International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 08 | Aug -2017 www.irjet.net p-ISSN: 2395-0072

Web Service Discovery Mechanisms Based On IR Models

Arockia Panimalar.S [1], Subhashri.K[2], Iniya. R[3], Sumithra. V[4]

[1, 2] *Assistant Professor, Department of BCA & M.Sc SS, Sri Krishna Arts and Science College, Coimbatore, India*
[3, 4] *III BCA, Department of BCA & M.Sc SS, Sri Krishna Arts and Science College, Coimbatore, India*

Abstract - *Web Service discovery is one of the real research push ranges in the field of computing environment. From a decade ago, a great number of specialists are contributing their considerations in finding the best available service from a pool of services that can satisfy the user's requirement. Diverse scientists have embraced distinctive philosophies and thoughts to envision their honourable considerations in the field of web service discovery. Information Retrieval (IR) methods are one of them. Usage of IR methods in service discovery approaches makes the discovery process efficient. This paper focuses on the service discovery approaches which employ information retrieval methods for the purpose of automatic discovery. It gives a study of how these methodologies contrast from each other while discovering a service.*

***Key Words*: Web Service Discovery, Information Retrieval, WordNet**

1. INTRODUCTION

Web Services are loosely coupled, distributed and independent application components that can be published, found and used on the web [1]. W3C defines web service as: "A web service is a software system designed to support interoperable machine to machine interaction over a network. It has an interface portrayed in a machine-processable configuration (particularly WSDL). Different frameworks connect with the Web benefit in a way endorsed by its depiction utilizing SOAP messages, regularly passed on utilizing HTTP with a XML serialization in conjunction with other Web-related standards"[2]. Web services runs on the technologies called WSDL (Web Service Description Language), UDDI (Universal Description, Discovery and Integration) and SOAP (Simple Object Access Protocol). WSDL[3] describes services a set of network endpoints. WSDL document provides the functionalities of a service in XML format. UDDI [4] is the universally accepted ML based standard service repository in which the services in WSDL format can be published. Services in Service Oriented Architecture (SOA) can communicate with each other through SOAP[5]. SOAP is a lightweight XML based protocol for exchange of information in a decentralized, distributed environment.

Web service discovery is one of the most popular domains in SOA. The fundamental concept is that web service providers publish their services in the service repository and web service consumers use those services; web service discovery is the process of finding the most appropriate service from the repository that satisfies the consumers' need. Automatic discovery of web services without human intervene is a popular research area for many researchers. Researchers have suggested many methods for the automatic discovery of web services. This paper focuses on the discovery processes which utilize information retrieval methods for the purpose of automatic discovery.

The paper is organized as follows:

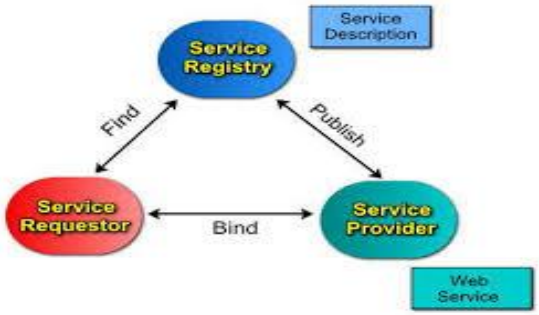
Section II provides overview of web service discovery mechanism. Section III discusses some of the information retrieval methods. Section IV deals with various approaches for web service discovery. Section V concludes the study.

2. Web Service Discovery

If a web service consumer wishes to avail a service but is unaware of the providers of the service, then the consumer must initiate the 'discovery' process. Discovery is "the act of locating a machine-processable description of a Web service that may have been previously unknown and that meets certain functional criteria"[1].

Web service providers publish their web service descriptions along with the associated functional descriptions of the services in the form WSLD in the service repository. In order to use some services available in the service repository, the service consumers need to provide the service requirements based on the associated functional descriptions.

Based on the consumer's service criteria, the discovery unit finds the appropriate service from the service repository that fulfills the specified criteria and returns the associated service description to the consumer. If the discovery unit returns more than one services for single query, then the consumer has to select one of them based on some additional criteria. Both service provider and consumer must agree on the service description and the semantics of the interaction. And then, the provider and consumer communicates by exchanging the SOAP messages[1].



