Towards Automatic Discovery of Web Portals
-Semantic Description of Web Portal Capabilities

Haibo Yu¹, Tsunenori Mine², and Makoto Amamiya²

¹ Graduate School of Information Science and Electrical Engineering
Kyushu University
6-1 Kasuga-koen, Kasuga, Fukuoka 816-8580, JAPAN
yu@al.is.kyushu-u.ac.jp

² Faculty of Information Science and Electrical Engineering
Kyushu University
6-1 Kasuga-koen, Kasuga, Fukuoka 816-8580, JAPAN
{mine, amamiya}@al.is.kyushu-u.ac.jp

Abstract. Due to the problem of information overload, locating relevant Web portals precisely based on user requirements is quite an essential task. As the need for application-to-application communication and interoperability grows, providing Web portal services that satisfy human as well as machine requirements is becoming a new challenge for Web portals. However, a Web portal capability expressing mechanism, which enables the precise location of Web portals as well as the automated discovery and invocation of Web portal services, is lacking. In this paper, we investigate how to incorporate Semantic Web technology with Web service technologies to describe the capabilities of Web portals. We also discuss the possibilities of using these descriptions for discovering and using the distributed existing portal resources.

1 Introduction

Web portals are information rich sites that gather a variety of useful information from different resources into a single “one stop” Web page and provide it in a compact and easily consumable form to an end-user [1]. However, locating relevant Web portals is quite a challenging task because of the problem of information overload. It caused us to reconsider user requirements on the Web. The following questions come to mind: (1) Are there any better ways to locate relevant Web portal resources precisely based on users’ requirements? (2) Is it possible to make use of the relevant heterogeneous Web portal resources automatically no matter what framework they are based on? (3) Furthermore, is it possible to build a user’s own personalized information warehouse (MyPortal) in her/his computer with these Web portal resources and use it conveniently, even sharing it with other people? In this paper, we are trying to answer the first question, as a preliminary step towards answering the latter two.

Locating Web portal resources should be based on a match between user requirements and Web portal capabilities. This requires a mechanism for expressing the capabilities of Web portals. Currently, there are some standards
used for describing the capabilities of Web sites [1] [2], however they are generally used for the aggregation of Web sites or portlets into a Web portal, not for application computing purposes. So an explicit description of Web portal capabilities, which can support automatic discovery of the Web portal as well as its services, is lacking.

Semantic Web [3] is an evolving technology which aims to tackle the information overload problem of the current Web. In the Semantic Web, the information is given well-defined meaning, better enabling computers and people to work in cooperation.

Web service mechanisms provide a good solution for application interoperability between heterogeneous environments. They are standard programmatic interfaces between applications that provide a new model which enables Web sites to exchange dynamic information on demand.

In this paper, we investigate how to incorporate Semantic Web technology with Web service technologies to support the description of Web portal capabilities, trying to enable the precise location of relevant distributed existing portal resources and their maximum reuse. The advantage of our approach is that we provide a mechanism for describing Web portal capabilities that not only enables precise and automatic discovery, but also enables the application to use the Web portal resources after they are located. Since we use standard ontology language and Web service technology, common existing applications, tools and resources can be used.

The rest of the paper is organized as follows. Section 2 briefly describes the basic technologies this research is concerned with. Section 3 describes a mechanism for the description of Web portal capabilities. Section 4 examines the semantic matching algorithm. In section 5 we discuss the possibilities of using these descriptions for discovering portal resources, to enable the maximum reuse of existing distributed portal resources. Related work is discussed in section 6 and the concluding remarks will be summarized in section 7.

2 The Basic Technologies

We next briefly introduce some of the technologies this research is concerned with.

2.1 Semantic Web

The Semantic Web [3] is an extension of the current Web. It is trying to change the current Web into a huge knowledge base with well-defined meaningful data enabling machines to cooperate with people to tackle the problem of information overload.

