学术论文

主动虚拟服务器和网上数据集成的新模式

何新贵[1]，杨良怀[2]，唐世渭[3]，杨冬青[2]，陈立军[2]，张　震[2]，林　斌[2]
(1. 北京系统工程研究所，北京　100101；　2. 北京大学 计算机科学与技术系，北京　100871；
3. 北京大学 视觉与听觉信息处理国家重点实验室，北京　100871)

[摘要]　针对 Internet 上信息搜索和数据集成存在的问题，提出了“主动虚拟服务器”的概念，并给出了基于这种虚拟服务器的数据集成模式。主动虚拟服务器可认为是网上海量数据统一的存储和处理器，而且具有主动服务的功能。在不改变服务器中原 DBMS 的前提下，只需用一种统一的可扩展语言 UXL 为各种服务器做一件外观统一的“主动服务外套”，就可构成 所需的“主动虚拟服务器”。上述“外套”对每种服务器只需实现一次，就能在整个网上统一使用，因此，它可在网上起到与 Java 虚拟机类似的作用。

[关键词]　主动虚拟服务器；数据集成；可扩展的标记语言 (XML)；统一的可扩展语言 (UXL)；信息代理

1　引言

随着 Internet 上站点的迅猛增长，各种在线数据源（信息源）不断涌现，网上信息发生了爆炸性增长。这些信息源各式各样，包括各种异构的数据库、对象存储 (object stores)、知识库、数字图书馆、电子邮箱等等，其中包含着大量宝贵的数据和信息。Internet 给人们带来便利的同时，也存在一个很大的问题，那就是面对这么多的数据源，人们发现越来越难找到自己所需的数据。用户纷纷抱怨他们不是经常在数据的海洋中迷失方向，就是得到许多无用的数据垃圾，Internet 的不断增长似乎反而使人感到越来越不好用。这种困境的解决已经迫在眉睫。

分析导致这种困境的原因，有多方面的因素。**a.** Internet 中日夜都在爆炸式地产生数据，目前已达到了海量，而且由于 Internet 的高度自治性，网上存在许多互相矛盾的数据和信息，要合理高效地处理和运用它们存在着客观上的困难。**b.** 非结构化、半结构化与传统的结构化数据并存，各数据源数据表示各异，数据互操作性很差，用户很难用统一的方法和工具来处理它们。**c.** Web 站点在提供 HTML 页面的同时，其实将许多数据隐藏在表单的背后，文献 [1] 指出有将近 80 % 的数据属于这种性质。这种情形加剧了信息检索和发现的困难。**d.** 现在各种数据源大都是“被动的”，不能主动为用户提供服务，再加上一些搜索引擎的质量较差，更导致大量数据垃圾的产生。

一个理想的信息检索和数据集成系统应能够对来自不同数据源的结构化的、半结构化的、或非结构化的数据进行统一处理、过滤、缩减、抽象、合并和归纳等工作，具有一定的智能；对于异构、分布的数据源，还必须解决信息表示与结构上的不匹配问题，对现有的 Internet 数据表示、交换和服务机制进行适当规范，并提供主动服务机制。Java 虚拟机已对 Internet 作了第一次“规范”，它使得 Java 小应用程序 (Applet) 得以在任意平台上运行，实现一处编程处处可用。类似地，现在已经很有必要对网上的数据源及其服务作第二次“规范”，这就是本文试图探讨的一个主要问题。

Web 已成为一个巨大的数据库，但到目前为止 Web 数据库的工作并不太多。一般，数据库仅

[收稿日期]　2000－11－08；修回日期　2001－2－16
[基金项目]　“九七三”国家重点基础研究发展规划资助项目 (G1999032705)
[作者简介]　何新贵 (1938－)，男，浙江浦江县人，北京系统工程研究所研究员，博士生导师

仅被当成 Web 环境中的外围角色，而没有把它当作构成 Web 结构整体的一部分。未来 Web 的内容应是动态的，而不再是静态的页面，服务应是主动的，而不再是用户提出服务时才得到服务。用户应能在网上提出问题，定义需求，并主动地得到个性化的服务。

