

基于分级结构的服务语义关系发现

赵安平 任 柯 王晓勇 邱玉辉

(西南大学计算机与信息科学学院 重庆 400715) (西南大学语义网格实验室 重庆 400715)

摘 要 为了以一种快速有效的方式从巨大的 Web 服务库中检索出满足用户请求的服务,探索服务语义关系发现的方法来增强服务搜索引擎的搜索能力是必不可少的。提出了一个基于分级结构的服务语义关系发现的方法,分级结构能够自然地表示不同粒度服务间的语义关系,根据服务相关度,不断调整分级结构,将功能上相关的服务汇集在一起,不同粒度服务间的语义关系就可以逐步被发现。实验结果表明了方法的有效性和可行性。

关键词 分级结构,语义关系,Web 服务,关系发现

中图法分类号 TP301 **文献标识码** A

Hierarchical Structure Based Service Semantic Relationship Discovery

ZHAO An-ping REN Ke WANG Xiao-yong QIU Yu-hui

(Faculty of Computer and Information Science, Southwest University, Chongqing 400715, China)

(Semantic Grid Laboratory, Southwest University, Chongqing 400715, China)

Abstract Increasingly, application developers seek the ability to search for existing Web services within large Internet-based repositories. The goal is to retrieve services that match the user's requirements in a fast and efficient way. It is necessary to explore an approach of service semantic relationship discovery to enhance search engine results with a list of related services for each hit. In this paper, an effective approach for service semantic relationship discovery based on the hierarchical structure was presented. Hierarchical structure is adjusted by moving related services together according to the relevance between services, and the relationships between services can be discovered gradually. Experimental results demonstrate the feasibility and effectiveness of the approach.

Keywords Hierarchical structure, Semantic relationship, Web service, Relationship discovery

1 引言

由于互联网上可获取的 Web 服务数量不断增长,使得在分布式环境下,电子商务中的服务过程变得更加复杂,其中涉及无数的商务对象交互。因此,如何从巨大的服务空间中快速有效地检索出一组功能上相关的服务集合来满足特定的服务请求成为面向服务的社区研究的重要内容。除了提供与特定查询具有直接关系的服务,还应提供与其语义功能上紧密相关的一组服务来实现服务间的自动搜索和协作。

因此,研究服务间的语义关系发现方法,旨在最大限度地提高服务网络的效用。本文着重介绍如何利用一个能够自然地表示不同粒度的服务之间的语义关系的分级结构来进一步发现服务间的语义关系的方法。

2 基本思想

为了发现服务间蕴含的语义关系,必须考虑以下问题:

(1)对于任意两个组件服务,它们的语义关系是直接的还是间接的,如何去发现组件服务间的语义关系?(2)在实际的应用

中,不仅组件服务间具有语义关联,由多个组件服务构成的高粒度服务间也具有语义关联。如何发现高粒度服务之间的语义关系?

分级对于服务的分类非常重要,并且分级的结构能够自然地表示不同粒度的服务关系,因此,利用一个分级的结构来发现服务间的语义关系。采用分级的结构 RT 来表示服务之间的相关性,RT 中的每一个叶结点表示一个组件服务,每一个有一组孩子结点的非叶结点表示功能上紧密相关的组件服务构成的高粒度服务。在每一对结点 s_i 和 s_j 间,有一条关联一个实值 $w_{i,j}$ 的无向边 (s_i, s_j) , $w_{i,j}$ 用来表示 s_i 和 s_j 的相关度。一个分级结构如图 1 所示。

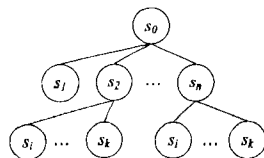


图 1 一个分级结构

基于一组功能上相关的服务集合来初始化 RT,由于没

到稿日期:2010-03-24 返修日期:2010-06-17 本文受 973 国家重大基础研究计划(2003CB317008),西南大学研究生科技创新基金(ky2009005)资助。

