第 36 卷 第 9 期 计 算 机 工 程 2010 年 5 月
Vol.36 No.9 Computer Engineering May 2010
·博士论文· 文章编号：1000—3428(2010)09—0023—03 文献标识码：A 中图分类号：TP393

基于 P2P 的分布式 Web 服务挖掘技术

朱红康[1,2]，余雪丽[1]
(1. 太原理工大学计算机与软件学院，太原 030024；2. 山西师范大学数学与计算机科学学院，临汾 041000)

摘 要：为对多个服务注册中心提供支持，方便服务访问日志的记录与挖掘，提出一种基于 P2P 的分布式服务执行挖掘框架。针对跨组织业务关联的需求，利用该框架构建服务注册联盟机制，设计基于日志库的 Web 服务关联规则挖掘算法进行组合服务频繁序列挖掘。仿真结果表明，该算法能有效挖掘日志库中的执行与交互信息，提高服务选择与组合效率。

关键词：组合服务；关联规则；序列模式；数据挖掘

Distributed Web Service Mining Technology Based on P2P

ZHU Hong-kang[1,2], YU Xue-li[1]
(1. School of Computer & Software, Taiyuan University of Technology, Taiyuan 030024;
2. School of Mathematics & Computer Science, Shanxi Normal University, Linfen 041000)

【Abstract】 In order to provide support for multiple service registry center, and convenient for recording and mining of service access log, this paper proposes a distributed service executive mining framework based on P2P. Aiming at the requirement of cross-organizational associated business, it constructs service registration union mechanism using this framework. In registration union, it designs Web service associated rule mining algorithm based on log base to progress frequency sequence mining of composite service. Simulation results show that this algorithm can mine executive and interactive information effectively, enhance efficiency of service selection and service composition.

【Key words】 composite service; association rule; sequence pattern; data mining

1 概述

近年来，Web 服务(Web service)技术作为实现面向服务体系结构的主要方式得到较大发展。企业使用 Web service 封装业务过程，通过注册被业务伙伴经由 Internet 自由访问，同时可以动态绑定单个服务来提供增值组合服务。这种新的软件开发模式避免了服务软件的重复开发，体现了 Web 服务支持快速应用集成的优势。然而，由于 Web 服务本质的自治与异构特性，因此实施组合服务面临的一个基本问题是如何确保其正确执行。

随着组合服务的大量运行，可以方便记录它们的执行轨迹，这些轨迹客观地反映了组合服务的运行情况[1]。通过挖掘这些有用信息，可以更好地分析、选择、监控、优化、改进组合模型。

针对组合服务选取问题，本文从服务执行信息入手，提出一种基于 P2P 的分布式服务使用模式挖掘框架，进一步提出基于注册联盟的服务关联规则挖掘算法(Service Associate Rule Mining based on Registry Federation, SARM-RF)。

2 基于P2P的分布式服务执行挖掘框架

组合服务选取是服务计算领域的一个研究难点。以往的选取方法大多基于难于准确获取的服务 QoS 信息，本文提出一种基于 P2P 的分布式服务执行挖掘构架(Distributed Service Execute Mining Framework, DSEMF)。

DSEMF 主要由 3 类 Peer 构成:RM-Peer(Registries Management-Peer), SL-Peer(Service Log-Peer) 和 C-Peer(Client-Peer)。其框架如图 1 所示。

RM-Peer 负责整个框架的管理，如为加入的服务注册中心生成 SL-Peer，构建与维护服务注册联盟等。SL-Peer 为每一个服务注册中心生成一个记录与挖掘执行信息，该类 Peer 功能如图 1 右半部分所示。

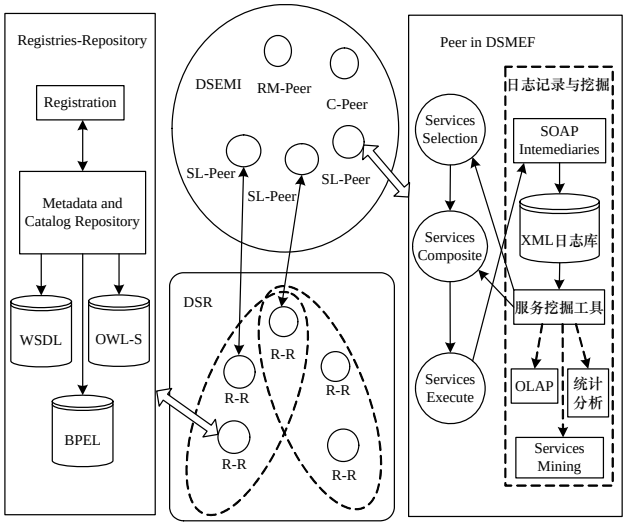

图 1 基于 P2P 的分布式服务执行挖掘框架

SOAP Intermediaries 负责服务执行信息的收集，经数据

基金项目：国家自然科学基金资助项目(60472093)；国家教育部博士点基金资助项目(200801120007)
作者简介：朱红康(1975－)，男，讲师、博士研究生，主研方向：面向服务计算；余雪丽，教授、博士生导师
收稿日期：2009-11-06 **E-mail：**zhuhkyx@126.com