本文将提出异构数据源的统一包装、主动虚拟服务器、信息代理等概念，为 Internet 上的信息检索和数据集成提供一种新模式。主动虚拟服务器可认为是网上海量数据统一的存储和处理器，而且具有主动服务的功能。在不改变服务器中原 DBMS 的前提下，只需用一种统一的可扩展语言 UXL (Unified eXtendable Language) 为各种服务器做一件外观统一的“主动服务外套”，就可构成所需的“主动虚拟服务器”。上述“外套”对每种服务器只需实现一次，就能在整个网上统一使用，因此，它可在网上起到与 Java 虚拟机类似的作用。

2 异构数据源的包装和虚拟服务器

从 Internet 上可以获得的信息源和信息服务正在爆炸式地增长。为了找到相关的信息，用户通常不得不手工浏览或查询各种各样的信息库、抽取相关的数据，然后融合成可用的形式。这是一件十分繁琐的任务，显然应当有一种更好的方法来完成这种任务，那就是“数据集成”。数据集成系统能根据用户定义的条件帮助用户收集信息，找到合适的方案。目前已有不少软件原型和设计，例如，元搜索引擎 (Metacrawler*, Google[2], Gloss[3], Savvy Search[4]) 和 Mediator (MIX[5], Garlic[6], TSIMMIS[7], Infomaster[8], Infosleuth[9], YAT[10])。它们在不同程度上解决了数据集成所要完成的部分任务。

构造这些软件代理时，存在两个关键问题：可扩展性和灵活性。Internet 是一个开放的、快速变化的环境。信息源、Internet 连接、信息代理本身都可能动态出现、变动或消失，各个站点的运行高度自主，任何改变都无需事先通知。因此软件代理在这样的环境中运行必须能够适应这种动态变化。类似地，为了满足用户新需求，为用户创建信息代理的技术必须支持自定义和演化。但是，目前实现的一些机制不能完全适应这种需求。为了集成网上数据，它们往往要求用户进行手工编程，并选定数据源进行检索、集成。应该说，这种工作方式在多变的 Internet 环境中是很脆弱的。

较好的方法是各个数据源以周知的格式发布数据，同时主动地提供相应的服务。近年来提出的 XML (可扩展的标记语言)** 可用来表示数据，并作为 Internet 上数据交换的格式，这为数据集成提供了便利。XML 是标准通用标记语言 SGML 的一个子集，具有比 HTML 强得多的功能。设计 XML 的目标是既要使它具有 SGML 的大部分功能，而且又要具有相对的简单性。客观上，XML 现已成为 Internet 上数据表示与交换的一个新标准。

首先，为了描述上的方便和直观，为一些终端用户提供更易用的语言，本文将用 XML 作为元语言来定义一种统一语言 UXL。其中，输入/输出的描述表示成 DTD，输入/输出数据表示成 XML 元素。由于 XML 本身的可扩展性，UXL 具有足够的灵活性用来描述各种数据源、Web 服务、数据库、乃至 Java 远程对象等。

UXL 通过定义一套标记来实现。这些标记的具体语义由专门设计的 UXL 解释器来解释，如图 1 所示。UXL 可针对用户查询需求、数据源包装、用户自定义过程等定义一组专门的标记，使用户可以直接使用。此外，UXL 还可进一步对信息代理的任务和功能等进行定义。

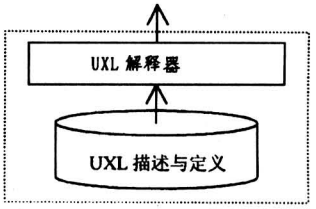

图 1 UXL 解释器

Fig.1 UXL interpreter

UXL 中的基本预定义标记包括：

·用户查询请求标记 该标记用来描述用户的查询请求。

·信息服务订购标记 通过这种标记用户可以向信息代理或中介系统订购自己想要的信息。

·用户自定义过程标记 当用户需要某些自己特殊要求的功能时，该标记可为用户提供自定义所需过程的能力。

---

* hhtp: //www. metacrawler. com

** http: //www. w3. org/TR/REC-xml

·控制流标记　为了定义并实现数据源的包装器及其代理的复杂功能，仅有描述还不够，还必须引入定义控制流的标记[11]。例如引入条件判断标记、循环标记以及变量标记等。

·外部过程调用　许多数据源不仅提供数据，而且提供访问数据的多种服务。但是，各个站点间的接口各异，语法、语义也不尽相同，要使用这些服务是一件十分烦琐的事。为此，在 UXL 中引入了外部过程调用，这将给数据源的包装和使用带来许多便利。为了描述数据源包装器的方便和其它用途，还可定义其它的标记。

