基于服务访问日志的服务发现研究

文俊浩 许超超 何 盼 冉 唯
(重庆大学软件学院 重庆 400044)

摘 要 如何在大量的服务集合中准确快速地发现服务是服务应用的关键。针对传统的基于关键字的服务发现方式存在召回率和查准率低等缺点,提出了基于服务访问日志的服务发现方法。该方法通过挖掘服务访问日志,对查询词进行扩展,选择最优服务,并查找尽可能相关的服务,以达到优化的目的。实验结果表明,该方法在准确率和召回率上相对于基于关键字的服务发现有明显提高。

关键词 服务访问日志,服务发现,日志挖掘
中图法分类号 TP311.5 **文献标识码** A

Research of Services Discovery Based on Service Access Log

WEN Jun-hao XU Chao-chao HE Pan RAN Wei
(Software College, Chongqing University, Chongqing 400044, China)

Abstract How to find service accurately and quickly is the key of the service application. Due to the low recall and low precision of the traditional service discovery based on keyword, this paper proposed the new service discovery based on service access log. This method optimizes the traditional service discovery by mining the service access log, expanding the query, choosing the best service and finding the possible related services. The experiment shows that this method has higher recall and higher precision compared with the service discovery based on keyword.

Keywords Service access log, Service discovery, Log mining

1 引言

Web服务是一种自包含、自描述、模块化的新的Web应用程序,可以发布、定位以及通过Web调用。Web服务通过一系列的XML的标准和协议,很好地执行异构平台上从简单请求到复杂商务处理的应用。近年来,Web服务标准的持续完善和支撑Web服务的企业级应用平台的不断成熟,使得越来越多的企业和商业组织将其业务功能包装成Web服务发布在网络上,从而实现快速、便捷地寻求合作伙伴,挖掘潜在客户和增值业务[6]。因此,如何在急速增长的Web服务中,快速、准确地找到满足用户需求的Web服务成为面向服务计算的关键问题。

传统的基于关键字的服务发现方式(如UDDI)存在以下两方面的缺点:一是服务注册方面,很难通过WSDL完全展现Web服务的能力,如服务的语义信息等;二是发现机制方面,基于关键字的服务发现方式的召回率和准确率都很低。

目前研究者从两个方面进行研究:一是为Web服务引入语义,借助本体的描述和推演实现Web服务的语义描述、匹配、发现和组合,这是目前主流的研究方法;二是挖掘Web服务的执行信息。文献[1]提出了服务挖掘的概念,认为Web服务挖掘将进一步推动Web服务的成功应用。本文的研究归类于第二种。

本文提出的服务发现方法旨在利用以往的服务访问日志来挖掘查询词与服务的内在联系,缩小查询空间,提高发现效率。通过对日志的分析,针对同一查询词,该服务被访问得越多,说明该服务越满足用户需求。本文首先从服务日志中挖掘出查询词与频繁使用服务的关联关系,以及查询词之间的关联。基于这两种关联,提出Web服务发现的实现途径。

2 基于服务访问日志的关系建模

本文定义的服务访问日志的格式如下。

定义1(服务访问日志) serviceLog={LogID, QueryString, ServiceID}。其中ServiceID标识了用户访问的服务。

设$WS=\{ws_1, ws_2, ws_3, \cdots, ws_n\}$表示被访问的服务集合;$Q=\{q_1, q_2, q_3, \cdots, q_n\}$表示查询词集合,且查询词是原子不可分割的。故查询词与被访问的服务之间的关联关系如图1所示。

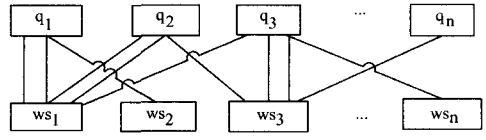

图1 查询词与服务的关系图

到稿日期:2011-10-25 返修日期:2012-02-12 本文受重庆市自然科学基金项目(CSTC,2010BB2244),中央高校基本科研业务费(CDJXS11181161)资助。

文俊浩(1969—),男,博士,教授,主要研究方向为服务计算与面向服务的软件工程,E-mail:jhwen@cqu.edu.cn;许超超(1986—),女,硕士生,主要研究方向为Web服务发现;何 盼(1984—),女,博士生,主要研究方向为可信服务计算;冉 唯(1986—),女,硕士生,主要研究方向为基于GIS的服务组合。