选择与过滤后以 XML 的形式存入日志库，服务挖掘工具使用日志库对服务选择与组合提供支持。C-Peer 负责收集客户的查询信息，返回 SL-Peer 的查询结果。

DSR(Distributed Service Registries)是分布于互联网的服务注册中心，每个 R-R 库(Registries-Repository)的组成如图 1 左半部分所示。每个 R-R 库除了包含发布服务注册信息的注册中心外，还提供服务实体信息库，如 WSDL 文档库、OWS-S 文档库、BPEL 文档库等。

3 服务注册联盟的构建

在跨组织、跨领域的业务集成中，分布的服务注册中心中的服务通常有密切关联，从逻辑上将这些注册中心划分为不同的注册联盟(Registry Federation, RF)，一方面有助于跨域服务契约的实现，另一方面，通过在不同联盟内进行服务执行信息的挖掘有利于缩小挖掘空间，提高挖掘精度。参照文献[2]定义如图 2 所示的注册本体(Registry Ontology, RO)。RO 由 RM-Peer 创建与管理，当有新的注册中心加入 DSEMF 时，RM-Peer 为该注册中心生成 SL-Peer，更新 RO，并把更新后的 RO 发布给联盟的其他 SL-Peer。

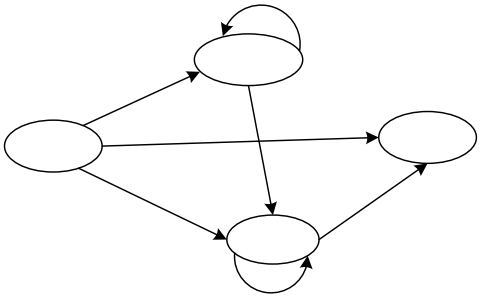

图 2 注册本体

4 基于注册联盟的Web服务关联规则挖掘算法

在注册联盟中，相应的 SL-Peer 日志库记录了各自的 R-R 库的执行信息，也包含不同注册中心的交互信息，有效地挖掘这些信息能提高服务选择及组合效率。本节研究目标是给出注册联盟中相关分布式日志库关联规则的挖掘方法。

4.1 日志库记录项的定义

为 SOAP Intermediaries 交互的每条信息记录定义给出的各项信息，要获得挖掘使用的日志数据库。为简化关联挖掘算法，假设各 SL-Peer 中日志库的记录格式相同。

定义 XML 日志项是日志库中为每个服务执行信息记录的一条数据，可用六元组(CompositeServicesID, InstanceID, ServicesID, Type, TimeStamp, Status)表示，其中，Composite ServicesID 为组合服务 ID，可以是基于 BPEL(Business Process Execution Language)文件等方式，表示一个抽象的组合服务过程；InstanceID 为组合服务执行实例 ID，唯一地标志组合服务的一个执行实例；ServicesID 为组合服务中所含抽象服务对应的服务实例，可由实例服务的 URI 唯一标识；Type 为 SOAP 信息的类型，表示请求或响应方式；TimeStamp 为当前 ServicesID 服务执行的时刻；Status 为服务请求/响应是否成功的状态。

4.2 算法构建

设注册联盟中共有 n 个 SL-Peer，相应的日志库分别为 $\{DB_1,DB_2,\cdots,DB_n\}$。对每个 DB_i 而言，其中的组合服务执行实例集(每个不同的执行实例由 InstanceID 唯一标识)可以看作是关联挖掘中的事务集。数据项集 $I_i=\{WS_{i(j)}\in DB_i,\ j=1,2,\cdots,m\}$，$WS_{i(j)}$为 DB_i 事务集中对应的具体服务。设 $DB=\bigcup_{i=1}^{n}DB_i$，则 SARM-RF 是在 P2P 的 DB_i 日志库中发现全局 DB 的频繁项集。

关联规则挖掘的经典算法是 Apriori[3-4]，但它仅适合集中式数据集。对于分布式数据集的关联规则挖掘，其关键问题是如何减少不同节点间的信息传输量，典型算法有 CD[5], FDM[6], D-Sampling[7]等。本文采用 FDM 为原型，构造 P2P 环境中的 SARM-RF 算法。SARM-RF 算法的标记及说明见表 1。

