

建立 workflow 数据仓库

任国珍 王海洋

(山东大学计算机科学与技术学院 济南250100)

摘要 workflow 管理系统(WfMS)是允许定义、执行、监视和管理商业过程的软件平台。一个 WfMS 记录了在过程执行中发生的每一个事件,因此,工作流的日志包含了相当多的重要信息。通过分析这些信息,可以评估工作流的性能,了解影响 workflow 性能高低的原因,从而对 workflow 进行调整。本文提出一种方案对 workflow 的日志信息进行加工、综合,建立 workflow 数据仓库(WDW),然后根据在 WDW 中进行 OLAP 的结果,调整 WfMS 的运行。

关键词 workflow 日志, workflow 数据仓库,星型模型,抽象维,ECA 规则

How to Build a Workflow Data Warehouse

REN Guo-Zhen WANG Hai-Yang

(School of Computer Science and Technology, Shandong University, Jinan 250061)

Abstract Workflow management systems (WfMSs) are software platforms that allow the definition, execution, monitoring, and management of business processes. WfMSs log every event that occurs during process execution. Therefore, workflow logs include a significant amount of information that can be used to analyze process executions, understand the causes of high- and low-quality process executions, and rate the performance of internal resources and business partners. In this paper we present a solution that how to build WDW and how to load the data from WfMS log to WDW, then we can adjust the WfMS based on the OLAP result in WDW.

Keywords WfMS log, WDW, Star model, Abstract dimension, ECA rule

1 简介

现在,越来越多企业正在使用 workflow 管理系统,旨在提高企业运行的效率,降低生产成本。一般的 WfMS 都会对过程执行中发生的事件记录日志^[1],例如每个活动的开始和完成时间,输入输出数据等等。这些日志信息包含许多用户非常感兴趣的信息,分析这些日志可以使 WfMS 管理员评估 workflow 的性能,并找出造成性能高低的原因。然而,大多数的 WfMS 只提供最基本的日志查询机制,例如显示指定时间段内过程实例完成的数目,以及它们的平均运行时间等。要得到更感性的报告,用户不得不配备商业报告工具写出对日志的查询以得到感兴趣的数据。于是, workflow 数据仓库^[2](WDW)的概念被引入进来。通过对 WfMS 的日志信息的综合、加工,建立 WDW,对 WDW 中的信息进行分析,得出用户感兴趣的信息,从而监测 WfMS 的运行,评估 WfMS 的效率,对 workflow 的运行进行调整,使 WfMS 更好地满足企业的需要。

建立数据仓库时采用何种模型是建立数据仓库首要的问题,数据仓库的结构应该满足多数用户 OLAP 的需要。另外, WfMS 的日志信息根据具体的应用采用了不同的存储格式,有的 WfMS 将日志信息存储在日志文件中,有的存储在关系表格中等等,如何针对不同的存储格式,对日志信息进行加工、整合是建立数据仓库的另一个问题。本文采用星型模型来建立 WDW,并建立了合适的维和数据以便于分析数据;为了能够从不同格式的日志信息中加载数据,本文采用了合适的方法加载数据仓库的数据。作为 OLAP 的一个应用,最后提出了一种基于 OLAP 的主动机制。

2 一个简单的工作流实例

为了方便直观地描述我们的方案,首先介绍一个简单的网上订购旅游的流程的工作流实例,如图1所示。一个顾客首先在网上选择自己满意的旅游线路,选择提交并确认后,顾客可以网上支付费用也可现场支付费用。在缴纳费用后,用户有两种旅游方式选择:自助旅游和跟团旅游。如果顾客是黄金顾客(已经订购旅游超过10次的顾客),顾客可以对不满意的地方提出异议,由旅行社酌情修改行程。旅游的过程中,由导游提交导游日志。旅游结束后,顾客提交旅游意见。旅行社根据导游日志和旅游意见考核导游业绩。

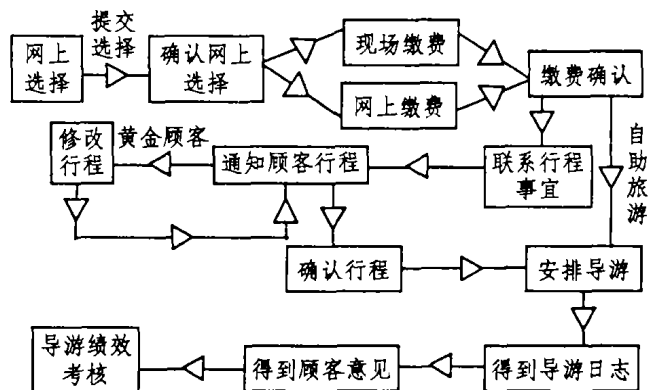


图1

3 建立 WDW

传统的数据仓库有多种模型,其中最为常用的是星型模

型^[3]。在星型模型中,数据分为维和事实。事实是用户感兴趣、需要分析的数据,维是用来分析事实的角度^[3]。星型模型有许多优点,主要优点如下:

