Journal of Theoretical and Applied Information Technology
31st December 2012. Vol. 46 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A MATCHING SCHEMA FOR SEMANTIC WEB SERVICE

CHUANGLU ZHU

Department of Mathematics and information science, Weinan Normal University, Weinan 714000, Shaanxi, China

ABSTRACT

Finding out useful Web service from the huge number of service is the target of Web service discovery. Matching algorithm of service are the main technologies to achieve Web service discovery. This dissertation studies how to effectively utilize the semantic information in domain ontology to enhance accuracy of Web service matching, and describes a variety of services matching algorithm in the semantic Web service discovery. By introducing the existing matching algorithm, analyzes the problems of the algorithm. To increase the accuracy of algorithm based on semantic distance, introduce the quantity of information which is defined combining the concept of level matching. Describes a service discovery model based on semantic method and verify the effectiveness of the proposed algorithm based on the comprehensive model through comparison with the existing framework of service discovery.

Keywords: *Semantic Web, Ontology, Semantic Similarity, Semantic Distance, Semantic Information Content*

1. INTRODUCTION

The semantic web becomes a research hotpot in information science after the concept of semantic web was put forward by Tim Berners-Lee in 1998. The semantic web service is the result of combining the semantic web and the web service, and OWL-S (web ontology language for services) works as the bridge which connects the two technologies. The web service discovery technology which supported by the OWL-S technology is faced with a main problem [1], that is how to improve the low accuracy rate of the matchmaking method of the semantic matching in order to achieve the goal of matching exactly.

Wu-Palmer et al[2] put forward improvement schema to a shortest path method, that schema can find the lowest node which are the common ancestor to measure their semantic similarity, the shortcomings of method are that matching is low accuracy and time complexity increases with more nodes. Resnik[3] put forward semantic similarity calculation method based on information theory model that compute semantic similarity through the degree of information sharing, but that can not distinguish concept with the same sharing information. In order to revise this problem, Lin et al [4] proposes an improved method that brings in information differences to semantic similarity. At present, the problems existing in the semantic web service discovery are the following aspects: Lacking of a proper description model of the semantic web service; Low accuracy rate of the service matching operated by the semantic matching method and so on. Those problems cause that the service discovery mechanism nowadays can't make effective management and exactly visit to a broad range of web services. And those problems have become the bottleneck for the development of the semantic web service technology.

This schema is mainly about the semantic web service matchmaking method that considers both semantic distance and information theory model, and is to solve the problem of the low accuracy rate of the semantic service matching based on this schema.

2. THE ALGORITHM OF THE SEMANTIC WEB SERVICE DISCOVERY

This part will analysis the present semantic matching algorithm considering the influence on the semantic matching which brought by the semantic distance and semantic information content. And an algorithm of comprehensive semantic matching degree is brought out.

2.1 The Algorithm Based on Semantic Distance Similarity

The algorithm based on the semantic distance similarity realizes the discovery of the semantic web service according to the semantic distance. And this method can achieve the goal of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

semantic matching to its greatest extent. Generally, the semantic distance refers to the correlations between two different concepts in a same ontology. And this correlations usually results from inheriting, contains and binary relations and so on [5]. The algorithm is mainly to figure out the semantic distance between concepts. When there is a big distance between 2 different concepts, the semantic similarity between them is small. The goal of this algorithm is to find out the connections between different concepts as far as possible and improve the similarity between them to realize semantic matching [6]. The premise of the algorithm is to establish a same ontology. That's to say, the computing of the semantic distance is based on a same ontology. For different ontology, the semantic distance can be supposed to be infinitely great because of that ontology in different domains convey different meanings. The semantic distance from one ontology concept to another may not be the only one, at this time, choose the shortest given semantic distance as the measurement. That's to say, there may be several paths between every two ontology concepts in the semantic structure. The path is usually means cyclic path. Every path may contain $n(n>0)$ edges [7, 8].

In the concept graphic, supposed that there have two different concepts C_1 and C_2 in the ontology and then simplify the weight of each edge to one. So the distance equals to the length of the shortest path. This is usually got through the quantitative values of the edge. It is obviously that the shorter the distance is between the 2 concepts, the more the similarity is between them. The formula (1) is got based on the ideas above.

$$Similarity(C_1,C_2)=1/(dist(C_1,C_2)+1) \quad (1)$$

In this formula, $Similarity(C_1,C_2)$ is the semantic similarity between concept C_1 and C_2, $dist(C_1,C_2)$ is the length of the path between C_1 and C_2 .The longer the distance is, the less the semantic similarity is; otherwise, the more the semantic similarity is. That is, the semantic similarity is determined by the semantic distance. The computational process below is to figure out the length of the path of the semantic.