RDF (Resource Description Framework) [4] is a metadata modeling language recommended by W3C (World Wide Web Consortium). It provides a common framework for expressing information so it can be exchanged between applications without loss of meaning. It uses XML as an interchange syntax.
Just as people need a common language to communicate with each other, machines also need one in order to share knowledge and to communicate with each other. Ontology is viewed as a dictionary that can satisfy this requirement. RDFS [5] is an ontology language which can formally describe the meaning of terms used in semantic material, and define their relationships and properties. In order to describe more complicated data relationships and perform useful reasoning tasks, OWL [6] was designed and recommended by W3C. These emerging foundation technologies of the Semantic Web have been accepted gradually, and tools for validation, annotation, authoring and editing have also been developed. The ontologies of certain domains have been developed too.

2.2 Web Services

A Web service is “a software system identified by a URI, whose public interfaces and bindings are defined and described using XML. Its definition can be discovered by other software systems. These systems may then interact with the Web service in a manner prescribed by its definition, using XML based messages conveyed by Internet protocols[7]”. Web services transform the Web from a collection of information into a distributed computation device.

The service will be instantiated between a requestor and a provider. Standards for description, registration and location, and accessing of Web services such as WSDL[8], UDDI[9] and SOAP[10] have emerged and are widely used currently. But they are based on keywords and lack semantics, and appropriate semantic description of Web service capabilities is necessary in order to enable automatic service discovery and invocation.

OWL-S[11] is “an OWL-based Web service ontology, which supplies Web service providers with a core set of markup language constructs for describing the properties and capabilities of their Web services in unambiguous, computer-interpretable form”. It enables automated Web service discovery, execution, composition and interoperation. OASIS\(^3\) has started to discuss UDDI support for semantic search[12]. Research on Semantic Service Matchmaker[13] with UDDI business registry is also ongoing.

3 The Description of Web Portal Capabilities

In this section, we explain our mechanism for semantically describing Web portal capabilities.

There are various kinds of Web portals for various purposes with different levels of functions. In this paper, we only focus on the research community Web portal and target the Semantic Web community domain as a starting point. For the purpose of future flexibility, we generalize the concept of Web portal to any information-rich Web sites that want to publish their resources, including individuals, group or project Web sites as well as community Web portals.

\(^3\) Organization for the Advancement of Structured Information Standards
3.1 The Structure of Web Portal Capability Description

Here we refer to a Web portal which has certain capabilities as “provider”, and the user or the application that is searching for Web portal resources as “requestor”.

The purpose of describing the capabilities of a provider is to facilitate the search and use of requestors. Let’s think about the human searching process: generally we match requests and capabilities hierarchically from general ideas to details. This observation can be applied to machine processing as well.

We describe the capabilities of the Web portal by layers. First, we semantically describe the general capabilities of the Web portal, and we call this a “site capability summary (SCS)”. Second, we describe its “service capabilities”. There is a link from the site capability summary to the service capability description. In order to semantically describe the capabilities and support the concrete realization of services, we express the service capability in two layers: “semantic Web service description” and “Web service description”. So the structure of capability description can be illustrated as seen in Figure 1.

```
layer 3 | site capability summary |
--------|--------------------------|
|        | -------------------------|
layer 2 |->| semantic Web service description |
        |-------------------------|
|        |        |-------------------------|
layer 1 |->| Web service description |
        |-------------------------|
```

**Fig. 1.** Structure of capability description.

This hierarchical capability-describing mechanism enables semantic capability-describing and matchmaking for different levels. There are links between these description layers, but the communication of each layer between provider and requestor can be done independently. The Web service description layer (layer 1) can be WSDL or any other service description method which is being used by the current Web service system. The semantic Web service description layer (layer 2) can be OWL-S or any other semantic service description method.

