

基于移动 Agent 及 RDF 的网络数据挖掘系统体系结构^{*}

A System Architecture of Data Mining in Network based on Mobile Agent and RDF

吴 跃 王 军 周明天 王湖南

(电子科技大学计算机科学与工程学院 成都610054)

Abstract In this paper, a system architecture of data mining in network-based on the technology of Mobile Agent and RDF is presented. This system architecture provides a standard and uniform structured data-platform for the concrete data-mining application. Mobile Agent implements the concrete data mining application. The relative technologies of this system architecture are introduced first, then two species of logic architecture of the network based application system are explained. And an accomplished prototyping of network-based data mining system architecture is proposed. The combination of the up-to-date XML/RDF and Mobile Agent technologies provides a new method for data mining in network.

Keywords XML, RDF, Mobile Agent, Data mining on network

1 引言

随着数据挖掘的研究与应用日益广泛和深入,基于 Web 的网络数据挖掘正受到人们的巨大关注。面向网络的移动 Agent 理论和技术,可以有效地克服网络环境状态问题,突破网络资源的限制,提升系统分布式计算的能力,并降低分布系统的复杂性。针对 Web 信息的描述和表达而发展起来的 XML/RDF 理论和技术,能够规范化处理异质异构的网络信息,同时,XML/RDF 理论和技术为 Agent 之间及 Agent 同外界的交互和通信提供了一种崭新而简洁的机制,从而也降低了 Agent 系统的复杂性。在基于移动 Agent 及 RDF 技术的网络数据挖掘应用系统中,移动 Agent 是具体挖掘应用的包装器(Wrapper)和承载体,XML/RDF 为具体的挖掘应用提供规范统一的结构化数据平台。两者的结合简化了网络数据挖掘系统的结构,并表现出巨大的发展潜力。

基于这两种技术,我们提出了一种网络数据挖掘系统的体系结构和网络应用系统的一般性逻辑结构,并以数据挖掘应用为例,给出了相应的原型化体系结构。

2 移动 Agent 和 RDF 技术

2.1 移动 Agent 技术

Agent 是在继承 Meta-Agent 基本属性(自治性、反应性、能动性、交互性)和基本框架(如层次结构)的基础上,根据实际的需要增添或嵌入新的属性、弱化或强化基本属性、添加或嵌入特定组件而形成的某类封装体。移动 Agent 就是具备移动性的一类强调自治移动能力的软件 Agent。

相对于静态 Agent 系统和其它基于常规技术的系统而言,移动 Agent 系统在应用中表现出以下明显的技术优势与特点^[4]:

(1)节省通信的开销。移动 Agent 根据需要动态迁移到目标信息源处执行,直接访问该信息源,同时移动 Agent 仅在自己发送和反馈目标任务情况与结果时,才需要网络连接,避免了大量数据的网络传输和长时间占用网络,从而显著节约网络带宽与网时,极大地减弱了对网络的依赖性。

(2)异步计算机自治性。移动 Agent 在目的信息源上自治地计算执行,独立于其创建进程,不需要统一的调度,更不需要与网络保持长期的、稳定的连接。用户在移动 Agent 工作时可离线,并可在不同的时间与空间,从特定的网上位置,用某种方式(如:从自己的邮箱中)取出所需的结果,这对移动设备而言是至关重要的。

(3)并行处理。针对某一任务的需求,可自动地而且适当地分解,并分配给(或克隆出)多个移动 Agents,在相同或不同的时间或地点并行完成任务。

(4)突破了本地资源的局限。移动 Agent 的应用可满足超出本地资源(如:计算机硬、软件系统、存贮系统)处理能力的用户需求。

(5)协议封装性。移动 Agent 可封装不同的已有网络协议或自己特定的协议,在网络上按需建立特定的(或用户化的)传输通道或动态升级网络。

(6)动态适应性。移动 Agent 能自治地对不理想或不良的环境状况与事件作出动态的、适当的反应,同时,通过监听、通告、智能化路由及群体任务动态协调等多种手段产生综合适应性。

(7)降低了分布式系统的复杂性。提供了灵活的功能划分机制,在一定程度上,克服了传统系统粒度过于细化的缺点;传统系统的交互机制过于僵化,移动 Agent 系统代之以灵活的、较简单的交互机制,同时,移动 Agent 系统具备了处理系统“组织结构”的完备机制,为分布式计算提供了一种独特的、灵活的新技术、新构架。

(8)理想的异构集成环境。移动 Agent 一般是独立于计算机和传输层的,为网络上异构的网络计算提供了无缝的系统集成的理想条件。同时,为软件分析、设计与实现提供了一种崭新的思路和较简便的方法。

