Grid ontology (Wolstencroft et al, 2007), a controlled vocabulary intended to define metadata of the web resources registered in BioCatalogue.

1.1. Related work

The problem of web resource discovery has been extensively studied (Garofalakis et al, 2006; Nair and Gopalakrishna, 2010). The applied techniques have varied according to the exploitable information, which has become ever richer. Originally, only low-level information on the operations of the web resources (such as the basic interface details like method names, types and parameters) was available, or non-functional criteria such as response time or usage data (Birukou et al, 2007); Al-Masri and Mahmoud, 2007). The development of registries based on standards such as UDDI allowed the addition of metadata-based techniques

(Dong et al, 2004; Chukmol, 2008; Crasso et al, 2008), initially based on traditional keyword search but eventually considering more advanced IR techniques (Plebani and Pernici, 2009) and, in line with the priorities of our work, a more user-oriented view (Rong and Liu, 2010), explicitly based on requirements (Hao and Zhang, 2007; Hao et al, 2010; Nazir et al, 2008). (Skoutas et al, 2010) also bases the discovery of web services on the requirements and, ranks and clusters the relevant services with objective measures based on dominance relationships among them. However, they assume well-described web services and well-defined inputs and outputs.

A natural choice in this setting is to incorporate semantic, ontology-based approaches. In (Cardoso, 2006; Skoutas et al, 2007), the similarity between the request and the candidate services is based on the semantics of the inputs and outputs. In principle, these techniques aim to provide higher search precision (Pilioura and Tsalgatidou, 2009), but at the cost of requiring higher development costs and requiring the users to provide a formal specification of their information needs, which is impractical in many cases (Wang et al, 2008). Web services in BioCatalogue itself are meant to be annotated with the ^{my}Grid ontology, which aims to provide a controlled vocabulary for the curation of data. However, in practice searches are hampered by poor documentation and most annotations are just free text, not conforming to the ^{my}Grid (or any) ontology at all. There are also proposals which aim to use semantic reasoning to discover and integrate heterogeneous data sources. For instance, SSWAP (Gessler et al, 2009) is an architecture, a protocol and a platform to semantically discover and integrate heterogeneous disparate resources on the web. Unfortunately, this approach heavily relies on the provided metadata, which is usually very poor. Other approaches focus on the development of interfaces to assist in the location of web resources; for example (Navas-Delgado et al, 2006) presents a client engine for the automatic and dynamic development of service interfaces built on top of the BioMoby standard.

A specific development in the Life Sciences field is the possibility to exploit web-based registries such as BioCatalogue in two new ways: as a social graph (already being exploited in the context of workflow repositories (Tan et al, 2010)) and, most importantly for this work, as the target of data enrichment and integration using the extensive repositories of information available for the Life Sciences, such as PubMed or the UMLS.

2. Requirement-driven discovery of web resources

While the number of web resources increases continuously, the Life Sciences community lacks user accepted standards for annotation, as well as common procedures to guide the resolution of complex tasks (Tran et al, 2004). It is usual that only a small group of researchers is aware of the existence of some web resources. For example, by checking the query logs of BioCatalogue we observed how most queries just ask for the name of some concrete service. This indicates that the catalogue is being used as an entry to the specifications of the services that the users already know, but not as a way to discover new relevant resources. From our point of view, the problem is that these catalogues are not oriented to the specification of user requirements and it is difficult to use them in order to find out appropriate resources for a given task. This problem is even harder for researchers looking for resources that are out of the scope of their

Fig. 1. Phases of the proposed approach

fields. For example, a biomedical researcher looking for new biomarkers for some disease could find out interesting information within biological databases about putative functions of proteins, but she does not know where they are located nor how to deal with them.

As a consequence, it is a pressing question how to help researchers to discover the best possible mapping between their needs and the available tools that may be useful to them. We have designed a new approach for the discovery of web resources as follows:

1. The user requirements are provided to the system in natural language. They consist of a brief description of both the user goals and the tasks that can be done to reach them.
2. The selection of the relevant web resources is made using semantic technologies based on knowledge stored in BioCatalogue and in well-accepted Life Sciences ontologies, such as UMLS and ^{my}Grid.
3. For each task, the system prompts to the user a short list of web resources that could be used to execute it.
4. In case the answer of the system does not meet the user requirements, it should be possible to modify the initial descriptions so that a new answer is returned.

From the point of view of the users, the main advantages of this approach are the use of free text to describe their requirements, and the ability to follow their own procedures in the specification of their goals and the tasks to reach them. Notice that different users can define different sets of tasks to reach the same goal. Moreover, with this approach, it is the user who finally chooses the sequence of web resources to be executed, and if the list proposed by the system is not considered to be good enough, the user can modify the initial goal and task descriptions to get a new answer. In this way, the web resource discovery method proposed by this approach is guided by the expertise of the user. Furthermore, semantic technologies facilitate the discovery of mappings between user requirements and web resources.

3. The overall process

Our approach consists in a process divided into three phases as depicted in Figure 1. The main purpose of this process is to normalize both the user requirements and the web resources metadata in order to compare them and to discover the web resources that best match the user needs. In this section we explain the methods and techniques applied in each phase.

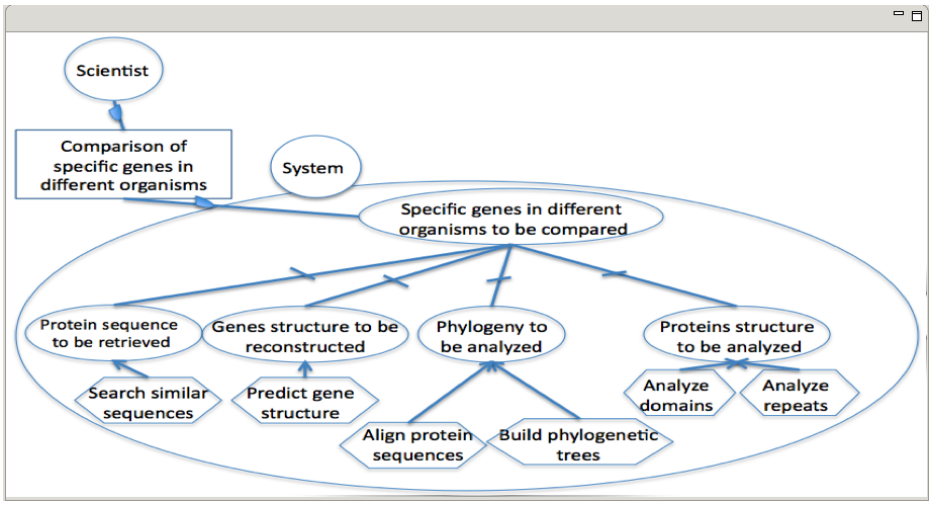

Fig. 2. i* Requirements model

3.1. User Requirements

Given the experience and the knowledge users have on their domains, they can easily provide natural language descriptions of their requirements and the tasks that would be manually performed to meet them. Natural language descriptions are easy to express but difficult to process in an automatic way. In this section we present the model we have adopted to specify user requirements in a formal way so that they can be automatically used in the subsequent phases of the resource discovery process.

