2009 International Conference on Advanced Information Networking and Applications Workshops

A Hybrid Service Metadata Clustering Methodology in the Digital Ecosystem Environment

Hai Dong, Farookh Khadeer Hussain, Elizabeth Chang
Digital Ecosystems and Business Intelligence Institute
Curtin University of Technology
Perth, Australia
{hai.dong, farookh.hussain, elizabeth.chang}@cbs.curtin.edu.au

Abstract— **Digital Ecosystem is defined as "*an open, loosely coupled, domain clustered, demand-driven, self-organizing and agent-based environment, in which each species is proactive and responsive for its own benefit and profit*" [1]. Species in the Digital Ecosystem can play dual roles, which are service requester (client) service provider (server). A service provider enters the Digital Ecosystem by publishing a service metadata in the service factory, in which the service metadata can be clustered by domain-specific ontologies provided by the Digital Ecosystem. Two issues emerge here. First of all, vast and heterogeneous service metadata are ubiquitous before the Digital Ecosystem technology emerges. It is a challenge for the Digital Ecosystem to organize these metadata. In order to solve this issue, an automatic service metadata clustering approach could be desired. However, this could educe the second issue – the automatic association between service concepts and service metadata could not agree with service providers' perceptions, as a result of the differences among individual understandings. To solve the two issues, in this paper, we present a hybrid ontology-based metadata clustering methodology comprising an extended case-based reasoning algorithm-based automatic concept-metadata association approach and a service provider-oriented concept-metadata association approach.**

Keywords-Digital Ecosystems; extended case-based reasoning algorithm; metadata clustering.

I. INTRODUCTION

Digital Ecosystem is defined as "*an open, loosely coupled, domain clustered, demand-driven, self-organizing and agent-based environment, in which each species is proactive and responsive for its own benefit and profit*" [1, 2]. The emergence of Digital Business Ecosystem can be attributed to the natural existence of business ecosystem, along with the evolution of business network and information technology. Species in the Digital Ecosystem can play dual roles, which are service requester (client) that needs services from other entities and service provider (server) that provides services. A service provider enters the Digital Ecosystem environment by publishing a service entity, which will be stored into distributed service knowledge base [1, 2]. Here these service entities are stored in the format of metadata [3]. Service factory is a group of functional components within the Digital Ecosystem environment, which allows a service provider to create and test a service metadata [11]. When a service provider publishes a service metadata, by means of the service factory, the service metadata can be clustered by domain-specific ontologies provided within the Digital Ecosystem, by referencing the IRI (Internationalized Resource Identifier) of the metadata to the ontological concepts [3, 4].It is crucial to note that, before the emergence of Digital Ecosystem, service metadata has already been ubiquitous without sufficient ontological support. Moreover, there is no existing methodology for clustering the ubiquitous metadata in the existing literature, which reduces the retrieval performance towards these metadata. Hence, owing to the huge amount of the less-semantic service metadata, there is an urgent need for an automatic ontology-based service metadata clustering methodology. Such a methodology could be used to cluster service metadata, thereby resulting in an improvement in service metadata retrieval. Nevertheless, the proposed methodology could educe another issue caused by the essence of ontology – an explicit specification of a conceptualization [5]. In other words, some ontologies provided by the Digital Ecosystem could represent subjective understandings towards certain domain knowledge (because ontology can be subjective [10]). However, in some cases, the knowledge (ontologies) may not be agreed to by every service provider since everyone has individual understanding about certain knowledge. As a result of this, the automatic clustering based on these ontologies could not satisfy some service providers, which educes another research question – how to make an agreement between the outcome of the automatic metadata clustering approach and service providers' individual perception towards the metadata clustering?

In this paper, we will present a hybrid metadata clustering methodology, in order to synchronously solve the two research issues identified above for metadata clustering in the Digital Ecosystem environment. In order to address the first research question, we use a novel set-theoretic model to associate service metadata with relevant service concepts; in order to address the second, we utilize a customized approach that enables service providers to customize the association between service metadata and service concept.

978-0-7695-3639-2/09 $25.00 © 2009 IEEE
DOI 10.1109/WAINA.2009.205

The rest of the paper is organized as follows: in Section 2, we will present the system framework of our proposed hybrid methodology. In Section 3, we will respectively introduce and explain each part of the system. In order to evaluate our methodology, we implement a prototype in a specific domain, and test the performance of the prototype based on several information retrieval benchmarks. The conclusion of the evaluation process and the direction for future work are presented in the final section.

