第35卷 第23期 计 算 机 工 程 2009年12月
Vol.35 No.23 Computer Engineering December 2009
·开发研究与设计技术· 文章编号：1000—3428(2009)23—0278—03 文献标识码：A 中图分类号：TP391

分布式专家行为信息系统

胡少华
(科学技术部信息中心，北京 100862)

摘 要：针对专家行为数据分布广、难以全面分析利用的问题，提出分布式专家行为信息系统。该系统建立安全可信的分布式数据采集体系，依据转换规则库、映射规则库、元数据对不同含义、不同格式的数据进行转换处理，形成统一的专家行为数据库，基于不同业务阶段的数据特点进行比对、关联，利用关联规则算法实现对专家行为模式的挖掘分析。结果证明该系统有助于及时、准确地得到专家行为信息。
关键词：分布式；数据处理；关联规则；行为分析

Distributed Expert Conduct Information System

HU Shao-hua
(Information Center, Ministry of Science and Technology, Beijing 100862)

【Abstract】Aiming at the problem that expert conduct data is widespread and difficult to get a complete analysis, this paper proposes a distributed expert conduct information system. This system builds secure and trustworthy distributed conduct data framework. Based on transforming rules, mapping rules, metadata references, expert conduct data with different types and meanings are transferred into a unified expert conduct database. After comparing, correlation, analyzing the data of different phrases, expert conduct patterns can be reached based on association rules mining algorithm. Results proves the system is helpful to get the expert conduct information timely and exactly.
【Key words】distributed; data processing; association rule; conduct analysis

1 概述

在目前的科研管理体制中，随着专业化、科学化管理要求的不断加强，不同专业领域的专家成为重要群体，他们在科研项目的建议征集、评估评审、总结验收过程中往往承担着主导性作用，建立专家信用信息的分析、管理体系成为目前科研活动的迫切需求，而准确地记录、分析专家的行为信息是其首要技术前提。科研活动一般具有周期长、过程多的特点，科研项目的建议征集、评估评审、总结验收往往由散布于不同网域的多个管理信息系统完成。散布于这些信息系统、数据平台中的专家行为信息，或多或少存在着数据含义、数据格式、存贮方式等方面的差异，难以集成利用，也难以进一步挖掘分析。

分布式专家行为信息系统以现有科研项目征集申报、评估评审、总结验收等管理信息系统、数据平台为基础，通过安全可信的数据采集通道，及时采集来自各类专用管理信息系统的相关数据资源。基于不同规则库建立多重自动处理、人工评价机制，对来自不同渠道的专家行为数据进行转换、映射、清洗、标引，形成规范可用的数据，并基于科研项目周期特点对数据进行关联、分析、挖掘。

2 总体应用框架

分布式专家行为信息系统涉及多类信息系统、数据资源，其中包括科研项目征集申报信息系统、科研项目专家评审系统、科研项目总结验收系统以及其他相关科研管理信息系统，这些系统记录了专家在不同科研活动阶段的行为信息，是专家行为分析的主要数据来源。

分布式专家行为信息系统首先通过分布式数据采集代理及数据汇集系统完成专家行为数据及相关信息的汇集，然后通过处理分析系统形成规范化行为数据库，并针对行为信息进行挖掘利用，各系统互相联系，分布运行于不同网域中。分布式专家行为信息系统总体结构见图1。

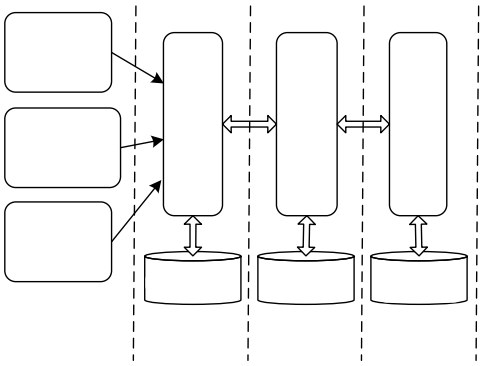

图1 分布式专家行为信息系统总体结构

2.1 分布式数据采集

数据采集是整个系统的关键环节，目前多数科技计划、科研活动均实现了信息化管理，并已逐步形成比较完备的数据库平台和资源更新机制，是采集专家行为数据的主要渠道。这些系统一般分为2类：(1)依托互联网运行，一般采用Web服务架构；(2)依托业务网运行，一些采用Web服务架构，少数仍采用传统的C/S方式运行管理。