赵安平(1978—),男,博士生,主要研究方向为服务计算、语义网格,E-mail:zap@swu.edu.cn;任 柯(1977—),男,博士生,讲师,主要研究方向为语义网格;王晓勇(1972—),男,博士生,工程师,主要研究方向为语义网格;邱玉辉(1938—),男,教授,博士生导师,主要研究方向为人工智能、语义网格。

有考虑服务内部的相关性,初始的 RT 不能捕获服务之间真正的相关关系。使用一个迭代的方法来调整 RT 的结构,将每一对兄弟叶结点的相关度存储在 RT 中,非兄弟叶节点的相关度通过估计它们的祖先结点间(高粒度服务间)的相关度获得。RT 的结构通过在每一次迭代过程中将相关的结点移到一起而不断调整。而两个非叶结点间的相关度是和它们相链接的结点间的平均相关度。基于分级的结构,根据服务间的相关度,不同粒度服务间的语义关系可以被逐渐发现。

3 语义关系发现

3.1 问题定义

我们定义分级的结构表示不同粒度的服务,一组具有共同的父结点的组件服务间具有相关的语义关系。由一组组件服务(叶结点)构成的高粒度服务(非叶子结点)中,父结点和孩子结点间具有合成的关系。以上是本文进一步讨论的基础。接下来,定义关于 Web 服务的语义模型。

定义 1(服务集群) 一个服务集群定义为 $S_C = \{S_1, S_2, \dots, S_n\}, 1 \leq n \leq M$, 其中 $S_1, \dots, S_M$ 表示在给定域中 M 个有序的服务, $S_n (1 \leq n \leq M)$ 是一个独立的服务,表示在对应的 WSDL 文档中的一个操作。

定义 2(服务引用链) 一个服务引用链 $S_l = (ID, S_1, S_2, \dots, S_n), 1 \leq n \leq M$, 其中 ID 是链 S_l 的唯一标识符, $(S_1, S_2, \dots, S_n)$ 表示一条服务执行过程中的引用路径。

定义 3(服务引用的关系矩阵) 服务引用的关系矩阵

$$M = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

使得

- (1) M 中的每一行对应于 S_C 中的一个 Web 服务 S_i ;
- (2) M 中的每一列对应于一条服务引用链 S_l ;
- (3) 如果服务 S_i 在第 j 条链 S_l 中被引用,则元素 $a_{ij} = 1$;

否则, $a_{ij} = 0$ 。

定义 4(高粒度服务) 一个高粒度的服务 $CS = \{S_1, S_2, \dots, S_n\}, 1 \leq n \leq M$ 。其中 $\{S_1, S_2, \dots, S_n\}$ 是 S_C 中不同的组件服务。

3.2 组件服务间的语义关系

组件服务实现单一的服务功能,简单定义组件服务间的语义关系为相关关系。

定义 5(相关关系) 若组件服务 S_A 以某种方式依赖于组件服务 S_B ,则称 S_A 和 S_B 具有相关关系。

分布式的历史服务引用信息对于发现服务间的内在语义关系是必不可少的。在基本的 Web 服务体系下,通过解析对应的 SOAP 消息能够获取历史的服务引用信息^[10]。在本文以前的相关工作中^[11],使用一个语义消息头结构 SSH 来收集服务间必要的语义交互信息,SSH 是一个封装在服务之间交互的消息流中的管理属性的集合。附带 SSH 的 SOAP 消息如图 2 所示。分析 SOAP 消息,从中可以看到服务之间的引用关系。因而可以获得服务引用链和它的关系矩阵。

基于服务引用关系矩阵,使用服务集合挖掘算法^[11]初始化 RT。初始化的 RT 尽管不能捕获服务间真正的相关度,但仍然能够将功能上相关的服务聚集在一起,以便提供良好的

