Advances in Computer Science Research, volume 91

International Conference on Modeling, Simulation and Big Data Analysis (MSBDA 2019)

The Method for Semantic Similarity Based on Concept Distance

Xinying Chen[1,2], Guanyu Li[1,*], Heng Chen[1,3], Yunhao Sun[1] and Wei Jiang[1]

[1]College of Information Science and Technology, Dalian Maritime University, Dalian Liaoning 116026, China

[2]School of Software Technology, Dalian Jiao Tong University, Dalian Liaoning 116028, China

[3]School of Software Technology, Dalian University of Foreign Languages, Dalian Liaoning 116044, China

*Corresponding author

Keywords: Semantic web of things, Semantic matching, Service discovery, Semantic concept similarity

Abstract. Semantic matching is an important problem of service discovery. In order to find effective services, a method for semantic concept similarity is proposed. The method calculates the concept similarity between the parameter concepts of services by directly using the distance relationship between the concept nodes in the classification tree. And uses the nonlinear function to calculate the similarity and redefine the concept-based similarity between concepts. The new method effectively solves problems in existing algorithms and further improves precision. Finally, theoretical analysis and experimental result reveals the validity of the proposed method.

Introduction

In 1999, Ashton first proposed the Internet of Things [1]. The introduction of semantic annotation and ontology can greatly improve the ability of agents to understand and reason related information, resulting in a qualitative improvement in the function of the Internet of Things. This article refers to this smart Internet of Things as the Semantic Web of Things (SWoT)[2].

Web services [3] use standard format information such as WSDL. Using Semantic Web technology to extend Web services, semantic Web services with semantic tag information can be obtained. Internet of Things services are generated on the basis of ontology, specification of web services, and the addition of electronic tags.

Semantic Web of Things services and other services are not completely divided. After the services in the network are mapped with physical entity devices after semantic annotation, they have the characteristics of Semantic Web of Things services and can be considered as Semantic Web of Things services (such as Figure I) [4].

The service combinations and discovery methods adopted documents are mostly directed to static Web service information. At present, most existing service interface matching relationships [5-11] are based on the traditional interface matching model [8]. However, because the interface matching model is too strict, it will often affect the validity of service discovery algorithms. As a result, the quality of service is declining, so that the returned service cannot truly meet the needs of users. To solve the problem, this paper proposes a concept distance-based semantic similarity calculation method and proves the validity of the algorithm with experiments.

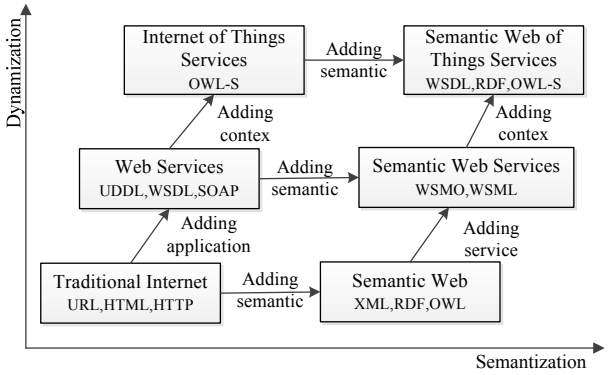

Figure 1. Semantic Web of Things services

The Method for Semantic Similarity Based on Concept Distance

The Semantic Web of Things service uses the domain ontology to describe the service, and the service description in the service registration system is a service profile form. At the same time, the service requirement description can also be considered as a profile.

Semantic matching is the problem of service composition among sub-services. It mainly solves the matching of attribute parameters among sub-services. It mainly deals with the semantic matching between the parameter concept pairs of subservices and the semantic matching between the parameter concept sets of subservices. Among them, the prerequisite for semantic matching any two sets of service parameters is to first solve the problem of semantic matching between arbitrary pairs of parameter concepts. At present, in order to solve the problem of semantic matching of parameter concept pairs, methods based on information sharing or tree (graph) are often used. In contrast, it is considered that the use of the latter (such as using a more intuitive conceptual tree approach) to calculate similarity is ideal [7-11].