II. System Architecture

The system architecture of our proposed hybrid metadata clustering methodology is shown in Fig. 1, which consists of two basic parts: 1) a service knowledge base, and 2) a hybrid association module. The service knowledge base is designed to store service metadata and service ontologies. The hybrid association module is used to associate the IRIs of metadata with the IRIs of the semantically relevant ontological concepts, in order to cluster metadata based on service ontologies. Detailed description about the service knowledge base and the hybrid association module is described in the following sections.

III. Service Knowledge Base

As described earlier, service knowledge base contains two categories of metadata, service metadata and service ontology metadata. In this section, we will respectively define the format of both categories of the metadata, as a prerequisite of the forthcoming hybrid metadata clustering approach.

A. Service Metadata Format

The major objective of a service metadata is to represent descriptive information about a service provided by a service provider in the Digital Ecosystem environment. It is important to note that a service metadata only belongs a service provider, Namely, even though two service providers provides same service in the real environment, the two services still need to be conceptualized as two service metadata owing to different service providers.

The Service metadata can be represented as a tuple where the elements of the tuple can be complex elements as defined below:

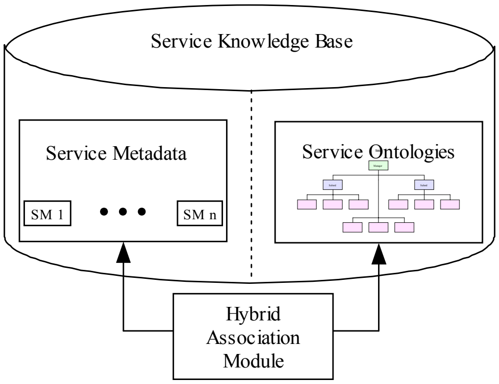

Figure 1. System architecture.

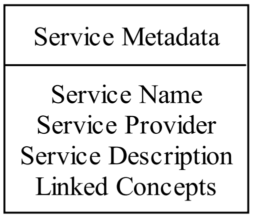

Figure 2. Service metadata format.

[Service Name, Service Provider, Service Description, Linked Concepts] (Fig. 2) where

Service Name refers to the identity used to uniquely identify a given service.

Service Provider refers to the identity of the person or organization that provides the service.

Service Description refers to the detailed text description about the service. This property can be used for matching it with a service concept. This process would be described later.

Linked Concepts refer to the IRIs of the semantically associated service concepts to the service metadata.

B. Service Concept Format

The service ontology is a hierarchical structure of services concepts that are related by generalization-specification relationships.

We present the Service Ontology as the combination of the ontology name and a tuple where the elements of the tuple can be complex elements as defined below:

Service [Concept Name, Concept Description, Linked Metadata] (Fig. 3) where

Concept Name refers to the identity used to uniquely identify a service concept.

Concept Description refers to the definitional descriptions of the service concept. The normal form of a service description is a set of words (noun, adjective, or adverb). A service concept may have many service descriptions. The purpose of setting the property of service description is to compute the semantic similarity value between the service concept and service metadata, which will be introduced later.

Linked Metadata refers to the IRIs of the semantically associated service metadata to the service concept.

IV. Hybrid Association Module

The hybrid association module combines an automatic association sub-module and subsequently a customized

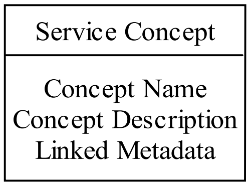

Figure 3. Service concept format.

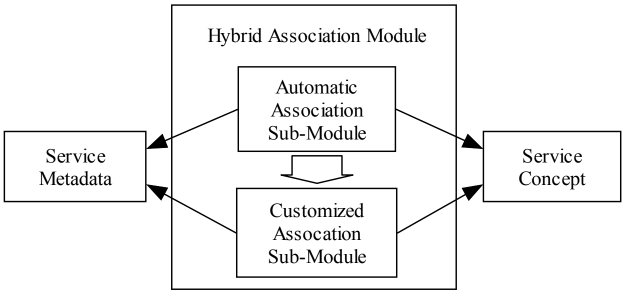
Hybrid Association Module
Automatic Association Sub-Module
Customized Assocation Sub-Module
Service Metadata
Service Concept

Figure 4. Hybrid association module.

association sub-module, in order to associate service metadata with service concepts with higher precision and user satisfaction (Fig. 4). In this section, we will explain the working mechanism of the two sub-modules in detail.

A. Automatic Association Sub-module

The automatic association sub-module is built upon an ECBR (Extended Case-Based Reasoning) model, which is an index term-based set-theoretic model [6].

