第 33 卷 第 5 期 计 算 机 工 程 2007 年 3 月
Vol.33 No.5 Computer Engineering March 2007
·软件技术与数据库· 文章编号：1000—3428(2006)05—0071—02 文献标识码：A 中图分类号：TP311

基于流数据技术的信息监测系统的研究与设计

刘 佳[1]，张 芳[1]，刘国华[2]，刘 琳[3]
(1. 中国环境管理干部学院信息工程系，秦皇岛 066004；2. 燕山大学信息工程学院，秦皇岛 066004；
3. 秦皇岛职业技术学院信息工程系，秦皇岛 066004)

摘 要：以开发的网上信息监测系统为应用背景，建立了流数据应用的原型系统。通过 Internet 互联网实现网络数据截获、监控、还原、查询及存储等技术，提出并应用了连续查询及动态存储策略，在原型系统中测试了系统的性能。
关键词：流数据；连续查询；动态存储；监测系统

Research and Design of Information Inspect System on Streaming Data Technology

LIU Jia[1], ZHANG Fang[1], LIU Guohua[2], LIU Lin[3]
(1. Information Engineering Department, Environmental Management College of China, Qinhuangdao 066004;
2. Information Engineering Department, Yanshan University, Qinhuangdao 066004;
3. Information Engineering Department, Qinhuangdao Institute of Technology, Qinhuangdao 066004)

【Abstract】The paper implements an antetype application system on streaming data background of Internet information inspection system. It realizes data capture, inspection, recover, query and store by Internet. The system provides and applies continuous query and dynamic store. The system is tested in experimentation.
【Key words】Streaming data; Continuous query; Dynamic store; Inspection system

网络的不断发展导致了数据规模成指数倍的增长，在这个海量数据空间里，要动态收集、处理、追踪某些数据是很困难的。传统的信息搜索是无数次的排查数据库信息，不断刷新结果，而想得到的数据仅仅是海量数据中极少的一部分，这样严重导致了系统资源的浪费。为了适应信息源的不断变化、不可预测的特性，必须以动态的查询与存储技术、以流动的数据来代替传统的静态的数据。

流数据的产生从根本上改变传统的信息收集和交换方式，使得网络数据逐步转向了一个新的方向——在线数据。例如电信系统、Web 日志、网站点击、股票交易和通信流量、工业控制中的检测信号、航空航天等。在这些应用中，数据需要被连续地选择、分析，存储。这就需要一个处理流数据的管理系统，即流数据管理系统(SDMS)。它是建立在“瞬间流”的数据集的概念上，而不是存储相对静态的数据记录，它要求连续地查询以及查询带来的动态的存储技术。

本文结合一个网上信息监测系统实例，引入了流数据技术，使得本系统成为一个实时数据流监测系统，同时满足历史数据的查询处理。

1 流数据相关概念及技术

1.1 基本概念

定义 1 流数据(Streaming Data) 一个多属性元组(属性个数与数据源相关)$d:(Source,StreamId,Value,Time\{...\})$，其中$d_{source}$是流数据的数据源，$d_{streamid}$是流数据的序列号，$d_{value}$是流数据的当前值，$d_{time}$是流数据的采样时间或生成时间，也成为实时数据的时标，$\{...\}$表示其他应用属性。

定义 2 实时数据(Real-time Data) 一个三元组$d:(value,time,avi)$，其中d_{value}是实时数据的当前值，d_{time}是实时数据的采样时间或生成时间，也成为实时数据的时标，d_{avi}是实时数据的绝对有效期限。设当前时间为t，当$(t-d_{time})\le d_{avi}$时，实时数据的当前值d_{value}有效[1]。