2.2 XML/RDF 技术

XML 把从 SGML 继承的可扩展性和强大功能与 HTML 易用性相结合,为基于 Web 的应用提供了新的格式平台。XML 用标记对文档中的数据进行描述,标记的含义由 DTD 文件或 XML Schema 定义,以实现互操作。XML 提供了一种

^{*}国家自然科学基金资助项目,基金号:69871005。吴 跃 教授,在职博士生,主要从事数据库系统、数据挖掘技术的研究。

自然的编码形式,把结构化数据组成一个文本文件,各个异质系统间,通过 XML 文档这个统一的中介格式,就可以进行相互的交流和信息传递。XML 作为一种元语言,可以定义新的格式描述语言,如 RDF,OSD 等。

RDF 模型作为描述和交换元数据(Meta-data)的框架,它是处理元数据的基础。RDF 为那些进行 Web 信息交互的应用提供互操作性,强调促进 Web 资源的自动处理。RDF 利用 XML 的名域空间引用一些 RDF Schema 中已定义的资源类型和属性类型等,以避免重复定义和识别冲突问题。不同数据源的数据(XML 文档和非 XML 文档)都可以用 RDF 来描述,RDF 为描述 Web 上的所有资源及其相互关系提供了一个统一的模型,在此基础上,基于 Web 的应用就可以对该 RDF 数据空间进行操作,实现资源的自动处理。XML 和 RDF 是互为所用的。RDF 依靠 XML 支持,RDF 借助 XML 的语法和 XML 的名域空间来实现 RDF 的模型。XML 依靠 RDF 这种应用提供的互操作性实现数据在不同应用间的交换。

XML,特别是 RDF 与 RDF Schema 的出现和发展为 Agent 通信,为各类同构或异构 Agent 之间的交互、Agent 与环境的交互,提供了一种新的机制、新的通信过程,提出了一种新的规范化的内容描述语言标准^[3],提供了新的消息格式和新的规范化的语言格式,以 XML/RDF 来定义或映射 ACL (Agent Communication Language)本体基本结构与原语,直接导致了基于 XML/RDF 的 ACL 的诞生。基于 XML/RDF 的通信交互体制、ACL 及内容描述语言具有普适、简洁、灵活、可靠、表达能力强、便于扩展、便于其它 ACL 和内容描述语言映射的特点。

3 网络应用系统逻辑结构

网络环境是异质异构的、动态变化的和不太可靠的,网络中各类信息源也同样具有异质异构和动态的特点,基于其上的应用系统必然是复杂的。我们在研究 Agent、XML/RDF 理论和技术的基础上,提出了简洁实用的网络应用系统一般性嵌套和层次逻辑结构。

3.1 网络应用系统嵌套逻辑结构

在图1中,Agents 作为应用的包装器和承载体,封装了具体应用及其数据或执行状态,自治地进行相互间的交互和与环境的交互,并依据自身状况和环境信息作出适当的反应,主动地协调相互间的行为、分配资源以完成所肩负的应用任务。Agents 为应用提供恰当的环境、资源和接口并替代应用对外进行交互,可视为具体应用的标准化中间件。每一种应用和 Agent 封装就产生了一种特定的应用 Agent。

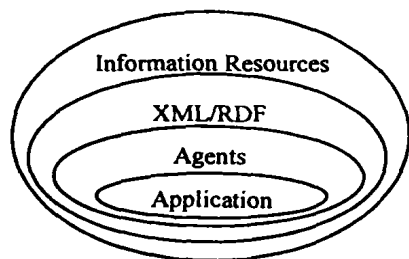


图1 基于 Agent 和 XML/RDF 的网络应用系统嵌套逻辑结构

XML/RDF 在各种异质异构的网络信息源基础上,为具体的应用建立起规范化、结构化的统一数据平台,并可灵活反

映所描述信息源的动态变化,使面向异质异构的多数据源的网络综合应用成为可能。各具体应用只需简单地通过 RDF API 就可在该平台上实现应用目标。同时 XML/RDF 也为各 Agent 之间和 Agent 同外界环境间的通信交互提供了新的平台。

3.2 网络应用系统层次逻辑结构

从逻辑层次上看,基于 Agent 和 XML/RDF 的网络应用系统,可分为信息层、中间件层和应用层,如图2所示。

应用层		Applications
中间件层	Agent 层	包装 Agents
		RDF API Agents
	统一描述层	RDF
		XML
信息层		Information Resources

