



一种基于数据服务超链进行情景数据集成的方法^{*}

王桂玲¹, 张 峰^{1,2}, 韩燕波¹

(1.北方工业大学云计算研究中心 北京 100144; 2.山东科技大学信息科学与工程学院 青岛 266590)

摘 要:借鉴 Web 网页超链的思想,提出利用“数据服务超链”的抽象辅助用户显式地描述数据源之间隐含的关联关系,在此基础上提出基于超链的数据服务即时构造方法,使用户能在进行数据服务导航过程的同时对数据进行即时集成。为了辅助用户在数据服务即时构造过程中进行临机决策,以应对变化的需求,还讨论了一种数据服务超链推荐算法以帮助用户选择合适的数据服务超链。结合案例进行了分析评价及实验,结果表明,本文提出的方法能够有效地支持情景数据集成。

关键词: 数据服务; 服务超链; 服务推荐; 情景数据集成

doi: 10.3969/j.issn.1000-0801.2014.02.006

An Approach to Situational Data Integration Based on Data Service Hyperlink

Wang Guiling¹, Zhang Feng^{1,2}, Han Yanbo¹

(1.Research Center for Cloud Computing, North China University of Technology, Beijing 100144, China;

2.College of Information Science and Engineering, Shandong University of Technology, Qingdao 266590, China)

Abstract: Enlightened by web page hyperlink, a novel abstraction “data service hyperlink” for linking and integrating data was proposed. With the help of “data service hyperlink”, users can explicitly describe the relationship between data sources. Based on data service hyperlink, the approach to just-in-time data service construction was proposed to enable users integrate data while navigating data services. A data service hyperlink recommendation algorithm was also discussed to help users select proper data service hyperlink to deal with the temporary requirements. A case-study and some experiments how to build a situational data integration application based on data service hyperlink efficiently were presented.

Key words: data service, service hyperlink, service recommendation, situational data integration

1 引言

随着 Internet 的快速发展和普及以及云计算的兴起,数据的采集、存储和传播的数量与日俱增,但这些数据往往存储在许多不同管理域的不同数据源之中。为更有效地利用这些信息,需要从多个分布、异构和自治的数据源中

集成数据。由于互联网环境的开放性及动态变化性^[1,2],数据集成面临的环境也越来越复杂。本文所探讨的数据集成与传统数据集成不同,其集成需求无法完全预先定义,而是即时出现,称为“情景数据集成”,根据即时出现的需求对不同来源、不同格式、不同语义、不同特点的数据源临时进行逻辑或物理上的有机集成^[3]。

^{*} 国家自然科学基金资助项目(No.61003294),北京市属高等学校创新团队建设与教师职业发展计划基金资助项目(No.IDHT20130502),北京市自然科学基金重点资助项目(No.4131001),北京市教委科技计划基金重点资助项目(No.KZ201310009009)



以公安系统面临的嫌疑逃犯抓捕场景为例,各级公安干警在案件侦破过程中,要分析来自不同部门的大量数据,试图寻找有用信息。传统的做法是通过查阅档案或者直接通过常年积累的直觉经验,判断众多事件和信息之间的关联,从中发现有用的信息。随着现代社会各部门信息系统的逐渐完善,公安干警逐渐可以从各类信息查询系统、视频监控系统中找寻案件线索。公安干警在定位嫌疑人的过程中,受害人信息往往易于明确,但进一步提取信息的需求会随着情景的变化而动态演化。例如,当受害人和受害地点确定,提取受害人的身份证信息后,可进一步调取此受害人的社会关系信息(如移动电话通话信息、亲属信息、QQ、微博联系人信息、银行卡购买记录等通信、网络多维度的相关信息)以及最近联系人的车辆信息,通过对比事发现场的车牌识别历史记录,发现与此联系人车牌一致。本案告破,此人正是凶手。

这个过程映射到软件系统中,是一个对多个异构数据源进行抽取、转换、查询、连接、合并、比对等的数据集成问题。与传统的数据集成不同,在此类场景中,公安干警无法事先确定提取、集成和分析哪些数据源,而只能根据当前数据集成的结果逐步明确下一步进行数据集成的需求。在这个过程中,数据的集成需求无法完全预先定义,是随集成结果的逐步显现而逐步明确的。用户可以根据动态变化的需求临时调整数据集成的过程,这是一种典型的“情景数据集成”案例。

在这个场景中,数据来源于多个不同管理域的部门。各种部门信息系统的数据访问方式、格式、语义异构,且各部门出于数据隐私保护、系统性能等考虑,往往不希望暴露数据源的数据模式及全部的查询访问信息,只允许有限制的查询访问。人们通常采用面向服务的数据集成技术^[4],首先使用服务的方式封装对各部门数据进行查询、访问的功能,一方面可满足部门对自身数据的保护等要求,另一方面,也有可能对异构的数据源提供一致的访问方式。

之前的研究提出并实现了一种用户主导的数据服务编程方法,面向无法预先定义、即时出现的数据集成需求,提供数据服务的基本抽象,提出基于嵌套表格和 Spreadsheet 结合的数据服务编程模型及环境,使不具备专业编程知识的普通用户可以即时进行服务聚合^[5]。以前数据服务的调用、结果的转换以及处理操作完全依靠用户手工进行,数据服务之间的关联关系隐含在应用的构造逻辑中,不利于其他用户重用。为了方便用户建立并重用数据