3. Information Retrieval Models

Information retrieval is the study of finding unstructured documents from a large collection of documents with respect to a given query. IR can be broadly classified in two categories: semantic and statistical. Semantic technique extends retrieval capability by adding both syntactical and semantic analysis. Semantic analysis utilizes the contextual meaning of the user query to produce more relevant result. In statistical IR techniques, the documents which match the user query most closely are retrieved based on some statistical measures[6].

Boolean, vector space and probabilistic methods are the widely used statistical approaches in IR. The documents and queries are often broken into words, generally referred to as terms. These terms need to undergo a number of pre-processing techniques before using them in any statistical measures [6].

Boolean IR is the oldest and simplest Information Retrieval model. In Boolean information retrieval system, any query can be expressed in terms of Boolean expression. The terms in the query are combined with the operators and, or, and not [7].

The result of a Boolean query is either true or false. A document is either relevant or irrelevant w.r.t. a query depending on whether the document satisfies the query or not[6]. For a query, this model retrieves that document for which it finds the exact match. Partial matches are not retrieved and also there is no ranking mechanism[10]. In Vector Space Model[8], each document is represented as vector of terms.

$dj = \{w1j , w2j , \ldots , wtj\}$

where, *t* represents the total number of terms in the document collection. If a term is present in the document, then the weight for the term in the document vector is non zero. The query is also represented as a vector in of terms in Vector Space Model. There are two different techniques for computing the term weights of the terms present in the document vector and query vector: term-frequency (tf) and inverse document frequency (idf). The next step is to measure the similarity between the query vector and each of the document vectors. The similarity is generally measured using the cosine similarity[9]. The documents whose similarity values are greater than the threshold value are retrieved as a set of relevant documents. In vector Space Model, the term weights are not binary; partial matching and ranking of documents are possible. The order of appearance of the terms in the document does not coincide when represented as vector space which is one of the limitations of this model. Another limitation is that this model does not capture the semantics of the query and document[10]. Probabilistic IR model retrieves documents based on the probability of the document relevant to the query.

The Probabilistic Ranking Principle[11] (PRP) is as follows: "On the off chance that a reference recovery framework's reaction to each demand is positioning of the reports in the gathering arranged by diminishing likelihood of importance to the client who presented the demand, where the probabilities are assessed as precisely as conceivable on the premise of whatever information have been made accessible to the framework for this reason, the general viability of the framework to its client will be the best that is realistic on the premise of that information".

4. Web Service Discovery Mechanisms

A. Singular Value Decomposition (SVD)

It is a linear algebra to locate the similar services for a service request. For the purpose of retrieving similar services, they first collect keywords from the service descriptions of web services available in UDDI registry. These words are passed through a pre-processing stage which involves the elimination of stop words and then obtaining stemmed words. In this way, they build a collection of unique words for all the services present in the UDDI. All the words in the collection are assigned weight which is the value of inverse document frequency of the corresponding word. In the next step, vectors for all services are created where the value of each vector element for a particular service is calculated based on the weight of each word multiplied by its binary occurrence in that service. A service matrix is constructed considering all these vectors and then the singular value decomposition method of linear algebra is applied on this service matrix to find the relationship between services, define threshold for retrieving similar services and filter out the irrelevant services. To find similar services for a given service, they apply the concept of cosine similarity measurement. They present a comparison table between keyword matching technique and their SVD based proposed model which shows better result for their model.

B. Analysis and Discovery of Web Services

They combine one of the information retrieval methods and the existing standards that describe the web services for the purpose of web service discovery. For a service request, their search engine finds the similar service/s from a

collection of web services based on the information like functionalities, descriptions etc. of the services. The distributed search engine employs the concept of vector space model of information retrieval system to discover the analogous service for a specific service request. The first step in their model deals with the extraction of keywords like endpoint URLs, types and their attribute names, message names, service names and XML comments from web services that are available in the WSDL file repository. Using these extracted keywords, a vector space is created where each dimension is represented by a term. Each document (WSDL file) is represented by a vector within this vector space. Vector elements of each vector are assigned normalized term weights calculated based on the term frequency and inverse document frequency of the term. They also extend their model to work with distributed vector space. The query processor extracts keywords from the given query string and also creates the query vector. The cosine similarity value is calculated between the query vector and each document vector present in the term space. Documents are then sorted based on their similarity rating. This model works fine in distributed environment.

