

Web 服务发现框架,将语义 Web 服务的发现过程分为 4 个阶段匹配:服务类别、服务功能、服务名称、服务文本语句等的匹配。本文的语义 Web 服务发现框架如图 1 所示。

上述语义 Web 服务发现框架可以分成下面几个模块:OWL-S 服务文件构造、OWL-S/UDDI 映射、UDDI 及本体库、Web 服务匹配模块。

①OWL-S 服务文件构造。根据服务提供商和服务请求者提供的服务相关信息,结合领域本体库,生成语义 Web 服务的描述文件 OWL-S。

②OWL-S/UDDI 映射。用于实现 UDDI 的语义扩展。应用 OWL-S 与 UDDI 一一映射机制将服务的 OWL-S 中的 Service Profile 内容置入 UDDI 中。

③UDDI 及本体库。UDDI 中心用于存储发布的 Web 服务^[8]。模型采用 UDDI 规范的 4 种模型 businessEntity, businessService, bindingTemplate 和 tModel 以及 IBM UDDI4J 提供的发布 API、查询 API。本体库则包含两种本体^[3]:一种是领域本体库,为服务功能描述提供一个共同认知基础;一种是语言学本体库,保证服务的文本描述内容能够被计算机认知和理解^[4]。

④Web 服务匹配。该模块是整个语义 Web 服务发现框架的核心,该模块将语义 Web 服务描述信息的匹配分成 4 个阶段,一方面能使语义 Web 服务的匹配更加全面,另一方面,允许服务请求者对匹配过程进行控制,及时剔除匹配相似度较低的服务。Web 服务匹配的过程可以用如下的几个步骤来描述。

第一步 提供服务有关信息。服务请求者根据发现框架中提供的应用程序接口,提供其需要服务的信息,如服务的名称、服务的类别、服务的输入/输出、服务的文本语句的描述、服务类别匹配相似度的阈值 Sim_1 、服务功能匹配相似度的阈值 Sim_2 、服务名称匹配相似度的阈值 Sim_3 、服务文本语句匹配相似度的阈值 Sim_4 、最终匹配相似度的阈值 Sim_0 以及 4 个阶段相似度所占的比重 a_1, a_2, a_3, a_4 。

第二步 构造 OWL-S。发现框架中的 OWL-S 服务文件构造模块,将服务提供者提供的相关信息,按照 OWL-S 的规范生成 OWL-S 描述文档。

第三步 获得服务集合。服务代理商获得服务请求者的请求后,从 UDDI 中心获得广告服务的集合,为执行服务 4 个阶段的匹配提供匹配的服务集合。

第四步 执行服务类别、功能、名称、文本语句匹配。从第三步中的 OWL-S 集合中取出一个服务描述信息,依次执行功能匹配、名称匹配、文本描述匹配,并获得相应阶段的相似度 $Sim_{cat}, Sim_{fun}, Sim_{nam}, Sim_{des}$ 。同样,如果该阶段的相似度值小于用户设定的阈值,则执行下一个服务的匹配。

第五步 计算两服务的最终相似度值 Sim ,其中, Sim 的计算公式如下:

$$Sim = a_1 Sim_{cat} + a_2 Sim_{fun} + a_3 Sim_{nam} + a_4 Sim_{des}$$

式中, $a_1 + a_2 + a_3 + a_4 = 1$ 。

如果 $Sim \geq Sim_0$,则将该服务加入满足条件的服务集合。当匹配完所有的广告服务集合后,将满足条件的服务集合按 Sim 排序,并将该服务集合返回给用户。

2 语义 Web 服务多阶段匹配过程

2.1 语义 Web 服务的类别匹配

在整个发现过程的第一个阶段,语义 Web 服务的类别是由 OWL-S Service Profile 中的 Service Category 进行描述的。本文中研究的服务采用的分类标准是 NAICS(北美工业分类系统),同时也采用其典型的相似度计算方法。