服务之间的关联关系,需要提供更加有效的手段辅助用户显式地描述和表达各种数据服务之间隐含的关联关系,并可向用户动态地推荐,辅助用户从各种看似无关的数据中建立关系、发现线索,进行临机决策和动态集成。

为了显式描述并使用服务之间的关系,可以借鉴 Web 页面中超级链接(简称“超链”)的思想。“超链”的思想源于网页超级链接,超链是构成万维网的关键要素,作为节点元素的一部分,却与节点内容松散耦合,具有结构简单、单向性、局部性^[7]、松耦合、可扩展能力和容错能力等很多优良的性质以及丰富的表达能力^[8]。在 Web 网页浏览中,用户通过超链的导航,能够自行逐步发现新的可用资源。而超链的作用并非仅限于网页浏览,其既是面向人的,又是机器可理解的,例如在 REST 风格的服务中,超链可用于描述应用的状态改变和转移^[9],从而有可能可靠地协调不同系统间的交互,实现完整的业务流程。受此启发,本文将“超链”的思想引入服务的设计中,希望能够通过超链的支持,显式地表达并存储数据服务之间的关联关系,使得用户在各种服务之间简便、快捷地导航,并同时完成复杂的数据集成操作,从而更好地支持人的临机决策以加速解题的过程。

在设计一个网页时,提供到其他网页的链接是事实上的标准做法。但在服务设计中,往往并不包含到其他服务或资源的链接,客户端必须根据服务调用的结果,决定其下一步的行为。本文将超链引入数据服务模型的设计中,提出了一种具有超链能力的数据服务模型,用数据服务超链描述服务之间的前驱后继关系,并作为建立数据服务之间有机关联关系的核心元素。在此基础上,探讨了一种基于超链进行数据服务即时组合的方法,并讨论了超链智能推荐关键问题,以辅助用户利用数据服务之间关联关系的历史,从各种看似无关的数据中建立关系,发现线索,进行临机决策。

最后,通过案例和实验对所提出的组合方法及其实现机制给予了分析和评价,本文提出的基于数据服务超链的情景数据集成方法继承了“超链”的优良性质,支持用户在数据服务导航的过程中动态完成数据服务的组合,同时灵活支持人机决策,可以有效支持情景数据集成。

2 基于数据服务超链的数据服务即时组合模型

2.1 数据服务及数据服务超链

本文提出的数据服务是以提供数据资源的访问为目

的软件服务,提供了对数据资源的抽象。数据服务之间是互相关联的,其关联关系通过服务超链显式表征。数据服务之间的关联关系可以有多种,本文以数据服务的输入/输出关联关系为例。假设一个数据服务 a 输出的一部分可作为另一个数据服务 b 的输入,则客户端可通过服务超链从 a 导航到 b ,即从应用的一个状态转移到其他状态。这里,应用的状态是指在某个特定的时刻系统中所有变量的值。例如,“受害人信息”数据服务的输出是该受害人的唯一身份标识 ID,而“亲属关系信息”数据服务需要的输入参数是某人的身份标识 ID,进而得到该人员的亲属关系信息。两个数据服务之间通过受害人的唯一身份标识 ID 建立了关联,客户端可以在调用“受害人信息”数据服务得到输出结果之后,通过显式建立的服务超链导航到“亲属关系信息”数据服务,从而得到受害人的亲属关系信息。下面给出几个关键的定义。

定义 1 (数据服务)数据服务是信息系统对外发布数据资源的软件服务,通过数据资源封装或组合得到,实现了数据资源在给定数据模型下的统一表示。它接收一组对该数据资源进行 HTTP 标准请求的参数,并返回相应的结果数据集或响应。一个数据服务用 URI 唯一标识,通过 HTTP 标准的方法获取或操作数据,对外暴露其输入参数以及输出的资源表述。其中,资源表述中包含了与其有单关联关系的数据服务超链及输出的数据模式。

数据服务可表示为一个七元组(为了简便起见,省略了对数据质量、服务质量以及其他非功能属性的描述): $ds = \langle name, id, desc, URI, method, params, representation \rangle$, 其中, $name$ 是服务的名字; id 为服务的全局唯一标识; $desc$ 是服务的描述; URI 可以包含变量、正则表达式或 XPath 表达式,如 $http://some.com/{victimID}/VictimInfo$ 、 $http://some.com/[a-zA-Z0-9\-_]/relative$ 等,这样可以使得数据服务支持到数据源查询操作的动态绑定,即客户端在选择超链时根据当前数据服务的表述动态决定链接的具体 URI; $method$ 为标准的 HTTP 方法 (get、put、delete 或 post); $params$ 是数据服务的输入参数,每个输入参数为一个不嵌套其他元素的原子元素, $params = \langle name, value \rangle$; $representation$ 是数据服务所对应的数据源查询操作输出的表述, $representation = \langle id, type, links, schema, output \rangle$ 。 $type$ 是该操作结果表述的格式,有纯文本格式、XML 格式、JSON 格式或嵌套表格式等; $schema$ 是数据服务输出结果的数据模式; $representation$ 中还包含指向其他数据服务的超级链接