图2 基于 Agent 和 XML/RDF 的网络应用系统层次逻辑结构

其中,中间件层可分为 Agent 层和统一描述层。Agent 层又可划分为包装 Agents 和 API Agents。包装 Agents 由用来封装具体应用及其数据或执行状态的包装 Agents 所组成;API Agents 指各种“Agent 化”的 RDF API,它们是嵌入 RDF API 功能属性的一类 Agent。RDF API 的“Agent 化”使系统具有了对具体的 RDF API 的相对独立性,具有了灵活的可扩展性和兼容性,即:自治的 RDF API Agents 可克服 RDF API 的“非标准化”问题,可为上层的不同应用提供灵活统一的智能化 API。在统一描述层中,XML 为各种信息源提供了一种统一的数据描述机制,为各个异质信息源提供一种统一的中介格式。RDF 可作为一种规范的消息传递模型,统一的消息格式和明确的语法、语义描述,实现各个应用之间的信息交互。这里,RDF 是用 XML 定义的一种格式描述语言,是一个描述元数据的框架,也是处理元数据的基础。在 RDF/XML 体制下,网上的所有资源和资源间复杂关系可统一用 RDF 模型来描述。RDF 模型提供了一个规范的资源描述框架,该框架的语法由 XML 这种自然的编码语言来表达,相应的语义由 RDF Schema 来定义。总之,在统一描述层中,XML 表达的 RDF 模型,通过 XML 这个网上的统一中介格式,为不同应用提供了一个规范化、结构化而又灵活的关于底层异质异构信息源的统一集成平台;同时,RDF 模型和 XML 提供描述和交换元数据的统一机制与格式,在 RDF Schema 的定义和描述支持下,也为系统中各类 Agent 提供了一种简洁实用的通信机制和交互场所。

4 网络数据挖掘系统的一种体系结构

在网络数据挖掘中最具典型意义的是 Web 数据挖掘,它不仅进行内容挖掘而且进行结构挖掘。目前,传统的数据挖掘算法和工具仅仅能处理单一的、结构化的和规范化的数据,创建融合各种异构数据的挖掘技术是网络数据挖掘的应用需求和技术难点。

XML/RDF 提供了一种明确描述网上各种信息源的统一的、结构化、规范化模型和机制。把网络环境中的对象视为资源,并制定了明确的描述性的语法和语义,既可对资源的数据结构也可对资源数据本身及它们之间的关系进行明确具体而且灵活简洁的描述。使我们可能在一个规范化结构化的统一数据层面上,在研究和开发新的网络挖掘技术的同时运用传

统的挖掘算法和工具对各种目标资源进行特定的多种层次的综合数据挖掘。

针对上述网络应用的一般性逻辑结构,我们提出了一个原型化的网络数据挖掘系统体系结构,如图3所示。

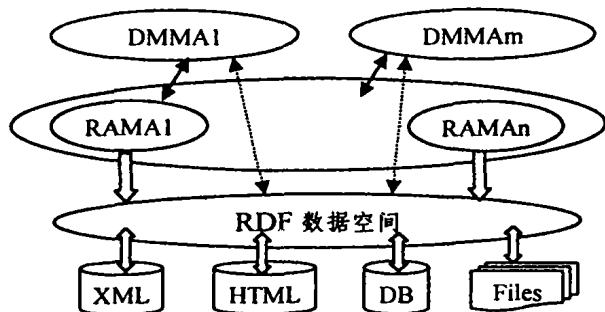


图3 基于 Agent 和 RDF 的网络数据挖掘系统体系

Web 上的各种类型的信息源组成信息层,如:结构化的 XML 文档,半结构化的 HTML,数据库,文本文件等。不同格式的数据源对 RDF 数据空间来说,只有资源一种类型。资源可用其 Web 地址即 URI(Uniform Resource Identifier)标识。

RDF 数据空间对应于统一描述层。RDF 数据空间是由遵循 RDF 模型和 RDF Schema,以及用 XML 来表达的描述资源、属性及其相互间关系的元数据所构成的。RDF 模型描述不同网上资源和资源间的关系,描述元数据的 RDF 模型中:用 URI 表示资源;用 Properties 表示资源的属性;用 Property-value 表示属性的取值。RDF 用 Statements 来表示 Resource, Property 和 Property-value 三者之间的关系,并用主语,谓语和宾语来称呼。Statement 的宾语可以是另一个 Resource 或一个 Literal;其中,Literal 是一个简单字符串或由 XML 定义的原始数据类型,不能做进一步处理;一个 Statement 中的宾语可以是另一个 Statement 的主语或其它 Statement 的宾语,如同 HTML 中的超链接,只不过这种超链接关系是双向和多个源多个目标的。因此,网上的所有资源都可以通过 Statement 对其相互间的关系进行描述而关联起来,这为 Web 资源的自动处理提供了依据。例如下面的元数据:

Ora Lassila is the creator of the resource <http://www.w3.org/Home/Lassila>.