进一步优化基础。初始化的 RT 如图 3 所示。

```
<SOAP-ENV:Envelope
  xmlns:SOAP-ENV = "http://schemas.xmlsoap.org/soap/envelope/"
  SOAP-ENV:encodingStyle = "http://schemas.xmlsoap.org/soap/encoding/"
  <SOAP-ENV:Header
    <SSH>
      <parent_SSH></parent_SSH>
      <message_id>1122-3551-5721</message_id>
      <message_name>GetZipcode</message_name>
      <source>zipcode.com</source>
      <target>officeplace.com</target>
    </SSH>
  </SOAP-ENV:Header>
  <SOAP-ENV:Body>
    <GetPostofficeByCity>
      <City xsi:type="xsd:string">Beijing</City>
    </GetPostofficeByCity>
  </SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

图 2 附带 SSH 的 SOAP 消息

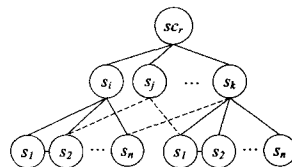


图 3 初始化的 RT

在初始化的 RT 中,每一个非叶结点是包含一组组件服务的高粒度服务。由于在用作初始化的服务集合中,一个服务可能会出现在不同的服务集合中,也就是说,在初始化的 RT 中,一个结点可能会有多个父结点,我们根据相关度的值来调整 RT 的结构,以便能够反映出准确的服务关系。

使用 TF/IDF (Term Frequency/Inverse Document Frequency) 来计算服务间的初始相关度。TF/IDF 是信息检索领域最重要的理论发现之一^[12],TF/IDF 权重通常用于信息检索和文本挖掘中,是用统计测量的权重来评价一个术语在一个集合或语料库文件中的重要程度。在搜索引擎中,给出了用户的查询,利用 TF/IDF 权值的变化来作为文档相关性排序的主要工具。

服务的相关性计算问题,是评价在一个由服务引用链构成的服务集群中任意两个服务间的链接的重要程度,即相关度。可以通过 TF/IDF 问题的方式解决在分级结构中服务的初始相关度的计算问题。映射服务相关问题到 TF/IDF 问题:

- 任意两个服务之间的链 L_i 映射为文档中的一个术语 t_i ;
- 每一个服务引用链 L_i' 映射为一个文档集合;
- 服务引用链的总数 N_i 映射为文档集中文档的总数 N 。

因此,服务的相关度计算如下:

$$w_{i,j} = \frac{f_{i,j} * f_i'}{\sqrt{\sum_{i,j}^M (f_{i,j} * f_i')^2}} \quad (1)$$

式中,

$$f_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (2)$$

是两个服务间的链出现的频率, $n_{i,j}$ 是特定的两个服务之间的链 L_i 在一个服务引用链 L_j' 中出现的次数。

$$f_i' = \log\left(\frac{N_j}{n_i}\right) \quad (3)$$

是可逆服务引用链频率。 n_i' 是包含了链 L_i 的服务引用链 L_i' 的数目。 N_j 是在服务集群中服务引用链 L_i' 的总数, M 是一个 L_i' 中包含的链的总数。

在分级结构中, 初始的服务间的相关度可以由式(1)获得, 每一对兄弟叶结点之间的相关度存储在 RT 中。同时两个非兄弟叶结点之间的相关度由估计它们的祖先结点之间的相关度得到。初始的 RT 不能捕获服务之间真正的相关度, 用一个迭代的方法来改善 RT 结构, 在每一次迭代的过程中将相关的结点移到一起。

自上而下的策略被用来进一步计算服务之间的相关度, 首先计算非叶结点间的相关度, 然后迭代地更新叶结点之间的相关度。最初的非叶结点是由一些组件服务构成的高粒度服务, 初始的两个非叶结点的相关度由它们所包含的孩子结点的交集大小决定:

$$rel(N_i, N_j) = \frac{k}{m+n} \quad (4)$$

式中, $rel(N_i, N_j)$ 表示两个非叶结点 N_i 和 N_j 间的相关度, m 是 N_i 的孩子结点数目, n 是 N_j 的孩子结点数目, k 是 N_i 和 N_j 共同的孩子结点数目。而两个非兄弟叶结点间的相关度通过估计它们的祖先结点的相关度来计算。假设在 RT 中两个非兄弟叶结点 N_1 和 N_k , 通过它们各自的祖先结点 N_i 和 N_j 链接, 其中 N_i 和 N_j 是兄弟结点, 因此, 在 N_1 和 N_k 间存在一条路径 $N_1 \rightarrow \dots \rightarrow N_i \rightarrow N_j \rightarrow \dots \rightarrow N_k$, 则非兄弟叶结点 N_1 和 N_k 之间的相关度为:

$$rel_p(N_1, N_k) = \prod_{j=1}^{k-1} rel(N_j, N_{j+1}) \quad (5)$$

式中, k 是 N_1 和 N_k 之间的路径上结点的数目。

相关度对于决定任意两个给定的组件服务是否在语义功能上相关是至关重要的, 因此, 通过计算组件服务间的相关度, 可以将具有相关关系的组件服务聚集在分级结构中的相应位置, 组件服务间的语义关系可以被逐渐发现。

3.3 高粒度服务间的语义关系

组件服务间的语义关系可以通过计算分级结构中叶子结点间的相关度来发现。高粒度的服务由多个组件服务构成, 在一个具体的领域中, 对高粒度服务的调用是通过调用不同的组件服务实现的。高粒度服务间除了具有相关关系外, 还具有合成关系。

定义 6(合成关系) 若组件服务 S_A 实现高粒度服务 S_B 的部分功能, 则称 S_A 和 S_B 具有合成关系。

可以基于组件服务的语义关系度量方法来度量高粒度服务的语义关系。因此, 要发现高粒度服务间的语义关系, 也可以通过计算高粒度服务间的相关度来实现。如前面所述, 两个非兄弟叶结点间的相关度可以通过估计它们的祖先(非叶结点)的相关度获得。两个非叶结点(高粒度服务)的初始相关度由它们的孩子结点的交集决定, 如式(4)所示。但是, 高粒度服务的初始相关度是不准确的, 需要用迭代的方法反复更新调整。更新高粒度服务间的相关度遵循以下的原则: 在分级结构 RT 中, 两个非叶结点的相关度是和它们分别链接

的孩子结点间的平均相关度, 即对两个非叶结点 N_i 和 N_j 间的相关度进行如下更新:

$$rel'(N_i, N_j) = \frac{1}{K} \sum_{k=1}^K rel_{pk}(N_i, N_k) \quad (6)$$

式中, $K=m \times n$ 是结点 N_i 的孩子和 N_j 的孩子间所存在的路径的数目。

由此, RT 是一个通过一种父子关系所表示的分级结构, 这种父子关系通过相关度的计算在每一次迭代的过程中更新, 使得语义功能上紧密相关的服务被移到一起。因而, 分级的结构 RT 自然地表示了组件服务间的相关关系以及高粒度服务和组件服务之间的组合关系, 在分级结构的基础上, 根据不同粒度服务间的相关度, 服务间的语义关系可以逐步被发现。

4 实验结果

设计基本的实验来验证本文提出的服务语义关系发现方法的有效性和可行性。

实验主要关注的是最终的分级结构 RT 是否能准确地表示 Web 服务之间真正的语义关系。因此, RT 所能表示的服务之间语义关系的准确度被用来评价语义关系发现方法的质量, 真实的服务数据集对于实验具有更现实的评价意义。

使用一个现实的数据集来进行实验, 数据集采用一个公开的可用服务数据集 QWS 2.0 版^[13-15]。QWS 服务数据集为 Web 服务的研究人员提供了一个实验基础, 更新的 QWS 2.0 版包括存在于互联网上的 2507 个一系列现实的 Web 服务, 其中大多数服务来自于互联网上的开发资源, 包括服务统一描述、发现和整合注册库 UDDI、搜索引擎和服务门户。从数据集中根据专家经验手工生成 200 条服务引用链, 覆盖数据集中 2357 个服务, 由于这部分数据由人工生成, 在规模上比较小, 但完全满足实验的目的要求。生成这些数据, 一个服务引用的关系矩阵也随之确定。

在实验中, 使用两种不同的方法来测试准确度评价方法的有效性。

方法一 将生成的分层结构 RT 看作一些服务聚类。在数据集中预定义一些服务类标签, 用杰卡德系数来评价聚类结果。在数据集中可能有多个服务类标签, 而一个服务只能出现在一个聚类中。如果在聚类中的两个服务至少具有一个共同的服务类标签, 就说它们的聚类是正确的。聚类的准确度定义为在同一聚类中正确聚类的服务对所占的比例。

$$acc_1 = \frac{|R_s|}{|P_s|} \quad (7)$$

式中, $|R_s|$ 是在同一个聚类中正确聚类的服务对的个数, $|P_s|$ 是服务聚类中所有的服务对的个数。我们的分析主要在于关注将一个特定服务领域中语义功能上紧密相关的服务聚为一类, 实验结果如图 4 所示。随着服务引用链数量的增加, 测试服务聚类的准确度, 可以看到, 有关服务聚类的准确性对于服务引用链的数量是不敏感的, 结果接近线性。

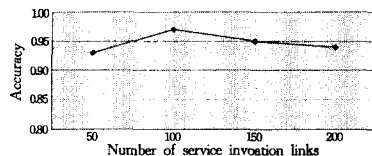


图4 聚类的精度

方法二 根据分级结构 RT, 通过搜索由几个组件服务

构成的合成服务来评价发现的服务关系的准确度。

从数据集中,根据经验人工生成 10 个合成服务作为测试集,每个合成服务包含的组件服务个数最少 5 个,最多 10 个。实验中,在生成的 RT 中根据发现的语义关系,分别搜索由几个组件服务组合而成的合成服务从而评价准确度。准确度被定义如下:

$$acc_2 = \frac{|R_s|}{|Q_s|} \quad (8)$$