The ECBR model is to calculate the similarity value of a service concept c_j to a service metadata m, as shown in (1) below:

$$sim(cd_j, sd) = \frac{\sum_{j=1}^{m} f(cd_{k_j}, sd)}{m} \tag{1}$$

$$f(cd_{k_j}, m) = \begin{cases} 1 \text{ if } \exists sd_k | (\forall k_i, g_i(cd_{k_j}) = g_i(sd_k)) \\ 0 \text{ otherwise} \end{cases}$$

where cd_j is the content of the concept description property regarding a service concept, k_i is an index term; $cd_j = (cd_{k1}, cd_{k2} \ldots cd_{km})$, where cd_k is the index terms that occurs with cd_j, m is the number of index terms that occur with cd_j; sd is the content of the service description property regarding a service metadata; $sd = (sd_{k1}, sd_{k2} \ldots sd_{kn})$, where sd_k is the index terms that occur with sd, n is the number of index terms that occur with sd; g_i is a function that returns weight associated with k_i.

The association procedure is described as follows:

```
Input: C = ($c_1$, $c_2$…$c_m$) is a sequence of service concepts in a service ontology, M = ($m_1$, $m_2$…$m_n$) is a sequence of service metadata.

Procedure:

For i = 1 to n
  Fetch the service description of $m_i$ and store it into $sd_i$
  For j = 1 to m
    Fetch the concept description of $c_j$ and store it into $cd_j$
    Compute the similarity value of $sd_i$ and $cd_j$ by ECBR and store the value into $s_{ij}$
    If $s_{ij}$ > threshold then
      Put IRI of $c_j$ into the linked concepts property of $m_i$
      Put IRI of $m_i$ into the linked metadata property of $c_j$
    End if
  End for
End for
```

The metadata-concept association follows a simple rule – many to many, which can be interpreted as follows: (1) a service concept can be associated with more than one service metadata; (2) a service metadata can also be associated with more than one service concepts. The reason behind (2) is that a service entity may relate to many service concepts in the real environment.

B. Customized Association Sub-module

As described earlier, the automatic association sub-module can be used to match service metadata with service concepts. A key issue that needs to be considered is whether the matching really reveals every service providers' perception to the relationship between their service metadata and the service domain knowledge (ontology)? The answer to this question would be no as in the real environment diversity of thought is one of the most essential characteristics of human being. Subsequently, the issue that needs to be addressed is how to keep the reliability of the metadata clustering? In other words, how can we enable personalized/customized semantic metadata clustering while at the same time taking account of end-user's perception? In the past few years, personalization/customization through information technology (specifically via the World Wide Web) has become an increasingly significant trend in multidiscipline [12], e.g., personalized search in Google^{TM} [7], personalized recommendation system in Amazon.com [13], etc. Here we design a customized association sub-module, in order to complement the potential defect of the automatic association sub-module.

The function of this sub-module is that after a service metadata has been clustered by the crawler, the service provider who owns the service metadata can modify (add/remove) the associations between the metadata and ontological concepts according to their individual point of view. It is believed that, by the personalized association, the service provider-oriented service metadata can be clustered in a more reasonable manner, due to the satisfaction of individual preference toward the concept-metadata association.

V. Implementation and Evaluation

In order to evaluate the hybrid service metadata clustering methodology that we present in this paper, we implement a prototype and carry out experiments based on two information retrieval benchmarks.

A. *Prototype Implementation*

The prototype implementation is divided into two steps according to the system architecture.

The first step is the construction of the service knowledge base. As mentioned before, service knowledge base consists of domain-specific service ontologies and service metadata. First of all, we choose transport service as the specific domain for ontology construction and we build a transport service ontology with Protégé-owl. Subsequently, in order to harvest the instances of service metadata, we design a semantic crawler by Java. The semantic crawler is able to generate service metadata by extracting service information from normal HTML documents and annotating the service information with OWL (Web Ontology Language). The in-depth information with regard to the transport service ontology and semantic crawler can be found from [6, 8]. Additionally, in order to harvest the transport service sources, we choose the business webpages under the transport service category of Australian Yellowpages® website as the domain where the semantic crawler works.

The second step is to build a hybrid service metadata clustering module. This module is implemented by Java. Fig. 5 is the screenshot of the customized association interface. In this example, the service provider "*Airport executive chauffer*" publishes a service metadata "*taxi cabs*". The automatic association module associates the metadata with the ontological concept "*taxi*". By making use of this interface, the service provider can publish a new service metadata or remove this service metadata by clicking the "*add*" or "*remove*" button. Furthermore, the service provider can modify the properties of the metadata by clicking "*edit*" button, including its service metadata name, service description and linked concepts assigned by the automatic association module.