User requirements are represented by means of goal and task elements in a formal specification called the *Requirements model*. This specification is made using the *i** formalism (Yu, 1995; Yu, 1997), which is both a goal-oriented and an agent-oriented language. We use this framework because it provides the functionality required to obtain a formal specification of the user's requirements without taking into account the characteristics of the system. The goal and the task elements of the *Strategic Rationale* (SR) model of the *i** framework capture the user's information requirements and the steps to achieve them. This model generalizes the work in (Pérez-Catalán et al, 2009) to allow the specification of the user requirements in the context of finding appropriate similarity measures for complex information.

As an example, Figure 2 shows the description of the goals and tasks specified by a user in demand of gene comparison as part of a study of the presence of the LRRK2 genes in the organism "N. Vectensis". Notice that this way of specifying user requirements by means of a hierarchy of goals and tasks resembles the approach followed by (Tran et al, 2004) to describe, analyse and classify the tasks undertaken by bioinformaticians.

3.2. Normalization

User requirements are described in natural language in the *Requirements model* but to be automatically processed, they have to be normalized. During this process, the normalized user specifications will be compared to the web resources metadata, which will be also normalized in the same way.

Semantic annotation (SA) can be seen as the process of linking the *entities* mentioned in a text to their *semantic descriptions*, which are usually stored in knowledge resources (KRs) such as thesauri and domain ontologies (Kiryakov et al, 2004). Former approaches to SA were mainly guided by users (e.g., (Kahan and Koivunen, 2001)) through seed documents, manually tagged examples or ad-hoc extraction patterns. However, in our scenario, we require the SA process to be fully automatic and unsupervised. This is because the volume of data to be processed is huge and the set of possible user requirements is unknown a priori. There are few approaches performing fully unsupervised SA, and many of them are based on dictionary look-up methods or ad-hoc extraction patterns and others are based on statistics on the web such as (Sánchez et al, 2011) (see (Uren et al, 2006) for a review of SA concepts and approaches).

Most SA approaches assume that the entities are *named entities* (e.g., people, locations, organizations, and so on); hence named entity recognition (NER) is the basic pillar of these approaches. However, the proliferation of comprehensive KRs has extended the notion of entity to any kind of object or concept. For example, in (Dánger and Berlanga, 2009; Berlanga et al, 2010), domain ontology concepts are considered potential entities whose instances occur in the texts. In the biomedical scenario, proteins, diseases and organs are considered entities to be identified from texts.

Our SA process consists of three main steps. In the first step, the KR is processed to generate a lexicon, which should contain the lexical variants with which each concept is expressed in the written texts. The second step consists in applying some *mapping function* between the text chunks likely to contain an entity and the KR's lexicon, in order to obtain the list of concepts that are potentially associated. Notice that entities usually appear in noun phrases; thus, the text chunks to be considered are restricted to these syntactic structures. Finally, in the third step, the concepts whose lexical forms best fit to each text chunk are selected to generate the corresponding semantic annotation. The final semantic annotation contains the unique references to the corresponding concepts and the text span matched by these concepts.

3.2.1. Knowledge resources

As our method relies on the SA of both the user requirement specifications and the web resource metadata, we need to establish the reference KRs from which concepts are brought. Unfortunately, a single comprehensive ontology for this application domain does not exist, and therefore we need to combine several existing resources. For this purpose, we have selected as the main KR the reference ontologies of Biocatalogue (i.e., ^{my}Grid ontologies) and EDAM Ontology (Pettifer et al, 2010), that improves the annotations of the ^{my}Grid ontologies. We have also used the UMLS Meta-thesaurus (version 2010AA) to cover the concepts about procedures, anatomy, diseases, proteomics and genomics. This metathesarus is an integrated resource that includes a great variety of thesauri

Source	Concept reference format	Comment
UMLS	`UMLS:C<number>:STypes`	`STypes` are the semantic types associated to UMLS concepts (e.g. Disease, Protein, etc.)
Wikipedia	`Wiki:W<number>:Categs`	`Categs` are the categories associated to the page entry of the referred concept.
myGRID	`myGR:D<number>:`	Concepts extracted from the ^{my}Grid ontologies.
EDAM	`EDAM_<number>:`	Concepts extracted from the EDAM ontology.

Table 1. Concept reference formats used for the different semantic sources. The generic format for a reference is `Source:ConceptID:SemanticTags`

and ontologies such as the Gene Ontology (GO) [1], the HUGO database [2], and many other related to the biomedical domain. Finally, in order to provide broad coverage for the names of the algorithms and methods involved in Bioinformatics, we have included the entries of the Wikipedia related to any sub-category of the *Bioinformatics* category. Currently, we are building other specific annotators based on other Wikipedia categories, since the Wikipedia is becoming the information hub for emerging semantic technologies (Hu, 2010) (Bizer et al, 2009).

For tagging purposes, all these KRs are loosely integrated into a concept repository which consists of an inventory of concepts, their taxonomical relationships (i.e. is_a relationship) and the lexical variants associated to each concept (e.g. alternative labels, synonyms, and so on) (Jimeno-Yepes et al, 2009). From now on, we denote the whole repository as KR, the taxonomical relationship between two concepts $C, C' \in KR$ as $C \preceq C'$, and the lexical variants of each concept $C \in KR$ as $lex(C)$.