2.1 查询词之间的关系建模

如何判定两个查询词之间存在的联系，主要依赖于哪些服务被访问以及访问的频率。

定义2 基于某个查询词的Web服务被访问的概率定义为：

$$P(ws,q)=\frac{F(ws,q)}{F(q)} \tag{1}$$

式中，ws 表示某一个Web服务，q 表示某一个查询词，$F(ws,q)$表示基于查询词 q 访问的Web服务 ws 的数量，$F(q)$表示基于查询词 q 访问的Web服务总数。

$$\begin{cases} P(ws,q_1)\geqslant\omega \\ P(ws,q_2)\geqslant\omega \end{cases} \tag{2}$$

式中，ω 表示查询词关联阈值，认为查询词 q_1 和 q_2 基于 ws 存在某些联系，该联系不一定是语义关联。

假设 ω 为0.7，当以 q_1 为条件访问 ws_1 的 $P(ws_1,q_1)$是0.7，访问 ws_2 的 $P(ws_2,q_1)$是0.3，以 q_2 为条件访问 ws_1 的 $P(ws_1,q_2)$是0.8，访问 ws_3 的 $P(ws_3,q_2)$是0.2时，则可以相信 q_1 和 q_2 基于 ws_1 存在联系。判定的关键在于如何确定两个查询词基于某个服务的概率值。

2.2 查询词与服务的关系建模

在服务访问日志里面，查询词和服务都存在关联，但这些关联有强有弱，强关系才是我们需要的。本文依据基于该查询词，访问该服务的概率 $P(ws,q)$来选择强关系，即当

$$P(ws,q)\geqslant\sigma \tag{3}$$

式中，σ 表示关联阈值，则认为查询词 q 和Web服务 ws 存在强关联，表示基于查询词 q 的服务访问行为中，Web服务 ws 是主要的访问对象之一。实验证明，该服务是当前最能满足用户需求的服务之一。

当用户输入 q_1 查询时，若服务访问日志中存在该查询词，则在日志中基于该查询词查询并计算出用户访问的服务以及这些服务被访问的概率是多少，从而返回概率最大的服务。另外，基于上节结论，假设 q_1 和 q_2 是具有关联的。当用户输入 q_1 查询时，q_2 将被当成 q_1 的扩展词加入到 q_1 的服务查询中，从而提高 q_1 的召回率和准确率。

3 Web服务发现模型

以往的基于关键字的服务发现改进的是语义，一般考虑将新的用户查询映射到语义库，从语义库中选择与其相近的词，而未将以往的服务访问日志考虑在内。服务访问日志是多个用户多次进行服务查询并寻求最优服务的多次“反馈”结果的集合，对其分析相当于使用大量的服务反馈信息。相对于传统的及时相关回馈，对服务访问日志的分析更具有普遍性，同时并不需要用户的额外操作。

3.1 实现途径

先对服务访问日志进行分析，得到查询词和查询词以及查询词和服务之间的关联集合，即扩展词集以及最优服务匹配集，再将关键词和扩展词集在服务集运用关键词匹配，将得到的匹配集和左右匹配集返回给用户。该过程描述如下：

```
输入：用户请求 q；服务访问日志 ServiceLog；Web 服务集 ServiceResource；
输出：与请求 q 匹配的服务集 ServiceEnd；
beigin
if q⊏Q，then
//判定 q 是否存在于 ServiceLog 里
QR=KeywordExpand(ServiceLog,q)
//获得与关键词关联的扩展词集
BestService=Match(ServiceLog,q)
//获得与 q 匹配的服务集
else then
QR=q；BestService=null；
ServiceMiddle=KeywordMatch(ServiceResource,QR)
//根据查询词在服务集中进行关键字匹配得到匹配集；
ServiceEnd=BestService∪ServiceMiddle //结果集作并集
return(ServiceEnd)
```

