Journal of Network and Computer Applications 49 (2015) 78–87

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Efficient distributed semantic based data and service unified discovery with one-dimensional semantic space

Ying Zhang [a,*], Houkuan Huang [b], Hui He [a], Jing Teng [a], Zhuxiao Wang [a]

[a] School of Control and Computer Engineering, North China Electric Power University, Beijing, China
[b] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

ARTICLE INFO

Article history:
Received 27 March 2014
Received in revised form
18 September 2014
Accepted 30 November 2014
Available online 13 December 2014

Keywords:
Unified discovery
Distributed system
Semantic space
OWL-S
JXTA

ABSTRACT

Data and service discoveries are two significant applications in the Internet, and almost all the network functions need their support. A lot of research work has focused on either service discovery or data discovery respectively, although they cannot be separated. The existing network, due to its decentralized nature and weak support for semantic, is still chaotic and lacks the ability to allow users to discover, extract and integrate the information of interest from heterogeneous resources. This paper proposes a scalable, high performance distributed system for data and service unified discovery. In order to exclude ambiguity, our unified discovery system adopts one-dimensional vector in semantic space to generally identify and locate data and services, and then represent specific services by OWL Web Ontology Language for Services (OWL-S). Moreover, we employ Juxtapose (JXTA) architecture as a proper foundation to organize network peers. Experimental results illustrate that, compared with other discovery systems, the semantic-based unified discovery system can improve accuracy and efficiency of the search results, and satisfy users much better.

1. Introduction

The World Wide Web is expanding with a broader variety of emerging resources that include data and services. Data includes various files such as HTML, plain text, music, image, XML and RDF, while services have their own features as Input, Output, Precondition, Effect (IOPE) and so on. Data and service discoveries are two significant applications in the Internet, and almost all the network functions need their support (Michael and Aphrodite, 2003; Tamer et al., 2013; Zaha et al., 2008).

Yet, current research work has been focused on either service discovery or data discovery without considering the relationship between them. Some systems only support data search (Tang et al., 2003; Crespo and Garcia-Molina, 2003; Shen et al., 2004; Gu et al., 2007) while others merely support service discovery (Cuiting et al., 2012; Arabshian and Schulzrinne, 2005; Carlo and Giuseppe, 2014; Pilioura et al., 2004; Kang et al, 2007; Li et al., 2004; Al-Oqily and Karmouch, 2009). The existing work did the match based on the content, they did not take IOPE (input, output, precondition and effect) of services into consideration. While the other work just deal with the services. There are various relationships between data and services, such as caller and callee. Namely, when the requesters search some interested data, they have to employ corresponding services. However, we cannot expect that users can be aware that which belongs to data, and which belongs to service. For example, one requester wants to find some songs which adapt to the sad mood, and it is not easy for the requester to determine where to go. Thus, the unified discovery system can facilitate users' searching process. In order to provide a convenient and unified discovery system for people and to decrease wastes created by two separate systems, this paper puts forward a semantic-based data and service unified discovery system through taking advantage of the relations between data and services.

Developing a semantic-based data and service unified discovery system poses several technical challenges. First, data and services need a universal description. There are some semantic description techniques such as Vector Space Model (VSM) (Wong et al., 1987; Berry et al., 1999) and Latent Semantic Indexing (LSI) (Papadimitriou et al., 1998) used in Tang et al. (2003) and Shen et al. (2004). LSI can overcome synonymy, polysemy and noise problems incurred by VSM. It can also discover the underlying semantic correlation among documents by building a concept space. However, it costs a lot of computation. Resources in Li et al. (2004) were represented as data objects, and the semantic features of such data objects were identified by a k-element vector, namely Semantic Vector (SV) (also

* Corresponding author. Tel.: +86 010 80767919.
E-mail addresses: dearzppzpp@gmail.com (Y. Zhang), hkhuang@bjtu.edu.cn (H. Huang), hh1012@gmail.com (H. He), jing0404@gmail.com (J. Teng), wzx0000@126.com (Z. Wang).

http://dx.doi.org/10.1016/j.jnca.2014.11.008
1084-8045/

called feature vector). It then uses adaptive space linearization to linearize the clusters of high dimensional space into a one-dimensional Semantic Small World. Papadimitriou et al. (1998) describe services as a sequence of keywords. In contrast, we adopt one-dimensional vector in semantic space to generally identify and locate data and services, and then represent specific services by OWL-S (Web ontology language for services) because services have peculiar characteristics compared to data (Martin, 2003; Zheng and Bouguettaya, 2009; Vassiliki et al., 2008; Kim and Lee, 2009; Chakraborty et al., 2006; Meditskos and Bassiliades, 2010).

The unified discovery system also needs to provide semantic queries in addition to the keyword-based queries to better support search accuracy. This mandates the employment of semantic-based searches (Cilibrasi and Vitanyi, 2007; Ferreira et al., 2008b; Ludwig and Reyhani, 2005b). We make use of three types of ontological data in our system to achieve this goal, namely, resource domain ontology, QoS (Quality of Service) ontology and service description ontology.

Also, the unified discovery system needs the unified matching algorithms to search resources. We present three algorithms to implement data and service unified discovery process, including getting semantic related group algorithm, locating resource algorithm and service matching algorithm.