定义 3 历史数据(History Data) 相对于实时数据而言，它是曾经的实时数据，已被存储在数据库中一段时间的数据。

定义 4 连续的查询(Continuous Query) 在某段时间间隔$[T_s,T_e]$，对流入系统中数据序列元组流$S_i,1\le i\le n$($n\in 1,2,\ldots,\infty$)，连续执行的系列查询$[Q_1,Q_2,\ldots,Q_n]$操作，即用新进来的数据连续探测查询处理系统中的查询，匹配查询条件，返回结果。

定义 5 时间戳(Timestamps) 在数据元组从数据源流出的时候，或者进入查询系统的时候，或者不同时间流出的多个数据源组成一个单一数据流的时候所设定的时间属性。其在数据元组里的表现形式为

$$T=(sid,tid,value,timestamp)$$

1.2 相关技术

在传统的 DBMS 中，数据的查询主要是基于“拉”的思想，在需要的时候触发相应的操作，从数据库中得到结果，数据库相对来说是被动的。相对比而言，流数据是从数据源“推”出数据到系统处理。这种思想在文献[2]中充分体现。文献[3]中 Tapestry system 将用户静态的查询转变为增加的查询，利用了仅增加数据库的单调性原理，提出了一个重写规

作者简介：刘 佳(1978－)，女，硕士生、讲师，主研方向：数据库；张 芳，讲师、硕士生；刘国华，教授、博士后；刘 琳，助教
收稿日期：2006-05-25 **E-mail：**liujia78928@126.com

则，将非单调的查询划分成单调的子表达式，能有效地找到新的匹配记录。为了提高查询效率，文献[4]中 NiagaraCQ 系统将具有相似结构的查询一并处理，最后再将结果分离。文献[5]提出了适应性的查询计划，其显著特点是利用了多种操作符，并用一个动态的路由模式代替了传统的查询计划，各种操作运算符作为独立的线程运算，增加了灵活性。文献[6]中提出了一个比较完善的流查询处理方案，但是没有对大量的数据如何存储加以说明。上述这些处理，要么查询计划是固定的，要么存储计划是静态不变的，仍然不能匹配数据流动的思想。

在本文监测系统实例中，引入了流数据技术、连续查询的思想来高效获取网络信息，提供动态存储策略，把静态的数据库变成动态的数据库。

2 基于流数据技术的 Internet 信息监测系统

2.1 连续查询系统

外网，内网截获的原始数据，通过 1~4 层协议，UDP、TCP 解析以后，形成完整的 TCP 会话数据、UDP 用户数据报。再通过 5~7 层协议正确连接、交换、表示数据，形成数据源。数据源发出的数据分为两种类型：(1)普通数据；(2)XML 数据。普通数据存到关系数据库，XML 数据存到 XML 数据库或者文件系统。

从图 1 中可以看到，查询流经两条路经：一条为普通查询，另一条为 XML 高级查询。普通查询遵循传统的查询原则，按照传统的查询方法处理，唯一不同在于此处处理流数据，即查询是在用户提交时间范围内连续执行，并且数据库中的数据不断更新，查询结果是个变化的过程。XML 数据处理部分作为一个发布系统，通过应用层部分完成 XML 到 SQL 查询的转换，查询到的结果再通过逆转换，将关系数据按照预先定义的 DTD 或者 schema 表示成用户要求的模式显示在窗口上，最终操作的仍是关系数据。而虚线部分则是独立的 XML 处理，不通过 SQL 转换。