3.2 模型描述

基于服务访问日志的服务发现模型如图2所示。

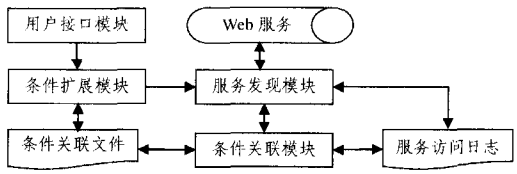

图2 服务发现模型图

如图2所示，基于服务访问日志的服务发现模型包括了用户接口模块、条件扩展模块、服务发现模块和条件关联模块这4部分。

其中，条件扩展模块是对用户输入的查询词进行扩展。当用户输入查询词时，该模块在条件关联文件中查找其是否存在：若存在，则将用户输入的条件作为主条件，并将条件关联文件里的关联词作为辅助条件；若不存在，则不对用户输入的条件做任何操作。然后，该模块将扩展后的条件集合提交给服务发现模块。

服务发现模块负责查找服务，并依据用户选择的服务对服务访问日志进行更新。其过程如下：

1）接收条件扩展模块发送的查询词；

2）依据主条件在服务访问日志中查询服务，并按照用户访问的频率放在前置部分返回给用户；

3）依据主条件在Web服务数据库中查找相关服务，并与2)的结果一起返回给用户，置于第二级；

4）依据辅助条件在Web服务数据库中查找相关服务，并与3)的结果一起返回给用户，置于第三级；

5）当用户选择了其中某一个服务时，更新服务访问日志。

条件关联模块负责对服务访问日志进行分析，是一个实时性模块。该模块通过对服务访问日志的分析，建立各个查询词之间的关联及查询词与服务之间的关联。查询词和服务是多对多的关系。首先分析查询词，每个查询词对应哪些服务，并以{Query，ServiceID，Times，Frequency}的形式存储。再次分析服务，针对每个服务对应哪些查询词设定查询关联因子 ω。当这些查询词针对该服务的访问频率Frequency大于或等于 ω 时，判定这两个查询词关联，关联度即为访问频率。

4 仿真实验分析

为了验证基于服务访问日志的服务发现方法的可行性和有效性，下面给出仿真实验结果。

4.1 实验数据的产生

1）软硬件环境

CPU为Intel Pentium(R) Dual-Core，内存为3.49GB，操作系统为Window XP，编程语言采用C#。

2) Web 服务集的产生

Web 服务集来自于真实的服务网站的 WSDL 文件，如 webxml 网站[8]，是真实的数据集，共采用 200 个服务进行实验分析。

3) 服务访问日志的产生

目前没有相关的 Web 服务访问日志的标准平台和标准测试数据集。本文采用模拟生成日志作为测试日志，即模拟 100 个用户随机访问服务，使其产生 2000 条服务访问日志。此外，将实验中产生的服务访问记录继续添加到实验用日志中。随着访问次数的逐渐增加，服务访问日志里的日志项也随之增加，从而使得依据访问日志所得的分析结果趋向普遍性。

4.2 衡量标准

本文采用准确率、召回率和匹配时间 3 个度量标准来衡量基于服务访问日志的服务发现方法和一般的服务发现方法。准确率和召回率的常规计算公式为：准确率=(检索出的相关正确的服务数/检索出的服务总数) * 100%；召回率=(检索出的相关正确的服务数/正确的服务总数) * 100%；

4.3 实验结果分析

本文通过对基于关键字的服务发现方法和基于服务访问日志的服务发现方法的结果进行综合比较，得到初步对比结果，如图 3－图 5 所示。

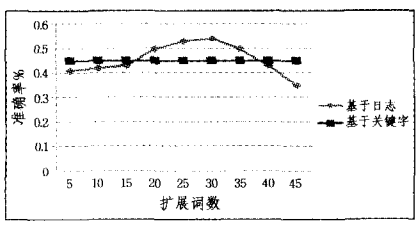

图 3 两种方法的准确率比较

由图 3 可知，基于服务访问日志的服务发现方式的准确率是否高于基于关键字的服务发现方式的关键是扩展词的选择。当扩展词数在 25～35 之间时，本文提出的服务发现方法的准确率明显高于基于关键字的方法。实验表明，在查询词中加入 20～30 个扩展词，查询性能达到更高，超过 30 个后查询性能将下降得很快。