B. *System Evaluation*

To evaluate the feasibility of the hybrid metadata clustering approach, we carry out experiments and determine the precision and recall of the automatic association process. In addition, for the ECBR model utilized by the automatic association module, a proper threshold needs to be decided to filter the irrelevant concepts for metadata. Hence, another purpose of the evaluation is to find an appropriate threshold value where the automatic association sub-module can deliver optimized performance.

Precision in information retrieval is used to measure the preciseness of a retrieval system [9]. Precision for a single concept is the proportion of associated and semantically relevant metadata in all associated metadata to the concept, which can be mathematically as shown below:

$$\text{Precision(S)} = \frac{\text{no. of associated \& semantically relevant metadata}}{\text{no. of associated metadata}} \quad (2)$$

With regard to the whole collection of concepts, the whole precision is the sum of precision for each concept normalized by the number of concepts in the collection, which can be mathematically as shown below:

$$\text{Precision(W)}=\frac{\sum_{i=1}^{n}\text{Precision}(\text{S}_i)}{n} \quad (3)$$

Recall in information retrieval refers to the measure of effectiveness of a query system [9]. Recall for a single concept is the proportion of associated and semantically relevant metadata in all semantically relevant metadata, which can be mathematically as shown below:

$$\text{Recall(S)} = \frac{\text{no. of associated and semantically relevant metadata}}{\text{no. of semantically relevant metadata}} \quad (4)$$

With regard to the whole collection of concepts, the whole recall is the sum of recall for each concept normalized by the number of concepts in the collection, which can be mathematically as shown below:

$$\text{Recall(W)}=\frac{\sum_{i=1}^{n}\text{Recall}(\text{S}_i)}{n} \quad (5)$$

We employ the semantic crawler to download 1000 transport service business webpages from the Australian Yellowpages® website. As a result of this process, we obtain 2261 service metadata. Then the automatic association sub-module is utilized to cluster the metadata by means of the transport service ontology. According to a peer-review approach, we obtain the performance of the automatic association sub-module on the indicators of precision and recall.

Fig. 6 shows the performance of the automatic association sub-module on precision with the variation of threshold from 0.5 to 1.0. As higher threshold can filter more semantically non-relevant metadata and reduce the number of associated metadata, the precision rises sharply during the

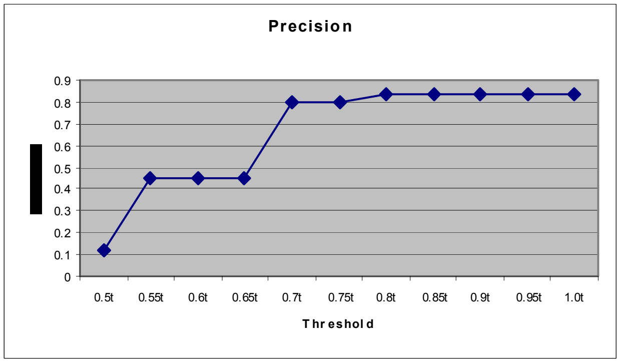

Figure 6. Precision at different thresholds.

period. Obviously the peak value is gained when threshold is higher than 0.8.

Fig. 7 shows the performance of the automatic association sub-module on recall with the threshold varying from 0.5 to 1.0. Since higher threshold probably filter more semantically relevant metadata, the number of associated and semantically relevant metadata may decrease analogously with the threshold. Thus, the recall curve experiences a digressive course. As shown in Fig. 7, this trend stops until threshold reaches to 0.7.

As the variation extent threshold value in this experiment is 0.8. At this point, the precision of the automatic association module is above 0.8, and its recall keeps beyond 0.98, which can be considered as sound performance. Therefore, this experiment primarily proves the feasibility of the automatic association sub-module. Since the variation of recall is relatively smaller than it of precision, we only need to pick the threshold point where precision reaches to the highest value and ignores recall. Finally it can be easily determined the prime threshold is 0.8.

