面向计算机审计的移动数据挖掘服务研究①

Research on the Mobile Data Mining Services for Computer Assisted Auditing

汪加才 （南京审计学院 计算机科学与技术系 南京 210029）
朱艺华 （浙江工业大学 信息智能与决策优化研究所 杭州 310014）

摘要：对智能化在线审计需求的增加驱使审计系统能够支持多种多样的挖掘算法、能够通过 Web 界面甚至移动用户的手持设备来使用数据挖掘服务。基于移动数据挖掘服务的审计系统可允许审计师在更广泛的范围内查询、定位，并能在本地完成其挖掘工作的 Web 服务。这除了可降低网络带宽的需求，还从仅需支付软件使用费而无须考虑软件的购置、配置及培训成本中获得好处。在分析了数据挖掘服务的构造、发现、合成、移动之后，给出了一个基于移动数据挖掘服务的计算机审计框架模型。

关键词：计算机辅助审计 数据挖掘 Web 服务 移动 Agent

1 引言

计算机辅助审计（Computer – Assisted Auditing，CAA）是指审计机关、审计人员将计算机作为辅助审计工具，对被审计单位的财政、财务收支及其计算机应用系统实施的审计。随着“金审工程”的推进，在全面改善审计机关信息化硬件条件的同时，更需要形成与时俱进的现代审计方式和手段，以提高审计质量、降低审计风险，真正实现从单一的事后审计转向事后审计与事中审计相结合、从单一的静态审计转变为静态审计与动态审计相结合、从单一的现场审计转变为现场审计与远程审计相结合。建立这样的审计监督信息化系统有利于增强审计机关在计算机环境下查错纠弊、规范管理、揭露腐败、打击犯罪的能力，维护经济秩序，促进廉洁高效政府的建设，更好地履行审计的法定监督职责。

目前的 CAA 普遍采用一种“数据集中和专家找问题”方式[1]，即要求将一些分布存储在被审计单位各种业务、财务系统中的数据通过网络收集到一个集中的地方，然后由审计专家在这些集中的数据里查找问题。同时，Web 技术及分布式计算技术的广泛应用也使“连续审计”（Continuous Audit）[2]的设想变为可能，可使审计人员实时地得到被审计单位的审计报告。

数据挖掘（Data Mining）是一个从数据中析取潜在有用的、先前未知的和最终可理解的知识的过程。在审计系统中采用数据挖掘技术可为现代化审计提供新的方法和思路，有利于提高审计质量、降低审计风险。数据挖掘技术为审计师反欺诈提供了新的强有力武器，在“连续审计”模式下，它允许计算机持续不断地监控活动、记录交易和实时提供审计线索以识别潜在的欺诈活动。

作为一个新兴的多学科交叉应用领域，数据挖掘正在各行各业的决策支持活动中扮演着越来越重要的角色。但由于数据挖掘是一个复杂的知识发现过程，因此，期望单一或几个数据挖掘系统来满足复杂的审计需求是不现实的。在 CAA 系统中如何结合 Web 服务（Web Services）技术、语义 Web（Semantic Web）技术、软件代理（Software Agent）技术来更有效地利用数据挖掘技术是本文研究的重点。面向 CAA 的数据挖掘服务研究涉及 DMS 的构建与发布，服务的查找、匹配与合成，服务的迁移、执行与监控等。论文在分析了数据挖掘服务的构建、发现、合成、移动之后，给出了一个基于移动数据挖掘服务的计算机辅助审计框架。

① 基金项目：国家自然科学基金（60473097）

2 数据挖掘服务的构建

Web 服务是当今 IT 业界的焦点所在，其主要目标是在现有的各种异构平台的基础上构筑一个通用的与平台无关、与开发语言无关的技术层，各种不同平台之上的应用依靠这个技术层来实施彼此的连接和集成。利用基于 Internet 的数据挖掘服务（Data Mining Services，DMS），审计部门可以从广大的服务供应商那里有目的地选择使用（租赁）所需要的挖掘工具，不仅节省了建置成本和维护成本，而且使从多个挖掘系统中找到最优服务者成为可能。