C. Flexible Service Discovery

It is a method to discover useful services. Textual elements of the service request are compared against textual elements of all available services in the repository to discover similar services and to rank them according to their similarity. To carry out this task, they use vector space model of information retrieval system. Next, a structure matching algorithm is projected on this ranked list of services to refine and evaluate the quality of the service set. For this purpose, they develop a heuristic, domain-specific tree-edit distance algorithm. WSDL specifications describing the web services based on XML syntax are hierarchical in nature. Therefore, the tree-edit distance algorithm can calculate the similarity between two tree structures as because minimum node modifications required to match them. Comparing two web services is based on comparison of service operations which, in turn, is based on comparison of service messages which again is based on comparing the data-types of the WSDL specifications. The algorithm matches the data-types, messages and operations of the WSDL specifications respectively. Each of these three steps results in individual matrix that evaluates the similarity score of all pair-wise combinations of source and target datatypes/messages/ operations. The final similarity score is the maximum score calculated as the sum of matching scores of all individual operation pairs. Their report includes experiments on service discovery with IR methods only, with structure matching only and with IR and structure matching combined. For service discovery with IR methods only, they report a precision of 51% at 95% recall on average with structure matching only, they report a precision of 20% at 72% recall on average. The retrieval system with IR and structure matching combined achieves a precision of 61.5% at 90% recall with an increase in precision by 10.5% and drop in recall by 5% as compared to retrieval technique with IR method only.

D. Web-Service Search Method

It is used to discover similar service operations given a natural language description of the desired service. Their work is almost similar to that of Wang & Stroulia[14], but they focus on semantic similarity rather than structural similarity. For a given service description, they first obtain a candidate set of service operations by using TF and IDF techniques of Information retrieval System. Web service data-types can be primitive or complex. Primitive data-types do not contain semantic information. Primitive data-types are converted to complex data-types by replacing them with their corresponding parameters to contain semantic meaning. XML schema is modelled as a tree of labelled nodes. The labelling is done as tag node and constraint node. Constraint node is further subdivided as sequence node, union node and multiplicity node. They apply bottom-up-transformation algorithm of time complexity O(n) in which they propose three transformation rules to transform all three constraint nodes to tag nodes. The service operation matching is done with the help of schema tree matching. On the candidate operations set, they use tree edit distance approach of schema matching algorithm to calculate the similarity among the operations. To measure tree edit distance between two schema trees, they introduce a new cost model to compute the cost of each edit operation which is based on weights and semantic connection of nodes. After identifying similarity of web service operations, the candidate set of service operations is clustered using agglomeration algorithm. For each cluster, operation with minimum cost will be output as search result. Since operations are associated with scores, they rank the search result based on scores of operations.

E. Ontology Linking and Latent Semantic Indexing(LSI)

The web service discovery combining the concepts of ontology linking and Latent Semantic Indexing (LSI) in combination. Ontology linking is achieved by mapping domain ontology against upper merged and mid-level ontology. LSI determines the relationship between query terms and the available documents to capture the domain semantics. After preparing the service request vector using domain ontology linking concept and service description vector from the selected WSDL documents using the LSI classifier, both the vectors are put together utilizing the cosine similarity technique to determine the similarity and to discover the relevant web services. After pre-processing of the service request, it is enhanced by associating the related upper concepts utilizing the upper ontology which helps in web service categorization. From these categorized service collections, the service descriptions are extracted and parsed to form the term-document matrix.

The SVD transformation applied on term-document matrix produces reduced dimension vector for each term and each document which helps to determine the appropriate web service. Because of the categorization of services, their experimental results produce a better result.

To avoid huge number of retrieved services w.r.t. a keyword and the inability of the keywords to express semantic concepts, Jiangang Ma et al[17] propose a clustering semantic algorithm to semantically discover the web services. First they prepare a working dataset by using a clustering semantic algorithm. For a particular query, the dataset is prepared based on the relevant web services whose contents are compatible with the query. They employ K-means clustering algorithm to eliminate the irrelevant services w.r.t. a query. They use the Vector Space Model to prepare the service transaction matrix for the services in the working dataset. The working dataset is then grouped into a number of semantically related clusters by employing the concept of Probabilistic Latent Semantic Analysis approach (PLSA). This PLSA technique is used to capture the semantic concepts of the terms of the query and also the descriptions of the services which help to determine the semantic similarity between the service and the query within the related cluster. The PLSA utilizes the Bayesian network model and an intermediate layer called hidden factor variable to associate the keywords to its corresponding documents. For each and every service of the service dataset, PLSA computes the probability w.r.t. each latent hidden variable. Then it determines the maximum probability for each service and puts the service in the semantically related cluster. By comparing the similarity between the query and related clusters, a set of relevant services are retrieved for the query.

