Decision Support Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Decision Support Systems

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

A provenance-based approach to semantic web service description and discovery

Tom Narock [a,*], Victoria Yoon [b], Sal March [c]

[a] Department of Information Technology and Management Science, Marymount University, Arlington, VA, USA
[b] Department of Information Systems, Virginia Commonwealth University, Richmond, VA, USA
[c] Owen Graduate School of Management, Vanderbilt University, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 19 August 2013
Received in revised form 3 April 2014
Accepted 23 April 2014
Available online xxxx

Keywords:
Web service discovery
Semantic web service
Ontology
Provenance

ABSTRACT

Web services have become common, if not essential, in the areas of business-to-business integration, distributed computing, and enterprise application integration. Yet the XML-based standards for web service descriptions encode only a syntactic representation of the service input and output. The actual meaning of these terms, their formal definitions, and their relationships to other concepts are not represented. This poses challenges for leveraging web services in the development of software capabilities. As the number of services grows and the specificity of users' needs increases, the ability to find an appropriate service for a specific application is strained. In order to overcome this challenge, semantic web services were proposed. For the discovery of web services, semantic web services use ontologies to find matches between user requirements and service capabilities. The computational reasoning afforded by ontologies enables users to find categorizations that weren't explicitly defined. However, there are a number of methodological variants on semantic web service discovery. Based on e-Science, an analog to e-Business, one methodology advocates deep and detailed semantic description of a web service's inputs and outputs. Yet, this methodology predates recent advances in semantic web and provenance research, and it is unclear the extent to which it applies outside of e-Science. We explore this question through a within-subjects experiment and we extend this methodology with current research in provenance, semantic web, and web service standards, developing and empirically evaluating an integrated approach to web service description and discovery. Implications for more advanced web service discovery algorithms and user interfaces are also presented.

1. Introduction

Service oriented architecture (SOA) defines a set of principles and methodologies for designing and developing software in the form of interoperable services. These services, referred to formally as *web services*, are applications that can be used automatically by a computer on behalf of a user. Frequently they are embedded within applications to enable rapid and reliable system development. Web services are described using the Web Service Description Language (WSDL) [1]. They are offered over the Web as functional software building blocks accessible via standard Internet protocols, independent of platforms and programming languages [2]. Web services have become common, if not essential, in the areas of business-to-business integration, distributed computing, and enterprise application integration due to their interoperability and extensibility [3]. Web services can be used together in a loosely coupled fashion and new services can be formed from the aggregation of existing services [2].

However, the XML-based standards for web service descriptions encode only a syntactic representation of what is expected by, and returned from, a service, i.e., its inputs and outputs. The actual meaning of the terms used, their formal definitions, and their relationships to other concepts are not represented. This poses several challenges for leveraging web services in the development of software capabilities. In particular, as the number of services grows and the specificity of users' needs increases, the ability to find an appropriate service for a specific application is strained [4]. Hendler [4] proposed using ontologies to provide a more powerful method for the discovery of web services. This work provided the foundations of what is now known as *semantic web services*. Semantic web services aim toward reducing the manual effort required for discovering and using web services [5].

One of the prime application areas for semantic web services is e-Science. An analog to e-business, e-Science [6] began as the application of computing to traditional science research. However, in recent years computers have become vital to scientific research and e-Science has transformed the way in which scientific research is performed [7]. Increasing data volumes within many scientific domains makes it no longer practical to copy data and perform local analysis. Instead, hypotheses are tested through online tools that combine and mine pools of data [8]. Web services are enabling such efforts, termed "industrial scale science" [7], by making data and algorithms programmatically accessible on the Web. Service oriented architectures have become the common distributed technology in e-Science [7,9] and web services allow an increasing volume of scientific analysis to take place on the web.

* Corresponding author at: Marymount University, 1000 Glebe Road, Arlington, VA, 22201.
E-mail address: tnarock@marymount.edu (T. Narock).

http://dx.doi.org/10.1016/j.dss.2014.04.007
0167-9236/

The most prominent type of web service discovery is known as *interface matching* [10]. It consists of services being discovered by matching descriptions of input and output semantics of the web service to the application requirements. However, e-Science requirements have challenged this notion. Semantic web service discovery in e-Science requires much more conceptual description than is present in a service's interface [11]. Domain information including descriptions of tasks the service will perform, additional information resources consulted in the performance of those tasks, and software algorithms utilized are frequently required for scientists to adequately match requirements to web service [11]. The "black-box" interpretation of web services, which is used in interface matching, assumes that each output depends on all inputs provided, and fails to model the internal state and execution processes of the service. Such coarse-grained approximations are rarely true and can be misleading in understanding and interpreting a service's output, particularly in e-Science applications [12].