式中, $|Q_s|$ 是预定义的合成服务 Q 包含的组件服务的个数, $|R_s|$ 是搜索 Q 时,返回的组件服务和预定义的 Q 中组件服务匹配的个数,实验的结果如图 5 所示。实验结果显示,通过设计的实验方法获取的准确度值较高,并且相对稳定,表明了提出的服务语义关系发现方法是有效的。

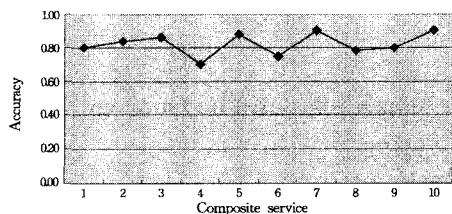


图 5 服务搜索的精度

因此,分级结构是一种能有效地表示不同粒度服务间语义关系的模型。这些实验结果验证了本文提出的发现不同粒度的服务语义关系的方法是有效的、可行的和实用的。

结束语 本文探讨了一个基于分级的结构发现服务语义关系的方法,使得具有相关和合成关系的服务被聚集在相应的位置。主要目的是方便服务发现,增强服务搜索引擎的能力,能自动搜索出语义功能上相关的服务以便执行协作的任务,最大限度地提高服务网络的效用。方法紧紧围绕分级结构所能表示的自然语义,即自然地表示了不同粒度的服务关系。简单介绍了 Web 服务相关的语义模型的简单定义,基于分级结构讨论了度量不同粒度服务间相关度的方法,并进一步根据服务相关度,将语义功能上相关的服务汇集在分级结构的不同层次,进而逐步发现不同粒度服务间的语义关系。

由于 Web 服务之间复杂的相互关系,服务之间的关系会不时地动态变化。如何捕获动态的服务关系是值得探讨的另一个重要问题,也是将来我们需进一步研究的工作。

参考文献

[1] Liu F, et al. Discovery of Web services based on Collaborated Se-

mantic Link Network [C]//Proc. IEEE International Workshop on Semantic Computing and Systems. 2008

[2] Zhuge H, et al. An automatic semantic relationships discovery approach [C]//Proc. WWW 2004. 2004;278-279

[3] Kun Y, et al. Discovering semantic associations among Web services based on the qualitative probabilistic network [J]. Expert Systems with Applications, 2009, 36(5): 9082-9094

[4] Platzer C, et al. Web service clustering using multidimensional angles as proximity measures [J]. ACM Transactions on Internet Technology, 2009, 9(3): 11-37

[5] Paliwal A V, et al. Web Service Discovery via Semantic Association Ranking and Hyperclique Pattern Discovery [C]//Proc. the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. 2006; 649-654

[6] Yin X, Han J, et al. LinkClus: Efficient Clustering via Heterogeneous Semantic Links [C]//Proc. VLDB '06. 2006

[7] Zheng Q. Collaborative Semantic Association Discovery from Linked Data [C]//Proc. IEEE International Conference on Information Reuse & Integration. 2009

[8] Luo X, et al. Automatic discovery of semantic relations Based on Association Rule [J]. Journal of software, 2008, 3(8): 11-18

[9] Zhuge H. Communities and Emerging Semantics in Semantic Link Network: Discovery and Learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(6): 785-799

[10] Rouached M, et al. Web service mining and verification of properties: An approach based on event calculus [C]//Proc. Coop-IS, LNCS. 2006, 4275: 408-425