·数据源内容的描述　采用 DSCD.DTD 文件描述。

通过以上描述和标记，可对各种异构数据源进行包装，定义数据源存取的抽象接口。这种定义不依赖于任何系统，因此可以构造包装器的分析器，对定义加以分析，并生成具体的包装器。包装器的框架如图 2 所示。

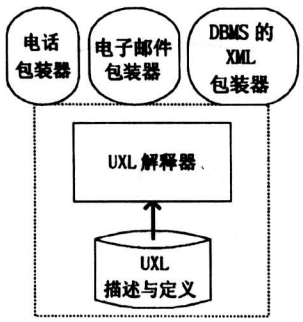

图 2　包装器的框架

Fig.2　Framework of wrapper

在这个框架中，数据源的包装（描述和定义）就是要为数据源外部的可观察行为进行描述，即描述和定义数据源所能接受的查询以及它将交付给用户的数据表示形式。该规范完全独立于数据源内部的行为。可在一个服务器上注册，而信息代理或中介器则可从注册的服务器上选择需要集成的服务，定义需要集成的逻辑描述。经定义后，信息代理就可接受复杂的查询，将查询分解为子查询，并将子查询返回的结果集成后交给用户。

因为各个数据源可用包装器来完成由原有数据，如 HTML 文档、DBMS 数据，到 UXL 的转换。对于它所能提供的服务也用统一的机制加以描述，从而使旧服务的删除，新服务的增加也变得很容易。

各种异构数据源（服务器）经 UXL 包装器的包装之后，就构成外观统一的“虚拟服务器”，它们能以一种统一的“面貌”出现在各种用户面前，提供所需的服务。很显然，上述包装工作对每种服务器只需实现一次。即每一类服务器只需实现一个包装器，就可实现一处编程，处处可用，产生类似 Java 虚拟机的效用。

3　主动的虚拟服务器

迄今，网上服务器的工作模式应该说绝大多数是“被动”服务的，即用户端需要服务器做何种服务就选定某个服务器被动地为它做相应的服务。被动的服务器丝毫不会根据具体情况自主地做些工作。其实，主动数据库的思想给我们不少启迪，提出能实现主动服务的“主动服务器”概念。主动服务器仿照主动数据库的工作模式，它能根据由用户设计的一个规则库，当发现某个事件发生，且某给定的条件满足时，就自主地激发执行规则库中指定的相应动作，进行主动服务。

一般实现主动服务，可在主动服务器中设一个“事件监视器”专门用来监视事件的发生，并根据规则库激发相应的主动服务。此外，采用触发器（trigger）也是实现主动服务的一个方法。事件监视器和触发器的语义及细节可参看相应文献［12～14］。但是，在此应该指出，由于上述的“事件”可表示十分广泛的含义，包括数据库中某个状态或数据的改变、某数据源或公共网站公告牌（事件队列）上发布的一个“事件”等等，而且主动服务器中的规则库可根据需求来定义和更改，因此主动服务器可以提供的服务是多种多样的，而且是动态可变的。仍然可以采用 UXL 来定义各种主动服务需求，可用 UXL 对主动服务进行编程，事件、条件和动作都可用 UXL 定义，因而规则库也就可用 UXL 来编制和定义。因此，在不改变服务器中原 DBMS 的前提下，用 UXL 给各种服务器做一件外观统一的“主动服务外套”，并与 UXL 虚拟服务器集成到一起，就可构成一种“主动的虚拟服务器”。可以指望，它将与 Java 虚拟机具有类似的作用，因为从外观上看，服务器中数据的表示、数据库操作、主动规则库的定义和编制都采用了 UXL，用户可以采用一种统一的方式来接受虚拟服务器的主动服务。此外，也很显然，上述“外套”对每种

