基于移动代理的层次优化挖掘模型

李成安[1]，吴铁军[2]

(1. 华南理工大学电子商务学院 广州 510006; 2. 浙江大学智能系统与决策研究所 杭州 310027)

【摘要】对于大规模分布式数据挖掘问题，提出一种基于移动代理的层次结构挖掘模型，该模型对OIKI DDM模型进行扩展，利用层次设计思想，基于移动代理和增量优化技术进行挖掘和增量集成。实验结果表明该模型对于数据站点大小具有更好的伸缩性，实现更加灵活，可根据网络特点有效降低通讯代价，特别适合于大规模分布式环境。

关 键 词 分布式挖掘; 层次模型; 增量集成; 移动代理

中图分类号 TP18 文献标识码 A

A New Hierarchical Optimization Model Based on Mobile Agent for Distributed Data Mining

LI Cheng-an[1]，WU Tie-jun[2]

(1. School of E-Business, South China University of Technology Guangzhou 510006;

2. Institute of Intelligent Systems & Decision Making, Zhejiang University Hangzhou 310027)

Abstract For large-scale distributed data sets, a new hierarchical mining model based on mobile agent is proposed to perform distributed mining tasks. Based on hierarchical idea, the proposed approach integrates multiple local results using mobile agent and incremental optimization. The experimental results demonstrate that the proposed approach is scalable and particularly suited to large-scale distributed environments. In addition, the proposed model can reduce dramatically communication cost based on network characteristics.

Key words distributed mining; hierarchical model; incremental integration; mobile agent

随着数据库和网络技术的发展，数据量成指数级增长，且数据物理地或地理地分布在不同的地点。传统的单机挖掘技术已经不能适应分布式环境下的挖掘需要，设计有效的分布式数据挖掘(Distributed Data Mining，DDM)算法及其系统是目前数据挖掘的研究热点之一[1]，其中算法伸缩性和网络通信负荷是分布式数据挖掘中非常重要的问题。

基于代理的模型(Agent-Based Model)是目前最流行的DDM模型[2]，如BODHI，JAM[3]等。该模型的设计思想是，在每个数据站点使用一个或多个代理，这些代理负责分析局部数据以及与其他代理通讯。其优点在于灵活性，但是大多数基于代理的挖掘模型仍需要将所有局部结果集中到一个中心站点，当站点数据庞大时，将造成很大的通讯负荷，同时中心站点的内存也不一定能满足要求。

文献[4]提出一种基于移动代理技术的优化增量知识集成分布式数据挖掘(Optimized Incremental Knowledge Integration，OIKI) DDM模型，根据站点结果集的大小优化地增量集成站点间的局部结果，对于站点数量具有良好的伸缩性。但是当站点数量非常庞大时，串行集成、响应效率比较低下，执行时间很长；此外，客户端需要与每个数据服务器通讯，通讯量比较大。由于该模型仅考虑了局部结果集的大小，而没有考虑站点的分布性，所以对于一些非常大的数据库如天文数据库，站点数量可能非常庞大，且站点一般并不是随机地分布在各个地方，往往集中分布在若干地点，具有局部集群性。

本文提出一种新的DDM模型，该模型基于移动代理和分治策略，利用层次思想，根据数据站点的网络分布等特性，先将系统分为若干子系统，对每个子系统，利用OIKI DDM模型进行学习，每个子系统生成一个局部结果。然后利用局部结果和OIKI DDM模型生成最后的全局结果。该模型具有更好的伸缩性，执行效率更高，并能有效降低网络通讯。

收稿日期：2006 – 06 – 05

作者简介：李成安(1969 –)，男，博士，工程师，主要从事分布式计算、数据挖掘、多智能体系统等方面的研究.

1 OIKI DDM模型

OIKI DDM模型[4]是一种基于移动代理的分布式挖掘模型，该模型利用优化和增量学习思想，根据局部结果集的大小，优化增量地集成两个站点间的结果，对任意n个的数据源都具有伸缩性。在该模型中使用两个移动代理：(1) 用移动代理数据挖掘器(MADMs)在局部站点执行数据挖掘任务；(2) 用移动代理知识集成器(MAKIs)增量地组合每两个站点的挖掘结果，使小的站点结果R_j迁移到更大的站点结果R_i，其中$1 \le i, j \le n, i \ne j$。该模型如图1所示。OIKI DDM模型分为三个执行阶段。