假设广告服务的服务类别为 $AScategory$,请求服务的服务类别为 $RScategory$ 。上述两个类别匹配的结果可以分为如下 4 种情况:情况 1, $AScategory = RScategory$,两者分类属性完全相同;情况 2, $AScategory \subseteq RScategory$,表示 $AScategory$ 是 $RScategory$ 的一个子节点,也就是提供服务类别是请求服务类别的一个子集;情况 3, $RScategory \subseteq AScategory$,表示 $RScategory$ 是 $AScategory$ 的一个子节点;情况 4,其他情况。定义服务类别匹配的相似度为 Sim_{cat} ,即

$$Sim_{cat} = \begin{cases} 1 & \text{情况 1, 情况 2} \\ \frac{Count(RScategory)}{Count(AScategory)} & \text{情况 3} \\ 0 & \text{情况 4} \end{cases}$$

式中, $Count(RScategory)$ 表示请求服务类别所包含的类别个数的总和, $Count(AScategory)$ 表示广告服务类别所包含的类别的个数总和。

2.2 语义 Web 服务的功能匹配

语义 Web 服务的功能匹配是整个发现过程中的第二个阶段。语义 Web 服务功能的内容是由 OWL-S Service Profile 中的 Input/Output 进行描述的。该阶段的任务是将请求服务与广告服务的 Input/Output 分别进行匹配,并可得出请求服务与广告服务在功能上的相似度值 Sim_{fun} 。

设请求服务的输入参数为 req_{in} ,请求服务的输出参数为 req_{out} ;广告服务的输入参数为 adv_{in} ,广告服务的输出参数为 adv_{out} 。设 IN_{req} 是请求服务的输入参数集合, OUT_{req} 是请求服务的输出参数集合,则输入/输出参数集合 $IN_{req} = \{req_{in1}, req_{in2}, req_{in3}, \dots, req_{inm}\}$; $OUT_{req} = \{req_{out1}, req_{out2}, req_{out3}, \dots, req_{outn}\}$ 。同理,广告服务的输入/输出参数集合可表示为 $IN_{adv} = \{adv_{in1}, adv_{in2}, \dots, adv_{inm}\}$; $OUT_{adv} = \{adv_{out1}, adv_{out2}, \dots, adv_{outn}\}$ 。在功能匹配时,需要计算请求服务和广告服务在输入/输出方面的相似度 Sim_{in}, Sim_{out} 。最后,语义 Web 服务在功能匹配阶段的相似度 Sim_{fun} 可表示为如下公式:

$$Sim_{fun}(req, adv) = a Sim_{in} + (1-a) Sim_{out}$$

式中, $a \in [0, 1]$ 。

在输入/输出方面的计算过程可描述为如下步骤:

第一步 计算两个领域本体概念之间的相似度。

针对领域本体概念相似度的计算,当前的研究都只是考虑了本体概念间的深度对概念相似度的影响,而对本体概念之间的密度考虑不足,如基于语义距离的改进 GCSM 算法。因此本文充分考虑本体库中概念之间密度和深度的因素,提出了基于信息量的改进 GCSM 算法来计算领域本体概念之间的相似度。

首先,我们给出如下定义。

定义 1 $IC(c) = -\log^p(c)$, c 为领域本体概念树中的任意概念, $p(c)$ 为任意一个实例属于概念 c 的概率, $IC(c)$ 为概念 c 的语义信息量。在本体概念层次树中, $P(c)$ 可认为是单调的,即如果存在概念 c_1 包含概念 c_2 ,则 $P(c_1) \geq P(c_2)$ 。

定义 2 $P(c) = P(p)/n$, $P(R) = 1$,其中概念 p 为概念 c 的父概念, n 为概念 p 直接子概念的数量,概念 R 为领域本体

中的根概念。

通过上述定义可以计算每个概念的信息容量,图2表示带信息容量的 Artifact 本体库片段的示例。在图2中,每个概念的表示如 Thing(1,0)所示,其中 Thing 表示概念的名称,括号中的二元组分别表示该概念出现的概率和该概念所含有的信息量,(1,0)则表示概念 Thing 出现的概率为1,其含有的信息量为0。

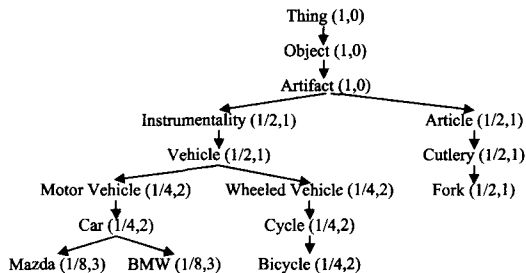


图2 带信息容量的 Artifact 本体库片段

改进后的 GCSM 的定义如下所示:

$$GCSM(con_1, con_2) = \frac{IC(con_1) + IC(con_2)}{2 \times IC(LCA(con_1, con_2))}$$