3.2.2. Normalization through semantic annotation

As previously explained, user descriptions of goals, tasks and resources are expressed in natural language. In order to reconcile the users requirements and the resources, we need to normalize their representation under a well-defined semantic space. This normalization process involves the annotation of all the descriptions with the concepts of the knowledge resource KR. As mentioned before, this process consists of a mapping function between each text chunk, denoted with T, and the lexical variants of each KR concept, denoted with $lex(c)$. This function is defined as follows:

$$sim(C,T) = max_{S \in lex(C)} \left[\frac{idf(S \cap T) - idf(S - T)}{idf(S)} \right]$$

This function measures the information coverage of T with respect to each lexical variant of a concept C. Notice that we assume that text chunks and lexical strings are represented as bags of words. Information is measured as usual with

[1] http://www.geneontology.org/
[2] http://www.genenames.org/

Task description

```
Build <e id="UMLS:C1519068:T062:1|Wiki:W149326;6555571;825200;976276:1,2|
UMLS:C9000005:T090:1,2|myGR:D9000400::1,2"> phylogenetic trees < /e>

T= {'C1519068': 11, 'W149326': 18, 'C9000005': 15, 'D9000400': 18}
```

Service description

```
<service id="2027">
<name>GlobPlot< /name>
<category><e id="UMLS:C1513868:T087">Domains< /category>
<tag>order< /tag><tag>globularity< /tag>
<tag><e id="UMLS:C0012634;T047">disorder< /e>< /tag>
<tag>EMBRACE< /tag><tag>EMBL< /tag>
<tag><e id="myGR:D9000419.15::1,2">protein sequence< /e>< /tag>
<description>Globplot <e id="UMLS:C1999219:T169">performs< /e>
<e id="UMLS:C0021699:T116:1,2|UMLS:C0549551:T046:2,3">intrinsic
protein disorder< /e><e id="UMLS:C1513868:T087">domain< /e>
...
< /description>
< /service>

S_1429 = {'C1519068': 11, 'W149326': 18, 'C9000005': 15, 'D9000400': 18,
'C0040811':11}
```

Table 2. Semantic Annotation of a task and a service description. Semantic vectors are shown below each annotated description.

the inverse document frequency (IDF), which is an estimation of the string words entropy. Thus, $idf(S)$ is defined as follows:

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

As background corpus for stating the word probabilities, we have used the whole Wikipedia.

All these definitions are inspired by the information-theoretic matching function presented in (Mottaz et al., 2008) and the word content evidence defined in (Couto et al, 2005).

The set of annotations associated to each text chunk T are those concepts that maximize both $sim(C, T)$ and the word coverage of T. That is, the system selects the top ranked concepts whose lexical variants best cover the text chunk T. In order to avoid spurious annotations, a minimum threshold for $sim(C, T)$ is required (usually above 0.7).

From the annotation set of each description, we define a semantic vector weighted by the $tf \times idf$ score, where $tf(C)$ is the frequency of the concept C in the description, and $idf(C)$ is calculated as follows:

$$idf(C) = max_{S \in lex(C)} idf(S)$$

Considering the concept reference formats of Table 1, the annotations generated for the example task described as *build phylogenetic trees* and the metadata of the web service *GlobPlot* are shown in Table 2. We have used the IeXML notation[3] to show the generated annotations. In the same table we also show the resulting concept vectors from these annotations.

[3] http://www.ebi.ac.uk/Rebholz-srv/IeXML/

3.3. Web resource discovery

The discovery of suitable web resources for the user's requirements is based on the matching between the annotations of the tasks and the metadata of the web resources. This matching can be now performed over the semantic vectors associated to them. For example, we could apply the cosine measure to calculate the similarity between web resource descriptions and user requirements, or a concept-based probabilistic model like that presented in (Jimeno-Yepes et al, 2010). However, this kind of measures does not take into account the relevance of each concept within the context of the task.

For example, in the queries "define structurally and functionally important domains of the membrane", "predict gene functions" and "compare functional relationships", the concept *function* does not have the same relevance. In the first query, *functionally* describes only a characteristic of the domain, in the second one, *function* is the key concept in the query, since it is the object that must be predicted and, finally in the third one, *functional* specifies the type of relationship that must be compared. Therefore, the relevance of the same concept in different queries varies depending on the context. To be able to exploit this contextual information, in this work we propose to use a *topic-based ranking model* (Steyvers and Griffiths, 2007).

The basic idea behind a topic-based model is the translation from a word-based statistical model to a topic-based one. Topic-based methods in the literature assume that the topics are hidden and they must be estimated somehow (e.g. *Latent Dirichlet Allocation* (Blei et al, 2003; Griffiths and Steyvers, 2004)). However, in our work, topics are the biomedical tasks underlying both web resources and user requirements. Thus, our method assumes that the topics are predefined (i.e. the target tasks), and we profit from existing annotations (e.g. tags) to automatically estimate the corresponding topic models. One limitation of LDA is that topics are biased to frequent co-ocurrences and, consequently, topics are dominated by frequent tasks. This is the reason why we have adopted a relevance-based model similar to (Pérez et al, 2009).

3.3.1. Topic-based model for web resources

In this paper we propose a topic-based model for resource retrieval where topics represent base user tasks. From now on, we use base task instead of topic.

Let $\{t_1, \cdots, t_N\}$ be the set of base tasks specified in the requirements and to be met by the web resources. Let RT_k be a set of web resource descriptions deemed relevant for the base task t_k.

The conditional probability of each concept $c_i \in KR$ for a base task t_k is estimated as follows:

$$P(c_i|t_k) = \sum_{ws_j \in RT_k} P(c_i|ws_j) \cdot P(ws_j|t_k)$$

That is, we use a mixture of two distributions to calculate the desired one. The distribution $P(c_i|ws_j)$ is estimated from the concept frequencies observed in each resource $tf(c_i, sw_j)$, and smoothing them with Dirichlet priors (Mackay and Peto, 1995) as follows:

$$P(c_i|ws_j) = \frac{tf(c_i, ws_j) + \mu \cdot P(c_i|G)}{\sum_{c_k \in ws_j} tf(c_k, ws_j) + \mu}$$

CUI	Concept	$P(c\|t)$
C1261322	Evaluation	0.0152
C0028811	Job	0.0096
C0080143	Sequence alignment	0.0087
C0002520	Amino acids	0.00607
C1514918	Retrieval	0.00557
W4066308	Multiple alignment	0.0047
C0002518	Protein sequence	0.0043
C1514562	Domain, region	0.0041
C0004793	Sequence	0.0039
C0033684	Proteins	0.0039

Table 3. Top-10 concepts for the base task "align sequences".

To calculate the frequency of concepts, we benefit from the concept taxonomical relationships of KR:

$$tf(c_i, ws_j) = \sum_{c' \in KR, c' \preceq c_i} tf(c', ws_j)$$

Consequently, concepts c_i can either be present in the semantic vector of ws_j or be an ancestor of some concept in ws_j.