表 1 SARM-RF 算法的标记及说明

记号	意义说明
D_i	日志库 DB_i 中的事务数
D	$\sum_{i=1}^{n}D_i$
S	支持度最小阈值
$L(k)$	全局频繁 k-项集
$X.Sup$	项集 X 的全局支持度计数
$CA(k)$	由 $L_{(k-1)}$产生的候选集，$CA(k)=Apriori\_gen(L(k\text{-}1))$
$GL_i(k)$	DB_i 的全局局部频繁 gl-k-项集
$CG_i(k)$	由 $GL_i(k\text{-}1)$产生的候选集，$CG_i(k)=Apriori\_gen(GL_i(k\text{-}1))$
$LL_i(k)$	$CG_i(k)$的局部频繁 k-项集
$X.Sup_i$	项集 X 在 DB_i 中的局部支持度计数

SARM-RF 算法的核心思想是：每个节点按照 Apriori 产生局部候选集，由局部候选集根据文献[6]定理产生全局候选集。为了减少节点间的信息传输，对候选集使用 2 类剪枝技术，即局部候选集剪枝与全局候选集剪枝。

定理 对任一 k>1，有如下公式成立：

$$L(k)\subseteq CG(k)=\bigcup_{i=1}^{n}CG_i(k)=\bigcup_{i=1}^{n}Apriori\_gen(GL_{i(i-1)})$$

证明略。

定理说明：全局候选集可由各节点局部候选集的并产生，且它远远小于 $CA(k)$的个数。若一个 k-项集为 DB 的全局频繁项集，则它一定是 DB_i 的局部频繁 k-项集，记为频繁 gl(globally and locally)-k-项集。

SARM-RF 算法可描述如下：

输入 事务数据库 DB_i，最小支持度计数 Sup_i

输出 所有的频繁项目集 L

在每个 Peer 迭代执行下列程序段，算法终止条件为 $L(k)=\varnothing$，或候选集 $CG(k)=\varnothing$。

```
Begin
 /*产生候选集*/
  If k=1 then
     T_i(1)=get_local_count(DB_i, ∅ ,1)
  Else {
     CG(k)= ⋃_{i=1}^{n} CG_i(k) = ⋃_{i=1}^{n} Apriori_gen(GL_{i(i-1)}) ;
     T_i(k)=get_local_count(DB_i,CG(k),i) }
/*局部及全局剪枝*/
 for all X ∈ T_i(k) do
   if X.Sup_i ≥ S×D_i then //局部剪枝
      if  X.Sup_i + Σ_{j=1,j≠i}^{n} MaxSup_j(X) ≥ S×D  then
//全局剪枝，其中 MaxSupi=min{Y.Sup_i | Y<X, and |y|=k-1}
          insert <X,X.Sup_i> into LL_i(k)
/*广播 LL_i(k)，计算 gl-k-项集*/
 for j=1 to n do send LL_i(k) to Peer S_j;
 Receive LL_j(k) from other Peer;
 for all X ∈ LL_i(k) do{
   X.Sup = Σ_{i=1}^{n} X.Sup_i ;
   if X.Sup>S×D then
```

insert X into $G_i(k)$; }

/*广播计算结果*/

broadcast $G_i(k)$

receive $G_i(k)$ from all other Peer $S_j(j \neq i)$;

$L_{(k)} = \cup_{i=1}^{n} G_{i(k)}$;

Divide L(k) into $GL_i(k)$;

Return L(k)

End

5 原型实现及算法分析

5.1 原型实现说明

DSEMF 使用 JXTA(Juxtapose)实现，任何一个 Peer 作为一个 JXTA Peer 实现一个或多个 JXTA 协议。JXTA 由一系列协议构成，本文主要使用 2 类协议：PDP(Peer Discovery Protocol)和 PBP(Pipe Binding Protocol)。PDP 使一个 Peer 可以发现其他 Peer，Pipe 是 JXTA 网络数据传输的主要方式，PBP 规范了对等 Pipe 的绑定、解析、响应等。

为了实现对注册联盟的创建与管理，本文定义一个控制交互协议(Control Initiation Protocol, CIP)。CIP 部署在 RM-Peer 上，用于创建新的 SL-Peer，使用新 SL-Peer 注册本体更新原注册联盟本体，并将更新后的联盟本体发送给相应 SL-Peer。DSEMF 实现见图 3。