F. Vector Space Model

The Vector Space Model of IR in combination with Vector Space Model is enhanced with lexical database WordNet to efficiently discover the web services. WordNet is utilized to capture the semantics of the WSDL elements that assist in deriving the accurate similarity. They also use Linear Discriminant Function of pattern classification technique to devise the relative similarity between the term and term's synonyms and calculate the optimized weights of the terms using batch perceptron algorithm. For both original vector and vector enhanced with WordNet, they calculated (i) the relative frequency of each word in each WSDL document, (ii) global importance of each word in all WSDL collection, (iii) relative importance of each word in each WSDL document. The same procedures are used for the query vector also. The local similarities between the WSDL and the query for both original and the enhanced vectors are calculated as the inner product of the relative importance of words in WSDL and relative importance of words in query. Finally, the global similarly between the WSDL and the query is calculated based on the local similarities in case of both original and enhanced vectors. They compared their algorithm with four other IR based WS discover algorithms and found that their algorithm is showing an improvement of 0.6% to 1.9% in precision and 0.7% to 3.1% in recall for the top 15 web services.

5. Conclusion

Finding the efficient service that satisfies the user's need is a most challenging task in the field of web service discovery mechanisms under SOA. Various techniques have been developed to answer this issue. Focus is on the information retrieval techniques related service discovery approaches. Some of the basic information retrieval methods are discussed in this paper. Some approaches are centralized in their nature while others are distributed. Also some approaches focus on semantic capability to make the discovery process an efficient one. It also agreed that, discovery units should address the semantic capability precisely in order to find the most appropriate service.

6. References

[1] Papazoglou, M. P., Georgakopoulos, D., "Service-oriented computing", Communications of the ACM, Vol. 46, No. 10, 2003, pp. 25–28.

[2] (2004, February) Web Services Architecture [Online].

https://www.w3.org/TR/2004/NOTE-ws-arch-20040211/

[3] (2001, March) Web Services Description Language (WSDL) 1.1 [Online]. https://www.w3.org/TR/wsdl

[4] (2004, October) UDDI Spec Technical Committee Draft [Online]. http://www.uddi.org/pubs/uddi_v3.htm

[5] (2000, May) Simple Object Access Protocol (SOAP) 1.1 [Online]. https://www.w3.org/TR/soap/

[6]Ed Greengrass, "Information Retrieval: A Survey", Available online at: www.csee.umbc.edu/csee/research/cadip/IR.report.120600.book.pdf

[7] Christopher D. Manning, Hinrich Schütze, and Prabhakar Raghavan, "Introduction to Information Retrieval", © Cambridge University Press 2008, ISBN-13 978-0-511-41405-3

[8] G. Salton, A. Wong and C. S. Yang, "A vector space model for automatic indexing", Communications of the ACM, v.18 n.11, p.613-620, Nov. 1975

[9] Sidorov, Grigori; Gelbukh, Alexander; Gomez-Adorno, Helena; Pinto, David. "Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model".

[10] Joydip Datta, "Ranking in Information Retrieval"

[11] S.E. Robertson "The Probability Ranking Principle in IR" http://www.emeraldinsight.com/10.1108/eb026647.

[12]Atul Sajjanhar, Jingyu Hou, and Yanchun Zhang, "Algorithm for Web Services Matching"

[13] Christian Platzer and Schahram Dustdar, "A Vector Space Search Engine for Web Services"

[14] Yiqiao Wang and Eleni Stroulia, "Flexible Interface Matching for Web- Service Discovery"

[15] Yanan Hao and Yanchun Zhang, "Web Services Discovery Based on Schema Matching"

[16] Aabhas V. Paliwal, Nabil R. Adam, Christof Bornhovd, "Web Service Discovery: Adding Semantics through Service Request Expansion and Latent Semantic Indexing"

[17] Jiangang Ma, Yanchun Zhang and Jing He, "Efficiently finding web services using a clustering semantic approach"

[18] Ricardo Sotolongo, Carlos Kobashikawa, Fangyan Dong, and Kaoru Hirota, "Algorithm for Web Service Discovery Based on Information Retrieval Using WordNet and Linear Discriminant Functions".