服务器也只需实现一次，就能在整个网上统一使用。

下面将把这种主动虚拟服务器应用于信息检索和数据集成，提出一种网上数据集成的新模式。

4 信息搜索和数据集成模式

基于上述主动虚拟服务器的信息搜索和数据集成模式可用图3表示。

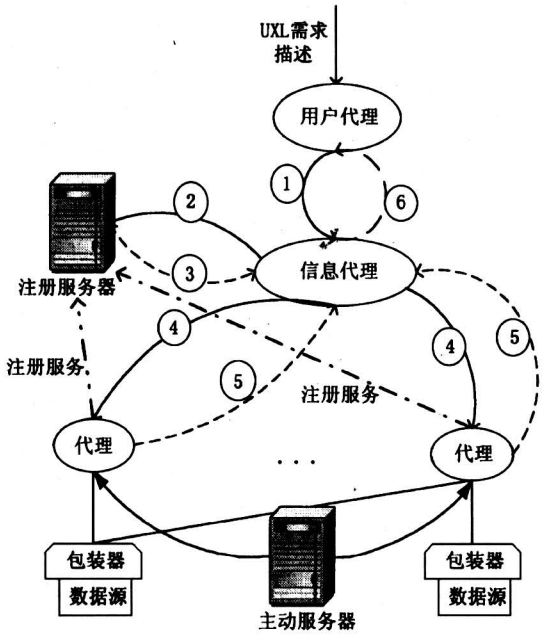

图3 数据检索或集成模式

Fig.3 Mechanism for information retrieval or integration

其中注册服务器用来管理各个数据源代理所能提供服务的注册登记工作，并完成服务查找和匹配等任务。注册服务器还为信息代理提供服务，管理各种元数据，对各个数据源代理进行分类。用户代理接受UXL用户需求描述，将其进行初步处理并转换成合适的请求；信息代理对用户请求进行查询规划，确定信息和服务的搜索范围，并把分解后的子任务转发给各个数据源代理去完成，最后，信息代理把各数据源代理返回的结果进行融合后返回给用户代理。

信息搜索和数据集成的流程如下：

步1 用户的需求描述连同用户标识一起交给用户代理；由用户代理对需求进行必要的预处理和粗略分类，选择合适的信息代理，并将用户请求交给信息代理（数据集成/检索代理）；该步骤对应于图3①。

步2 信息代理对用户请求进行分析后，请求注册服务器（对应于图3②）根据各已注册数据源代理的服务功能选择合适的代理，进行任务分配，然后注册服务器返回选中的数据源代理的标识给信息代理（对应于图3③），由信息代理完成查询分解与优化，并把子查询交由各选中的数据源代理去执行（对应于图3④）。另一种方案是：不采用注册服务器，而由信息代理自主确定粗略搜索范围（通过分类/聚类，搜索引擎或查询引擎粗查等手段确定一个数据源的集合）；信息代理根据粗略搜索范围所指分发需求给数据源代理，“挂牌”应征（即制造一个事件放到相应数据源代理的事件队列中）；在“牌”上包括需求、信息代理的标识和邮箱地址等信息；信息代理为用户代理对该查询需求建立存放结果的邮箱。

步3 实现主动服务：在主动虚拟服务器端（数据源代理处）用一个优先级适当的进程来完成主动服务。该进程根据事件队列依次检查主动服务规则库，开启相应的邮箱，按照其中描述的需求获得（或/和处理）所需数据或信息，并向相应的信息代理（有偿或无偿）提供，即将所获数据放入信息代理存放结果的相应邮箱，然后“摘牌（即从事件队列中去掉已被处理的事件）”；这种服务允许多层嵌套。

步4 信息代理将相应结果邮箱中来自不同数据源的数据（对应于图3⑤）进行融合和集成，将集成（或检索）的数据根据用户代理标识分发给相应的用户信箱（对应于图3⑥），关闭相应的结果邮箱。

可见，上述处理可分为三层：第一层由步1构成，由浏览器或用户代理完成；第二层由步2和步4构成，由信息代理和注册服务器完成；第三层由步3构成，由主动虚拟服务器（数据源代理）完成；在网上统一使用UXL作为数据表示和数据操作的工具；这样，在继Java之后实现了Internet应用的又一次规范化。