式中, $LCA(con_1, con_2)$ 是概念 con_1 和概念 con_2 的共同祖先中深度最大的祖先节点。IC 函数代表本体库中一个概念所含的信息量。同样,改进后的相似度可表示为:

$$sim(con_1, con_2) = \frac{1}{GCSM(con_1, con_2)}$$

通过上述公式就可以计算本体库中两个概念之间的相似度。

第二步 根据第一步中的相似度值,计算请求服务的输入参数集合 IN_{req} 和广告服务的输入参数集合 IN_{adv} 之间的相似度 Sim_{in} ,同理可以得出广告服务和请求服务在输出参数集合上的相似度 Sim_{out} 。

在计算两个本体概念的相似度的基础上,可以通过概念间的相似度进一步计算请求服务和广告服务在功能上的相似度。由于输入/输出参数集合中含有多个输入/输出的参数,因此在该步骤中,采用了参数扩展的方式进行处理,步骤如下:

①若请求服务的输入参数和广告服务的输出参数个数都为0,则请求服务和广告服务在输入参数上是完全匹配的。 $Sim_m(req, adv) = 1$,计算结束。

②若请求服务的输入参数或广告服务的输入参数个数为0,则两服务在输入参数上是不匹配的。 $Sim_m(req, adv) = 0$,计算结束。

③如果 m 和 a 均不为0,则可以将请求服务和广告服务的输入条件建模成一个二部图 $G = (IN_{req}, IN_{adv}, Sim_{in})$,其中 IN_{req} 表示请求服务的输入集合, IN_{adv} 表示广告服务的输入集合, Sim_{in} 表示请求服务的输入参数与广告服务的输入参数的相似度集合。在这种情况下,将参数个数较少的服务的参数扩展到与另一个服务的参数个数相等,并且扩展的参数与另一个服务的每个参数间的相似度为0,这样就形成了一个完备二部图,进而采用 KM 算法进行处理,最终便可以计算广告服务与请求服务在输入上的相似度,同理,也可以求出它们在输出上的相似度。最终,可得请求服务和广告服务在功能上的相似度。

2.3 语义 Web 服务的名称和文本语句匹配

语义 Web 服务的名称和文本语句的内容分别是由 OWL-S Service Profile 中的 Service Name 和 Text Description 描述的。本文研究的语义 Web 服务是采用中文来描述的,因此,考虑语义 Web 服务名称和文本语句的匹配,就是考虑两个简短的中文词汇组和两个中文语句的相似程度。这两个方面的匹配步骤中,本文主要是对词义消歧进行重点考虑,其他的都比较典型的方式进行处理。因此,词义消歧的质量直接影响服务名称匹配的好坏。

知网中的每一个义项都含有两个重要的内容:实例、基本义原。实例用来表达当前词语的搭配关系;基本义原表示当前词语的基本属性。本文则提出基于实例和基本义原的综合消歧策略。在采用知网消歧之前必须要对知网中的数据进行提取并形成相应的知识库,为消歧提供基础。针对实例、基本义原,本文为多义词建立两个知识库:实例搭配库、基本义原库。

①实例搭配库

实例搭配库的数据结构定义如下:

$\langle No, Word, SenseNo, Tie_W_No, Tie_W, Tie_W_C, Tag \rangle$

其中, No 表示该条实例搭配在实例搭配库中的序号; $Word$ 表示多义词 W 的名称; $SenseNo$ 表示该义项的编号; Tie_W_No 表示该实例搭配在 W 的实例搭配中的序号; Tie_W 表示该实例搭配的名称; Tie_W_C 表示该实例搭配的词性; Tag 表示该实例搭配与多义词 W 的相对位置,其取值为左或者右。

②基本义原库

基本义原库的建立过程:从知网中,获取多义词 W 的 DEF 属性的内容,提取 DEF 中的基本义原,将其保存到结构化的数据结构中。其数据结构的定义如下:

$\langle No, Word, SenseNo, BSense_No, BSense_Word \rangle$

其中, No 表示该基本义原在基本义原库中的序号; $Word$ 表示多义词 W 的名称; $SenseNo$ 表示该义项的编号; $BSense_No$ 表示该基本义原在义项中的序号; $BSense_Word$ 表示该基本义原的名称。