$links$,由多个服务超链(见定义 2)构成; $output$ 是数据服务输出结果表述的内容字符串。

定义 2 (服务超链)一个数据服务超链可表示为如下六元组: $link = \langle id, desc, target, selector, generate-uri, protocol, constraint \rangle$, 其中, id 是该服务超链的全局唯一标识; $desc$ 是其具体描述; $target$ 是链接指向的下一个数据服务的 ID,并不是下一个数据服务的 URI,以避免链接和数据服务 URI 的紧耦合; 服务超链是通过 $selector$ 从数据服务输出的资源表述 $representation$ 中获取的, $selector$ 用于从当前数据服务输出的资源表述中选取相应的部分,以作为下一个数据服务的输入,其值与资源表述的格式有关,若资源表述格式为 XML,则 $selector$ 是一个 XPath 表达式; $generate-uri$ 用于根据 $selector$ 的值 (即前一个数据服务的输出)生成下一个数据服务的超链,是服务超链对数据服务的输入/输出关系进行显式描述的关键; $protocol$ 是服务超链的访问协议,描述了其标准的 HTTP 方法 (get、put、delete 或 post) 及返回数据的格式; $constraint$ 是链接的约束,用于表达链接起作用的条件。

定义 3 (原子数据服务)原子数据服务是一类基本的数据服务,一个原子数据服务即直接包装数据资源得到的服务,原子数据服务封装的技术和过程不在本文的探讨范围之内,具体请参见参考文献[6]。

定义 4 (复合数据服务)复合数据服务由多个原子数据服务遵照第 3 节的方法即时组合而成,一个复合数据服务由多个通过数据服务超链关联起来的原子数据服务构成,对应多个数据源查询操作。

定义 5 (数据服务空间)数据服务空间是数据服务的集合,数据服务之间可能通过服务超链互相关联。在一个系统中,新加入的原子数据服务、新生成的复合数据服务均需添加到该集合中,用户在构造应用时所使用的服务全部来自于该集合。

数据服务的描述模型如图 1 所示。实线框表示元素,虚线框表示属性。表 1 列出了本文场景中所涉及的数据服务。

上述数据服务模型的设计思想如下。

融合了 REST 服务的设计思想,显式地使用标准 HTTP 方法,通过数据服务超链完成数据服务之间的访问,将应用从一个状态改变为另一个状态。在本文提出的数据服务模型中,超链为什么不单独表示,而是表示到数据服务中?这是借鉴了 REST 服务“超媒体作为应用状态引擎(hypermedia as the engine of application state,HATEOAS)”

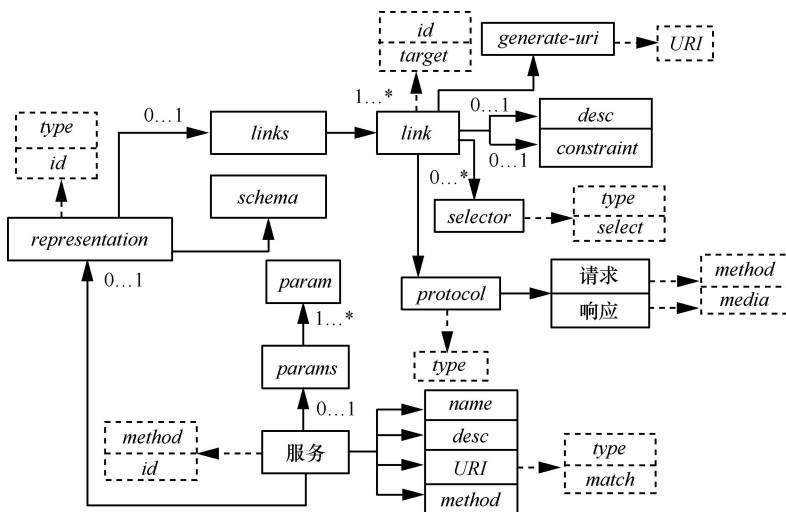


图1 数据服务描述模型

表1 “嫌疑犯确定”场景中涉及的数据服务

数据服务	方法及 URI	输入参数	描述	服务超链
VictimInfo Service(受害人信息)	get/victim	victimID	获取受害人详细身份信息	get/relatives、get/contacts
Relatives Service(亲属信息)	get/relatives	ID	获取某人的亲属信息	无
Contacts Service(最近联系人信息)	get/contacts	ID	获取某人的联系人列表	post/union
Vehicle Service(车辆信息)	get/vehicle	IDList	获取某组人的车辆注册信息	无
MonitorVehicle Service(现场监测车信息)	get/monitorVehicle	crossRoadID, timeInterval	获取某时间段某路口的车牌号	post/intersect
Union Service(合并信息)	post/union	left, right	对输入进行合并运算	get/vehicle
Intersect Service(交叉信息)	post/intersect	left, right	对输入进行交叉运算	无

的原则^[10]。当前客户端对数据服务调用后,服务器以对资源的表述作为响应,其中包含一个或多个超级链接,用来描述下一个被允许的与服务器交互的集合。客户端通过对这些超级链接的选择,将应用从一个状态改变为另一个状态。服务器端不保存应用状态,提供一系列的链接,状态的转移完全由客户端通过链接的选择驱动。借鉴这种设计思想,本文在数据服务中包含相应的服务超链,通过超链的支持,客户端能通过超链的选择驱动数据的操作与流转。