In reality, an output may depend only on a small subset of the input and on the internal state of the web service and its processing algorithm. For example, output from a service utilizing a learning algorithm depends on both current and historical inputs and on the specific learning algorithm itself [12]. In addition, documenting the assumptions or decisions made during web service execution give a context in which the results can be reused and enables proper crediting of the scientists involved [13]. Describing a web service at a granularity that includes algorithms and assumptions should ensure proper interpretation and validation of the related scientific claims made. These notions led Wroe et al. [11] to create semantic descriptions of web service processes and incorporate them in the discovery process. While a seminal work in this area, the proposed approach has several limitations on today's semantic web:

1. It is based on technology that is no longer utilized in the semantic web.
2. Current research in provenance and provenance ontologies can be leveraged to better model web service execution processes, and
3. it has never been evaluated for generality and impact on end-users.

Thus, it is unclear the extent to which this approach can be used outside of e-Science. Furthermore, it is not clear how end-users perceive similarity when presented with the additional information provided by this approach.

We address these limitations by providing an experiment to assess the generality of web service discovery based on the semantics of web service processes. Further, we provide an update to Wroe's methodology [11] through the creation of a new ontology based in provenance research and current semantic web standards. We show enhancements to the methodology enabled by this ontology and describe how this ontology can be linked to the emerging W3C web service discovery standard Semantic Annotations for WSDL (SAWSDL) [14]. Finally, we present implications for semantic web service discovery interfaces based on web service processes.

The remainder of this paper is organized as follows. First, we present a brief motivating example of two e-Science web services that appear identical, but function quite differently and thus would be appropriate for different applications. We then discuss related work on web service discovery. Next, we introduce our notion of service provenance, the role it can play in web service discovery, and our provenance-based web services ontology. This is followed by a description of our empirical study and experimental design. We then present an analysis of our results and conclude with a discussion of our contributions and a presentation of an agenda for future research.

2. Motivating example from e-Science

The Jena Geography Dataset[1] is a collection of 200 geographic services that have been collected from around the web. These services focus primarily on geocoding — the process of finding geographic coordinates from other geographic data, such as street addresses and zip codes. Consider the two Jena services shown in Table 1. Both services have input *Location* and output *Distance*. Within the Jena dataset from which these services originated, there are 18 services that have input *Location* and output *Distance*. These 18 services would all be considered to be similar by current web service discovery techniques. Interface matching techniques provide no means of distinguishing services by execution details and the process of matching task to service is left to the user. In this particular example, the results differ in precision, which limits the applicability of each web service. Inadvertently choosing a service that is not applicable can lead to incorrect conclusions and erroneous decisions, which can have far reaching consequences [15].

[1] http://fusion.cs.uni-jena.de/professur/jgd.

3. Related work

The majority of semantic web service discovery algorithms operate on the assumption that explicitly defined service semantics can be exploited to match available services with user requests [16]. This type of discovery is known as interface matching [10] and consists of services being discovered by matching input and output semantics. Interface matching techniques require users to specify desired web service inputs and outputs using concepts from an ontology. SAWSDL [14], the current state of the art technology for interface matching, provides a means of encoding the linkage between ontology concepts and a service's inputs and outputs.

Discovery algorithms then identify exact, more general (superclass concept), and more specific (subclass concept) matches using the ontology, user input, and SAWSDL annotations. This classification scheme was developed by Zaremski and Wing [17] and is widely used in practice. However, interface matching has three main drawbacks [18]. First is low recall due to the rigid hierarchy that is required. Matches are missed due to discovery being limited to only exact, subclass, and superclass relationships. Although additional matches could be determined through relationships created by the ontology developers, the usage of such relationships in discovery algorithms is just now beginning to be studied. Second, depending on the context, more specific and more general results may not be suitable replacements, leading to false positives. Third, discovery is often a manual and iterative process in which a user progressively narrows down the set of candidate services. Users need to take into account the functions the service carry out and the resources it uses to accomplish its goals [11]. There is often insufficient information to make an informed decision leading to significant manual effort to identify an appropriate web service.

So-called hybrid approaches have been devised to overcome some of these limitations. These approaches combine interface matching with information retrieval techniques. Research has shown that hybrid approaches outperform ontology-based approaches by increasing precision and recall during service discovery [19,20]. However, regardless of whether it is hybrid or ontology-based, interface matching techniques lack information describing how services operate.

These deficiencies led Martin-Recuerda and Robertson [21] to request finer-grained classifications of web services. Chen and Jiao [22] met this challenge in a UK e-Science project [22] in which they developed a number of insights regarding service discovery. In particular, they note the need for *provenance* in service discovery. Historically, provenance referred to the history and lineage of an object. In this context, provenance refers to the creation and specification of the web service and includes such information as runtime environment, settings, and algorithms used. Chen and Jiao [22] highlight the need to find services with specific or equivalent algorithms for specific types of applications. Such information is not readily available in most web service descriptions and thus is not included in many discovery applications. They conclude that service discovery needs to include provenance information at multiple levels of abstraction and over multiple facets. While they offer a potential solution to this problem, their solution was

Please cite this article as: T. Narock, et al., A provenance-based approach to semantic web service description and discovery, Decision Support Systems (2014), http://dx.doi.org/10.1016/j.dss.2014.04.007