The distribution is smoothed in order to avoid probabilities being zero. In this case we use Dirichlet prior smoothing, which is regulated by both the parameter μ (set to 50 in our experiments) and a background corpus G where concept probabilities can be estimated.

The second distribution $P(ws_j|t_k)$ of the conditional probability represents the chance of retrieving a relevant web resource in the context of given base task. This probability is estimated by sampling instances of t_k and counting how many times each web resource of RT_k is retrieved. Thus, the probability is calculated as follows:

$$P(ws_j|t_k) = \frac{n(ws_j, t_k)}{\sum_{ws_i \in RT_k} n(ws_i, t_k)}$$

where $n(ws, t)$ returns the number of times ws is retrieved with t's instances. We consider that an instance of a task is an example description of the base task.

Table 3 shows the top-10 ranked concepts for the base task *align sequences*. *Evaluation*, *job* and *retrieval* are common concepts in web resources descriptions and, therefore, are highly ranked in almost all the base tasks. Then, the other concepts are true representatives of the base task, like *sequence alignment* and *multiple alignment*.

3.3.2. Web resource ranking

Once topic models are built from the set of relevant web resources RT_k, we can define the similarity between a description q and a web resource ws_j as the probability given by the mixture of topic models[4]:

$$P(q|ws_j) = \prod_{c_i \in q} \sum_{t_k} P(c_i|t_k) \cdot P(t_k|ws_j)$$

[4] We also assume the independence of concepts $c_i \in q$.

However, now we do not know the probability $P(t_k|ws_j)$ because ws_j can be out of RT_k. Applying Bayes, we can calculate $P(t_k|ws_j)$ as:

$$P(t_k|ws_j) = \frac{P(ws_j, tk)}{P(ws_j)}$$

If we assume that all the web resources have the same chance to be retrieved, then $P(ws_j)$ is an unknown constant (we do not know how many web resources exist) for all the web resources. Thus, we can rewrite the formulas above as follows:

$$P(t_k|ws_j)\ \alpha\ P(ws_j, t_k)$$
$$P(q|ws_j)\ \alpha\ \prod_{c_i \in q} \sum_{t_k} P(c_i|t_k) \cdot P(ws_j, t_k)$$

Thus, the joint probability of web resources and base tasks can be estimated as follows:

$$P(ws_j, t_k) = \sum_{c_i \in t_k \bigcap ws_j} P(c_i|t_k) \cdot P(c_i|ws_j)$$

4. Evaluation

The effectiveness of a discovery system can be evaluated by measuring the quality of the results obtained for a representative set of heterogeneous queries, considering as the relevant results those included in a well-designed gold standard. In this section, we explain how to use this method in order to evaluate our approach. The evaluation is focused on the Bioinformatics domain and we have selected BioCatalogue[5] as the reference web resource catalogue. At the end of the section, there is a discussion of the results of this evaluation.

BioCatalogue contains 2081 registered services (as of November 2011). Although some services are described through a set of predefined categories, most of them have no metadata and just provide a free text description and/or the web service documentation. Some services do not provide any kind of information but just the URL to their web sites. For these cases, we have downloaded the web site main pages and used them as the service descriptions. We remark that these limitations motivate the use of our approach.

To build the topic based model, we have defined 13 base tasks and, for each base task t_k, we have specified a series of key concepts with which the relevant resources for the RT_k sets are selected. For example, the concept *phylogeny* is deemed relevant for task T_{11} shown in Table 4. It is worth mentioning that these concepts can be automatically gathered from existing documents such as Wikipedia pages related to each base task. With these concepts we have automatically retrieved the top-10 ranked resources by using the cosine measure over their $tf \times idf$ semantic vectors (see Section 3.2.2).

Table 4 shows the cardinality of each RT_k and the most frequent Biocatalogue categories assigned to their resources. Notice that there are highly frequent base

[5] http://www.biocatalogue.org/

	Task	$\|RT_k\|$	Top BioCatalogue categories
T_1	Search proteins with a functional domain	58	Domains
T_2	Localize protein expression	3	N/A
T_3	Search similar sequences	71	Protein sequence similarity Nucleotide sequence similarity
T_4	Identify and characterize genes linked to a phenotype	21	N/A
T_5	Analyze transgenic model organism	44	Microarrays, Biostatistics Data retrieval
T_6	Find genes with functional relationships	60	Pathways, protein interaction
T_7	Find common motifs in genes	9	Function prediction, motifs
T_8	Predict structure	103	Protein secondary structure Protein tertiary structure Protein structure prediction
T_9	Identify putative function of gene	33	Functional genomics Function prediction Domains
T_{10}	Gene prediction	18	Genomics Sequence analysis Gene Prediction
T_{11}	Analyze phylogeny	99	Phylogeny
T_{12}	Align sequences	229	Protein sequence alignment Nucleotide multiple alignment Protein multiple alignment Nucleotide sequence alingment...
T_{13}	Protein identification and characterization	12	Chemoinformatics

Table 4. Number of different resources in the RT_k of each base task and the most frequent BioCatalogue categories.

tasks like *align sequences* and *predict structure*, whereas others are hardly covered by BioCatalogue, like *localize protein expression*. Notice also that the proposed model finds out non trivial associations between the base tasks and the BioCatalogue categories. For example, the base task *analyze transgenic model organism* is related to the category *data retrieval*.

Once the topic-based model is created, the evaluation of the approach is carried out by executing a set of heterogeneous queries (i.e. task description examples) that captures different ways to describe bioinformatics tasks, thus reflecting the variability in the users' information needs. To create the query pool[6], we have selected more than 250 short descriptions extracted from other Life Sciences resource catalogues such as OBRC [7] and ExPaSy [8].

These queries are evaluated over a gold standard due to the difficulties to determine the whole set of relevant results for each query. The gold standard[9] has been built with 443 resources (out of 2081 registered resources), but only for the 7 base tasks that can be unambiguously related to BioCatalogue categories. Additionally, we have manually revised the gold standard in order to ensure the quality of the final set.