So far, the researchers have conducted in-depth research on the problem of semantic matching and semantic similarity between the concepts in the domain ontology database of Web services matching, and proposed various calculation methods [8-11]. Among them, the more classic algorithm was proposed by Paolucci [8]. The algorithm associates the input and output of Web service with the upper and lower relations between concepts, and then classifies the semantic matching relations between the semantic Web service interface concepts into four levels: exact, subsume, plugin, and fail. This algorithm has become a classic algorithm for matching semantic web services. Although this algorithm is simple and easy to implement, it can't perform effective semantic quantification on services.

The literatures [9-10] are all based on the concept matching principle [8], and have improved the semantic-based service discovery methods. However, they still cannot accurately express the similarity between concepts, and they cannot perform semantic quantitative analysis on the overall composition service. Therefore, the similarity should be calculated by means of the related function [11].

If the relationship between the concepts in the ontology is expressed as a tree structure relationship, then when evaluating the similarity between two service parameter concepts, the relationship between the concepts in the WordNet classification tree structure can be referred to. WordNet classification tree structure can refer to Figure II. Replace the concept node of WordNet with the parameter concept node of service, and replace the WordNet classification tree with the domain ontology classification tree of the corresponding service. Then, the concept similarity between the parameter concepts of the two services can be calculated by directly using the distance relationship between the concept nodes in the classification tree.

In addition, we should consider a series of relationship parameters between two concept nodes to design the similarity function [11]. On this basis, the main idea of conceptual similarity calculation method based on concept distance is given. For any two attribute parameter concepts x and y, the

following factors can be considered, including $D(x,y)$ (the semantic distance between x and y), $Lmin(x,y)$ (the minimum distance between x and y), $H(x)$ (the node depth of the concept node x in the tree structure), $H(\text{Root})=0$, and $Bnum(x)$ (the concept density of the concept node x).

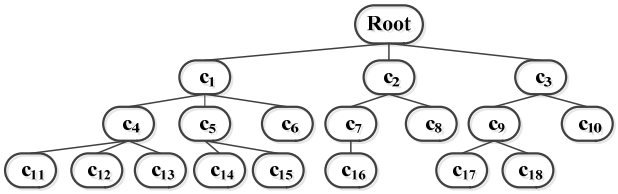

Figure 2. The case of the WordNet classification tree structure

These parameters should have the following attribute characteristics. $D(x,y)\in[0,1]$, $D(x,y)=1$ means that x and y match exactly. $D(x,y)=0$ indicates that x and y match failed. $D(x,x)=1$ indicates that the matching degree function is reflexive. $D(x,y)=D(y,x)$, indicates that the matching degree function has symmetry.

In addition, $D(x,y)$ should be inversely proportional to $Lmin(x,y)$, and the smaller $Lmin(x,y)$ is, the higher the similarity between concept nodes is. For example, given concept nodes c_1, c_4 and c_{11}, it can be seen from Figure II that $Lmin(c_4,c_1)=1$, and $Lmin(c_{11},c_1)=2$. It is easy to see that the conceptual relationship between c_4 and c_1 is more closely related to the concept relationship between c_{11} and c_1. That is, $D(c_4,c_1)>D(c_{11},c_1)$ holds.

If $Lmin(x,y)$ and $|H(x)-H(y)|$ are determined (that is, in the same minimum distance and the same hierarchy difference), $D(x,y)$ should be proportional to $H(x)+H(y)$, because the lower the level is, the higher the density of concepts is, the smaller the difference between concepts is. For example, given concept nodes c_2, c_4, c_5 and c_{11}, it can be seen from Figure II that $Lmin(c_4,c_2)=3$, $|H(c_4)-H(c_2)|=1$, $Lmin(c_{11},c_5)=3$ and $|H(c_{11})-H(c_5)|=1$. Since $H(c_4)+H(c_2)=3<H(c_{11})+H(c_5)=5$, $D(c_4,c_2)<D(c_{11},c_5)$ holds.