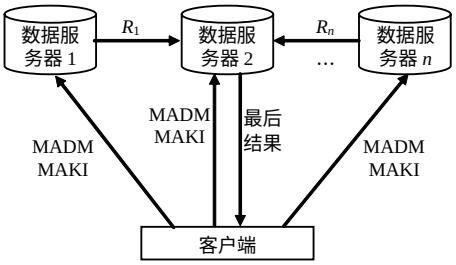

图1 OIKI DDM模型

1.1 准备阶段

客户端广播MADMs和MAKIs给需要执行挖掘任务的数据服务器i，且$1 \le i \le n$。

1.2 数据挖掘阶段

数据挖掘任务在每个数据服务器上局部执行，生成该站点的挖掘结果。

1.3 知识集成阶段

一种优化增量集成技术在数据服务器上执行，其中小的结果被迁移到更大的结果中以优化数据服务器间的结果迁移代价。

该模型对任意数量的数据站点具有伸缩性。然而，由于客户端可能需要与所有数据服务器通讯，客户端将挖掘代理和集成代理分发给所有数据服务器。此外，所有服务器也需要将结果集大小传递给客户端，当站点数量非常庞大时，通讯开销也非常大，其串行集成模式也将使响应效率可能大大降低。

2 层次优化增量知识集成模型

在许多实际应用中，大量的数据站点集中地分布在若干个物理地点，同时，站点间的连接方式也不同，分布在同一地点的站点用高速局域网连接，网络传输和通讯非常方便，通讯代价非常低。而分布在不同地点间的站点用远程网、互联网甚至无线网连接，相互之间的通讯代价比局域网大得多，但网络传输的安全性较局域网也要低得多。由于它们之间的计算和通讯代价是不同的，因此尽量减少广域网之间的网络传输，加强局域网之间的通讯和计算，安全性提高，可有效提高整体系统的性能。

本文将分布在一个地点的所有站点看成一个整体，按照一定准则将整个系统分解为若干个子系统，如将每个地点的站点或若干相近地点的站点组成一个子系统。利用一种分治策略，将一个复杂大系统分解成若干易于求解的小系统，然后对每个小系统独立进行求解，最后合成局部结果得到整个系统的求解结果。

例如对S个站点组成的分布式大系统，可形式化地假设所有站点分布在g个地点，其中第i个地点的站点数为n_i，则$\sum_{i=1}^{g} n_i = S$，其中$1 \le n_i \le S$；$1 \le g \le S$。将每个地点的站点组成一个子系统，因此该系统被分为g个子系统。当然，也可根据实际需要将一个地点的站点分为多个子系统，或多个地点的站点组成一个子系统。

对于由g个子系统组成的大系统，利用层次结构对系统中的数据进行挖掘，得到一种新的挖掘模型，即层次优化增量知识集成(Hierarchical Optimized Incremental Knowledge Integration，HOIKI)模型，其基本思想是：(1) 在第一层，根据每个子系统i的所有数据站点，利用OIKI DDM模型生成一个子系统的优化结果R_i。(2) 在第二层，对第一层生成的结果，再利用OIKI DDM模型生成整个系统的优化结果。

HOIKI模型过程分为四个阶段：

(1) 准备阶段。客户端将MADMs和MAKIs广播给每个子系统。(2) 数据挖掘阶段。每个数据服务器站点执行挖掘任务，生成局部挖掘结果。(3) 子系统知识集成阶段。MAKI利用优化技术在子系统进行结果集成，生成子系统的挖掘结果。(4) 全局知识集成阶段。对各子系统生成的挖掘结果，MAKI利用优化思想进行集成，生成全局模型并传送给客户端。

HOIKI模型的结构如图2所示。

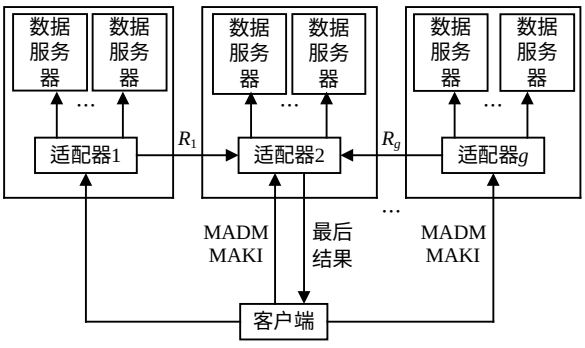

图2 HOIKI模型

图2中，每个适配器对于其子系统来说相当于一个客户端，将挖掘任务传送给该子系统的所有数据服务器。各数据服务器执行挖掘任务后将挖掘结果传送给适配器。对于由小系统组成的大系统，适配器又充当集成服务器，各子系统的结果在适配器之间传递和集成，最后由适配器将挖掘结果送回给客户端。