[6] `http://krono.act.uji.es/KAIS/pool_queries.xml`
[7] http://www.hsls.pitt.edu/obrc/
[8] http://expasy.org/
[9] `http://krono.act.uji.es/KAIS/gold_standard.xml`

	Task	Number of queries	P@5	P@10	P@20	P	R
T_1	Search proteins with a functional domain	21	0.75	0.65	0.54	0.34	0.93
T_3	Search similar sequences	27	0.89	0.91	0.92	0.57	0.74
T_5	Analyze transgenic model organism	38	0.77	0.75	0.71	0.53	0.57
T_6	Find genes with functional relationships	51	0.77	0.75	0.71	0.45	0.57
T_8	Predict structure	38	0.71	0.63	0.61	0.39	0.7
T_{11}	Analyze phylogeny	12	0.96	0.92	0.9	0.57	0.99
T_{12}	Align sequences	33	0.98	0.98	0.95	0.59	0.84
	Average	31.4	0.85	0.82	0.79	0.49	0.81

Table 5. Precision (P) and recall (R) for the gold standard, including the precision for the top-5, top-10 and top-20 results

With this gold standard, we have evaluated the results obtained for each one of the queries from our query pool with the traditional precision and recall quality measures:

$$precision = \frac{|relevant\_resources \cap retrieved\_resources|}{|retrieved\_resources|}$$

$$recall = \frac{|relevant\_retrieved\_resources|}{|relevant\_resources|}$$

Table 5 shows the precision and recall of the results obtained for the queries, taken from the above mentioned query pool, associated to the 7 base tasks of the gold standard. The results show that the discovered web resources are in most cases adequate matches for the user requirements.

In Table 6, we show the precision at top-10 for some queries randomly selected from the gold standard results, calculated over the whole BioCatalogue.

The results of this evaluation indicate that our approach is an improvement over searching web resources by category, because it discovers web resources that are not categorized in BioCatalogue. The reason is that discovery is based on all the available metadata of the web resources. For example, Table 7 shows the top-10 resources for the query *calculate the maximum likelihood phylogenies from nucleotide* and it can be shown that only one is categorized in BioCatalogue. All of them except *INB:inb.bsc.es:runPhylipDnapars* and *ClustalW2 Phylogeny* calculate phylogenies using the maximum likelihood algorithm over nucleotide sequences.

Moreover, if a user sends these queries to BioCatalogue, most of them do not produce any answer. The reason is that BioCatalogue implements the search as a string matching procedure instead of using an information retrieval engine. Therefore, one of the main advantages of our approach is that it allows users to express their requirements in free text. It is also important to note that some queries involve vocabulary related to more than one base task, which is problematic for the discovery of the most appropriate resources. However, thanks to the topic-based model, the most relevant concepts are the ones that guide the discovery.

Task	Query	P@10 GS	P@10 All
T_1	Search for information on protein domain familiy	0.8	1.0
T_1	Localize protein domains situated at the surface of virus particles	0.8	1.0
T_1	Find information about protein domains in protein function and interaction evolution	0.8	1.0
T_3	The sequence similarity of the mandarin fish domain with those of higher vertebrates	1.0 1.0	1.0 1.0
T_5	Analyze microarray gene expression and other functional genomics-related data	1.0	1.0
T_5	Interpret microarray gene expression data	1.0	1.0
T_5	Analyze gene expression profiles	1.0	1.0
T_5	Interpret gene expression data obtained from microarray experiments	1.0	1.0
T_5	Search for microarray data of Arabidopsis	1.0	1.0
T_6	Find similar protein interaction networks between species	0.8	1.0
T_8	Predict transmembrane protein helices	0.7	0.9
T_8	Perform secondary structure predictions on protein sequences	1.0	1.0
T_{11}	Perform pipeline phylogenetic analysis of protein or DNA sequences	0.6	1.0
T_{12}	Conduct protein alignment analysis	1.0	1.0
T_{12}	Align multiple DNA or protein sequences	1.0	1.0

Table 6. Precision at top-10 in the gold standard (GS) and regarding all the resources.

Resource	BioCatalogue categories
BuildPhylogeneticTreeFromFastaAlignmentService	N/A
INB:inb.bsc.es:runPhylipDnamlk	N/A
PhylipService	N/A
INB:inb.bsc.es:runPhylipDnapars	N/A
INB:inb.bsc.es:runPhylipDnaml	N/A
INB:inb.bsc.es:runPhylipDnapenny	N/A
ClustalW2 Phylogeny	Phylogeny
EMBOSS fdnamlk	N/A
EMBOSS fdnaml	N/A
EMBOSS fproml	N/A

Table 7. Top-10 results of the query "Calculate maximum likelihood phylogenies from nucleotide sequences".

From these results, we can conclude that using semantic technologies provides a satisfactory solution to the discovery of web resources in the context of the Life Sciences. This assertion is based on the following observations. First of all, the emergence of metadata-based repositories such as BioCatalogue. Second, the existence of broad semantic resources, such as UMLS, that can be used to characterize the researcher's requirements with a high degree of precision. However, it is not realistic to assume that Life Sciences researchers will manually create formalized requirement specifications, which could be a costly and tedious process. Instead, our approach proposes the automatic semantic annotation of free-text description of resources and users requirements.

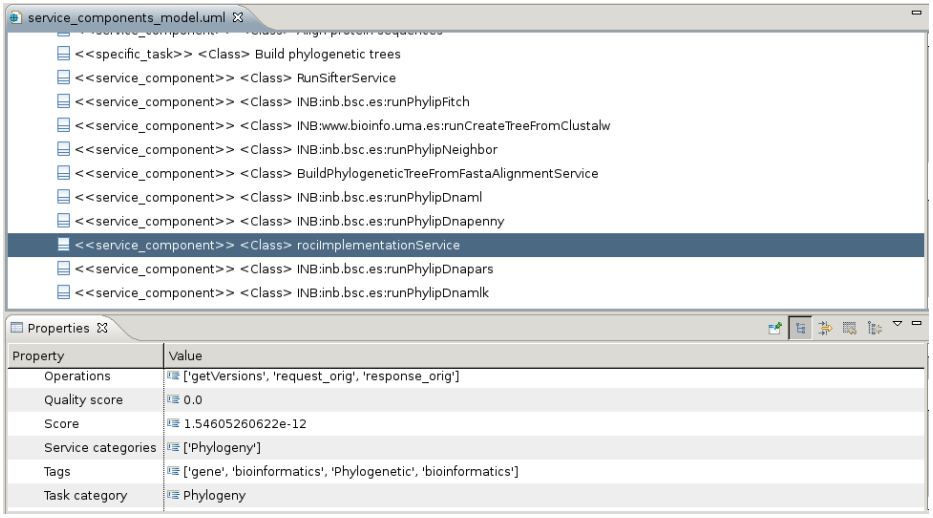

Fig. 3. Screen showing the ranked list of web resources discovered by the prototype for the task *build phylogenetic trees*.