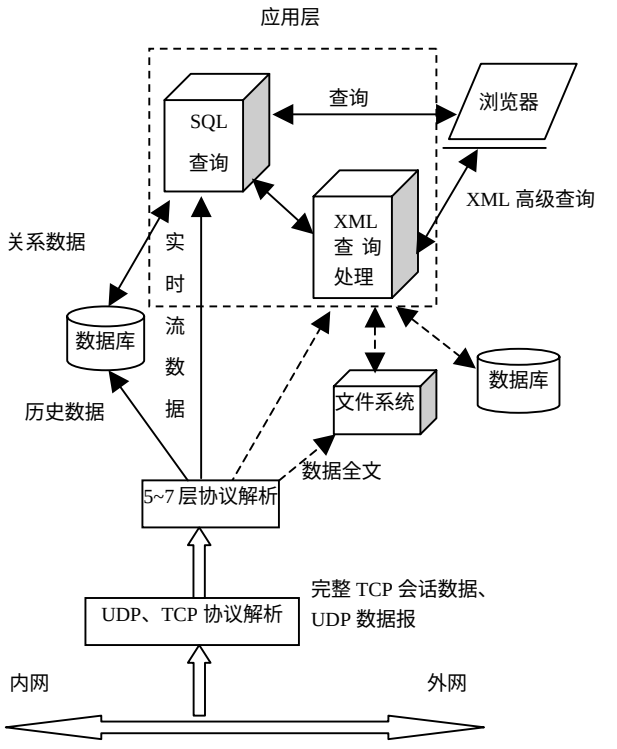

图 1 流数据管理系统结构

2.2 系统分析

Stream：输入数据源与输出流。

Query：查询流。

Store：外存数据库，存储历史数据。

Trash：回收站，根据数据元组的时间戳，收容过期的数据。

一级 DB：存储首次经过查询流的流数据，并通过一次过滤删除无用信息以后留下的数据。

二级 DB：存储一级 DB 中，按时间戳规律再经过动态存储策略淘汰旧的数据。

图 2 为 Internet 信息监测系统结构。

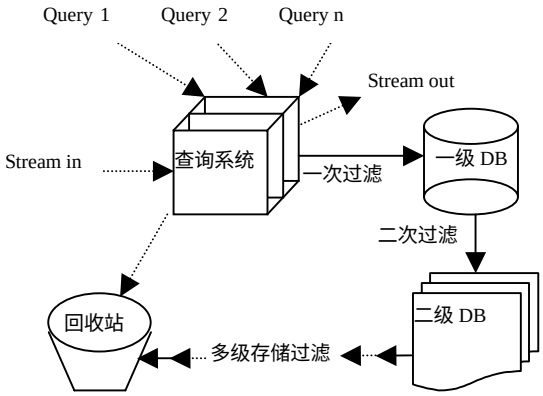

图 2 Internet 信息监测系统结构

Query 1 到 Query n 的查询流不断流经查询系统，与 Stream 数据流相遇，并判断查询条件，输出查询结果流，同时保存流数据到数据库中。DB 仍然是关系数据库，只是内容不再是静止的数据。大量的流数据经查询系统完成第 1 次数据过滤，因为查询的连续性，流数据的动态流经每个查询，基本上在经过整个查询系统后，不满足条件的数据可以先被过滤掉。被一次过滤后的数据存到一级 DB 中，当对一级 DB 存储空间造成威胁的时候，实行动态存储策略，将比较不常用的信息转存到二级 DB 中，以此类推，直到数据不具备实际应用意义的时候，直接删除。

3 连续的查询与传统的查询系统分析与比较

3.1 数据量

本文取了从 5 000~100 000 条数据来对比传统的查询与流数据查询的性能差异，观察随着数据量的增加，响应时间的变化如图 3 所示，Trad 代表传统的查询，Tstm 代表流数据查询。

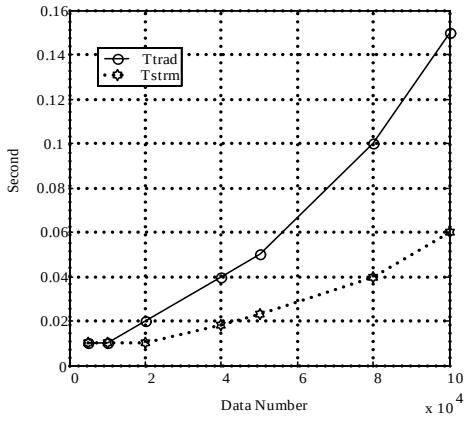

图 3 基于查询数量的连续的查询响应时间

（下转第 75 页）