[11] Zhao A, et al. Semantic message link based service set mining for service composition [C]//Proc. 5th International Conference on Semantics, Knowledge, and Grid. 2009; 338-341

[12] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval [J]. Information processing & Management, 1988, 24(5): 513-523

[13] Al-Masri E, Mahmoud Q H. The qws dataset. Web page[OL]. Available: <http://www.uoguelph.ca/~qmahmoud/qws/index.html>

[14] Al-Masri E, Mahmoud Q H. Qos-based discovery and ranking of web services [C]//Proc. the IEEE International Conference on Computer Communications and Networks. 2007

[15] Al-Masri E, Mahmoud Q H. Investigating web services on the world wide web [C]//Proc. the International World Wide Web Conference. 2008

(上接第 186 页)

[7] Saitta L, Zucker J. Semantic Abstraction for Concept Representation and Learning [C]//Proc. SARA. 1998; 103-120

[8] Saitta L, Zucker J-D. A Model of Abstraction in Visual Perception [J]. Applied Artificial Intelligence, 2001, 15(8): 761-776

[9] 孙善武, 王楠, 欧阳丹彤. 广义 KRA 抽象模型 [J]. 吉林大学学报: 理学版, 2009, 47(3): 537-542

[10] Neches R, Fikes R, Finin T, et al. Enabling technology for knowledge sharing [J]. AI Magazine, 1991, 12: 36-56

[11] Uschold M. Knowledge level modelling: concepts and terminology [J]. The Knowledge Engineering Review, 1998, 13(1): 5-29

[12] Guarino N, Giarretta P. Ontologies and Knowledge Bases: Towards a Terminological Clarification [M]//N. Mars, ed. Towards Very Large Knowledge Bases: Knowledge Building and

Knowledge Sharing. Amsterdam: IOS Press, 1995: 25-32

[13] Gruber T. Towards principles for the design of ontologies used for knowledge sharing [J]. International Journal of Human-Computer Studies, 1995, 43(5/6): 907-928

[14] William S, Austin T. Ontologies [J]. IEEE Intelligent Systems, Jan/Feb 1999; 18-19

[15] Chandrasekaran J, Josephson R, Benjamins V R. What Are Ontologies, and Why Do We Need Them? [J]. IEEE Intelligent Systems, Jan/Feb 1999; 20-25

[16] Wang Nan, Ouyang Dan-tong, Sun Shan-wu, et al. Formalizing the Modeling Process of Physical Systems in MBD [C]//The 2009 International Conference on Web Information Systems and Mining (WISM'09) and the 2009 International Conference on Artificial Intelligence and Computational Intelligence (AICI '09). 2009; 685-695