5. Prototype

In order to demonstrate the usefulness of the proposed web resource discovery system, we have implemented a prototype which consists of several components. First, the requirement specification step is supported by an Eclipse-based add-in, built on the Eclipse EMF modelling framework [10], which supports $i*$-based task specification. As an example, Figure 3 shows the ranked list of web resources discovered by the prototype for the task *build phylogenetic trees*. The user is provided with relevant metadata, such as the BioCatalogue categories, the tags or the name of the operations of the resource. While the current implementation provides access to the full modeling capabilities of the Eclipse EMF, a simplified graphical editor is under development.

The core matching functions between web resources and requirements are implemented as a set of Python modules. The associated semantic resources are available either as off-the-shelf databases (e.g. MySQL) or, when required for performance, as customized indexes. For easy programmatic access, this has been encapsulated as a web service. A simple search interface for testing and evaluation purposes built on top of this web service is publicly available [11]. As an example, Figure 4 shows the results of a matching operation for a simple task. In addition to the ranked list of web services, additional information is presented: the annotation of the input task, and the probabilities that each service is related to each of the 13 base tasks on which our topic model is built.

[10] `http://www.eclipse.org/modeling/emf/`
[11] `http://krono.act.uji.es/biocat/BioCatClient.html`

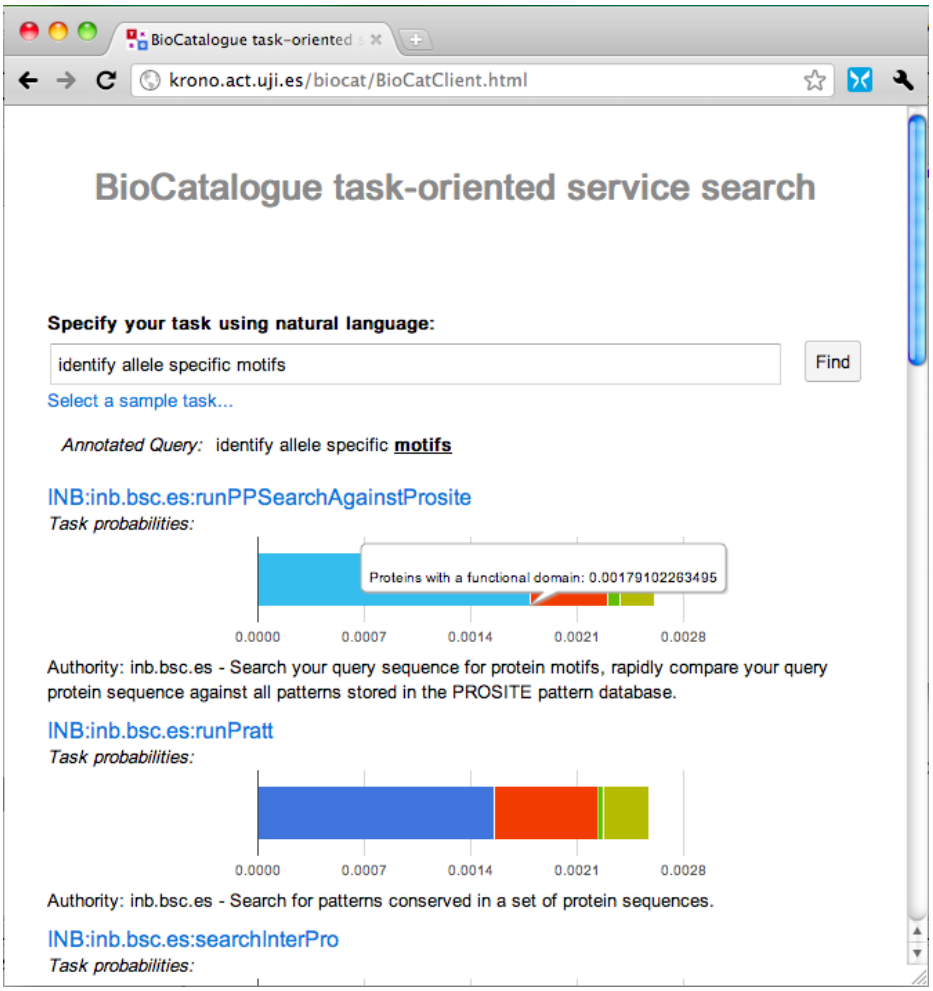

Fig. 4. The results of matching of a user-specified task with the annotated BioCatalogue-indexed web resources.

6. Conclusions

We have presented a semi-automatic approach that guides researchers in the Life Sciences in the discovery of the most adequte web resources for their well-defined requirements. With this approach, users can easily find out web resources that were previously unknown to them because fell out the scope of their main field of interest, or were poorly categorized with existing tags. We exploit two specific characteristics of the domain: the availability of broad semantic resources that allow the annotation of plain text, and the emergence of web-based repositories of web resources metadata.

Due to the importance of the semantic normalization, we have considered it appropriate to annotate the available information of all the resources registered in BioCatalogue. To achieve this, we have used a biomedical semantic annotator

that applies several ontologies to normalize biomedical texts. 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 resources. However, we have noticed that not all the concepts have the same relevance in different contexts and, therefore, we have defined a topic-based model for web resources to take into account the relevance of a concept in each task. The result of the process is that users are provided with a set of ranked lists of web resources that are relevant for their stated information needs.

The intended end user of our approach is assumed to be an expert on her field, with the ability to recognise which web resources are the most appropriate for a task once they have been discovered and characterised with enough metainformation. In this context, one of the main benefits of our approach is that it is a semi-automatic process. In practice, we consider it necessary to allow the user to be able to change parameters, annotations or automatic selections in each of the phases, based on her own knowledge and experience, or on previous results. Thus, the process described in this paper is an instance of *exploratory search*, advising the user in each step, and taking advantage of her previous knowledge.