本文提出的数据服务将提供相应的描述文件,支持以声明性的方式描述数据源的基本信息以及数据服务之间的导航机制。以嵌套关系作为数据服务的中立数据结构,为了方便数据可视化操作,将数据资源可视化嵌套表格的形式,具体分析见参考文献[5]。

描述模型中的数据服务超链可以由用户手工创建,也可以由机器自动创建(如基于历史记录的方法以及信息抽取和自动匹配的方法等),只需要将相应的数据服务超链信息增加到数据服务的描述文件中即可。数据服务超链具体的各种自动建立方法不在本文的探讨范围之内。

数据服务空间提供了领域中的所有数据服务,这些数据服务之间通过数据服务超链互相连接,形成一个庞大的服务网络。当数据服务空间中的数据服务达到一定规模时,将显现一定的性质,如同通过网页超链连接起来形成的 Web 一样。分析数据服务空间的性质是一个有意义的工作,但不在本文的探讨范围之内。

2.2 基于数据服务超链的数据服务即时组合过程

本文设计的情景数据集过程是一个复合数据服务的构造过程,通过基于数据服务超链的数据服务即时组合完成。其设计思想如下:用户选择并实例化当前数据服务后,从该数据服务的数据服务超链集合中选择一个数据服务超链,从而实例化该服务超链指向的下一个数据服务。以此循环,直到用户发现满意的集成结果,过程终止,系统输出一个复合数据服务。其形式化定义如下。

定义 5 (复合数据服务的构造过程)一个复合数据服务的构造过程可以表示为一个四元组 $\langle DS, States, OP, \delta \rangle$, 各变量介绍如下。

- $DS=\{ds_i|i=1,\dots,n\}$ 是一个有限集,由此次构造过程中产生的所有数据服务实例构成。该集合中的元素满足: $\exists \text{init}DS \in DS$ 为应用的起始数据服务实例; $\forall a \in DS, \exists b \in DS$, 使得 $(a, b) \in \text{links}$ 或 $(b, a) \in \text{links}$ 。每个数据服务实例有确定的状态,以 $dsstate$ 表示, $dsstate \in \{\text{ready}, \text{active}, \text{finished}, \text{terminated}\}$, 分别表示服务实例准备运行、已启动、已成功完成和已终止运行。已成功完成和已终止运行的数据服务返回码为标准的 HTTP 请求返回码。
- $States$ 表示复合数据服务构造和执行过程中应用状态的集合, $state \in States$, 是当前复合数据服务构造与运行过程中应用的状态,体现在该复合数据服务构造和运行过程中当前已经成功执行的数据服务实例及其状态上。
- OP 是用户操作集合, $OP=\{\text{selectLink}, \text{send}, \text{terminate}\}$, 表示从当前数据服务返回结果的资源表述中选择数据服务超链、向数据服务超链指向的数据服务发送 HTTP 请求 ($\text{get}, \text{put}, \text{delete}, \text{post}$ 等标准操作)、终止复合数据服务的构造过程。 $op \in OP$, 表示当前用户操作。
- δ 表示应用状态的转移。 $\delta(ds_i, op)=ds_k$, 表示操作 op 作用于数据服务实例 ds_i , 产生了新的数据服务 ds_k 实例。状态的转移体现在该复合数据服务构造和运行过程中,当前执行的数据服务实例及其状态的改变。

以上述定义为基础,用状态机图对复合数据服务的构造和运行过程进行描述,如图 2 所示,简便起见,省略了异常情况下的应用状态。复合数据服务构造和运行过程中应用的状态集合为 $\{\text{Start}, \text{VictimInfoReady}, \text{RelativesReady}, \text{ContactsReady}, \text{RelatedPersonReady}, \text{VehicleReady}, \text{MonitorVehicleReady}, \text{GotSuspicious}, \text{End}\}$ 。初始的数据服务实例为表 1 中

数据服务“受害人信息”或“现场监测车信息”的实例。当“受害人信息”成功执行后,该数据服务实例的状态为“finished”,而相应的应用状态为“VictimInfoReady”,标志着当前应用已经为下一个数据服务实例准备好相应的输入。用户此时从其资源表述中选择数据服务超级链接“/relatives”,并向其发送 HTTP get 请求,触发相应的数据服务实例,应用状态改变为“RelativesReady”。

一旦复合数据服务构造完毕,将自动生成一个新的复合数据服务并对外发布,同时生成该复合数据服务构造过程的描述文件。当用户访问并执行该复合数据服务时,数据服务执行引擎将根据该描述文件自动化复合数据服务的执行。

3 数据服务超链的即时推荐机制

为了方便数据服务关联关系的建立和重用,除了提供有效的手段辅助用户手工建立数据服务超链之外,还可以向用户主动推荐数据服务超链,从而辅助用户进行临机决策,即时进行数据集成。