构建 DMS 的关键是对数据挖掘数据、结果、过程的标准化。随着数据挖掘功能逐渐成为智能分析业务处理中的重要组成部分，首先导致了 CRISP－DM（Cross－Industry Standard Process for Data Mining）标准的开发。CRISP－DM 标准定义了知识发现过程中的关键步骤，包括：业务理解、数据理解、数据准备、数据挖掘、结果的解释，分析和评价、发现知识的部署和利用。后来，许多针对数据挖掘特定过程的标准相继出现，如 DMG（Data Mining Group）的预测模型标记语言 PMML（Predictive Modeling Markup Language），Microsoft 的 OLE DB for DM 用户接口规范、OMG（Object Management Group）的元数据标准 CWM（Common Warehouse Metamodel），面向 Java 平台的数据挖掘应用程序接口规范 JSR－073 等。目前，已存在两个与 DMS 有关的标准规范，即 JDM API Web 服务扩展（JDM API Web Services extensions，JDMWS），Microsoft 及 Hyperion 的 XML 分析规范（XML for Analysis specification，XMLA）。XMLA 重用了 OLE DB for DM 中的模式行集（Schema Rowsets）概念，是专为 Web 客户应用与数据分析提供者（如 OLAP、数据挖掘应用）之间数据访问标准化而设计的基于 SOAP 的 XML API。XMLA APIs 支持任意平台上的客户、服务器以不同的语言进行分析数据的交换。这些标准为提供开放性的数据挖掘服务提供了研究基础和方向。

3 数据挖掘服务的查找与合成

Web 服务的中心思想是围绕服务的发布（Publish）、查找（Find）和绑定（Bind）展开的。因此，Web 服务的查找是整个 Web 服务生命周期中的一个关键环节。对于 Web 服务的检索一般称为匹配（Matching），并将执行匹配的功能模块称为 MatchMaking，其主要功能就是检索出那些服务能力足够满足服务请求者需求的服务提供者。

目前，Web 服务的搜索与发现采用的普遍方法是由客户使用搜索引擎找到服务，或者在相关的 Web 页面中考查其是否满足服务请求的要求。由于用于描述 Web 服务的 WSDL（Web Service Description Language）并不能很好地表达 Web 服务的语义信息，并且它所描述的是静态的 Web 服务，不包含任何有关服务执行过程的信息，文[3]提出了在 WSDL 中加入以 XPath 语法形式表示的语义标注，从而对 WSDL 进行扩充；文[4]提出了一种基于语义 Web，利用过程本体论（process ontologies）的 Web 服务发现技术。该技术将服务的功能作为过程模型（process model）定义了过程本体论，并将 Web 服务通过索引建立到本体论上（即用过程本体论表示服务）；DAML－S/OWL－S 则是使用 OWL、DAML（DARPA Agent Markup Language）＋OIL（Ontology Inference Layer）定义的一种 Web 服务的本体，旨在支持实现 Web 服务的自动发现、调用、合成和运行监测。总之，Web 服务发现的自动化、基于语义和约束的 Web 服务发现，特定应用环境中的 Web 服务发现技术仍是重要的研究方向。

基于语义约束的 DMS 发现要求解决三个问题，一是数据挖掘任务的描述，二是 DMS 语义描述，三是挖掘请求与挖掘服务的匹配算法。数据挖掘任务通常包括数据源定义，挖掘数据的提取、转换，挖掘数据、功能、算法的设置，挖掘模型的构建和应用，PMML 预测模型的导入和输出等过程。数据挖掘任务描述可能包括的信息有：（1）任务类型，如属性重要性分析任务、预测模型建立（分类、回归分析、聚类分析、关联分析、奇异分析）任务、预测模型运用任务、预测模型可视化表示任务等；（2）数据描述，如数据类型（纯数值型、纯标量型、混合型）、数据量（记录长度、记录数目）、数据组织形式（传统的平面文件、XML 格式、数据库类型）等；（3）挖掘算法声明；（4）要求的预测模型所遵循的标准（如 PMML）；（5）响应时间约束等。DMS 语义描述主要包括服务的服务类型、使用约束条件、输入输出参数等。为此，需要建立面向 DMS 发现的本体库（可采用 DAML＋OIL），设计基于 DAML－S/OWL－S 的描述文件和语义匹配算法。