If $Lmin(x,y)$ and $H(x)+H(y)$ are determined, $D(x,y)$ should be inversely proportional to $|H(x)-H(y)|$, namely the greater the level difference between the concept, the smaller the $D(x,y)$ is. For example, given concept nodes c_{11}, c_1, c_5, and c_4, it can be seen from Figure II that $Lmin(c_{11},c_1)=2$, $H(c_{11})+H(c_1)=4$, $Lmin(c_5,c_4)=2$ and $H(c_5)+H(c_4)=4$. Since $|H(c_{11})-H(c_1)|=2>|H(c_5)-H(c_4)|=0$, $D(c_{11},c_1)<D(c_5,c_4)$ holds.

If $Lmin(x,y)$, $H(x)$ and $H(y)$ are determined, $D(x,y)$ is proportional to $Bnum(x)+Bnum(y)$. For example, given concept nodes c_{11}, c_1, c_5, and c_4, it can be seen from Figure II that $Lmin(c_{11},c_1)=2$, $H(c_{11})+H(c_1)=4$, $Lmin(c_5,c_4)=2$ and $H(c_5)+H(c_4)=4$. Since $Bnum(c_{11})+Bnum(c_1)=11<Bnum(c_5)+Bnum(c_4)=14$, $D(c_{11},c_1)<D(c_5,c_4)$ holds.

Based on the above analysis, this paper will use the nonlinear function [11] to calculate the similarity and redefine the concept-based similarity between concepts as:

$$D(x,y)=\begin{cases}1 & x=y\\ e^{-\alpha w_1}*\dfrac{e^{\beta w_2}-e^{-\beta w_2}}{e^{\beta w_2}+e^{-\beta w_2}} & x\neq y\end{cases} \tag{1}$$

$$w_1=Lmin(x,y)+\left|H(x)-H(y)\right| \tag{2}$$

$$w_2=\frac{\left(H(x)+H(y)\right)}{2*n}+\frac{Bnum(x)-Bnum_{\min}}{Bnum_{\max}-Bnum_{\min}}+\frac{Bnum(y)-Bnum_{\min}}{Bnum_{\max}-Bnum_{\min}} \tag{3}$$

In the above formula, n is the tree height; $\alpha\geq 0$ and $\beta>0$ are smoothing factors, which are used to control the influence of w_1 and w_2 on semantic similarity. This formula guarantees $D(x,y)\in[0,1]$. The value of α, β depends to a large extent on the structure of the domain ontology tree used. Based

on formula (1), the calculation method of semantic matching degree between two concept sets is given.

Definition 1 (Parameter Dependencies between Concepts). Given two sets of ontology concepts X and Y, and semantic similarity based on concept distance $D(x,y)$, $\forall x \in X$ and $\forall y \in Y$ satisfy $D(x,y) \in [0,1]$. If $\exists x \in X$, $\exists y \in Y$, satisfy the similarity between the two parameters $D(x,y) \geq \sigma$. It is considered that there is a parameter dependency between x and y, expressed as $x \leftrightarrow y$. Where $\sigma \in [0,1]$ represents the system's conceptual parameter similarity threshold.

Definition 2 (Semantic Matching between Concept Sets). Given two concept sets X and Y, $\forall x \in X$ and $\forall y \in Y$ satisfy $D(x,y) \in [0,1]$. $M(X,Y)$ is a semantic matching between X and Y if and only if the following conditions are true:

(1) $\forall \langle x_1,y_1\rangle, \langle x_2,y_2\rangle \in M(X,Y), x_1 \neq x_2, y_1 \neq y_2$.

(2) $\bigcup_{\forall x_i,y_i \in M(X,Y)} x_i = X, \bigcup_{\forall x_i,y_i \in M(X,Y)} y_i = Y$.

(3) $\forall \langle x,y\rangle \in M(X,Y), x \leftrightarrow y$.