基金项目：国家科技基础条件平台基金资助项目(2005DKA33401)
作者简介：胡少华(1969－)，男，高级工程师、硕士，主研方向：复杂信息系统架构，数据集成，Web服务
收稿日期：2009-03-05 **E-mail：**hsh@most.cn

分布式数据采集系统为星型结构，包括一个控制中心和多个数据采集代理。控制中心用于配置、调度分布于不同信息系统的数据采集代理，并与数据采集代理协商，完成采集任务管理、数据安全传输管理、数据完整性验证等功能。考虑到部署管理需求，对于 Web 服务架构的系统及 C/S 方式的系统，均采用统一的数据采集代理。数据采集代理包括配置调度模块、数据操作接口、安全验证模块。数据操作接口直接架构在本地数据层上，用于数据采集、缓存、发送、校验；安全验证模块用于身份验证、数据安全传输。分布式数据采集体系见图 2。

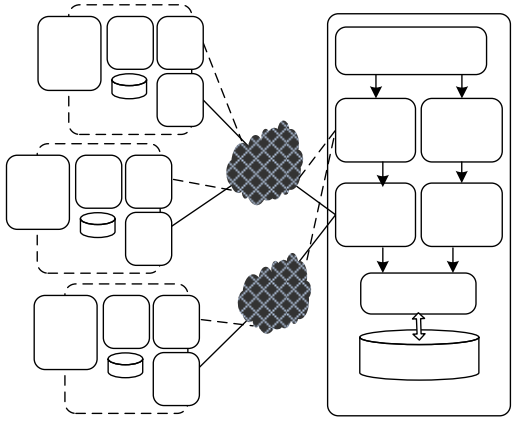

图 2　分布式数据采集体系

数据采集的方式包括 3 种：(1)基于发布/订阅机制，数据采集代理向控制中心发布数据计划，控制中心依据订阅计划采集数据。发布/订阅机制适于数据更新具有明显周期的采集行为。(2)自动触发机制，对于不具备前期特点的数据源，基于数据更新总量或其他需求设定触发条件，一旦制定条件由数据采集代理发起向控制中心推送数据。(3)对于部分无法自动采集的数据，可以离线采集数据，然后将数据文件集中导入。

2.2　数据处理分析

集成应用多源数据时往往要消除模式冲突、清洗数据，并进行抽取、转化和装载(Extract, Transform, Load, ETL)操作以提高数据质量[1]。由于本系统采集数据均源自严格管理运行的业务系统，因此数据质量较高，内在关联强，在数据处理分析过程中，更关注消除模式冲突，对数据进行规范化处理、融合。在实现过程中，主要依据多类规则库实现对数据含义、数据格式、数据存贮方式的自动处理，并结合人工审核评价不断调整规则库，优化处理结果，实现对不同渠道采集数据的规范化处理。在此基础上形成面向研究建议、评审评估、检查验收、成果发布应用等不同业务阶段专家行为数据的分析、关联，为进一步评价挖掘提供信息支持。数据处理分析的技术框架主要包括如下模块：

(1)数据汇集及转换模块。对所有接收数据进行登记，在此基础上依据转换规则库进行格式方式、存贮方式等转换工作。

(2)自动映射关联及人工关联审核模块。自动映射关联模块基于映射规则库完成对来自不同渠道的数据内容的关联，映射规则库包括针对人员描述、单位描述、活动主体、项目或课题描述、活动阶段等要素的映射规则。通过人工关联审核对处理结果进行评价，对错误结果进行修正，如错误率高于一定阈值，通过规则库管理模块修改映射规则、映射模板，或创建新的映射模板，重新进行自动处理、审核、修正循环。图 3 描述了通过映射规则库对人员编号、证件编号、专家 ID 进行映射，依据基本信息库实现数据记录关联的过程。

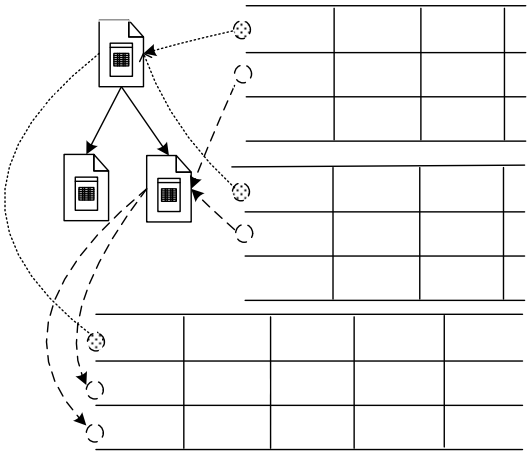

图 3　基于映射规则和基本信息库的数据映射

(3)自动清洗标引及人工分析审核模块。自动清洗标引模块依据标引规范库、元数据参照对当前数据进行清洗、分类、标注以备进一步分析评价。通过人工审核分析对结果进行评价，对错误结果进行修正，如错误率高于一定阈值，通过规则库管理模块修改标引规则库，并重新进行自动处理、审核、修正循环。