向用户推荐数据服务超链的依据有两种:基于规则或基于统计。基于规则的方法是领域专家需要事先制定并维护推荐规则,通过逻辑推理得知用户可能需要的算子;基于统计的方法是根据用户构造复合数据服务的历史提取服务超链的选择模式,并把不同模式出现的概率作为推荐服务超链的依据。相比于基于规则的方法,基于统计的方法无需事先制定推荐规则,可以随着用户构造的复合数据服务中新型数据服务超链的出现模式进行动态更新,当“数据服务空间”中积累起足够多的数据服务超链,出现模式样例时,可实现相对准确的推荐。因此,本文基于统计方法进行数据服务超链的推荐。

数据服务超链的推荐可以转化为条件概率的比较问

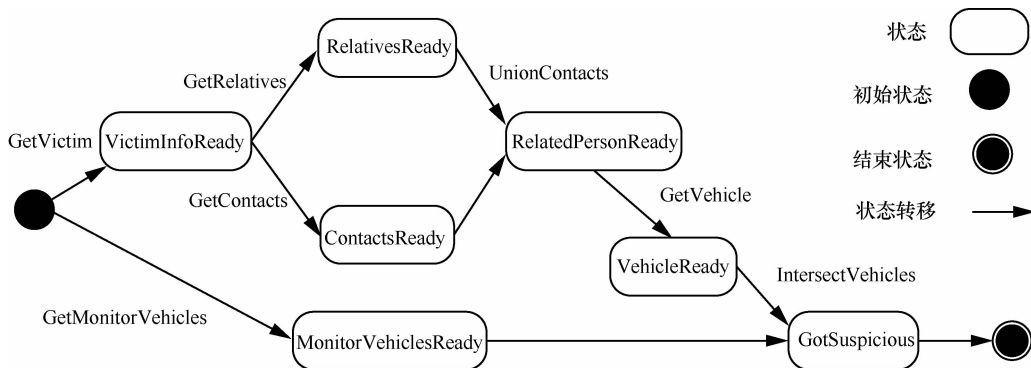


图 2 可疑数据服务对应的状态机示意



题。将从数据服务 a 指向 b 的数据服务超链记为 $a \rightarrow b$, 其中, a 为该数据服务超链的起点, 记为 a^s , b 为该数据服务超链的终点, 记为 b^t 。对数据服务 a , 选择数据服务超链 $a \rightarrow b$ 的条件概率为:

$$P(b^t|a^s) = P(a \rightarrow b) / P(a^s) \quad (1)$$

选择数据服务超链 $a \rightarrow c$ 的条件概率为:

$$P(c^t|a^s) = P(a \rightarrow c) / P(a^s) \quad (2)$$

式(1)、式(2)相除, 得:

$$\frac{P(b^t|a^s)}{P(c^t|a^s)} = \frac{P(a \rightarrow b)}{P(a \rightarrow c)} \quad (3)$$

用 $Count(x \rightarrow y)$ 表示数据服务空间中数据服务超链 $x \rightarrow y$ 的出现次数, 则:

$$P(a \rightarrow b) = \frac{Count(a \rightarrow b)}{\sum_{(x,y)} Count(x \rightarrow y)} \quad (4)$$

同理, 可得:

$$P(a \rightarrow c) = \frac{Count(a \rightarrow c)}{\sum_{(x,y)} Count(x \rightarrow y)} \quad (5)$$

将式(4)、式(5)代入式(3), 得:

$$\frac{P(b^t|a^s)}{P(c^t|a^s)} = \frac{Count(a \rightarrow b)}{Count(a \rightarrow c)} \quad (6)$$

因此, 为数据服务 a 推荐某数据服务超链 $a \rightarrow b$ 的出现概率, 等比于数据服务空间的所有超链中 $a \rightarrow b$ 出现的个数。根据这个结论, 在推荐超链时, 通过统计数据服务空间中数据服务超链的个数对其进行排序。

此外, 还可以利用复合数据服务构造的一些性质提高推荐的准确度。例如, 复合数据服务的内部存在数据服务超链组合片断的自相似性。例如, 在图2中, 利用从“受害人信息”数据服务指向“最近联系人信息”和“亲属信息”两个数据服务的数据服务超链, 分别获得受害人的亲属信息和联系人信息, 然后通过指向“合并”数据服务的超链, 发送“post/union”请求将联系人信息和亲属信息合并。在这里, 由于“union”自身的特点, 使得产生“/union”的输入的数据服务超链组合片断高度相似。

将复合数据服务内部的构造性质称为数据服务超链的局部特征, 而通过整个数据服务空间内部的数据服务超链计算得出的超链出现概率则体现了超链通用的使用方式, 是数据服务超链的全局特征。

结合数据服务超链的局部特征和全局特征, 用如下算法生成数据服务超链的推荐列表。

算法1 数据服务超链推荐算法

输入: 当前复合数据服务, 当前数据服务空间, 需要推荐的服务超链个数 k 。

输出: 长度为 k 的数据服务超链推荐列表。

算法流程介绍如下。

(1) 利用当前复合数据服务构造过程中已有的数据服务超链片段, 根据上述自相似性建立的简化推荐规则, 生成数据服务超链的局部推荐列表, 记列表中的超链个数为 m 。

(2) 若 $m \geq k$, 则取前 k 个, 算法完成; 若 $m < k$, 则进入步骤(3)。

(3) 利用整个数据服务空间的数据服务超链, 根据式(6)的概率比值计算方法, 生成超链的全局推荐列表, 从全局推荐列表中取前 $(k-m)$ 个与局部推荐列表不同的数据服务超链, 追加至局部推荐列表的末尾, 该列表即最终数据服务超链推荐列表。