Definition 3 (Optimal Semantic Matching between Concept Sets). Given two concept sets X and Y, $\forall x \in X$ and $\forall y \in Y$ satisfy $D(x,y) \in [0,1]$. $Max(X,Y)$ is a optimal semantic matching between X and Y if and only if: $Max(X,Y)$ is a semantic matching between X and Y, and for any semantic matching $M(X,Y)$, satisfy $\sum_{D(x,y) \in Max(X,Y)} D(x,y) \geq \sum_{D'(x,y) \in M(X,Y)} D'(x,y)$. It is also called $\left(\sum_{D(x,y) \in Max(X,Y)} D(x,y)\right)/n$ the optimal semantic matching degree between concept sets X and Y.

To facilitate the follow-up discussion and analysis, the following related concepts are proposed.

Definition 4 (Function Operation) A function operation *Fun* can be formalized as a 7-tuple $Fun=(Inf,In,Out,F_{IO},Pre,Post,Qos)$:

Inf represents the feature description information of *Fun*. $In=\{in_1,in_2,\ldots,in_m\}$ represents the input parameter set of *Fun*. $\forall in_t \in In$, $1 \leq t \leq m$, in_t represents a specific input parameter. $Out=\{out_1,out_2,\ldots,out_n\}$ represents the output parameter set of *Fun*. $\forall out_t \in Out$, $1 \leq t \leq n$, out_t represents a specific output parameter. $Pre=\{pre_1,pre_2,\ldots,pre_l\}$ represents a set of preconditions. $Post=\{post_1,post_2,\ldots,post_k\}$ represents a set of post-conditions indicating the effect on the current state after the execution of *Fun*. F_{IO} represents the inter-parameter dependency function of Fun's output input interface. $Qos=\{Qos_1,Qos_2,Qos_3,Qos_4\}$ represents a set of non-functional attributes of *Fun*.

Definition 5 (Service) An abstract service S can be formalized as a 2-tuple $S=(INF,FUN)$. *INF* represents the feature description information of S. $FUN=\{Fun_1,Fun_2,\ldots,Fun_w\}$ represents a set of function operations provided by S.

Definition 6 (Service Request) A service request R can be formalized as a 5-tuple $R=(In^R,Out^R,Pre^R,Post^R,W)$. $\forall w \in W$, $w \in [0,1]$, represents a specific threshold set by the service demander.

Based on the above method for semantic similarity based on concept distance, the semantic matching method between service parameter concept sets is as follows:

Step 1. Set the smoothing factors α and β and set similarity threshold k.

Step 2. Parse the user's service request and get the five-tuple form of service request $R=(In^R,Out^R,Pre^R,Post^R,W)$.

Step 3. Complete the semantic matching of each parameter concept set between candidate services and the service request R using Equation (1).

Step 4. Sort the results by user requirements and context weights. Select the best result in the sort results.

Experiment Analysis

This paper automatically generates a service test dataset based on the simulation data design platform (WS-Ben) [3]. In the same way, a service demand dataset is generated using the method of simulation generation and semantically annotated with the domain ontology source.

In order to verify the feasibility and validity of the method proposed in this paper, the simulation experiments were conducted on the Intel Core 1.00GHz, 4.00GB RAM, Windows XP and Matlab 7.1 environment for comparison experiments on algorithms.

This paper chooses the approach for measuring semantic similarity between words using multiple information sources (SSMIS) [11] to conduct comparative experiments. Figure III shows comparative analysis of time performance between methods, and Figure IV shows comparative analysis of precision between methods.

After analysis, it is found that the method MSSCD proposed in this paper can obtain better results and its ratio of precision all more than 93%, which is higher than the comparative method SSMIS. And the method MSSCD has the better time performance than the comparative method SSMIS. Therefore, the method MSSCD can provide more reasonable solutions on the premise of lower cost.