Web 服务使分布在网络中的资源构成了一个虚拟的计算机系统。为了分散和简化应用逻辑，提高服务的可重用性，单个 Web 服务都不可能做得非常复杂，因此需要按照一定的规则将多个简单的 Web 服务合成为合成服务（Composite Service）。合成服务的合成方式可以有静态的和动态的两种：静态合成服务是在设计时由设计者根据实际的需要事先选择组件服务的提供者，并按照一定的顺序合成服务；动态合成服务是指在运行的时候，系统根据客户的请求，自动选择组件服务的提供者，并与其他的 Web 服务合成，产生新的服务。因此，合成服务系统需要一种描述合成服务方案构成和执行过程的方法。目前，许多组织和企业提出了面向合成服务的定义语言，如 Microsoft 的 XLANG、IBM 的 WSFL（Web services Flow Language）、IBM 和 Microsoft 共同提出的 BPEL4WS（Business Process Execution Language for Web services）等。WSFL 和 XLANG 扩展了 WSDL 以支持 Web 服务的合成，前者支持面向图形的过程说明，后者为服务合成提供结构化过程算子（structural constructs for processes）；而 BPEL4WS 则结合了 WSFL 和 XLANG 的这两个特征。为此，需要根据挖掘任务的特点，采用上述的 Web 服务合成语言，设计与开发相应的合成服务定义生成器、解析器和执行引擎。

4　基于移动 Agent 的数据挖掘服务

无论是采用“数据集中”审计还是“连续”审计，都需要将被审计单位的业务数据、资金往来数据传输到远端的 Web 服务执行服务器，无疑会给被审计单位的经济情报等方面的保护工作带来巨大的威胁。移动 Agent（Mobile Agent，MA）是一个能在异构网络中自主地从一台主机迁移到另一台主机，并与其他 Agent 或资源交互的程序。MA 较少依赖网络传输这一中间环节而直接面对要访问的服务器资源，避免了大量数据传输；MA 不需要统一的调度，由用户创建的 MA 可以异步地在不同节点上运行，待任务完成后再将结果传递给用户，降低了系统对带宽的依赖，减少了网络冲突，在低带宽、非稳定连接的网络环境下依然能够保持稳定的工作。典型的 Agent 架构具有比 Web 服务更丰富的含义[5]，两者具有较强的互补性。但目前这方面的研究成果并不多。文[6]用 MA 处理在 UDDI（Universal Description，Discovery and Integration）中的 Web 服务搜索：服务请求者根据自身需求生成并发送 MA 到 UDDI 注册中心进行服务目录查询，然后自主地到服务提供者本地执行，最后 MA 携带执行结果返回；文[7]给出了一种将 MA 和 Web 服务技术融合为一种智能 Agenice 的方案，其目的是要将 Web 服务的静态远程调用方式变为一种动态的、可移动的、智能的调用。

在 CAA 中将 DMS 与 MA 结合，可以充分利用 MA 的许多优点，如审计部门可以将 DMS 提供者所提供的 Web 服务包装为 MA 动态移动到本地或被审计单位，避免了大量业务数据或中间数据的网络传输，有效防止了敏感数据的泄露。

5　基于移动数据挖掘服务的 CAA 框架设计

基于前文分析，我们给出了如图 1 所示的“基于移动数据挖掘服务的 CAA 框架”。框架系统由五个部分所组成：审计应用系统，审计部门的审计服务器，被审计单位的应用服务器、数据库服务器或 Web 服务器、DMS 提供中心和注册中心。其中：