(4)行为对比提取及行为评价审核模块。基于行为规则库，利用审核标引数据抽取或更新相关人员、单位等活动主体的行为记录。通过人工审核对结果进行评价，对错误结果进行修正，如错误率高于一定阈值，通过规则库管理模块修改行为规则库。

(5)专家评价审核模块。利用专家及相关科研活动主体的行为记录及相关数据，引入评价专家进行多轮评价、审核，并将评价结果作为参考记录。

(6)审计及操作日志管理模块记录所有批量自动操作及人工操作，用于后续分析、审计。

数据处理及分析体系如图 4 所示。

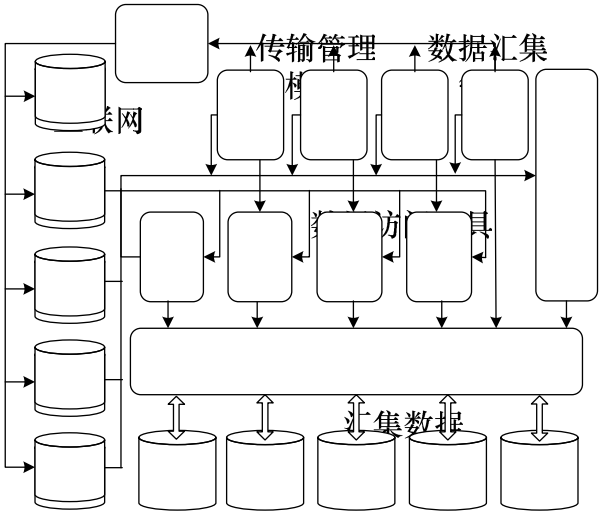

图 4　数据处理及分析体系

所有自动处理均依据转换规则库、映射规则库、标引规则库、元数据参照、行为规则库进行，通过人工审核对自动

处理结果进行判断，并利用规则库管理模块调整相关规则及算法。

所有模块均通过数据访问层及数据工具层完成数据操作，数据资源包括汇集数据及其缓冲区，相关应用逻辑视图包括基本信息、行为记录、评价记录等。

2.3 信息服务

专家行为数据、评价参考数据在经过多轮处理审核后成为静态数据，可依据授权面向不同用户提供服务，服务内容包括统计查询、决策管理信息服务、挖掘分析等。一方面这些数据可以通过 Web 服务、数据接口作为其他管理系统的数据源；另一方面，用户可以直接通过用户界面进行查询、分析。

3 关键技术

3.1 可信数据采集

可信计算组织(Trusted Computing Group, TCG)在可信计算框架中指出：信任是实体针对某一目的按特定方式行动的预期，可信至少包括保护性、完整性及可测量性[2]。在本系统中，主要数据源、数据汇集点均处于较强的安全控制域中，可信数据采集更关注于数据采集、数据传输、数据收发过程中的可信性，主要包括 3 个方面的要求：(1)身份可信性，包括数据发送、接收主体身份的确认和验证；(2)数据传输过程的可信性，包括数据的完整性、防篡改、防截获等；(3)行为可信性，数据提供者对所提供数据的不可抵赖性、数据接收者对接收数据的不可抵赖性。这些要求通过以下技术实现：

(1)基于 PKI 数字证书和动态密码的身份验证。对所有数据发送方、接收方颁发 X.509 数字证书[3]，数据采集中心运行证书目录，相关数据收发方存贮对方证书。对所有数据发送方、接收方部署基于时钟的动态密码生成模块，在每次身份验证时发送方用私钥加密当前动态密码，接收方在预定时限内用对方公钥解密并利用本地当前动态密码验证。

(2)双通道可信传输。综合考虑安全性、计算效率、部署难度、管理方式等因素，在安全验证模块中分别采用安全验证通道、数据通道完成验证和传输，并利用 SHA-1 单向哈希算法、RSA 算法及 3DES 算法[4]实现相关计算。安全验证通道用于数据量少、计算量较大的身份验证、密钥商定、安全收据确认等任务，主要利用 RSA 算法加密传送信息。数据通道用于大批量数据传输，并基于一次一密的原则利用 3DES 等对称算法加密传送信息。