Some direct follow-ups of this work are the refinement of the particular details of the semantic techniques used in our approach, and the creation of a GUI to facilitate its application. In this context, one future improvement is the integration of the different knowledge resources used for semantic annotation. Ontology matching techniques will be studied to perform such an integration (Martinez and Aldana, 2011). This will reduce the number of generated annotations as well as the ambiguity of some concepts. Moreover, we are also considering the use of non-functional requirements to evaluate the quality of the candidate resources. From a broader perspective, the study of the emerging metadata repositories shows opportunities for further research. Some of the promising avenues we are beginning to work on are the exploitation of as many different sources of metadata as possible, in order to improve the assessment of the relevance of a given web resource. For example, we are exploring the use of external data (bibliographical information, referenced web pages) to complete the limited user-provided metadata; another option is the automatic sampling of the output of web services in order to obtain accurate information on their semantic coverage. We are also considering the possibility of using faceted search to exploit the available information of the resources; in this respect, (Pérez-Catalán et al, 2011) shows some promising preliminary results. Finally, we intend to explore how to exploit the social data available in web-based repositories of web services (e.g. BioCatalogue's social networking and crowdsourcing aspects), by incorporating techniques such as social network analysis into our approach.

Acknowledgments

This work has been partially funded by the "Ministerio de Economía y Competitividad" with contract number TIN2011-24147, and the Fundació Caixa Castelló project P1-1B2010-49. María Pérez has been supported by Universitat Jaume I predoctoral grant PREDOC/2007/41. We thank the BioCatalogue team for providing us with statistical information about their queries, and also the reviewers for their constructive and detailed comments.

References

Al-Masri E, Mahmoud Q (2007) QoS-based discovery and ranking of web services. In: Proceedings of 16th International Conference on Computer Communications and Networks. ICCCN 2007, pp 529–534. doi: 10.1109/ICCCN.2007.4317873.

Berlanga R, Nebot V, Jimenez-Ruiz E (2010) Semantic annotation of biomedical texts through concept retrieval. *Procesamiento del Lenguaje Natural*, 45:247–250.

Bhagat J, Tanoh F, Nzuobontane E et al. (2010) BioCatalogue: a universal catalogue of web services for the life sciences. *NAR* 38 (suppl 2):W689–W694 doi: 10.1093/nar/gkq394.

Birukou A, Blanzieri E, D'Andrea V et al. (2007) Improving web service discovery with usage data. *IEEE Software*, 24(6):47–54, 2007. doi: 10.1109/MS.2007.169.

Bizer C, Lehmann J, Kobilarov G et al. (2009) DBpedia - A crystallization point for the web of data. *Web Semant.* 7(3):154–165. doi: 10.1016/j.websem.2009.07.002.

Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3(4–5):993–1022. doi: 10.1162/jmlr.2003.3.4-5.993.

Burgun A, Bodenreider O (2008) Accessing and integrating data and knowledge for biomedical research. *Med. Inform. Yearb.* 2008:91–101.

Cardoso G (2006) Discovering semantic web services with and without a common ontology commitment. In: Proceedings of the IEEE Services Computing Workshops 2006, SCW'06, pp 183–190 doi: 10.1109/SCW.2006.12.

Chukmol U (2008) A framework for web service discovery. In: Proceedings of the 2nd SIGMOD PhD workshop on Innovative database research – IDAR '08, pp 13–18. doi: 10.1145/1410308.1410313.

Cochrane G, Galperin M (2010) The 2010 Nucleic Acids Research database issue and online database collection: a community of data resources. *Nucleic Acids Research*, 38:D1–D4.

Couto F, Silva M, Coutinho P (2005) Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, 6(S-1):S21. doi: 10.1186/1471-2105-6-S1-S21.

Crasso M, Zunino A, Campo M (2008) Easy web service discovery: a query-by-example approach. *Science of Computer Programming*, 71(2):144–164. doi: 10.1016/j.scico.2008.02.002.

Dánger R, Berlanga R (2009) Generating complex ontology instances from documents. *J. Algorithms*, 64(1):16–30.

Dong X, Halevy A, Madhavan J et al. (2004) Similarity search for web services. In: VLDB '04, Proceedings of the Thirtieth international conference on Very Large Data Bases, pp 372–383.

Garofalakis J, Panagis Y, Sakkopoulos E et al. (2006) Contemporary web service discovery mechanisms. *Journal of Web Engineering*, 5(3):265–290.

Gessler D, Schiltz G, May G et al. (2009) SSWAP: A simple Semantic Web architecture and protocol for semantic web services. *BMC Bioinformatics*, 10:309.

Goble C, Stevens R, Hull H et al. (2008) Data curation + process curation = data integration + science. *Briefings in Bioinformatics*, 9:(6)506–517.

Griffiths T, Steyvers M (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235. doi: 10.1073/pnas.0307752101.

Hao Y, Zhang Y (2007) Web services discovery based on schema matching. In: Proceedings of the thirtieth Australasian conference on Computer Science, ACSC '07, pp 107–113.

Hao Y, Zhang Y, Cao J (2010) Web services discovery and rank: An information retrieval approach. *Future Generation Computer Systems*, 26(8):1053–1062. doi: 10.1016/j.future.2010.04.012.

Hu, B (2010) WiKi'mantics: interpretig ontologies with Wikipedia. *Knowledge and Information Systems*, 25(3): 445–472.

Jimeno-Yepes A, Jiménez-Ruiz E, Berlanga R, Rebholz-Schuhmann D (2009) Reuse of terminological resources for efficient ontological engineering in life sciences. *BMC Bioinformatics*, 10(S-10):4.

Jimeno-Yepes A, Berlanga R, Rebholz-Schuhmann D (2010) Ontology refinement for improved information retrieval. *Inf. Process. Manage.*, 46(4):426–435.

Kahan J, Koivunen M (2001) Annotea: an open RDF infrastructure for shared web annotations. In: Proceedings of the 10th international conference on World Wide Web, WWW '01, pp 623–632. doi: 10.1145/371920.372166.

Kiryakov A, Popov B, Terziev I et al. (2004) Semantic annotation, indexing, and retrieval. *J. Web Sem.*, 2(1):49–79.

Anlia Loureno, Carneiro S, Rocha M, Ferreira E et al. (2010) Challenges in integrating Escherichia coli molecular biology data *Briefings in Bioinformatics*,12(2):91–103.

Mackay D, Bauman Peto L (1995) A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):1–19.

Martinez-Gil J, Aldana-Montes J (2011) Evaluation of two heuristic approaches to solve the ontology meta-matching problem. *Knowledge and Information Systems*, 26(2):225–247.

Mesiti M, Jiménez-Ruiz E, Sanz I, Berlanga R et al. (2009) XML-based approaches for the integration of heterogeneous bio-molecular data. *BMC Bioinformatics*, 10(S-12):7.

