A Probabilistic Semantic Approach for Discovering Web Services

Jiangang Ma
School of Computer Science & Mathematics, Victoria University
Australia
ma@csm.vu.edu.au

Jinli Cao
Dept. Comp. Sci. and Comp. Eng.
La Trobe University,
Australia
jinli@cs.latrobe.edu.au.

Yanchun Zhang
School of Computer Science & Mathematics, Victoria University
Australia
yzhang@csm.vu.edu.au

ABSTRACT

Service discovery is one of challenging issues in Service-Oriented computing. Currently, most of the existing service discovering and matching approaches are based on keywords-based strategy. However, this method is inefficient and time-consuming. In this paper, we present a novel approach for discovering web services. Based on the current dominating mechanisms of discovering and describing Web Services with UDDI and WSDL, the proposed approach utilizes Probabilistic Latent Semantic Analysis (PLSA) to capture semantic concepts hidden behind words in the query and advertisements in services so that services matching is expected to carry out at concept level. We also present related algorithms and preliminary experiments to evaluate the effectiveness of our approach.

Categories and Subject Descriptors: H. 3.5.Online Information Services: Web-based services, Miscellaneous

General Terms: Algorithms, Design

Keywords: Web Service, Web Services Matching

1. INTRODUCTION

Web Service discovery is one of challenging issues in Service-Oriented computing. This discovery process mainly involves locating desired services either published in a registry like UDDI or scattered in P2P systems, matching users' requirements to a set of services and returning relevant ones to the consumers. With the ever-increasing number of services published in Internet, finding desired services is just similar to looking for a needle in a haystack[1].

Although there are different kinds of approaches dealing with discovery issue, three of commonly used methods are identified in the paper: keywords-based, SVD-based and Ontology-based approaches. First, keywords-based mechanism is one of the dominating techniques for web services discovery and matching. However, this approach based on term frequency analysis is insufficient in the majority of the cases because it fails to contemplate the semantic concepts hidden behind the service descriptions. An alternative to the keywords-based approach is to find common semantic concepts between the terms in the query and the services advertisements. In [4], a mathematics approach called Singular value decomposition (SVD) has been implemented to match web services. This method employs SVD of the service matrix to discover associated patterns between the word descriptions and the concepts hidden behind the description. Although this approach shows some advantages compared to the keywords-based one, its lacking of completely reasonable probabilistic interpretation [3] might limit its further application. More recently, ontology-based approach has been seeking to use ontology to annotate elements in web services, which aims to not only capture the information on the structure and semantics of a domain, but facilitate software agents to make inference at the level of concept. Nevertheless, creating and maintaining ontology may involve huge amount of human effort.

We propose to extend the SVD of matrix approach of matching web services [4] with a different methodology called Probabilistic Latent Semantics Analysis (PLSA) [2, 3], which turns out to have sound probabilistic interpretation and better performance. Our main contributions consist of two aspects. The first contribution is that the technique using probabilistic semantic approach is introduced to discover and match web service. According to our knowledge, no such similar methods have been found. Second, we present algorithms to discover and select web services.

2. PLSA INTRODUCTION

Our probabilistic semantic approach is based on the PLSA model that is called *aspect model* [2]. PLSA utilizes the Bayesian Network to model an observed event of two random objects with a set of probabilistic distributions. In the text context, an observed event corresponds to occurrence of a word *w* occurring in a document d. The model indirectly associates keywords to its corresponding documents through introducing an intermediate layer called hidden factor variable $Z = \{z_1, z_2, ..., z_k\}$. Based on the assumption that a document and a word are conditionally independent when the latent concept is given, the joint probability of an observed pair (d_i, w_j) obtained from the probabilistic model is shown as following:

$$P(d_i, w_j) = P(d_i)P(w_j \mid d_i), \quad (1)$$

Where

$$P(w_j \mid d_i) = \sum_{f=1}^{k} P(z_f \mid d_i)P(w_j \mid z_f) \quad (2)$$

From formula 2, we can see that the aspect model expresses dimensionality reduction by mapping a high dimensional term document matrix into the lower dimensional one (k dimension) in latent semantic space.

PLSA was originally used in text context for information retrieval and now has been used in web data mining [5]. In this paper, we utilize PLSA for discovering and matching web services.

3. PSDA—PROBABILISTIC SEMANTIC DISCOVERING APPROACH

Our approach is based on our observation on uncertainty on the usage of web services in the Web environment. This uncertainty is reflected in two aspects. For client side, a service user may not have specific goal in his mind while he browses web service categories in the web. Second, web services are priori unknown. Based on this observation, we introduce probabilistic approach to deal with service discovery.