5 信息代理

尼古拉斯·尼格罗庞帝（Nicholas Negroponte）——MIT媒体实验室的头领声称web每50天信息翻番。面对如此巨大的信息资源的确急需合适的工具来对其加以管理。能否创建一个简单的接

口对海量信息进行搜索与检索，就像使用电话或阅读报纸那样方便？上述主动虚拟服务器的概念正是适应了这种需要。在上述信息搜索和数据集成的过程中，信息代理是一个关键。信息代理的任务是对各分布的数据源进行管理、操纵和集成。要做到这一点，信息代理必须具备这样的知识：去哪里找信息，如何找到它们，以及如何集成这些信息。其实，信息代理还应具有社会性。一个代理可以选择合作伙伴来共同完成一个特定任务，这个过程称为选择服务[15]。在 Internet 上，不仅数据、信息会形成市场，服务也将形成一个市场。信息代理在这个市场上必须能够竞价、协商、达成交易。这是很值得进一步研究的论题。

在文献 [16] 中提出信息商务概念，例如消费者可向数据服务器订阅金融信息、天气信息、文档、报告以及消费者感兴趣的大量其他信息。信息商务也可用上述主动虚拟服务器和信息代理来实现。

信息代理的业务可由信息代理商设立信息代理站点来进行，也可作为 Internet 上的“公益服务”。用户向信息代理商交付服务费和有偿数据费；代理商再向服务器站点转交有偿数据费。结果数据和费用的交换可采用电子商务方式进行。

此外，信息代理站点可按行业分类或学科分类，信息代理商还可不断积累有用数据，建库（但需经常买进信息，不断更新库的内容）服务，进而实施数据挖掘、发现知识、开展更高级的服务。

6　结语

本文提出了以主动虚拟服务器为中心的信息搜索和数据集成的体系结构，用 UXL 来统一描述数据源外部的可观察行为，包括数据源所能进行的查询或处理，以及用户看得见的数据表示形式等。其中必需解决的关键技术包括：

1）检索范围的粗略分类和划分以及每个站点页面内容都说明本站点的服务范围，这可化为求一个粗糙集的上近似集或计算模糊语义距离的问题[17]。

2）信息代理站点可能数据堵塞，信息代理站点多元化并按行业或学科分类，问题会有所缓和。

3）应该为终端用户提供一种更方便的“需求描述语言”，以便用户更加方便地描述其需求，并将结果数据转换为终端用户熟悉的描述语言。

4）主动服务技术在主动数据库领域已做了很多研究，但如何达到高效服务仍是一个问题。

5）从 XML 到 UXL 的扩充必需认真进行，一方面应使用户使用 UXL 十分方便，另一方面，UXL 还必需得到国际上的认同，最终成为国际标准。

6）更远一点的目标是实现 UXL 与 Java 的融合，实现一种将 Java 虚拟机和主动虚拟服务器功能综合到一起的虚拟机。

上述数据集成模式的优点：

1）主动服务调动众多服务器站点的“积极性”，而且是并行服务，效率比串行的搜索查询高。

2）UXL 一处编程，处处可用。异构站点服务器间的互操作问题得到较好解决。

3）对不同服务器 DBMS 包装从“数据集成系统”搬到各服务器上，对每种服务器只需实现一次，就可被用户统一使用，便于网上的数据集成。

4）DBMS 的 UXL 包装和主动服务进程一起，使原来的服务器变成一个基于 UXL 的虚拟主动服务器，这是类似 Java 虚拟机的又一种虚拟站点概念。如把 UXL 与 Java 联系（统一）起来，形成一种更一般的虚拟站点概念。这种虚拟站点可能为 Internet 的发展起到积极的促进作用。

5）提倡数据持有者主动服务比由人家来搜索查询应该容易得多，提供的数据也可能准确得多，全面得多。从而既可避免大量数据垃圾的产生，提高查准率，又有可能达到较高的查全率。

6）如果网上的费用支付问题得以解决，各种信息代理商就可能蓬勃发展，乃至形成一个信息代理产业。

对于更进一步的工作，首先根据实际需求具体定义一个统一的可扩展语言 UXL。向网上终端用户或应用提供“主动”或“推式”服务是当前 Internet 上信息服务的一个热点论题。如何实现理想的事件服务设施还需要做进一步的研究，对 CORBA 等必须克服其事件选择能力有限等缺陷。这些都是很值得进一步研究的论题。