The semantic distance $dist(C_1,C_2)$ between node C_1 and C_2 can be figured out according to the different relations between concept C_1 and C_2 .There are 3 situations shown below:

1) If C_1 and C_2 are the same node, $dist(c_1,c_2)=0$. According to formula (1), $similarity(c_1,c_2)=1$.

2) If there are no path between C_1 and C_2 , $dist(c_1,c_2)=\infty$. According to formula (1), $similarity(c_1,c_2)=0$.

3) If there is a path between C_1 and C_2 , $dist(c_1,c_2)$ equals to the number of the path's edge. And the semantic similarity equals to $1/(dist(C_1,C_2)+1)$.

The algorithm based on the semantic distance similarity is the first choice algorithm in the area of the semantic web service. Its main advantage is that it just needs to design a hierarchy of the semantic relations. So its requirements of computing are few. When using this algorithm, it just needs to design the hierarchy of the concepts but needs no other extra information, for example, the amount of information in the concepts and the frequency of its appearance. This algorithm also has some disadvantages, that is: computing the shortest path between concepts by using hierarchy model can not differentiate the semantic between concepts completely. Especially when the ontology base is large, the depth of the concepts is deep and both the time complexity and the space complexity of the algorithms are increasing.

2.2 The Algorithm of Semantic Similarity Based on Information Content

The principle of this algorithm is that the more information that the 2 concepts share, the more similarity that their semantics do [9]. Every concept in the ontology is got by refining its ancestor node. The concept includes all the information content of its ancestor node. If the two concepts are sharing a same ancestor node, they are sharing the information content of the ancestor node [10]. According to the principle mentioned above, the semantic similarity of the two concepts can be measured by the information content of the closest ancestor node. The statistics methods can help with computing the information content of the concepts. Prepare enough data of the domain, the more frequent that the concept appears in the data base of this domain, the more abstract that the concept does and the less that its information content has [11]. Conversely, it has more information content. The information content of concepts can be computing by formula (2).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$I(C_1) = \log\left(\frac{1}{P(C_1)}\right) \quad (2)$$

In this formula, $P(C_1)$ means the probability that C_1 appears in the file; $I(C_1)$ stands for the information content of the concept. With the level up of the node hierarchy, $P(C_1)$ increases and $I(C_1)$ decreases. The semantic association of two concepts can be computing by formula (3).

$$\begin{aligned} &Similarity(C_1, C_2) \\ &= Max\left(I(C_f)\right) \\ &= Max\left(\log\left(P(C_f)\right)\right) \end{aligned} \quad (3)$$

In this formula, C_f is the ancestor node shared by C_1 and C_2. The semantic association of concept C_1 and C_2 is determined by their same ancestor node that has the most information content; in the ontology expression methods of the concept's hierarchy model, the semantic association reflects the information content of the closest sharing ancestor node. In the situation that inheriting happens many times, choose the closest node in the ancestor nodes and this node has the most information content.

This algorithm doesn't depend on the characteristic of the concept's hierarchy structure, using concept statistic to get the information content of each concept. The computing of semantic similarity needn't take the hierarchy relations of the concept and the distance of its appearance positions into consideration. The specific computing process is to figure out the appearance probability in one ontology area. This normal probability reflects its information content. That is, the accuracy of this algorithm depends on the statistic method used by the files. Different statistic method could get different result.

2.3 The Comprehensive Algorithm of Semantic Similarity

The two algorithms introduced above define the similarity between concepts with different method and structure. The algorithm based on the semantic distance similarity makes use of the ontology classification concept tree and computes the distance between concepts to figure out the semantic similarity matching degree. But only computing the distance will cause some error. For example, how to compute the shortest distance determines the accuracy of computing. This algorithm's accuracy is determined by the quality of the ontology construction. The increase of the ontology concept will cause big uncertainty on distance computing and cause its instability on time and space complexity. The algorithm based on the information content makes use of the information content of the concept in the ontology hierarchy to compute the semantic similarity. The matchmaking accuracy is determined by the information content of the concept in the ontology space and its computing result depends on the computing method of the information content. Different computing method will get different similarity result.