在完成上述两个知识库的构造之后,便能开始进行语义消歧。语义消歧的过程可以分为4个步骤。

①实例搭配计算:采用实例搭配计算主要是用来反映多义词 W 与周围词汇的搭配关系。根据实例搭配库找出多义词 W 的一个义项 W_i 的搭配词汇组 $\{Tie_W_1, Tie_W_2, \dots, Tie_W_n\}$ 。计算搭配词汇 Tie_W_a 与多义词 W 周围词汇的相似度可采取如下策略:

首先从多义词 W 的 Tag 方向找出第一个与搭配词 Tie_W_a 词性相同的词汇 $Word_Tag$,计算 Tie_W_a 与 $Word_Tag$ 的词语相似度 Sim_{tie_a} 。

如果在 Tag 方向没有找到词性相同的词汇,则从 Tag 的反方向找出第一个与搭配词 Tie_W_a 词性相同的词汇 $Word_Optag$ 。计算 $Word_Optag$ 与 $Word_Tag$ 的词语相似度 Sim_{tie_a} 。

如果在两个方向均没有找到词性相同的词汇,则计算下一个搭配词汇。

通过上述策略可以计算出义项 W_i 的实例搭配相似度数组 $\{Sim_{tie_1}, Sim_{tie_2}, \dots, Sim_{tie_m}\}$,其中 $m \leq n$,所以义项 W_i 实例搭配的相似度:

$$Sim_Tie_i = \frac{\sum_{a=1}^m Sim_tie_a}{m}$$

②基本义原计算:采用基本义原计算,主要是用来反映多义词 W 与句子语境之间的关系。根据基本义原库可知多义词 W 的基本义原组 $\{BSense_1, BSense_2, \dots, BSense_k\}$ 和语句除去多义词外的实词组 $\{ConWord_1, ConWord_2, \dots, ConWord_j\}$ 。则基本义原与实词组之间的相似度构成一个相似度矩阵:

$$\begin{matrix} & CWord_1 & CWord_2 & \cdots & CWord_j \\ \begin{matrix} BSense_1 \\ BSense_2 \\ \vdots \\ BSense_k \end{matrix} & \begin{bmatrix} BSim_{11} & BSim_{12} & \cdots & BSim_{1j} \\ BSim_{21} & BSim_{22} & \cdots & BSim_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ BSim_{k1} & BSim_{k2} & \cdots & BSim_{kj} \end{bmatrix} \end{matrix}$$

通过该相似度矩阵可得义项 W_i 的基本义原与语句的相似度:

$$Sim_Base_i = \frac{\sum_{a=1, b=1}^{k, j} BSim_{ab}}{j \times k}$$

③计算义项 W_i 的综合相似度。

$$Sim_Com_i = \frac{Sim_Tie_i + Sim_Base_i}{2}$$

④对多义词的每一个义项执行上述三步计算,假设 Sim_Com_x 为其中最大的值,则 W_x 所代表的意思即为本语句中的含义。

为验证本文提出的消歧策略的可行性和有效性,本文选取人民日报语料库中的文档作为词义消歧的对象。通过与文献[5]中的消歧方法进行比较,可以得出本文消歧策略的可行性和有效性,具体数据如表1所列。

表1 选定文档的词义消歧结果

文档标号	多义词数	文献[5]		本文消歧策略	
		正确数	准确率	正确数	准确率
980101-01-01	164	126	76.83%	131	79.88%
980101-01-02	152	126	82.89%	132	86.84%
980101-01-03	38	30	78.95%	31	81.58%
980101-01-04	88	70	79.55%	73	82.95%
980101-02-01	66	49	74.24%	50	75.76%
980101-02-03	39	23	58.97%	30	76.92%
980101-02-06	41	34	82.93%	35	85.37%
980101-02-07	103	84	81.55%	89	86.41%
980101-01-08	34	26	76.47%	29	85.29%
980101-03-01	65	54	83.08%	58	89.23%