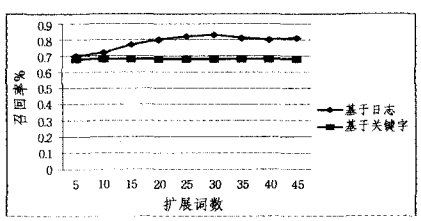

图 4 两种方法的召回率比较

由图 4 可知，基于服务访问日志的服务发现方式在召回率上明显高于基于关键字的服务发现方式。

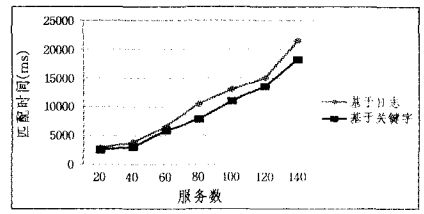

图 5 两种方法的匹配时间比较

由图 5 可知，基于不同的服务数进行服务匹配，本文提出的服务发现方法花费了较多时间。本文所提出的方法是建立在对大量用户长期查询访问服务行为的基础上的，它在选择扩展词汇时更具有针对性，避免扩展用词扩散，不影响后续查询。

结束语 高效的服务发现方法是服务应用的关键。以往的研究主要集中在语义描述上，本文通过挖掘服务访问日志，建立查询词之间以及查询词和服务之间的联系，从而在保证服务发现质量的前提下，提高服务发现的准确率和召回率。本文提出的方法中，如何更好地确定查询词之间的关系，取消扩展词的歧义以及如何提高查询效率，将是下一步的研究重点。

参 考 文 献

[1] Liang A, Miller S, Chung J Y. Service mining for Web service composition[C]// Proc of IEEE International Conference on Information Reuse and Integration. Las Vegas: s. nl, 2005: 470-475

[2] 朱红康，余雪丽. 基于用户和服务协同聚类的 Web 服务发现研究[J]. 计算机应用研究，2010，27(3)：986-988

[3] Wen J R, Nie J Y, Zhang H J. Clustering user queries of a search engine[C]// Proceedings of the 10th International World Wide Web Conference (WWW10). New York: ACM Press, 2001: 162-168

[4] Paolucci M, Kawamura T, Payne T, et al. Semantic matching of Web services capabilities[C]// Proc of the 1st International Semantic Web Conference. London, UK: Springer-Verlag, 2002: 333-347

[5] Syeda-Mahmood T, Shah G, Akkiraju R, et al. Searching service repositories by combining semantic and ontological matching[C]// Proc of IEEE International Conference on Web Services. Washington DC: IEEE Computer Society, 2005: 13-20

[6] 崔航，文继荣，李敏强. 基于用户日志的查询扩展统计模型[J]. 软件学报，2003，14(9)

[7] Wu Jian, Wu Zhao-hui. Similarity-based Web service matching [C]// Proceedings of the 2005 IEEE International Conference on Service Computing (SCC'05). IEEE Computer Society press, Orlando, Florida, USA, 2005: 287- 294

[8] Web Service(Web 服务)[BP/OL]. http://www.webxml.com

(上接第 132 页)

[9] 丁维龙，熊范纶. 植物模拟组件的设计与实现[J]. 中国科学技术大学学报，2003，33(6)：742-748

[10] 姜海燕，朱艳，曹卫星，等. 作物模型资源构造平台(CMRCP)的构建研究[J]. 农业工程学报，2008，24(2)：170-175

[11] 王忠芝，胡逊之. 基于 Xfrog 的树木建模及生长模拟[J]. 北京林业大学学报，2009，31(增刊 2)：64-68

[12] 赵海燕，张伟，麻志毅. 面向复用的需求建模[M]. 北京：清华大学出版社，2008

[13] 黄罡，张路，周明伟，等. 构件化软件设计与实现[M]. 北京：清华大学出版社，2008

[14] 谢冰，王亚沙，李戈，等. 面向复用的软件资产与过程管理[M]. 北京：清华大学出版社，2008

[15] 计算机软件构件复用属性规范[EB/OL]. http://www.sawin.cn/doc/Document/DocCase/blueski1126.htm，2010-12-23

[16] 王晓光，冯耀东，梅宏. ABC/ADL：一种基于 XML 的软件体系结构描述语言[J]. 计算机研究与发展，2004，41(9)：1521-1531