- 技术成熟。其基于传统的关系数据库理论,拥有多种成熟的产品作为其底层平台;
- 可以使用多种已有的查询优化技术和查询工具,快速地进行数据分析;
- 拥有众多的优秀的可供选择的报表工具,以便有效地进行多维的分析。

基于这些优点,我们的 WDW 也采用了星型模型。

3.1 WDW 的事实和维

数据仓库的事实是用户感兴趣的数据。在 WfMS 中,用户感兴趣的通常是某种状态的变化,例如 workflow 运行过程中不同流程之间的变化、同一流程中不同节点间的变化等等。所以,我们的数据仓库的事实就是过程和节点状态的变化。为了方便地进行用户自定义的分析过程,我们的方案也允许用户定义自己的事实。

通常按照用户的需求确定 WDW 的事实和维之间的关系,建立数据仓库,以便于优化许多典型分析过程。我们除了定义了一些常用的、必不可少的维之外,还定义了一些抽象的维,用以描述一些复杂的分析数据的视角。

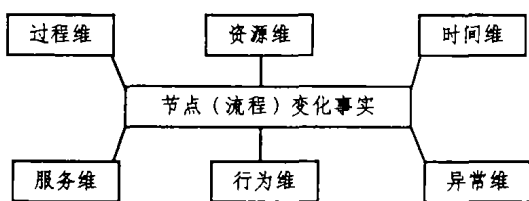


图2 WDW 的模型

我们的模型如图2所示,它描述了以节点或流程的状态变化为事实而建立的一个数据仓库的模型。这个模型包含了对大多数的用户分析来说必不可少的维,如时间维、资源维、过程维和服务维,这些维是大多数 OLAP 都需要的,跟具体的 workflow 模型无关。我们可以在这些维中对节点的变化进行分析。例如可以查看在某个时间段内发生了哪些节点的变化,这些变化使用了哪些资源,分析这些变化的特点和发生这些变化的原因。

在图2中,我们还建立了两个抽象维——行为维^[2]和异常维。这些维能从一个更为抽象的层次来分析数据。

行为是提取出来到达一个更抽象的层次的用户感兴趣的数据,它对应 workflow 中某个满足特定条件的活动。很多情况下分析者不只要分析节点的改变,他们更想分析过程实例以展示特定的行为,并找到发生这种行为的原因。例如上面的流程中,如果一个旅游过程实例在最近30天持续运行,这就是一个用户很感兴趣的行为;又如,有超过80%的黄金用户在一个网上选购流程中经过一个网上缴费的节点,这也是一个用户很关心的行为。

由于 WDW 的关于感兴趣的行为的目的是不可知的,因此我们允许用户自定义行为。行为类型可用与流程或节点有关的布尔表达式来定义,其形式如下:

$$Condition1 \wedge Condition2 \wedge \dots \wedge ConditionN$$

其中 $ConditionN = Function1 \ operator \ Function2$, $operator$ 是比较运算符, $Function$ 是节点或流程属性的集合到节点或流程属性的集合的映射。

异常是 workflow 运行过程中出现的不正常的行为,它是一种特殊的行为。在 workflow 中,异常有多种含义。我们把异常定义为背离优化的或可以接受的流程运行,阻止服务高质量的分配^[4]。当流程进入异常处理流程时,通常用户会非常关心 workflow 的运行,用户很希望了解异常出现的时间、条件、状态、占用的资源等,希望得出异常发生的原因,以便采取措施避免异常的出现。例如一个比较频繁出现的异常一般说明 workflow 模型的设计存在某种问题。异常同行为一样,通过流程执行数据的条件来定义。

通过定义抽象维,用户可以进行更为复杂的数据分析,而定义异常维,则可以使用户专门针对异常进行数据分析,得到有价值的信息。下面是我们的网上旅游 WfMS 中定义的几种典型的行为和异常:

Behavior A: 超过80%的黄金顾客(网上订购旅游超过10次的顾客)经过;