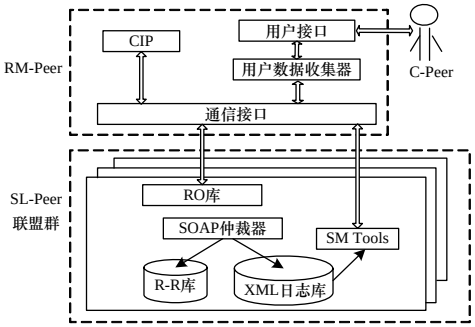

图 3 DSEMF 实现

5.2 仿真实验分析

本节通过仿真实验评估 SARM-RF 算法的有效性。实验基于DSEMF,每个Peer部署在一台Pentium(R) 4.3 GHz, 1 GB 内存的计算机上，它们通过带宽为 100 Mb/s 局域网互联。

第 1 组仿真实验的目的是为了验证服务联盟的节点数对算法性能的影响。实验仿真设置 6 个不同节点(3 个~8 个)，每个节点随机生成 40 个组合服务的 10 000 条实例数据，节点间的服务交叉访问率为 10%(服务交叉访问率=节点交叉访问的服务个数/节点的访问服务总数)，在最长序列长度为 7，不同最小支持度情况下，节点数对算法 SARM-RF 效率的影响如图 4 所示。

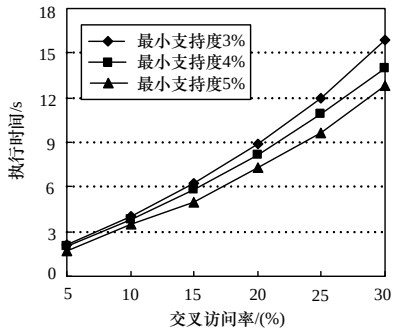

图 4 节点数对算法 SARM-RF 效率的影响

第 2 组仿真实验的目的是为了验证服务交互访问率对算法性能的影响。实验仿真节点数为 4 固定不变，节点间的服务交互访问率分别为 5%~30%(按 5%递增),每个节点生成 40 个组合服务的、满足测试的 10 000 条实例数据，在最长序列长度为 7，不同最小支持率下，交叉访问率对算法 SARM-RF 效率的影响如图 5 所示。

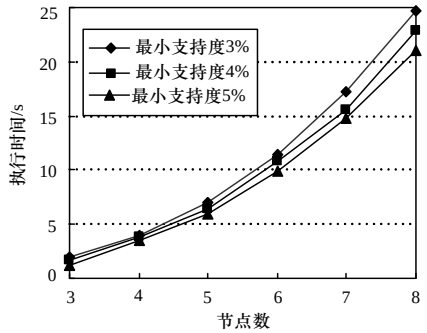

图 5 交叉访问率对算法 SARM-RF 效率的影响

实验结果表明，随着服务联盟节点数及交叉访问率的增加，SARM-RF 算法运行效率在线性可接受范围内。由于目前没有标准的测试平台和测试用例，笔者也未查到类似的分布式服务挖掘算法，因此无法进一步比较。实验结果只能初步说明本文 DSEMF 模型及其上 SARM-RF 算法的合理性。

6 结束语

本文 SARM-RF 算法与其他分布式序列挖掘算法相比，特别是应用于大规模服务信息挖掘时，其有效性还需进一步验证。下一步研究的重点为：如何将本文算法与传统组合服务过程建模方法相结合提高选取组合服务的质量。

参考文献

[1] 张明卫, 魏伟杰, 张 斌, 等. 基于组合服务执行信息的服务选取方法研究[J]. 计算机学报, 2008, 31(8): 1398-1411.

[2] Verma K, Sivashanmugam K, Sheth A, et al. METEOR-S WSDI: A Scalable P2P Infrastructure of Registries for Semantic Publication and Discovery of Web Services[J]. Information Technology and Management, 2005, 6(1): 17-39.

[3] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules[C]//Proc. of the 20th International Conference on Very Large Data Bases. Santiago, Chile: [s. n.], 1994.

[4] 钱光超, 贾瑞玉, 张 然, 等. Apriori 算法的一种优化方法[J]. 计算机工程, 2008, 34(23): 196-198.

[5] Agrawal R, Shafer J. Parallel Mining of Association Rules[J]. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 962-969.

[6] Cheung D, Han Jiawei, Vincent T N, et al. A Fast Distributed Algorithm for Mining Association Rules[C]//Proc. of the 4th International Conference on Parallel and Distributed Information Systems. Miami Beach, Florida, USA: [s. n.], 1996.

[7] Schuster A, Wolff R, Trock D. A High-performance Distributed Algorithm for Mining Association Rules[J]. Knowledge and Information Systems, 2005, 7(4): 458-475.

编辑 陆燕菲