3.1 Overview of Our PSDA Approach

Figure 1 illustrates the outline of the proposed probabilistic latent semantic approach. The approach first filters out those web services whose types are not compatible to a user's query, which will lead to a smaller size of services available (N-m). Next, PLSA is used to match semantic similarity between query and web services. Finally, the Quality of Services (QoS) is combined with the proposed semantic method to produce a final score that reflects how semantically close the query is to available services.

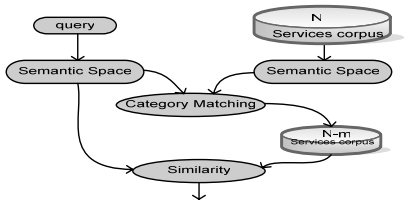

Figure 1 outline of PSDA approach

3.2 Web Services Information Processing

The overall process of discovering web services includes: information collection, data processing, data representation and similarity matching. ***The information collecting:*** As each web service has its associated WSDL file describing its functions, we firstly extract the service interface information such as name and textual description in the WSDL file. ***The data processing*** stage consists of transforming raw web service information into appropriate format of data suitable for model learning. For this purpose, commonly used approaches for words processing such as word stemming and stopwords removing are applied. ***The data representation:*** In our probabilistic model, the pre-processing information would be represented by the bag of words. We define a service document for each service and a service matrix for a service corpus for probabilistic semantic matching algorithm.

3.3 Probabilistic Semantic Matching Algorithms

The key of our discovering approach is first to cluster the services into a group of learned latent variables, which can be achieved by computing probability $P(latent - variable \mid service)$ for each latent variable using formula 3. The rationale for this is that in the dimension-reduced semantic space, each web service can be represented as a mixture of latent variables and the services with similar semantic concepts are projected to be close to each other. With the maximum value of the computation used for the class label for a service, we can categorize services into their corresponding group.

$$P(z \mid d_{new}) = \frac{P(d_{new} \mid z)P(z)}{P(d_{new})} = \frac{P(d_{new} \mid z)P(z)}{\sum_{z_f \in Z} P(d_{new} \mid z_f)} \quad (3)$$

After removing irrelevant services, the similarity of a query in respect to the services in the relevant group can be computed with formula 4. As a query may be outside the model, we use Expectation Maximization algorithm to fold the query in the model.

$$sim_{PLSA}(d_i, q) = \frac{\sum_{z_f \in Z} P(z_f \mid q)P(z_f \mid d_i)}{\sqrt{\sum_{z_f \in Z} P(z_f \mid q)^2} \sqrt{\sum_{z_f \in Z} P(z_f \mid d_i)^2}} \quad (4)$$

Finally, Quality of Services (QoS) can be combined with PLSA similarity score to produce a final ranking that reflects how semantically close the query is to available services

$$Sim(d_i, q) = \lambda \bullet S_{QoS} + (1 - \lambda) \bullet Sim_{PLSA}(d_i, q), \quad (5)$$

4. PRELIMINARY EVALUATION

The preliminary experiments are carried out on the collection of web services whose WSDL files can be accessed via service collection published in XMethods. In our case, we identify the services of four categories: General Information, Location Finder, Translation Services and Business Services. Thus, we obtain a corpus of services consisting of 77 services that are divided into two data sets: training data (44) and testing data (33).

We train the model with 6 aspects, which is slightly greater than the number of selected four service categories. In order to evaluate the outcome, we compute precision and recall and apply traditional vector-based similarity baseline used in information retrieval approach to compare to the proposed algorithm in this paper. As it turns out, our probabilistic semantic discovery method increases the overall recall because the approach considers semantic concepts hidden behind words in the query and advertisements in services.

5. REFERENCES

[1] John Garofalakis, Y. Panagis, E. Sakkopoulo and A. Tsakalidis. Web Service Discovery Mechanisms: Looking for a Needle in a Haystack? In *International Workshop on Web Engineering, August 10,* 2004.

[2] Thomas. Hofmann. Probabilistic Latent Semantic Analysis In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval. Berkeley, California, pages: 50-57, ACM Press,* 1999.

[3] Thomas. Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval*,1999.

[4] Atul Sajjanhar, Jingyu Hou and Yanchun Zhang. Algorithm for Web Services Matching. In *Proceedings of the 6th Asia-Pacific Web Conference, APWeb 2004, Hangzhou, China, April,14-17,*2004.

[5] Guandong Xu, Yanchun Zhang, Jiangang Ma and Xiaofang Zhou. Discovering User Access Pattern Based on Probabilistic Latent Factor Model. In *Proceedings of the 16th Australasian Database Conference – Volume: 39 pages: 27 – 35, Newcastle, Australia,* 2005.