VI. CONCLSUIONS

In this paper, we presented a hybrid service metadata clustering methodology. The goals of this methodology are to assist the Digital Ecosystem in the process of clustering the ubiquitous service metadata and to enable personalized service concepts association with service metadata. The system architecture consists of two main parts – a service knowledge base and a hybrid service metadata clustering module. The service knowledge base stores the service metadata and domain-specific service ontologies. In order to simplify the task of computing the similarity between service metadata and service concepts, we unify the formats of service metadata and service concepts. Based on these formats, we design an ECBR model that is grounded in set theory. The hybrid service metadata clustering module combines an automatic association sub-module and a customized association sub-module. With the ECBR model, the automatic association sub-module can automatically cluster service metadata based on the ontologies stored in the service knowledge base; additionally, the customized association module allows service providers to modify the associations between their metadata and automatically associated concepts, apart from the external function of publishing and modifying metadata. To evaluate the feasibility of our proposed approach, we implement a prototype in the domain of transport service. By means of a semantic crawler and Protégé-owl, we create a transport service knowledge base containing transport service metadata and transport service ontology. Next, we implement the prototype of the hybrid service metadata clustering module in Java and carry out the process of experimentation. Finally, we evaluate the performance of the ECBR model based on the indicators of precision and recall. The experiment results primarily indicate the feasibility of the automatic association sub-module.

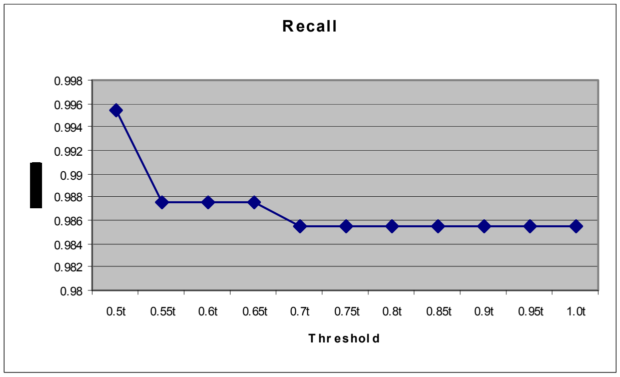

Figure 7. Recall at different thresholds.

In future, we plan to build a server-based service searching, publishing and managing prototype extended from the conceptual framework presented in this paper, and subsequently we intend to encourage numerous service providers to publish and edit services on this web-based platform, in order to thoroughly validate the effectiveness and efficiency of the combination of the automatic and customized metadata-concept association module. In addition, we will attempt to adopt other information retrieval models for the metadata clustering.

ACKNOWLEDGMENT

We would like to appreciate our programmer – Wei Liu who is responsible for the system implementation and evaluation. We also would like to express our gratitude to any reviewers of our manuscript for their valuable suggestions in order to improve this paper.

REFERENCES

[1] E. Chang, M. Quaddus, and R. Ramaseshan, "The vision of DEBI Institute: digital ecosystems and business intelligence: Digital Ecosystem and Business Intelligence Institute," DEBII, Perth 2006.

[2] E. Chang and M. West, "Digital Ecosystem - A next generation of the collaborative environment," in iiWAS2006, Yogyakarta, 2006.

[3] H. Boley and E. Chang, "Digital Ecosystems: Principles and semantics," in IEEE DEST 2007, Cairns, 2007.

[4] P.Malone, "Digital Ecosystem services in Ecosystem Oriented Architectures," in Digital Business Ecosystems, F. Nachira, P. Dini, A.Nicolai, M. L. Louarn, and L. R. Lèon, Eds.: European Commission, 2007.

[5] T. Gruber, "A translation approach to portable ontology specifications," Knowledge Acquisition, vol. 5, pp. 199-220, 1995.

[6] H. Dong, F. K. Hussain, and E. Chang, "A semantic crawler based on an extended CBR algorithm," in Lecture Notes in Computer Science: OTM 2008 Workshops. vol. 5333, R. Meersman, Z. Tari, and P. Herrero, Eds. Heidelberg: Springer-Verlag, Berlin, 2008, pp. 1084-1093.

[7] C. Sherman, "Google personalized search leaves Google labs ": Search Engine Watch, 2005.

[8] H. Dong, F. K. Hussain, and E. Chang, "Transport service ontology and its application in the field of semantic search," in 2008 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI 2008), Beijing, 2008, pp. 820-824.

[9] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. New York: Addison-Wesley, 1999.

[10] D. Hooijmaijers and M. Stumptner, "Improving integration with subjective combining of ontology mappings," in Foundations of Intelligent Systems. vol. 4994, A. An, Ed.: Springer Berlin, 2008, pp. 552-562.

[11] P. Dini, N. Rathbone, M. Vidal, P. Hernandez, P. Ferronato, G. Briscoe, and S. Hendryx, "The digital ecosystems research vision: 2010 and beyond," Creative Commons 2005.

[12] J. P. Bowen and S. Filippini-Fantoni, "Personalization and the web from a museum perspective," in Museums and the Web 2004, Arlington, 2004, pp. 63-78.

[13] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: item-to-item collaborative filtering " Internet Computing, vol. 7, pp. 76-80, 2003.