Mottaz A, Yip Y, Ruch P, Veuthey A (2008) Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC Bioinformatics* 9(S-5):S3.

Nair M, Gopalakrishna V (2010) Look before you leap: a survey of web service discovery. *International Journal of Computer Applications*, 7(5):5–11.

Navas-Delgado I, Rojano-Muñoz M, Ramírez S et al. (2006) Intelligent client for integrating bioinformatics services. *Bioinformatics*, 22(1):106–111.

Nazir S, Sapkota B, Vitvar T (2008) Improving web service discovery with personalized goal. In: 4th International Conference on Web Information Systems and Technologies, pp 266–277.

Pérez JM, Berlanga R, Aramburu MJ (2009) A relevance model for a data warehouse contextualized with documents. *Information Processing and Management*, 5(3):356–367.

Pérez-Catalán M, Casteleyn S, Sanz I, and Aramburu MJ (2009) Requirements gathering in a model-based approach for the design of multi-similarity systems. In: International Workshop on Model-Driven Service Engineering and Data Quality and Security, MOSE+DSQ'09. doi: 10.1145/1651415.1651425

Pérez-Catalán M, Berlanga R, Sanz I, Aramburu MJ (2011) Exploiting text-rich descriptions for faceted discovery of web resources. In: *Semantic Web Applications and Tools for the Life Sciences , SWAT4LS'11*.

Pettifer S, Thorne D, McDermott P et al. (2010) An active registry for bioinformatics web services *Bioinformatics* 25(16), 2090–2091.

Pilioura T, Tsalgatidou A (2009) Unified publication and discovery of semantic web services. *ACM Transactions on the Web*, 3(3):1–44. doi: 10.1145/1541822.1541826.

Plebani P, Pernici B (2009) URBE: web service retrieval based on similarity evaluation. *IEEE Transactions on Knowledge and Data Engineering*, 21(11):1629–1642.doi: 10.1109/TKDE.2009.35.

Rong W, Liu K (2010) *A survey of context aware web service discovery: from user's perspective*. In: Fifth IEEE International Symposium on Service Oriented System Engineering (SOSE) .doi: 10.1109/SOSE.2010.54.

Sánchez D, Isern D, Millan M (2011) Content annotation for the semantic web: an automatic web-based approach. *Knowledge and Information Systems*, 27(3):393–418.

Skoutas D, Sacharidis D, Simitsis A, Sellis T (2010) Ranking and clustering web services using multicriteria dominance relationships. *IEEE Transactions on Services Computing*, 3(3): 163 –177.doi: 10.1109/TSC.2010.14.

Skoutas D, Simitsis A, Sellis T (2007) A ranking mechanism for semantic web service discovery. In: IEEE Congress on Services, pp 41–48.

Smedley D, Schofield P, Chen C et al. (2010)Finding and sharing: new approaches to registries of databases and services for the biomedical sciences. *Database: the journal of biological databases and curation*, 2010(0):baq014.doi: 10.1093/database/baq014.

Stevens R, Goble C, Baker P, Brass A (2011) A classification of tasks in bioinformatics. *Bioinformatics*, 17(2):180–188.doi: 10.1093/bioinformatics/17.2.180.

Steyvers M, Griffiths T (2007) Probabilistic Topic Models. In: Landauer T, McNamara DS, Dennis S, Kintsch W (eds) Handbook of Latent Semantic Analysis Lawrence Erlbaum Associates.Hillsdale, NJ. ISBN 1410615340.

Tan W, Zhang J, Foster I (2010) Network analysis of scientific workflows: a gateway to reuse. *Computer*, 43(9):54–61.doi: 10.1109/MC.2010.262.

Tran D, Dubay C, Gorman P, Hersh W (2004) Applying task analysis to describe and facilitate Bioinformatics tasks. *Stud Health Technol Inform.* 107(Pt 2):818–822.

Uren V, Cimiano P, Iria J et al. (2006) Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1):14–28.doi: 10.1016/j.websem.2005.10.002.

Wang X, Hauswirth M, Vitvar T, Zaremba M (2008) Semantic Web Services selection improved by application ontology with multiple concept relations. In: Proceedings of the 2008 ACM symposium on Applied computing - SAC '08. doi: 10.1145/1363686.1364222.

Wolstencroft K, Alper P, Hull D et al. (2007) The myGrid ontology: Bioinformatics service discovery. *International Journal of Bioinformatics Research and Applications* 3(3):303-325.

Yu E (1995) *Modelling strategic Relationships for process reenginering.* PhD thesis, University of Toronto, Canada.

Yu E (1997) Towards modelling and reasoning support for early-phase requirements engineering. In: 3rd IEEE International Symposium on Requirements Engineering (RE'97) pp 2444–2448.

Author Biographies

María Pérez-Catalán received a BS and a MSc in Computer Science from Universitat Jaume I (UJI), Spain in 2007 and 2008 respectively. In 2008 she joined the department of Computer Science and Engineering at UJI as a predoctoral researcher. Her research interests include semantic technologies, information retrieval and web services discovery.

Rafael Berlanga is an associate professor in the Computer Science career at Universitat Jaume I, Spain since 1999. He received the B.S. degree from Universidad de Valencia in Physics, and the Ph.D. degree in Computer Science in 1996 from the same university. He is author of more than twenty articles in relevant international journals, as well as numerous communications in international conferences. His current research interests are text mining, knowledge bases and information retrieval.

Ismael Sanz is a lecturer (*Contratado Doctor*) at the Computer Science and Engineering department at Universitat Jaume I (UJI) in Spain since 2009. He received a BS, DEA and PhD from UJI in 1997, 2006 and 2007 respectively. His research interests include the processing of semi- and unstructured data, semantic technologies and information retrieval.

María José Aramburu received the BS degree in computer science from the Universidad Politcnica de Valencia in 1991 and the PhD degree from the School of Computer Science, University of Birmingham, United Kingdom, in 1998. She is associate professor of Computer Science at Universitat Jaume I, Spain. Author of articles in international journals such as Information Processing & Management, Decision Support Systems, IEEE Transactions on Knowledge and Data Engineering and numerous communications in international conferences such as ICDE, DEXA, and ECIR, her main research interests include knowledge repositories, decision support systems and integration of information.

Correspondence and offprint requests to: María Pérez-Catalán, Department of Computer Science and Engineering (DICC), Universitat Jaume I, E-12071 Castelló de la Plana, Spain. Email: mcatalan@uji.es