该元数据用 RDF 模型进行描述,如下:

主语(Resource) <http://www.w3.org/Home/Lassila>
谓语(Properties) Creator
宾语(literal) "Ora Lassila"

上述 RDF 模型用 XML 表达如下:

```
<rdf:RDF
  xmlns = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:s="http://description.org/schema/"
  <Description about="http://www.w3.org/Home/Lassila"
    <s:Creator>Ora Lassila</s:Creator>
  </Description>
</rdf:RDF>
```

RDF 模型中 Properties 的声明和相应语义由 RDF Schema 定义。RDF Schema 不仅定义 RDF 模型中资源的各种属性,还定义描述的资源类型,RDF Schema 为 RDF 模型中 Statements 的翻译或解释提供信息,同时为这些数据模型指定应该遵循的系统参数。在上述 RDF 模型的 XML 表达中,RDF 模型中 Properties 的含义由一些标准的 RDF Schema 文档或自定义 Schema 文件来定义,RDF 模型与 RDF Schema 的关联借用 XML 的名域空间来实现,避免重复定义

和识别冲突。

这里,RDF 通过 RDF 模型这种方式对元数据进行描述,表达的语法用 XML 来实现。元数据的该描述信息可以存在于所描述的资源文件中,也可以存在于一个单独文件中,还可以随着 HTTP 协议一起传输。RDF 模型中资源、属性类型和相关语义由 RDF Schema 来定义说明,但是,RDF 并无内建的行为(Actions)概念,由于 Schemas 可依同样语法来相互交换,从而,我们可依特定的需要对 Schemas 进行细致的修改或扩充,使其允许有关行为、行为结果和行为状态的明确编码^[3]。因此,本层之上的 Agents 可以在 RDF Schemas 中定义有关通信交互的新行为以及设置或规约已有行为,而不必预先在彼此间取得一致。遵循 RDF 模型及 XML 所提供的描述机制和语法表达格式,不同 Agent 通过 Schema 文件中对新行为的语义定义或对行为的规约,就可识别该新行为并确知其它 Agents 所提供的行为状态或结果,据自己和环境的状态自治地作出适当的反应,最终完成相互间的协商或协调工作,并协作完成特定数据挖掘应用。

所有这些元数据描述信息所组成的 RDF 数据空间中,由 XML 表达 RDF 模型,提供结构化的数据描述格式,语义由 RDF Schema 说明,使 RDF 数据空间中的元数据能够被机器识别和自动处理,从而为移动 Agent 实现网络数据挖掘提供了统一的数据空间和技术支持。

考虑到网络环境的具体特点,我们采用了移动 Agent 来作为具体的数据挖掘应用的封装器,得到了 DMMA(Date Mining Mobile Agent);用移动 Agent 来“Agent 化”RDF API,得到了 RAMA(RDF API Mobile Agent),以适应 RDF API 尚未统一规格化的现状,简化 RDF API 的构建,使系统具有对特定 RDF API 的相对独立性、兼容性和可扩展性。它们共同构成了移动代理层。不同的数据挖掘应用产生不同的 DMMA_i($i=1,2,\dots,m$),不同的 RDF API 产生不同的 RAMA_j($j=1,2,\dots,n$)。DMMA_i 与 RAMA_j 之间的匹配由各移动 Agents 主动灵活的自治交互所完成(也可以强制匹配)。而各移动 Agents 的交互在 XML/RDF 所提供的统一的灵活简洁的通信机制(图3中虚线所示)支撑下自治地完成。DMMA 所携带的具体的数据挖掘应用必须通过 RAMA 和 RDF 数据空间的协作来完成。因此,DMMA 往往是和相匹配的 RAMA 捆绑在一起在网络结点间迁移以完成特定的挖掘任务。

结论 把 Agent 和 XML/RDF 理论与技术引入到网络数据挖掘中,可解决常规分布式计算系统所难以克服的一系列问题。本文提出的基于 Agent 和 XML/RDF 技术的网络数据挖掘的体系结构,在一定程度上可以解决网络环境的异构性、动态性、连接的不可靠性、资源的局限性,以及异构信息源等问题。

参考文献

- 1 W3C. Extensible Markup Language (XML), <http://www.w3.org/xml/>
- 2 W3C Resource Description Framework (RDF) Schema Specification 1.0, <http://www.w3.org/TR/2000/CR-rdf-schema-20000327>, March 27, 2000
- 3 Bauwens B. XML-based Agent Communication: VPN Provisioning as a Case Study. <http://www.infoloom.com/gcaconfs/WEB/granada99/www.alcatel.com>, May 1999
- 4 Wu Y(吴跃), Tian Y.-H, Zhou W.-L. The Design of a Mobile Decision Support System. In: 2000 4th Intl. Conf. on Algorithms and Architectures for Parallel Processing, Hong Kong, Dec. 2000
- 5 Bray T. RDF and Meta data. <http://www.w3.org/RDF/>, June 9, 1998
- 6 Glushki R J, Tenenbaum J M, Meltzer B. A XML-Framework for Agent-based E-commerce. Communications of the ACM, 1999, 42 (3)