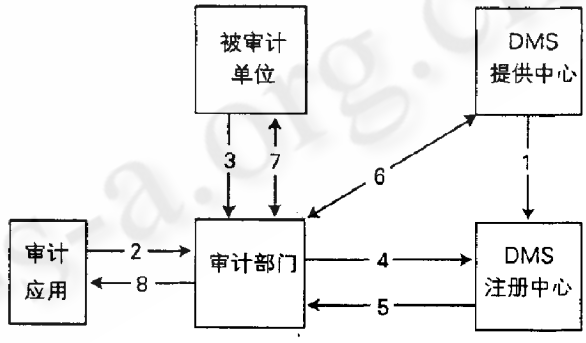

图 1　基于移动 DMS 的 CAA 框架

（1）DMS 服务提供中心负责构造 DMS 及其语义描述文件（OWL－S based Semantic Web Service Profile，SWSP），并将其发布到 DMS 注册中心的 UDDI 扩展服务器端（图中的 1）。其中，本体构造器负责构造 SWSP 用到的数据挖掘本体信息；SWSP 构造部件负责建立 DMS 的语义描述文件；

（2）审计应用以交互方式通过可视化界面构造挖掘任务，并传递给（图中的 2）驻留于审计服务器的任务解析器生成基于 BPEL4WS（或 XLANG、WSFL 等）的数据挖掘任务描述文件（BPEL4WS based Data Mining

Task Profile,DMTP);

(3) 审计服务器生成携带被审单位环境描述文件(图中的3)和DMTP等信息的DMS请求Agent并发送到DMS注册中心的UDDI扩展服务器端(图中的4)。后者根据请求查找UDDI的tModel,得到满足条件的服务列表,之后再对列表中的服务进行基于OWL-S的语义匹配,最后将筛选之后得到到的候选服务信息添加到DMTP并返回给审计服务器(图中的5);

(4) 审计服务器根据DMTP从DMS提供中心下载相应服务(图中的6),并包装为一批相互协作的服务移动Agent(Mobile DMS Agents,MDAs)发送到被审计单位。最终的计算结果由MDAs带回审计服务器(图中的7)和审计应用(图中的8)。

6 结论

以提高Web应用系统的健壮性、高性能计算能力为出发点的新型分布计算模型——Web服务技术,通过进行Web上已有网络计算组件的集成、基于现有协议提高Web应用的互操作能力及服务质量等手段,进而解决现实Web应用中"应用到应用(application-to-application)"及"点对点(peer-to-peer)"的核心问题,使当前Web应用适应全球化和复杂商务处理的需求。本文有机结合了数据挖掘、Web服务、语义Web、移动Agent等前沿技术来实现面向计算机审计及网上审计的移动数据挖掘服务。利用移动数据挖掘服务,审计部门可以从广大的服务提供商那里有目的地选择(租赁)所需要的挖掘功能并内嵌到相应的审计决策支持系统中,不仅可节省系统的建置成本和运行成本,而且使从多个数据挖掘服务中心中找到最优服务者成为可能,从而实现真正意义上的网上审计目标。

参考文献

1 董化礼、刘汝焯,计算机审计——数据采集与分析技术[M],清华大学出版社,2002年。

2 Groomer,S. M. and Murthy, U. S. ,Continuous Auditing of Database Applications:An Embedded Audit Module Approach[J]. Journal of Information Systems,Spring,1989.

3 Peer J. Bringing together semantic Web and Web services[C]. In: Horrocks, ed. Proc. of the Int'l Semantic Web Conf. Sardinia: Springer-Verlag, 2002. 279-291.

4 Klein M, Bernstein A. Searching services on the semantic Web using process ontologies[C]. In: Isabel C, ed. Proc. of the Int'l Semantic Web Working Symp. (SWWS2001). Amsterdam: IOS Press, 159-172.

5 Hendler, J. Agents and the Semantic Web[J]. IEEE Intelligent Systems,2001,16(2):30-37.

6 Sheng-Tzong Cheng; Jian-Pei Liu; Jian-Lun Kao. A new framework for mobile Web services[C]. Proc. of Symposium on Applications and the Internet (SAINT) Workshops,2002,218-222.

7 孙凡、景广军,Agenice:一种新的基于Web Service的Mobile Agent的理论框架[J],计算机工程,2004,30(3):85-88.