从实验数据来分析,本文所采用的消歧策略相对来说有比较高的准确率,可以达到较好的词义消歧效果,进而保证了语义 Web 服务的名称匹配的正确性。

4 实验结果和性能分析

为了验证上述策略的可行性和有效性,做了相应的测试实验。在对 a_i 的选取为 $a_1=0.25, a_2=0.25, a_3=0.25, a_4=0.25$, 并设定各阶段相似度的阈值为 0.60, 且最终相似度的阈值为 0.90 的情况下,对测试集中的 175 个广告服务进行 40 次实验测试,并将 40 次的实验数据平均分为 10 组。表2所列为本文的发现框架和文献[6]的发现策略在这 10 组实验中的查全率和查准率^[7]。

表2 两种发现框架的查全率和查准率

实验组序号	查全率		查准率	
	文献[6]	本文	文献[6]	本文
1	88.7%	92.5%	82.6%	89.2%
2	89.3%	93.3%	82.1%	90.0%
3	88.8%	92.6%	81.7%	89.4%
4	89.5%	93.0%	82.0%	90.2%
5	88.1%	93.1%	82.5%	89.3%
6	88.3%	92.9%	81.9%	90.2%
7	87.6%	93.0%	81.8%	89.6%
8	89.5%	91.7%	82.7%	90.0%
9	88.3%	93.1%	82.0%	89.7%
10	89.4%	92.6%	82.0%	89.6%

通过上述实验结果可以分析出,本文的语义 Web 服务发现策略在查全率/查准率方面高于文献[6],因此,认为本文提出的多阶段匹配的语义 Web 服务发现框架结构是可行的、有效的。

参考文献

- [1] Massimo P, Takahiro K, Payne Terry R, et al. Importing the semantic Web in UDDI[C]//Proceedings of Web Services, E-business and Semantic Web Workshop. Toronto, Canada, 2002: 225-236
- [2] Martin D, et al. OWL-S: Semantic Markup for Web Services[R]. Damlconsortium, 2009-3-9. <http://www.daml.org/services/owl-s/1.0/owl-s.pdf>
- [3] 杜小勇, 李曼, 王珊. 本体学习研究综述[J]. 软件学报, 2006, 17(9): 1837-1847
- [4] 吴健, 吴朝辉, 李莹, 等. 基于本体论和词汇语义相似度的 Web 服务发现[J]. 计算机科学, 2005, 28(4): 595-602
- [5] 张明宝, 马静. 一种基于知网的中文词义消歧算法[J]. 计算机技术与发展, 2009, 2(19): 9-15
- [6] 胡建强, 邹鹏, 王怀民, 等. Web 服务描述语言 QWSDL 和服务匹配模型研究[J]. 计算机学报, 2005, 28(4): 505-513
- [7] 邓汉成, 王瑛, 王敏芳. 从检索实例看查全率与查准率之间的关系[J]. 情报学报, 2000, 19(3): 237-241
- [8] 田晓珍, 尚冬娟. Web 的个性化服务[J]. 重庆工学院学报: 自然科学版, 2008, 22(7): 76-80
- [9] works[C]//The 2008 International Computer Symposium(ICS 2008). 2008: 503-509
- [8] Ramachandran C, Misra S, Obaidat M S. FORK: a novel two-pronged strategy for an agent-based intrusion detection scheme in ad-hoc networks[J]. Computer Communications, 2008; 31(16): 3855-3869
- [9] Zhou B, Shi Q, Merabti M. Balancing intrusion detection resources in ubiquitous computing networks[J]. Computer Communications, 2008, 31(15): 3643-3653
- [10] Vigna G, Gwalani S, Srinivasan K, et al. An intrusion detection tool for AODV-based ad hoc wireless networks[C]//Proceedings of the 20th Annual Computer Security Applications Conference. 2004: 16-27

(上接第 108 页)

- [4] Yi P, Jiang X H, Wu Y, et al. Distributed Intrusion Detection for Mobile Ad Hoc Networks [J]. Journal of Systems Engineering and Electronics, 2008, 19(4): 851-859
- [5] Sen S, John A C. A grammatical evolution approach to intrusion detection on mobile ad hoc networks[C]//Proceedings of the Second ACM Conference on Wireless Network Security. 2009: 95-102
- [6] Liu Y, Li Y, Man H, et al. A hybrid data mining anomaly detection technique in ad hoc networks[J]. International Journal of Wireless and Mobile Computing, 2007, 2(10): 37-46
- [7] Wang C H, Chin S. Reputation based intrusion detection system with threshold cryptography for wireless mobile ad hoc net-