Behavior B: 顾客满意度超过90%;

Behavior C: 低于10%顾客经过;

Exception A: 最近10天撤销(回退)超过10次。

Exception B: 选择好旅游地点方式并提交后却没有缴费

3.3 数据的抽取、转换和加载

通常 workflow 管理系统根据具体的需要记录日志信息,在将日志信息加载到 WDW 之前,首先要对数据进行抽取。一个运行过一段时间的 WfMS 日志信息数量是非常巨大的,不是所有的日志信息都要进入数据仓库,在加载之前,首先需要对日志信息进行必要的抽取,将用户感兴趣的日志信息提取出来。

通常数据仓库的日志信息按照具体的需要采用不同的存储格式,如文本文件存储,关系数据库表格存储,基于 web-services 的 workflow 产品的 XML 格式存储。同时,跨组织的 WfMS 的日志还可能采用异构的存储格式。在加载之前,需要将日志信息按照必要的规则转换成统一规则的数据。

我们使用 WDW 中定义的一组 ETL 脚本加载数据仓库信息,将转换好的日志信息执行一系列的检查、清洗操作,但过程数据的结构不作改变。清洗过的数据被插入到一个缓冲数据库中,在缓冲数据库中做一些更为复杂的操作(跟 WfMS 无关的清洗、检测某些特殊行为和异常等)后,数据被装载进了 WDW,整个装载过程结束。整个过程如图3所示。

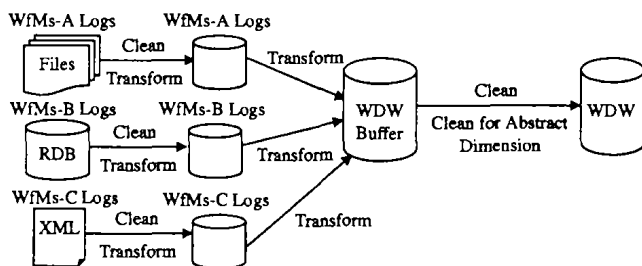


图3

3.4 使用 WDW 进行 OLAP

建立了 WDW 之后,就可以进行各种复杂的 OLAP 了。分析的结果可以用来评估 WfMS 的性能和效率,监测控制 workflow 运行,可以作为 workflow 状态转移的条件,还可以用来进行更为复杂的操作,例如作为主动机制的条件来主动地触发 workflow 的某些动作,这些动作可能是修改调整流程、从 WDW 中抽取知识为将来自定义新的流程作准备。这种分析属于工

作流智能的一部分^[5]。

4 利用 WDW 建立主动机制

主动规则的思想已经在多种环境中得到了成功的应用,包括主动数据库、工作流管理、网络管理及 XML^[6]等。主动规则一般包括三个部分:事件(Event)、条件(Condition)和动作(Action),因此又称为 ECA-规则^[6]。其基本执行模式为:

ON event IF condition DO action

它采用事件触发(Event-triggered)的思想,即当定义的事件发生,如果条件得到满足,即执行定义的活动。在我们的方案中,我们的 ECA 规则针对我们的 WDW 重新做了解释。

事件 很多事件都可以作为主动规则的触发事件,如时间事件、数据操作、异常、抽象事件等。事件既可以定义在 WDW 中,也可以定义在 WfMS 中。当预先制定的事件发生时,检查主动条件,判断是否触发主动动作 Action。

最常用的事件是时间事件,可以定义每隔一段时间检查一下条件。还可以将事件定义在 WfMS 中,例如是一个过程中从一个状态转移到下个状态时。还可以定义在 WDW 中,例如某个用户感兴趣的行为发生时或是预先定义的异常发生时。

条件 即用 OLAP 操作定义的布尔表达式,当某个主动规则中的事件被触发时,主动机制会判断是否满足定义的条件来决定是否执行操作。如果条件缺省,则 ECA-规则退化为 EA-规则,即只要事件发生就执行指定的动作。我们的条件都定义在 WDW 中,形如:

$Condition1 \wedge Condition2 \wedge \dots \wedge ConditionN$

其中 $Condition1 = OLAPResult1 \text{ operator } OLAPResult2$, operator 是比较运算符,OLAPResult 是 OLAP 操作的结果,OLAPResult 也可以是常数。