With this layered description mechanism, providers can describe simple or complicated capabilities and requestors can discover and invoke potential portal resources according to their preferences. For example, a simple Web site which only provides browsing content capability can be described with site capability summary without the service capability descriptions. A requestor can locate the Web portal without using its layer 1 and 2 services or only use the services of layer 1 without semantic capability. So it’s flexible and robust, and can satisfy all the possible semantic and non-semantic uses. Here we use the well-known service
Semantic Description of Web Portal Capabilities

### Table 1. The site capability summary

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
<td>The type is an important characteristic which can identify the Web portal capability. It can be “individual”, “organization” or “verticalPortal”. The organization can be an “institution”, “enterprise”, “government”, “community”, “association”, “project”, “group”, or “others”. The vertical portal can be a “digitalLibrary”, “e-Journal”, “openSource”, “event”, “news” or “others” for specific interests. This information would be useful when user can explicitly express his requirement for certain types of Web portals.</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>This represents the languages supported by the Web portal. We can use existing standard categories of languages to represent it.</td>
</tr>
<tr>
<td><strong>Scale</strong></td>
<td>The scale is a rough estimation result. It can be “small”, “medium” or “large” to help user identify proximate scale.</td>
</tr>
<tr>
<td><strong>Audience</strong></td>
<td>This is used to identify potential user level. It can be “elementary”, “intermediate” or “expert”. The user with a specific level can make use of this information to locate the most relevant Web portal.</td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td>This will define the location of the Web portal. It would be useful when considering network delay or usability. We can use existing standards such as ubr-uddi-org:iso-ch:3166-2003 to describe country and city.</td>
</tr>
<tr>
<td><strong>HomePageLink</strong></td>
<td>This is the URL of the Web portal itself. When a user just wants to locate the URL of a Web portal or when a portal does not support Web services, it will be used to connect the user to the page of the Web portal.</td>
</tr>
<tr>
<td><strong>ServiceLink</strong></td>
<td>The service link is an URL, used as a connection from the site capability summary to the Semantic Web service descriptions. When a user has located a Web portal and wants to use the Semantic Web services, it will be used to reach the Semantic Web services.</td>
</tr>
<tr>
<td><strong>Security</strong></td>
<td>It describes the security rule of the Web portal. It can be “private”, “community-Only” or “public”.</td>
</tr>
<tr>
<td><strong>Functionalities</strong></td>
<td>The functionalities is a list of all functions that the Web portal supports. From this list, one can check if the Web portal can satisfy certain requirements or not.</td>
</tr>
</tbody>
</table>

description methods OWL-S and WSDL for the service description of layer 2 and layer 1 respectively in order to maximize the reuse of current resources. For the details of WSDL and OWL-S, one can refer to the relevant documents [8, 11].

In order to semantically describe Web portal capabilities, we need to construct relevant ontologies to model the real-world concepts, create a WSDL document to describe the Web services, use OWL-S to semantically describe the Web services, and create the site capability summary.

We’ll define the description of layer 3 “site capability summary” in next sub-section, discuss Web portal functionalities after that, and then discuss the relevant ontologies.

#### 3.2 The Web Portal Site Capability Summary (SCS)

We argue that some explicit general ideas (site summary) are strongly required in order to precisely locate Web portals based on user preferences. So a brief capability summary of the Web portal is necessary. The site capability summary gives an explicit overview of the Web portal capabilities, and can be used as the initial filter for judging congruence with user preferences. The detailed semantic representations of the contents such as topics, projects, publications, can be found from the contents metadata which is constructed based on domain specific ontology. We will omit this part in this paper because of the space limitation. We define the items of the site capability summary in Table 1.
Table 2. The functionalities of Web portals