按上面的算法结合数据服务超链的全局特征和局部特征, 可以提高服务超链推荐算法的准确性。

4 案例和实验

以第1节介绍的场景为例, 说明本文提出的基于数据服务超链的数据服务即时组合方法能够满足情景数据集成的特定需求。

公安干警小王在进行此案件的分析时, 首先尝试选用“受害人信息”数据服务并发送服务调用请求, 该服务返回调用结果后, 小王根据自己的办案经验, 尝试选择“最近联系人信息”数据服务, 并配置其输入参数, 建立起二者之间参数的数据关联, 系统根据这些信息自动建立起“受害人信息”数据服务和“最近联系人信息”数据服务之间的数据服务超链。小王得到“最近联系人信息”数据服务返回结果后, 根据经验进一步选择“亲属信息”数据服务, 系统建立起从“受害人信息”数据服务指向“亲属信息”数据服务的超链。此时, 根据前文所述的推荐算法, 系统向其推荐“post/union”、“post/intersect”等数据服务超链列表, 小王根据自己的办案经验, 从中选择“post/union”数据服务超链, 并配置相应的输入参数。

在上述过程中, 用户每一步都可根据当前服务执行请求返回的结果或系统推荐的数据服务超链进行临机决策, 选择下一步要请求调用的服务, 逐步得到满意的结果。每一步服务执行返回的结果若不满意, 用户都可以选择其他数据服务, 重新建立数据服务超链。若用户认为得到满意结果, 可终止构造过程, 同时保存该应用。此时, 系统将根

据上述过程自动生成一个新的复合数据服务“嫌疑犯确定”,并可对外发布,从而方便其他公安干警访问。

情景数据集成参照参考文献[14]基于对 BPEL(business process execution language, 业务流程执行语言)引擎的扩展实现,图 3 用时序图描述了上述复合数据服务的构造过程。值得指出的是,基于数据服务超链进行数据服务即时组合具有即时性,可能在出现多次交互后才能获得一个满意的复合数据服务,另外,不掌握编程知识的业务用户或最终用户在构造复合数据服务过程中,也很少考虑路径的优化问题。因此,即便系统并不保存用户回退的过程,在输出的复合数据服务中也难免会包含一些冗余和非优化的步骤。解决此问题可以从两个方面进行:一方面,由专家或系统管理员对复合数据服务进行求精处理;另一方面,利用自动化的方法,由系统利用其他用户的构造历史、专家知识库等对复合数据服务的结构进行分析求精。

为了定量验证本文数据服务超链推荐算法的有效性,实验数据集来自 Yahoo! Pipes,由于其覆盖不同的数据格式和数据来源,具有较强的代表性。从 Yahoo! Pipes 的应用库中取出 620 个含有 4 857 个数据服务超链的应用,其中 600 个(含 4 549 个数据服务超链)作为样本集(即作为支

持推荐算法的数据服务空间),20 个(含 308 个数据服务超链)作为测试集(即接受数据服务超链推荐的对象)。Yahoo! Pipes 中的应用是以专有的图形化方式表示的,不能直接用于推荐算法中,因此利用本文的数据服务及超链模型对其进行转换。

假设推荐算法给出的候选数据服务超链列表的大小为 k ,需要进行数据服务超链推荐的复合数据服务的个数为 n ,第 i 个复合数据服务用户需要的数据服务超链在推荐列表中的排名为 R_i ($0 \leq R_i \leq k, R_i=0$ 表示不在列表中),用 $score(i)$ 表示对第 i 个复合数据服务推荐数据服务超链时的评分,满分为 k ,不在列表中评分为 0。用式(7)计算:

$$score(i) = \begin{cases} k - R_i + 1, & 0 < R_i < k \\ 0, & R_i = 0 \end{cases} \quad (7)$$

推荐算法的准确性可用 n 个应用的评分总和与满分($k \times n$)的比值衡量,如式(8)所示:

$$\tau = \frac{\sum_{i=1}^n score(i)}{k \times n} \quad (8)$$

由于 k 是直接给用户推荐的超链个数,取值不宜过大。本文取 $k=4,5,\dots,12$,在样本集上分别取 200,300, $\dots$,

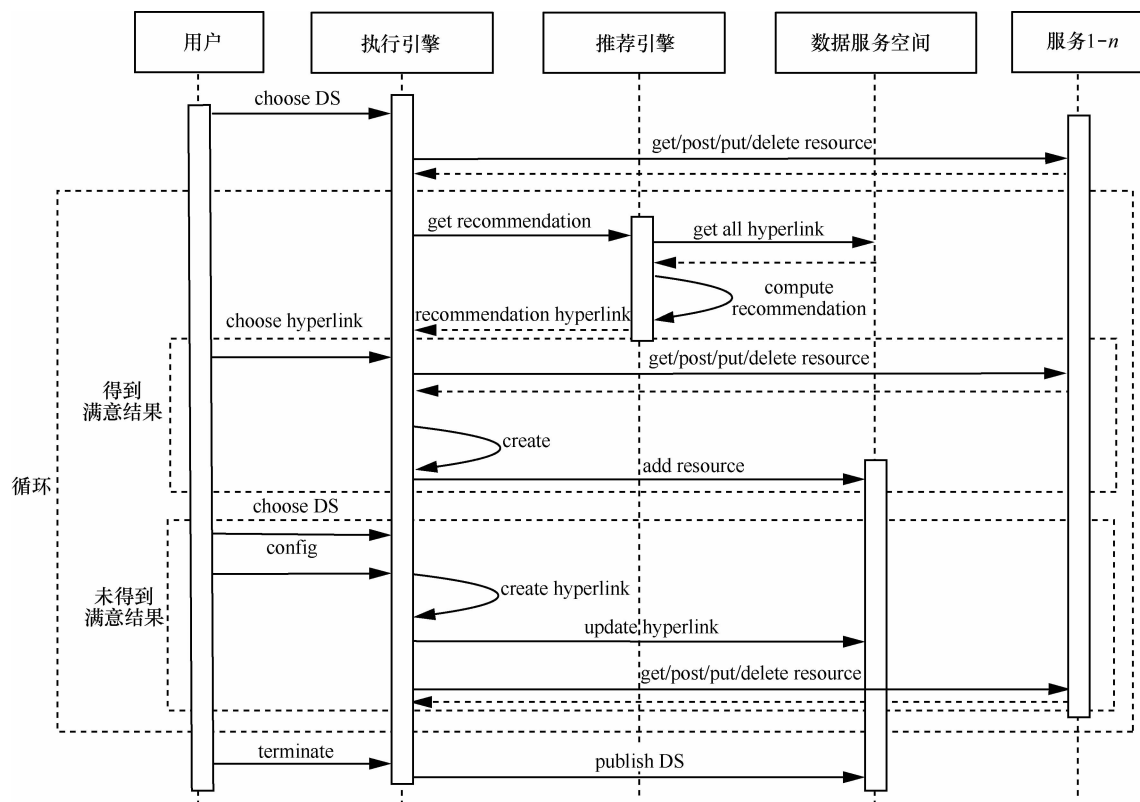


图 3 复合数据服务的构造过程



600 个复合数据服务对测试集进行推荐, 对不考虑自相似与考虑自相似两种情况进行实验, 发现 $k=5$ 时能够得到最佳的实验结果。图 4 给出了 $k=5$ 的实验结果。

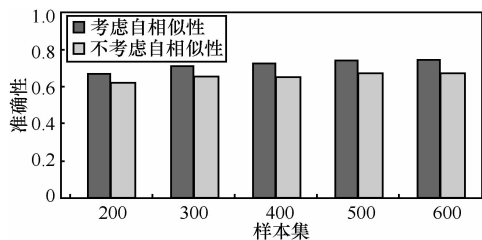


图 4 $k=5$ 时数据服务超链推荐的准确性

从图 4 中可以看出, 随着样本集集成应用数量的增多, 推荐算法的准确性不断提升, 这说明在实际应用中, 随着数据服务空间的演化, 复合数据服务不断增多, 推荐算法的准确性逐步得到提升。还可以看出, 考虑数据服务超链组合片断的自相似性的推荐算法准确性普遍高于不考虑自相似性的情况, 这说明还需要充分利用复合数据服务构造的内部性质提高推荐的准确度。

5 相关工作

本节从如下 3 个方面对本文提出的方法与相关工作进行比较。

(1) 借鉴超链对服务关联关系进行建模的方法

比较典型的工作有韩燕波等提出的服务超链^[11]以及 IBM 的研究者提出的“超级服务 (HyperService)”^[12]。但目前这些研究主要用来描述服务之间在业务逻辑上的相关性, 即服务行为约束关系, 而并不完全适用于描述数据服务之间的相关性。本文提出的数据服务超链基于 HATEOAS 原则, 在用来解决本文场景中的情景数据集成问题时, 能够更好地继承超链所具有的结构简单、客户端和服务端松耦合性更好等性质。

(2) 借鉴 HATEOAS 设计原则对服务进行组合的方法

近年来, 随着 REST 服务在更广范围内被使用, 有研究者提出借鉴 HATEOAS 原则的服务组合方法^[13-15]。这些工作的主要聚焦点在于如何扩展现有服务组合引擎, 使之支持 REST 服务的组合, 其方法主要适合于用基于 HATEOAS 原则的 REST 服务组合方法解决传统服务组合问题。本文与这些工作的相同之处在于融合了 REST 风格的服务模型以及 HATEOAS 的设计原则, 目的是更好地继承“超链”的优良性质, 但与这类工作不同的是, 本文工作的主要贡献不在于扩展现有服务组合引擎使之支持基于 HATEOAS 的

REST 服务组合, 而是探讨了一种通过超链导航结合超链的智能推荐等手段支持数据服务构造的新方法。

(3) 服务之间关联关系的即时推荐方法

Advisor M^[16]基于 Mashup 应用构造历史记录进行服务推荐, 根据服务之间关联关系的使用频率进行推荐, 并结合 AI 规划的方法为用户推荐最可能生成用户期望输出的 Mashup 路径。MARIO^[17]则基于标签进行关联关系的推荐, 生成用户期望的流程。这些方法以生成用户期望的整个流程为目标, 比较适合于终点已知的场景; 而本文方法更适合需求不确定、在每一步操作均需要人为参与进行决策的情景应用构造。