The two algorithms discussed above have both advantages and disadvantages. According to the aim of matchmaking of the semantic web service, a more efficient semantic similarity algorithm is brought out. This algorithm combines the two algorithm discussed above and make use of the semantic information content of the concept and the construction information of the concept hierarchy tree to compute the semantic similarity. And then use the shortest path to compute the semantic distance between concepts and measure their semantic similarity according to their semantic distance. The specific definition shows blow:

Definition 1: the appearance probability of concept C.

$$P(C) = \frac{\sum F(C_i)}{N} \quad (4)$$

C_i stands for one concept of the sub-concept set of C, the range of $\sum$ is the sub-concept set of C and N means the maximum times that the concept would appear. N is usually the appearance times of the top concept and its sub-concept.

Definition 2: the information content of concept C.

$$I(C) = \log\left(\frac{1}{P(C)}\right) \quad (5)$$

The similarity between one sub-concept C_s and its father concept C_f is determined by the weight of the related edge between C_s and C_f. According to different weight, the connection degree can be distinguished by the conditional probability C_s and C_f.

$$R(C_s, C_f) = \log\left(\frac{P(C_s \cap C_f)}{P(C_f)}\right) \quad (6)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Definition 3: the total connection degree of edge is a function that related to the depth of node and the connection degree of the edge.

$$LR(C_s, C_f) = F\left(\log\left(\frac{P(C_s \cap C_f)}{P(C_f)}\right), depth(C_f)\right) \quad (7)$$

The function F can be chosen according to different application environment. Choose $F(x,y) = \sqrt{x * y}$ in the prototype system of next section.

Definition 4: the distance between any two concepts C_1 and C_2:

$$dist(C_1, C_2) = \sum_{C_s \in (P(C_1,C_2) - S(C_1,C_2))} LR(C_s, C_f) \quad (8)$$

$P(C_1, C_2)$ is the set of all nodes in the shortest path between C_1 and C_2 . $S(C_1, C_2)$ is the nearest shared ancestor node.

At last, compute the similarity according to the semantic distance. Take formula (1) as reference:

$$Similarity(C_1, C_2) = 1 \Big/ \left(\sum_{C_s \in (P(C_1,C_2) - S(C_1,C_2))} LR(Cs, Cf) + 1 \right) \quad (9)$$

3. ANALYSIS ON THE PROTOTYPE SYSTEM OF THE SEMANTIC WEB SERVICE DISCOVERY

In order to test the validity of the semantic matching algorithm, a semantic web service discovery prototype system and a test data set is designed. The prototype consists of UDDI, ontology base and service discovery engine. The domain ontology is stored in the ontology base. Amount of web services (described by OWL-S) is registered in the service registry. According to the particular service requirement of the user and the ontology chosen from the ontology base, the service discovery engine uses the semantic matching algorithm discussed above to conduct service matching on the service registry, and then return the service which meets the user's requirement. The prototype develops and deploys the web service under J2ee l.4 platform. Build one ontology base resources in Protégé and a description file of the semantic web service (OWL-S); take Jena inference engine as the plug-in of the Protégé to make inference; and store the data of experiment in the database. The main development tools includes: Protégé, OWL-S Editor, OWL-S API, Jena and OWL-S/UDDI Matchmaker. The structure chart of the prototype system of the semantic web service discovery is showing in Figure 1. The core module is to compute the semantic similarity. It is used to realize the service matching methods introduced in the fist part. The semantic web service usually uses OWL-S description as the description language of the web service requirement and publishing. And it parses the web service described with OWL-S which comes from the service sender and the service publisher by the OWL-S parser. It combines Protégé with OWL-S API to get the input and output parameters of the service and then use the description logic inference engine Jena to make inference on the ontology concept of the input and output parameters. And then, compute the semantic distance between concepts according to the inference result.

The experiment of the prototype system builds 180 web services. Most of them are published by UDDI server. Besides, the system still designs some related web services specifically. During the course of the experiment, these web services are registered to the test collection; use Protégé and OWL-S Editor and take use of OWL-S API to transform the WSDL description file of the 180 services in the Protégé into OWL-S profile description files. Then, register the transformed files into Se-UDDI server with the three matching algorithms mentioned in this paper (In Figure 2, Match_D: The Algorithm Based on Semantic Distance Similarity; Match_I: The Algorithm of Semantic Similarity Based on Information Content; Match_N: The Comprehensive Algorithm of Semantic Similarity.). Suppose that the similarity values of the service matching in the prototype system are 10 grades which are equally disturbed from 0.1 to 1. Then, make matching on the 180 semantic web services with the three algorithms respectively and compare the matchmaking results. The final results are shown in Figure 2.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