<table>
<thead>
<tr>
<th>Function Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Browsing</td>
<td>This function is used for browsing the contents on the Web portal based on topics, projects, years and so on.</td>
</tr>
<tr>
<td>Searching</td>
<td>This function is used for searching relevant resources based on user requirements. It can search from internal or external resources. The searching function can provide simple or advanced searches such as match all in any order, match any, match as a phrase etc. It can search materials based on year(single/a range of year), topic etc. The powerful search function will support search bibliographic database, search citation databases, search e-journals and search events such as conferences.</td>
</tr>
<tr>
<td>Personalization</td>
<td>The personalized Web portal will enable users to change various aspects to suit themselves. For example, the user can control the appearance of information on her screen, and can change the alert item too. Personalization is strongly required by users as they want to avoid information overload.</td>
</tr>
<tr>
<td>Utilities</td>
<td>Some utilities such as “address book”, “calendar”, “people finder” can be used to help users.</td>
</tr>
<tr>
<td>News/Newsfeeds</td>
<td>The world or domestic news, technology news, research subject-specific news, job advertisements and alerts can be communicated by email or other means or be aggregated into the Web portal.</td>
</tr>
<tr>
<td>Community Communi-</td>
<td>The communication between members of a community or group can be realized by chat, newsletters, message boards or newsgroups.</td>
</tr>
<tr>
<td>Advertising</td>
<td>Announcements such as conferences, non-leisure events can be advertised through this function.</td>
</tr>
<tr>
<td>Teaching and Learn-</td>
<td>Web-based learning resources, courses or course announcements can be provided by this function.</td>
</tr>
<tr>
<td>ing Information</td>
<td></td>
</tr>
<tr>
<td>Assistance With Site Use</td>
<td>This function will support and guide users in using Web sites. It can be an immediate help, a help page or a feedback option.</td>
</tr>
<tr>
<td>Additional Features</td>
<td>The Web portal can provide other features such as job searching, online resource submission, event schedule management or registration into the portal.</td>
</tr>
</tbody>
</table>

3.3 The Functionalities of the Web Portals

As a result of thorough investigation by the JISC Subject Portals [14] and PORTAL [15] projects, the features of a portal (including institutional and commercial ones) has been summarized. We referenced their results and examined some typical Web portals, then extracted the main functionalities that we think should be included in a community Web portal. These are described in Table 2.

3.4 Relevant Ontologies

The description of Web portal capabilities must be based on formally defined vocabularies in order to make it machine-understandable and processable. Ontology is used to formally define terms and the relationships between them. The Web portal capability ontology should also include the following component ontologies as well as the ontology of Semantic Web services.

1) The Site Capability Summary Ontology: In this ontology component, the terms used for the site capability summary such as “type”, “location”, and the relationship between and restrictions on them are formally defined.

2) The Semantic Web Research Community Ontology: There are some existing ontologies for the computer science research domain such as KA2 [16] and CS AKTive Space [17]. We reuse these ontologies and modify them for
our purpose to construct our ontology of research community domain. The ontology defines terms such as Organization, Person, Publication, Events, Topics etc. The relationships between them such as the subclass, subproperty, restriction of range and domain are also defined. The Semantic Web involves various research fields, such as Databases, AI, networks, telecommunication, information retrieval, data mining, programming languages, logic, security, Web services etc. So the Semantic Web community involves a broad range of topics. It includes the foundation languages and framework, tools, knowledge sources, applications, technologies, tutorials and so on.

3) The Web Portal Functionality Ontology: This ontology component defines all the terms, relationships, and restrictions concerning the Web portal functionalities such as “browsing”, “searching”.

4 The Semantic Matching Algorithm

In this section, we explain our semantic matching algorithm used for the matching between user requirements and Web portal capabilities.

Locating a Web portal and its web services is a process of semantic matching between the requirements of the user and the capability description of a Web portal. The capabilities of a Web portal can be simple or complicated and the requirements of a user can also be quite varied. The matchmaking needs to deal with all possible situations including non-semantic use. For brevity, we use “R” to represent requestor, “P” to represent provider.

4.1 The Provider Capabilities

A Web portal may only support contents which can be used through a browsing interface or may also support Web services. So the capability description of a Web portal can be summarized as follows based on its capability.

P provides contents only:
- P contains capability summary and contents metadata

P provides contents and services:
- P contains capability summary, contents metadata, semantic service description, and service description

4.2 The User Requests

A requestor(user) can be semantic-capable or not. She may only want to locate the Web portal or also want to locate and use the Web portal services. So the request from the requestor can be summarized as follows based on her purpose.