(3)发送接收行为的不可抵赖性。数据传输完成后，数据发送方通过数据通道传送数据目录分组摘要，数据接收方对比通过安全验证通道得到的目录分组摘要，确认无误后增加操作记录与目录摘要一起签名发送至数据发送方，数据发送方将上述内容作为收方安全收据保存，再用私钥签名后返回，数据接收方以此作为发方安全收据，从而形成双向收据。

设数据发送方为 Sed，数据汇集中心为 Rec，数据通道标识为 DataCh，安全验证通道为 SecCh，可信数据采集的基本通信流程如下：

步骤 1 Sed 发起通信，计算本地当前动态密码，利用私钥加密后通过 SecCh 传送至 Rec。

步骤 2 Rec 查询本信任域目录得到 Sed 公钥，利用 Sed 公钥解密并对比本地当前动态密码，验证对方身份。验证成功后生成随机数种子，利用 Rec 私钥加密通过 SecCh 发出。

步骤 3 Sed 利用 Rec 公钥解密得到随机数种子，利用随机数种子计算出对称密钥组 *f*，同时将待发送数据目录分组摘要信息通过 SecCh 发送至 Rec。

步骤 4 Rec 利用随机数种子计算出同一密钥组 *f*，利用 Sed 公钥解密接收数据得到数据目录分组摘要信息。

步骤 5 Sed 利用对称密钥组 *f* 加密数据并通过 DataCh 传输至 Rec。

步骤 6 传输完成后，Rec 对比接收数据目录分组摘要，增加操作日志后签名发送 Sed 作为收方安全收据，Sed 接收后再签名返回作为发方安全收据。

3.2 关联行为分析

专家行为分析包括直接行为分析和隐含模式分析，直接行为分析通过规则比对、搜索即可确定，如评审专家与受评主体间关系。隐含模式需对较长周期的大量数据进行挖掘分析，如不同专业领域专家组合对项目评价的影响、不同单位背景专家评价的关联规则等，通过数据挖掘技术中的关联规则算法能够实现对隐含模式的挖掘。

关联规则是数据挖掘的主要方法之一，用于研究数据项集彼此间的关联性。在事务 T 中，项集 A 和项集 B 间的关联规则可以用蕴涵式 $A \Rightarrow B$ 表示，其支持度 $support(A \Rightarrow B) = P(A \cup B)$，表示在所有事务中包含 A 或 B 的概率；其置信度 $confidence(A \Rightarrow B) = P(B \mid A)$，表示 B 和 A 在事务 D 中同时出现的概率。设定最小支持度 min$S$，最小置信度 min$C$，如果项集支持度不小于 min$S$，则称之为频繁项集。挖掘关联规则就是针对事务 T，找出支持度不小于 min$S$，置信度不小于 min$C$ 的规则[5]。目前计算频繁项集的主要算法包括：Aprior 算法，基于频繁项集的 FP 树，基于粗集的 PC 树以及多种 Apriori 改进算法，综合考虑计算强度及对更新数据的计算，本系统综合采用 FP 树及 PFUP[6]。

将专家的专业领域、年龄段、单位类别、受评单位类别、专项经费渠道、支持经费、项目评价分值(或评价结论)等作为项，并将年龄段、支持经费、评价分值做分段离散化处理，计算其中关联规则的主要步骤如下：

步骤 1 对所有专家信息、评审事项进行预处理，形成规范化评审行为事务库 T。

步骤 2 设定最小支持度 min$S$，最小置信度 min$C$。

步骤 3 在 T 中求解满足最小支持度 min$S$ 的项集(频繁项集)。

步骤 4 生成满足最小置信度 min$C$ 的项集 R，即关联规则备选集。

步骤 5 筛选 R，删除无意义规则形成关联规则集 R'，若删除比率较大，则返回步骤 2。

步骤 6 解释 R'，生成关联规则集。

4 结束语

建立分布式专家行为信息系统包括数据采集、转换、处理、分析等多个环节，涉及多类信息系统、数据平台。

在设计开发过程中本系统着重解决 2 个技术环节：(1)数据资源的安全可信采集，综合考虑加密强度与计算效率，本系统利用 RSA 算法、动态密码实现身份验证、协商对称密钥、交换可信收据，利用对称密码分组传输数据，较好地实现了传输过程、传输行为的安全可信；(2)数据的处理和分析，本系统采用自动处理和人工评价相结合的反馈框架并进行多次循环，基本保障了数据快速处理和各类规则库的优化，并在此基础上利用关联规则算法实现了对行为隐含模式的分析。

下一步工作是进一步优化关联规则算法，提升专家行为信息的挖掘深度。

(下转第 283 页)