这样,主动规则的条件是一个以 WDW 中的数据操作的比较为元素的布尔表达式。将 N 个数据操作的结果进行比较后再进行一系列的布尔运算,得到的结果就是条件,当条件满足时执行主动规则定义的动作。即判断一个数据仓库中查询是否符合给定的条件。我们的数据仓库采用经典的雪花模型,所以我们的条件是关系中的布尔表达式。数据仓库中的操作结果是否符合条件最终表示为关系代数的运算。

活动 主动的活动多种多样。活动就是指一定的时间发生后,满足了预先制定的条件后所发生的动作。活动可能是触发某种工作流异常处理,可能是产生某种报表,也可能是将一

定的数据存入 WDW,还可能是仅仅发出一条消息触发外界的某种操作,甚至可以修改我们定义的 ECA 规则。可以制作复杂的或者是第三方的软件来响应这个消息,从而触发复杂的活动,所以活动的定义非常灵活。

如图4所示,用户事先定义好一些主动规则,存放在主动规则库中。首先按照主动规则库初始化主动处理模块。当 WfMS 或 WDW 运行过程中主动规则中定义的事件被触发时,主动处理模块会检查相应的条件是否满足。通过在 WDW 作某些数据操作,检查操作结果是否满足条件。如果满足条件,执行主动规则中定义的动作。

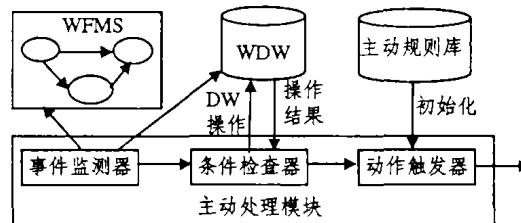


图4

结论 首先提出了 WDW 的概念,然后针对如何建立 WDW 提出了一个初步的模型。在我们的模型中提出了抽象维的概念,抽象维的建立对用户 OLAP 有非常重要的意义。本文还对如何加载异构的复杂的 WfMS 日志信息进行了描述。最后对 WDW 中 OLAP 应用举出了一个具体的例子,利用 OLAP 操作结果作为 WDW 主动机制的条件,建立 WDW 的 ECA 规则。

在今后的工作中,将对如何使用 OLAP 结果进行复杂的操作做进一步的研究。

参考文献

(上接第131页)

织间松散耦合的协作。封装、继承和抽象的特征使工作流能够灵活、敏捷地适应业务的变化。

进一步的工作包括对异构工作流的适应性研究,可适应工作流的长事务的研究,以及可适应工作流的异常处理的研究。

参考文献

- Adams M, Edmond D, ter Hofstede A. The Application of Activity Theory to Dynamic Workflow Adaptation. In: 7th Pacific Asia Conf. on Information Systems, Adelaide, South Australia, July 2003
- Luo Z W, Sheth A. Knowledge Sharing, Coordinated Exception Handling, and Intelligent Problem Solving for Cross-Organizational Business Processes. [PhD Dissertation]. University of Geor-

- Workflow Management Coalition. Workflow Reference Model
- Bonifati A, Casati F, Dayal U, et al. Warehousing Workflow Data-Challenges and Opportunities. In: Proc. of VLDB'01 2001
- Inmon W H. Building the Data Warehouse. John Wiley & Sons, Inc. 1996
- Grigori D, Casati F, Dayal U, et al. Improving Business Process Quality through Exception. In: Proc. of VLDB'01 2001
- Berry M, Linoff G. Mastering Data Mining. Wiley, 2000
- Paton N W, Daz O. Active Database Systems. ACM Computing
- gia, Jan. 2001
- Purvis M, Purvis M, Lemalu S. An Adaptive Distributed Workflow System Framework, University of Otago. <http://www.otago.ac.nz/informationscience/pubs/publications.htm>
- Hollingsworth D. Workflow Management Coalition. The Workflow Reference Model, Document No. TC00-1003, Issue 1. 1, 1995. <http://www.wfmc.org/standards/docs/tc003v11.pdf>
- Kwak M, Han D, Shim J. A Framework Supporting Dynamic Workflow Interoperation and Enterprise Application Integration. In: Proc. of the 35th Hawaii Intl. Conf. on System Sciences, Hawaii, 2002
- Bachmendo B, Unland R. Aspect-Based Workflow Evolution. Department of Mathematics and Computer Science. <http://www.comp.lancs.ac.uk/computing/users/marash/aopws2001/papers/bachmendo.pdf>, 20 July 2001
- Manolescu D A, Johnson R E. Dynamic Object Model and Adaptive Workflow. University of Illinois at Urbana-Champaign. <http://micro-workflow.com/PDF/domaw.pdf>, Oct. 1999