R wants to locate Web portal only:
- The request contains user preferences and contents query

R wants to locate and use Web portal services only:
- The request only contains service request
R wants to semantically locate and use Web services only:
The request only contains semantic service request
R wants to locate Web portal and semantically locate and use Web portal services:
The request contains user preferences, contents query and semantic service request

4.3 The Matching Process

As we use common standard Semantic Web service mechanisms, the matching
of the Web service capabilities can use the same methods proposed by other
research projects such as [18]. Non-semantic Web portal services can also make
use of existing matching methods such as UDDI. The semantic request and
response will use RDF format.

The matchmaking process of provider can be briefly summarized as follows.

Case 1: P provides contents only
(1) R wants to locate Web portal only
   Comparing the preferences of requestor and the capability summary
   of Web portal
   if match then compare contents metadata with contents query
   if match then return relevant information
   else end
   else end
(2) R wants to locate Web services only
   end
(3) R wants to semantically locate Web portal services only
   end
(4) R wants to locate Web portal and semantically locate Web portal services
   Comparing the preferences of requestor and the capability summary
   of Web portal
   if match then compare semantic service description and service request
   if match then invoke relevant services
   else response HomePageLink only
   else end

Case 2: P provides contents and services
(1) R wants to locate Web portal only
   Comparing the preferences of requestor and the capability summary
   of Web portal
   if match then compare contents metadata with contents query
   if match then return relevant information
   else end
   else end
(2) R wants to locate non-semantic Web services only
   Comparing non-semantic Web service descriptions and service request
   if match then invoke relevant services
   else end
(3) R wants to semantically locate Web portal services only
   Comparing semantic service description and semantic service request
   if match then invoke relevant services
5 The Possible Uses of the Description of Community Web Portal Capability

The location of Web portals, and the discovery and invocation of Web portal services can be realized in various ways. The matchmaker can be a distributed P2P system or a centralized system. Here we discuss the possible ways that we think reasonable and mainly focus on the usage of a multi-agent community-based P2P information retrieval system.

5.1 Centralized Solution

UDDI is a standard for registration, navigation and location of Web services. It was widely accepted by the industry and there are many applications that are based on it. But it lacks semantic capability and is limited to value sets and direct matching of values. Though it is in its early stages, and there are no solutions yet, work has begun on supporting semantic capabilities for categorization [12]. And experimental research [13] on semantic service search with the public UDDI registry is also ongoing. Our Semantic Web community Web portal services can be registered in this registry server and can be discovered and invoked based on a semantic-enabled UDDI standard. The semantic description in OWL-S and the service description in WSDL of Web portal services will all be registered into the UDDI registry. So non-semantic service requests can also be processed. We will trace the research situation of semantic UDDI standards and investigate the realization details in the future.

5.2 Peer to Peer Solution

Research community members are generally loosely coupled people, distributed across many locations and organizations. They are the community resource consumers and providers simultaneously. The peer to peer model is very natural for modeling community information concerns. Here we discuss a community-based multi-agent P2P retrieval system designed to discover community Web portal services.

In our system, we use KODAMA [19] multi-agent system to model the research community. Each community member is modeled as a peer and there is one or more than one agents to serve each community member. They help the
community member to locate the relevant resources and make use of the services that they need.

Location of the Web portal resources should be based on a semantic match between the explicit description of the user requirements and the description of the capability of the Web portal. The capabilities of the Web portal are described in the way we explained in section 3. The requestor and the provider share the same ontology to describe their preferences and capabilities respectively.

If a user wants to locate the Semantic Web community Web portal, the contents to be reasonable for a beginner, and wants to locate and use the semantic service of “Searching”, then his agent will add his searching preferences type=“community”, Audience=“elementary”, the contents request of “SemanticWeb” and semantic service request into his request. The Semantic Web community Web portal will find that its capability matches the request preferences and the contents metadata of topics match the contents request. It will continue to do semantic matchmaking of services, and invoke the relevant searching services according to the request.