下面简要说明HOIKI模型的执行过程。

2.1 准备阶段

客户端根据挖掘任务选择挖掘算法，将MADM和MAKI模块发送给需要挖掘的子系统的适配器。

2.2 数据挖掘阶段

(1) 适配器将MADM和MAKI发送给该子系统需要挖掘的数据服务器。

(2) 每个数据服务器利用MADM来执行挖掘任务。

2.3 子系统知识集成阶段

对于每个子系统：

(1) 所有数据服务器将挖掘结果大小传递给相应的适配器。

(2) MAKI集成每两个数据服务器的挖掘结果，利用优化思想将小的挖掘结果集成到更大的服务器。

(3) 最后一个数据服务器将最后结果传递给适配器。

2.4 全局知识集成阶段

(1) 所有适配器将该子系统的集成结果大小传递给客户端。

(2) MAKI集成每两个适配器的结果，利用同样的优化思想将小的结果集成到更大结果的适配器。

(3) 最后一个适配器将最后一次集成的结果返回给客户端。

在HOIKI模型中，子系统采用OIKI DDM模型进行挖掘，也可以采用其他挖掘模式，如传统的分布式挖掘模式：(1) 各数据站点先独立执行挖掘任务。(2) 然后将挖掘结果传递到适配器，由适配器集成这些局部结果生成该子系统的挖掘结果。

从结果精度上看，HOIKI模型与OIKI DDM模型是一致的。从通讯量优化上看，HOIKI模型是一种层次优化，但不是全局最优，而是次最优。每两个站点间网络传输的代价不一定相同。总体传输代价不仅与传输量有关，还与传输距离、传输方式等有关。因此从整体上看，HOIKI模型在通讯上较OIKI DDM模型更有优势，特别当数据站点成集群分布时，情况更是如此。由于客户端仅与子系统的适配器通讯，而不是与每个数据服务器站点通讯，通讯负荷可能大大降低。

由于层次优化降低了全局集成的复杂度，因而执行效率更高，响应速度更快。此外，利用分而治之的思想，将复杂的大系统分解为简单易于解决的小系统，甚至还可以根据需要将小系统分解成更小的系统，使该模型具有更好的伸缩性，并支持并行操作，从而有效提高响应速度和降低通讯代价。该模型实际上是一种模块化设计模型，因此具有更好的灵活性和健壮性。

可以发现，OIKI DDM模型是HOIKI模型的特例，即当每个子系统只包含一个数据源时，HOIKI模型就是OIKI DDM模型。

3 实　验

本文以数量巨大、分布存储在多个观测站的天文数据库为例说明HOIKI模型的性能[5]。为不失一般性，假定每个观测站有相同数量的数据库，每个数据库由两个数值属性和大小为10 K的记录组成。每个观测站内的数据库之间通过高速局域网连接，可以无限制通讯，而两个观测站之间通过远程网连接，带宽有限，两地点之间的通讯代价比较高。客户利用互联网进行访问，且客户到每个观测站一次通讯的代价为5个单位；观测站之间的一次通讯的代价为3个单位；观测站内数据库之间一次通讯的代价为0.5个单位。

挖掘任务是利用聚类方法将分布的数据划分为三类。在本例中，MADM采用应用最广泛的K平均算法；MAKI采用簇中心一致化方法进行局部聚类结果集成，将最后的全局簇中心返回给客户。

很容易验证两种模型的聚类结果基本相同，但这不是本文讨论的重点，不作详细讨论。本文的重点在于模型的伸缩性和通讯代价降低，分两种情况进行实验。

(1) 数据库数量的性能评价。

设两个观测站L_1和L_2共有的数据库数量分别为10、20、50、100个，其通讯代价的实验结果如图3所示。从图3可以看出，数据库数量的增加对HOIKI模型的通讯代价影响很小，而对OIKI DDM模型的通讯代价影响却很大。