6 结束语

在开放和动态变化的互联网及云环境下, 数据集成的需求往往即时出现而无法完全预先定义, 如何使用户根据当前业务需求以一种便于重用的方式建立数据源之间的关联关系, 并支持用户进行临机决策, 是传统数据集成技术面临的新挑战。“网页超链”作为 Web 导航与发现 Web 网页的关键元素提供了启发, 本文的新思路是借鉴网页超链的思想, 使得用户在数据服务导航过程的同时完成数据的即时集成。本文提出利用“数据服务超链”的抽象辅助用户显式地描述和表达各种数据服务之间隐含的关联关系, 同时融合 REST 服务和 HATEOAS 的设计思想, 在数据服务中包含相应的服务超级链接, 通过超链的支持, 使客户端能通过超级链接的选择驱动复合数据服务的构造与运行。

本文提出基于超链导航的情景数据集成方法和状态机描述, 并讨论了基于统计的数据服务超链推荐算法, 帮助用户从各种看似无关的数据中建立关系, 进行临机决策。结合案例进行了分析评价及实验, 结果表明, 本文提出的方法能够有效支持情景数据集成。进一步工作包括探讨具体的数据服务超链的自动生成方法, 进一步研究数据服务超链的性质 (如层次性、局部性等) 并给出验证, 在此基础上研究数据服务空间的构建、演化、性质以及基于数据服务空间的智能服务超链推荐算法等。

参考文献

- 1 Lv J, Ma X X, Tao X P, et al. Research and progress on internetware. Science in China (Series E), 2006, 36 (10): 1037~1080

2 Han Y B, Wang G L, Liu C, *et al.* Internet Computing: Theory and Practice. Beijing: Science Press, 2010

3 Jhingran. Enterprise information mashups: integrating information, simply. Proceedings of the 32nd International Conference on Very Large Databases, Seoul, Korea, 2006:3~4

4 Schahram D, Pichler R, Savenkov V, *et al.* Quality-aware service-oriented data integration: requirements, state of the art and open challenges. ACM SIGMOD Record 41, 2012(1): 11~19

5 Wang G L, Yang S H, Han Y B. Mashroom: end-user mashup programming using nested tables. Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 2009: 861~870

6 Han Y B, Wang G L, Ji G, *et al.* Situational data integration with data services and nested table. Service Oriented Computing and Applications, 2013, 7(2): 129~150

7 Nadav E, McCurley K S. Locality, hierarchy, and bidirectionality in the web. Proceedings of Workshop on Algorithms and Models for the Web Graph, Budapest, 2003

8 Luc M, Hall W. On the expressiveness of links in hypertext systems. The Computer Journal 41, 1998(7): 459~473

9 Thomas F R. Architectural styles and the design of network-based software architectures. University of California, 2000

10 Jim W, Parastatidis S, Robinson I. REST in Practice: Hypermedia and Systems Architecture. O'Reilly Media, Incorporated, 2010

11 Yan S, Han Y, Wang J, *et al.* Service hyperlink for exploratory service composition. Proceedings of IEEE International Conference on e-Business Engineering, ICEBE 2007, Hong Kong, China, 2007

12 Zhao C, Chun'e M, Zhang J, *et al.* HyperService: linking and exploring services on the web. Proceedings of 2010 IEEE International Conference on Web Services (ICWS), Miami, Florida, USA, 2010

13 Rosa A, Wilde E, Bellido J. Hypermedia-driven restful service composition. Proceedings of Service-Oriented Computing, Springer Berlin Heidelberg, 2011:111~120

14 Cesare P. Restful web service composition with BPEL for REST. Data & Knowledge Engineering 68, 2009(9): 851~866

15 Cesare P. Composing restful services with jopera. Proceedings of Software Composition, Springer Berlin Heidelberg, 2009: 142~159

16 Hazem E, Ivan A, Akkiraju R, *et al.* Mashup advisor: a recommendation tool for mashup development. Proceedings of IEEE International Conference on Web Services (ICWS'08), Beijing, China, 2008:337~344

17 Anton V R, Boillet E, Feblowitz M D, *et al.* Wishful search: interactive composition of data mashups. Proceedings of the 17th International Conference on World Wide Web, Beijing, China, 2008:775~784

【作者简介】



王桂玲,女,博士,北方工业大学副研究员,CCF 会员,主要研究方向为云计算、服务计算、数据集成等。

张峰,男,北方工业大学博士研究生,山东科技大学讲师,主要研究方向为服务计算、数据集成等。

韩燕波,男,博士,北方工业大学教授,主要研究方向为分布式系统、云计算和互联网服务等。

(收稿日期:2013-12-30)

·简讯·

2013 年第 4 季度通信业主要通信能力

指标名称	单位	本季末到达	比上年末新增
光缆线路长度	公里	17 451 180	2 645 473
其中:长途光缆线路长度	公里	897 772	29 329
固定长途电话交换机容量	万路端	1 270	-338.6
局用交换机容量	万门	41 052.2	-2 854.2
其中:接入网设备容量	万门	22 523.7	-878.3
移动电话交换机容量	万户	196 545.3	13 675.5
互联网宽带接入端口	万个	35 972.4	9 137.0

(来源:工业和信息化部运行监测协调局)