In order to reduce the message traffic of the P2P network, we make use of the historic records of queries sent to and received from other agents[20].

6 Related Work

In this section, we discuss some related work that is directly or indirectly a concern of our research work.

OASIS released the Web services for Remote Portlets Specification Version 1.0 [1] in August 2003. It defines a web service interface for accessing and interacting with interactive presentation-oriented web services between producer and consumer. Its goal is to enable an application designer or administrator to pick from a rich choice of compliant remote content and application providers, and integrate them conveniently. They use keywords and function calls which lack the semantics that is essential for automatic machine processing. Our description of Web portal services is based on standard semantic Web service ontology OWL-S, which enables automatic machine processing.

RSS[2] is a lightweight multipurpose extensible metadata description and syndication format. It is an XML application and conforms to the RDF specification. A brief description of the Web site capability can be summarized with RSS and the summary can be used for online publication, retrieval and further transmission or aggregation. But the resources of the summarized Web site cannot be used as a computational part of the application. Our description is based on Web service technology, the resource of the portal can not only be located but also used as a computational part of the application.

The Subject Portals project [14] is a project founded by JISC\(^4\). It builds portals targeted at human users through a Web browsing interface. They support back-end office services for portals but not for portal services targeted at application computing usage, as far as we understand.

\(^4\) Joint Information Systems Committee
There are Web portals based on Semantic Web technology, such as KA2[16] and SEAL[21], which support a semantic portal solution including ontology-based contents construction and maintenance, but they are generally based on human navigation and searching. SEAL provided an interface for a software agent but only for a crawler. None of them supports Semantic Web portal services at present, as far as we know.

Francisco et al. presented an architecture for an infrastructure to provide interoperability using trusted portals[22] and implemented such an infrastructure based on Thematic Portals. The searching portals use semantic access points based on metadata for more precise searching of the resources associated with the potential sources of information. The proposed architecture support specific and cross domain searching, but only provide semantic representation for the capabilities of Web portals not for their services as we understand. We are concentrating on the Semantic Web community domain specific searching but with a strong semantic describing ability not only for the capabilities of Web portals but also for their services.

OWL-S is an ontology of services which provides a mechanism for semantically expressing the capability of Web services. In our approach, we use OWL-S to describe the Web portal service capabilities, and we also add another “site capability summary” layer above it. This will help in the precise location of Web portals as well as the discovery and invocation of Web services.

Some other Semantic Web service solutions [23] are also proposed but none of them for Web portal services as far as we know.

7 Conclusion and Future Work

In this paper, we proposed a mechanism for semantically describing the capabilities of community Web portals, enabling automatic discovery of Web portals as well as Web portal services. We also discussed the possible use of these descriptions to discover Web portal resources. In our future work, we would like to realize the details and implement a prototype to reveal the possibility and effectiveness of our proposed solution, and reuse the resources of the automatically discovered portals to aggregate them and construct user-personalized warehouses (Myportal) based on a multi-agent P2P information retrieval system. Currently, we assume that all the portals, users and agents in a community agree on a common ontology and use it to represent the semantics of Web portal capabilities and Web services, but it’s not easy to get this agreement in reality. We need to give further consideration to these ontology-mapping issues in the future.

References

2. RDF Site Summary (RSS)1.0, http://web.resource.org/rss/1.0/
11. OWL-S: Semantic Markup for Web Services
15. PORTAL project, http://www.fair-portal.hull.ac.uk/
17. Nigel R.Shadbolt, monica m.c. schraefel, Nicholas Gibbins, Stephen Harris, CS AKTive Space: or How We Stopped Worrying and Learned to Love the Semantic Web, Proceeding of International Semantic Web Conference(ISWC) 2003
21. SEmanuel portAL (SEAL), http://ontobroker.semanticweb.org/ontos/aib.html
27. http://www.semanticweb.org/