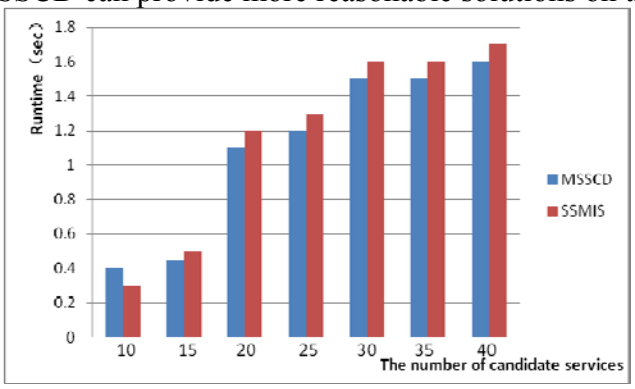

Figure 3. Analysis of time performance between methods

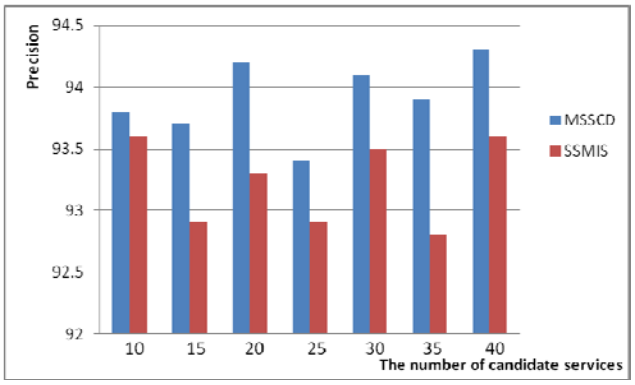

Figure 4. Analysis of precision between methods

Conclusion

In this paper, a method for semantic concept similarity is proposed based on the concept distance relationship. And related equations and definitions are drawn out. The method calculates the concept similarity between the parameter concepts of services by directly using the distance relationship between the concept nodes in the classification tree. And uses the nonlinear function to calculate the similarity and redefine the concept-based similarity between concepts. Theoretical analysis and experimental result reveals the validity of the proposed algorithm. The new method has better time performance and higher accuracy.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant (No.61371090, No.61602076 and No.61702072), the China Postdoctoral Science Foundation Funded Project (2017M621122 and 2017M611211), the Natural Science Foundation of Liao-Ning

Province (No.20170540144 and No.20170540232) and the Fundamental Research Funds for Central Universities (No.3132017118, No.3132017121 and No.3132017123).

References

[1] Aggarwal C C, Ashish N, Sheth A, "The internet of things: A survey from the data-centric perspective,"Managing and Mining Sensor Data. Springer US, 2013, pp.383-428.

[2] Wu X H, Shi Y M, Li G Y, et al, "Method of multi-domain information interoperability in semantic Web of things," Application Research of Computers, 2016, 33(09), pp.2726-2730.

[3] Deng S G, Yin J W, Li Y, et al, "A method of semantic web service discovery based on bipartite graph matching," Chinese Journal of Computers, 2008, 31(8), pp.1364-1375.

[4] Wang H, Zhu B, Li G, et al, "A Fusion of Granulation and Artificial Neural Network: A New Service Selection Method[C]// International Symposium on Computational Intelligence and Design. IEEE, 2017, pp.342-347.

[5] Bruijn J, "Logics for the semantic web," Semantic Web Services Theory Tools & Applications, 2007, 16(4), pp.4--9.

[6] Bianchini D, Antonellis V D, Melchiori M, "Flexible semantic-based service matchmaking and discovery," World Wide Web-internet & Web Information Systems, 2008, 11(2), pp.227-251.

[7] Yang H R, Liu S S, Yin B C, et al, "Matching algorithm of web services based on semantic distance," Journal of Beijing University of Technology, 2011, 37(4), pp.591-595.

[8] Paolucci M, Kawamura T, Payne T R, et al, "Semantic matching of web services capabilities," International Semantic Web Conference on the Semantic Web. Springer-Verlag, 2002, pp.333-347.

[9] Peng H, Shi Z Z, Qiu L R, et al, "Matching algorithm of semantic web service based on similarity of ontology concepts," Computer Engineering, 2008, 34 (15), pp.51-53.

[10] Wu J, Wu Z H, Li Y, et al, "Web service discovery based on ontology and similarity of words,"Chinese Journal of Computers, 2005, 28(4), pp.595-602.

[11] i Y, Bandar Z A, McLean D, "An approach for measuring semantic similarity between words using multiple information sources," IEEE Transactions on knowledge and data engineering, 2003,15(4), pp.871-882