(2) 子系统数量的性能评价。

将100个数据库平均分布在2、5、10、50和10个子系统中，其通讯代价的实验结果如图4所示。图4的结果显示，数据库的群集性越好，HOIKI模型的

通讯代价越低，当数据库完全分布在不同地点时，其通讯代价与OIKI模型相同。

当子系统数g分别为1、5、10、50和100个，而数据库数量为100时，HOIKI模型的执行效率结果如图5所示，T(HOIKI)/T(OIKI)表示两种模型的执行时间比例。

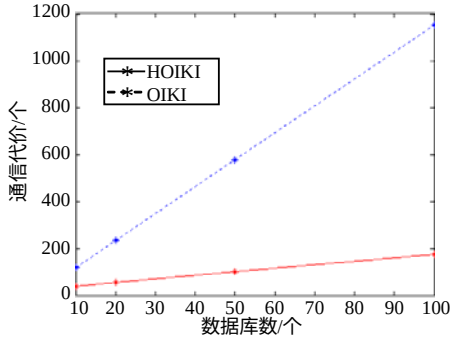

图3 不同数据库数量的通讯代价比较

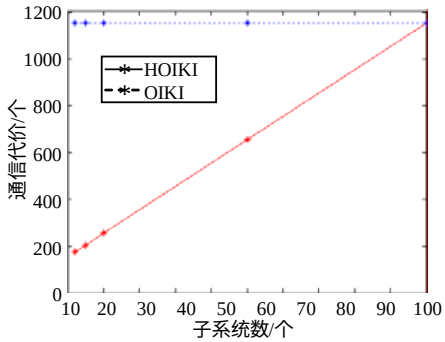

图4 不同地点数的通讯代价比较

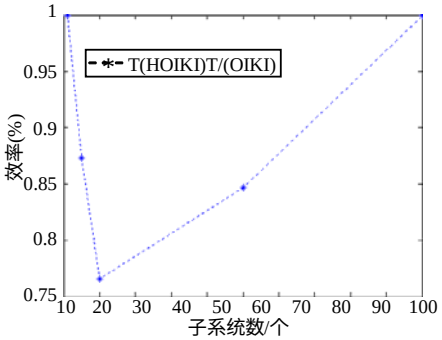

图5 执行效率比较

图5的实验结果充分表明，HOIKI模型的执行效率较OIKI DDM模型有显著提高，当$S = g^2$时执行效果最好。

从上面的实验结果可以发现，HOIKI模型与OIKI DDM模型执行效率提高，响应速度加快，具有更好的伸缩性，通讯量也明显降低；数据库的局部群集性越好，性能越好。由于采用层次结构，子系统可以采用不同的挖掘算法或模型，实现更灵活。

4 结 论

本文对于大规模分布式数据挖掘问题，提出一种HOIKI挖掘模型，该模型基于数据服务器的分布特性等原则，将规模庞大的系统分解成若干小的子系统，利用分而治之的策略，层次性地利用优化增量挖掘模型执行挖掘。通过理论分析和实验表明，较OIKI DDM模型，HOIKI模型具有更好的伸缩性，执行效率更高，通讯代价更小。因为对于各子系统可以采用不同的挖掘模式，因而实现更加灵活。此外，本文所提出的是一个两层次的HOIKI模型，还可以很灵活地拓展为多层次的HOIKI模型。

本文提出的是一个通用的分布式数据挖掘理论模型，该模型可以用来解决实际中大量存在的、大规模分布式数据库的分类、回归等问题，这也是下一步将要深入研究的课题。

参 考 文 献

[1] KARGUPTA H. An introduction to distributed data mining [EB/OL]. http://www.eecs.wsu.edu/~hillol, 2005-03-04.

[2]YE N. The handbook of data mining[M]. Metairie: LEA Inc., 2003.

[3] HLUSCH M, LODI S, MORO G. The role of agent in distributed data mining: issues and benefits[C]// Proceedings of the IEEE/WIC International Conference on Intelligent Agent Technology (IAT'03). Halifax: IEEE Press, 2003.

[4] SENOUSY M, MEDHAT M. A proposed model for distributed data mining using mobile agents[C]// Proceedings of the 11th Annual BIT2001 Conference. Manchester: Manchester Metropolitan University Press, 2001.

[5] CHEN R, GIANNELLA C, SIVAKUMAR K, et al. Distributed data mining for earth and space science applications[C]// Proceedings of the NASA Earth Science Technology Conference. Palo Alto: ACTA Press, 2004.

编 辑 熊思亮