Finally, the unified discovery system needs to have high scalability, so it is not a good choice to introduce the centralized approach. The centralized approach does not suit large numbers of resources, as it is prone to introduce a single point of failure and expose vulnerability to malicious attacks. This disadvantage is fatal for the evolving trend of Internet.

In order to achieve high scalability, we focus on developing a decentralized discovery approach (Nima et al., 2014). There are several distributed systems available, such as Gnutella (2000) and Napster (2001). However, most of them are intended for one specific application, such as file sharing. Therefore, our current research makes use of the distributed infrastructure JXTA (short for "juxtapose"). The JXTA Search discovery and access model exposes content unavailable through traditional indexing and catching search engines using a query routing protocol (QRP) for distributed information retrieval (Waterhouse et al., 2002). JXTA Search occupies the middle ground between the decentralized Gnutella and centralized Napster geometries. It is independent of platform, network transport and programming language.

The rest of this paper is organized as follows: Section 2 provides some related work. In Section 3, we introduce our unified discovery system architecture. Based on the proposed architecture, we explore the processes for semantic-based unified registry and discovery in Section 4. Next, we address cost issues in Section 5. The experimental results are discussed in Section 6 and, finally, we conclude with directions for the future work in Section 7.

2. Related work

Lots of research efforts are focused on improving search efficiency and accuracy by designing good routing and discovery protocols. However, current systems support either data or service discovery. Existing approaches to resource discovery can be broadly classified as centralized and decentralized. The centralized data discoveries include RDFStore (2008) and Jena 2 (2012), which are RDF repositories and lookup systems. UDDI (2000) or UDDI-based discovery systems (Ran, 2003; Maximilien and Singh, 2004; Verma et al., 2005) are centralized fashions for services. Although these centralized approaches can provide a fast response to a query, it is difficult for them to keep data or service description up to date. In addition, these approaches are not appropriate for large-scale environments because they suffer from the traditional limitations of the centralized approaches, such as single processing bottleneck and single point of failure.

To address the problems of the centralized approaches, a number of distributed systems have been proposed, where data slices and service indexes are distributed to multiple resolver nodes, and queries are routed to the appropriate resolvers (Rostami et al., 2008; Ferreira et al., 2008a; Ludwig and Reyhani, 2005a). There are two types of p2p systems depending on the way that resources are located in the network (Javad Akbari, 2012). In unstructured p2p systems, resources are placed at random points and each peer searching for a resource contacts all its neighboring peers. Gnutella (2000), Napster (2001) and Freenet (2000) are examples of such approaches. In contrast, resources are located at the specified peers in the structured p2p systems (Javad Akbari, 2012). Most resource discovery procedures in structured p2p systems such as Chord (Stoica et al., 2001), CAN (Ratnasamy et al., 2001), Taperstry (Zhao et al., 2004) and Pastry (Rowstron and Druschel, 2001) build a distributed hash table (DHT). Within the distributed hashing approach, each resource is assigned a key and each peer is associated with a range of keys. The unstructured P2P systems lead to long response time and do not scale well, while the DHT-based structured approaches lack flexible search capabilities, because they support lookup only by using a unique identifier and show deficiency of semantics.

In order to present semantic-based scalable resource discovery systems, many researchers devote to combining distributed systems with semantic techniques (Hai and Xiang, 2007; Zhuge and Feng, 2008). pSearch (Tang et al., 2003) is a decentralized non-flooding P2P data retrieval system. It uses CAN to organize search engines into an overlay and distributes document indices through the p2p network based on their semantic information generated by Latent Semantic Indexing (LSI) (Deerwester et al., 1990). INS/Twine (Balazinska et al., 2002) maps strands of the hierarchically partitioned data to a structured peer-to-peer system such as Chord. GloServ (Martin, 2003; Zheng and Bouguettaya, 2009) is a global service discovery architecture that uses the Web Ontology Language (OWL) for service classification and maps knowledge to a structured p2p network-CAN. Although combining semantic, these DHT-based data discovery systems (such as Tang et al., 2003; Balazinska et al., 2002) and service discovery systems (such as Zheng and Bouguettaya, 2009) still support only one kind of resources. It cannot resolve the problems brought by the single system which are mentioned in Section 1. Therefore, we present the semantic-based unified discovery system which can support both data and services.

Besides the DHT-based resource discovery systems, there are many other distributed approaches such as Semantic Overlay Networks (SONs) based data discovery system (Crespo and Garcia-Molina, 2003; Gu et al., 2007), in which nodes with semantically similar content are clustered together. In the hierarchical service discovery systems (Carlo and Giuseppe, 2014; Pilioura et al., 2004), resolvers are organized in a tree structure, where parent resolvers are in charge of all the domains managed by their child resolvers. The higher level resolvers in the hierarchical systems, especially the root resolver, are not immune to the bottleneck as registration and query rates increase. UbiSearch (Kang et al, 2007) is a semantic service discovery network based on Semantic Vector Space (SVS), in which semantically close services are mapped to the nearby positions. The peer nodes in Li et al. (2004) are clustered in a semantic space based on the semantic contents of their data, and then the clusters are organized into a Small World Network. JXTA is used by Xu and Chen (2007) to develop a prototype system for the P2P-based Web service discovery.

The above resource discovery systems focus on either data or service discovery. The work closest to ours is the Bloom Filters based system presented by Koloniari and Pitoura (2004). It assumes that data