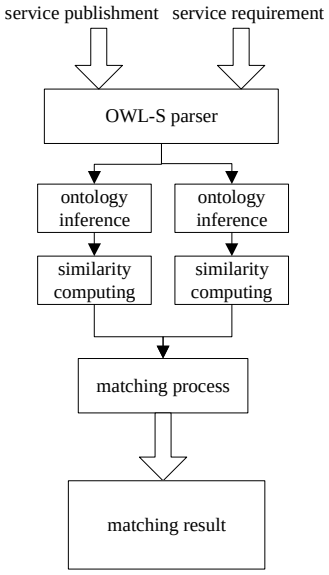

Figure 1: Matching Of The Semantic Web Service

4. CONCLUSION

The research is focus on the semantic service matching methods. It analysis the existing two traditional semantic matching algorithms and combines with the conception of level matching to bring in a new matching method based on the semantic distance algorithm and the semantic information content algorithm. This new method can improve the accuracy rate of the semantic matching to a certain extent. In order to test the accuracy rate of the new method, a web service discovery prototype system is developed. The system makes simulated experiments on the three semantic matching methods. And analysis the test data, make comparison between the new method and the traditional two matching methods. The results show that the new semantic matching method is more accurate than the traditional methods. Even so, there are still some inadequacies in this research and it needs to be further improved. Its main inadequacies are the following aspects: first, the application function of the prototype system is not very complete; second, it takes no consideration on the similarity matching of the properties during the process of the semantic matching, that will be gradually improved in the following researches; third, the logic description of OWL-S can be further extended to supply basis for further inference.

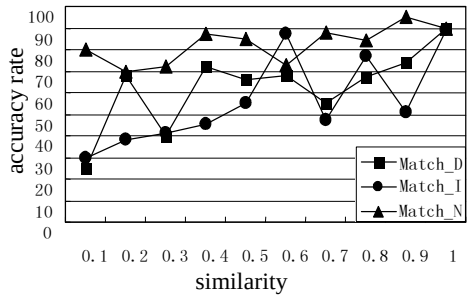

Figure2: Accuracy Rate Of The Matching Algorithms

ACKNOWLEDGEMENTS

This work was supported by Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No. 12JK0746).

REFERENCES:

[1] V. Cross, Y. Wang, "Semantic relatedness measures in ontologies using information content and fuzzy set theory", *FUZZ'05 the 14th IEEE International Conference On Fuzzy System*, May 25-25, 2005, pp.114-119.

[2] Z. Wu, M. Palmer, "Verb semantics and lexical selection", *In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, June 1994, pp. 133-138.

[3] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy", *In Proceedings of LJCAI-95*, 1995, pp. 448-453.

[4] D. Lin, "An information-theoretic definition of similarity", *In Proceedings of the 15th International Conference on Machine Learning*, 1998, pp. 296–304.

[5] Roddick, J. K. Hornsby, D. de Vries, "A unifying semantic distance model for determining the similarity of attribute values", *Proceedings of the 26th Australasian Computer Science Conference*, 2003, pp. 111-118.

[6] A. Budanitsky, G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness", *Computational Linguistics*, Vol. 32, No. 5, 2006, pp. 13-47.

[7] S. Miralaei, A. A. Ghorbani, "Category-based similarity algorithm for semantic similarity in multi-agent information sharing systems", *In Proceedings of IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, Sept, 2005, pp. 242-245.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[8] ZhangXianLi, ZhouJunLi, MengJun, "An approach and implementation to semantic Web service matchmaking", *Computer Science*, Vol. 34, No. 5, 2007, pp. 99-103.

[9] P. Resnik, "Semantic similarity in a taxonomy: an information-based measure and its application intelligence to problems of ambiguity and natural language", *Journal of Artificial Intelligence Research*, Vol. 11, 1999, pp. 95-130.

[10] J. Hau, W. Lee, J. Darlington, "A semantic similarity measure for semantic Web services", *The 4th International World Wide Web Conference*, Chiba, Japan, May 10-14, 2005.

[11] Fangfang Liu, Yuliang Shi, Jie Yu et al, "Measuring similarity of Web services based on WSDL", *IEEE 8th International Conference on Web Services*, July 5-10, 2010, pp. 155-162.