参考文献

[1] Lawrence S, Giles C L. Searching the world wide web [J]. Science, 1998, 280 (4): 98～100

[2] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine [A]. Proceedings of the Seventh World Wide Web Conference [C], April 1998

[3] Gravano L, Garcia-Molina H. Generalizing GLOSS to vector-space databases and broker hierarchies [R]. VLDB Conference, 1995. 78～89

[4] Howe A, Dreilinger D. SavvySearch:A meta-search engine that learns which search engines to query [J]. AI Magazine, Summer 1997,18 (2): 19～25

[5] Baru C, Gupta A, Ludaescher B, et al. XML-based information mediation with MIX [Z]. Exhibitions Program of ACM SIGMOD 99,1999. 579～599

[6] Cody W F, Haas L M, Niblack W, et al. Querying multimedia data from multimedia repositories by content: the garlic project [A]. Proceedings of Visual Database Systems (VDB－3) [C], 1995

[7] Garcia-Molina H, Papakonstantinou Y, Quass D, et al. The TSIMMIS project: Integration of heterogenous information sources [A]. Proceedings 10th Meeting of the Information Processing Society of Japan [C], 1994

[8] Duschka O M, Genesereth M R. Query planning in infomaster [A]. Proceedings of the ACM Symposium on Applied Computing [C], San Jose, CA, 1997

[9] Woelk D, Bohrer B, Jacobs N, et al. Carnot and infosleuth: Database technology and the world wide web [A]. Proc of ACM SIGMOD Conf on Management of Data[C], San Jose, CA, 1995. 443～444

[10] Cluet S, Delobel C, Simeon J, et al. Your mediators need data conversion! [A]. Proceedings of International Conference on Management of Data, ACM-SIGMOD [C], 1998. 177～188

[11] Lange D B, Hill T, Oshima M. A new internet agent scripting language using XML [R]. AAAI-99 Workshop on AI in Electronic Commerce, July 1999

[12] 何新贵. 特种数据库技术 [M]. 北京:清华大学出版社, 2000

[13] 何新贵. 事件代数和主动知识库系统 [J]. 软件学报, 1994, 5 (9): 23～29

[14] 何新贵. 具有主动功能的程序设计语言及其实现技术 [J]. 计算机学报, 1996, 19 (3): 221～229

[15] Weinstein P, Birmingham W P. Runtime classification of agent services [A]. Proceedings of the AAAI-97 Spring Symposium on Ontological Engineering [C], Stanford, Palo Alto, CA, March 1997

[16] Bernstein D, Sibert O, van Wie D, et al. Information commerce: launching online content [A]. Proc International Online Information Meeting Proceedings [C], Perugia, Italy, May 1995

[17] 何新贵. 模糊知识处理的理论与技术 [M]. 第2版. 北京:国防工业出版社, 1998

Active Virtual Server and a Novel Mechanism for Data Integration over Internet

He Xingui[1], Yang Lianghuai[2], Tang Shiwei[3], Yang Dongqing[2], Chen Lijun[2], Zhang Zhen[2], Lin Bin[2]

(1. *Beijing Institute of System Engineering*, *Beijing* 100101, *China*;
2. *Dept. of Computer Science and Technology*, *Peking University*, *Beijing* 100871, *China*;
3. *National Key Laboratory on Machine Perception*, *Peking University*, *Beijing* 100871, *China*)

[**Abstract**] Aiming at solving the problems of information retrieval and data integration over Internet, the idea of "active virtual server" is proposed, and a novel mechanism for data integration based on the virtual server is presented. Active virtual servers can be regarded as the uniform storage and processor of massive data over Internet, with the function of active services. By only using a unified extendable language (UXL) to wrap each kind of server with a uniform interface "active service coat", the "active virtual server" can be constructed as needed without changing the original DBMS on the server. The above "coat" needs to be done only once for each kind of server, and can be exploited uniformly all over the Internet, hence it plays the role as Java Virtual Machine does.

[**Key words**] active virtual server; data integration; extendable markup language (XML); unified extendable (UXL); information agent