流数据和传统数据存储及管理方法比较研究

李子杰[1,2],郑 诚[1,2]
(1.安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039;
2.安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:传统数据或静态数据是指来自关系数据库、数据仓库和事务数据库里面的数据,与之对应的流数据,是一种具有实时、快速和连续到达特点的动态数据。由于流数据的上述特点,使得应用于传统数据挖掘的技术和方法不能很好地适应流数据。对传统数据进行存储、查询和管理,使用成熟的DBMS完成,对流数据的类似操作,必须开发出具体的DSMS加以实现。提出了一个对流数据进行管理的系统框架,并在管理系统和存储方式两方面对两种数据进行综合比较。

关键词:数据挖掘;流数据;传统数据
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)04-0101-04

Comparative Study on Methods of Storage and Management of Stream and Traditional Data

LI Zi-jie[1,2], ZHENG Cheng[1,2]
(1. Ministry of Education Key Laboratory of Intelligent Computing & Signal Processing, Anhui University, Hefei 230039, China;
2. School of Computer Science and Technology of Anhui University, Hefei 230039, China)

Abstract: The traditional data or static data indicates the data from the relational database, data warehouse and transaction database. But the stream data is a real time, quick and continuously transported dynamic data. Based on the above features of the stream data, the technologies and methods applied into the traditional data mining can not be well adapted to the stream. Use mature DBMS to store, inquire and manage the traditional data. For the similar operation on the stream data, the specific DSMS must be developed out. Proposes a system framework for management of the stream data and makes a comprehensive comparison in the aspects of management system and storage mode.

Key words: data mining; stream data; traditional data

0 引 言

数据挖掘的对象一般是来自关系数据库、数据仓库和事务数据库里面的数据,称其为传统数据或静态数据。另一种形式的数据最早出现于银行和股票交易领域,现在广泛应用于Web服务器的日志记录、传感器网络数据监控与分析和电子商务等方面,这类数据具有连续快速、短暂易逝和不可预测的特点[1],称其为流数据或动态数据。

流数据是由Henzinger等人于1998年在论文"Computing on Data Streams"中首次提出[2]。自21世纪以来,流数据已经成为一个新的研究领域,对流数据的查询、分析及挖掘是在快速和大量的流数据上进行的,由于数据收集时间和分析处理速度的限制,使得应用于传统数据挖掘的一些技术和方法不能很好的适应流数据。文中将从管理系统和存储方式两方面着手对流数据和传统数据进行比较。

1 流数据和传统数据的比较

1.1 管理系统

和流数据相比,传统数据是静止的且规模较小,可以完全存储在数据库中,对传统数据的管理可以利用成熟的数据库管理系统(DBMS)如Oracle, Sybase, SQL Server, Access和Visual FoxPro等实现。

不同于DBMS,流数据管理系统(DSMS)处理的数

收稿日期:2008-07-24
基金项目:安徽省自然科学基金资助项目(050420204);安徽省高校自然科学研究项目(2006kj055B)
作者简介:李子杰(1978-),男,安徽阜南人,硕士研究生,研究方向为数据挖掘、语义web;郑 诚,博士,副教授,主要从事数据挖掘、机器学习方向研究。

据是一种实时到达的连续数据序列，它们具有到达速度快、顺序确定、单位时间数据到达量不均匀和数据量庞大等特点，这就要求 DSMS 必须具有一次存取、持续处理、有限存储、结果近似和快速响应的功能。目前比较成熟的 DSMS 有 STREAM(Stanford)、Aurora(Brandeis/Brown/MIT)、Gigascope(AT&T)和 Tribeca(Bellcore)等[3]。下面从数据模型、查询语言和系统框架三个方面对 DBMS 和 DSMS 进行对比。

1.1.1 数据模型

DBMS 按出现的先后顺序可分为层次型、网状型、关系型和面向对象型[4]四种类型，其基于的数据模型也相应地经历了层次模型、网状模型、关系模型和面向对象模型四个发展阶段。流数据的产生是一个长期、持续的过程，一些算法不是将所有的流数据作为处理对象，而是根据实际需要选取某个时间段，然后对时间段内的流数据进行处理。按算法处理流数据时所选取的时序范围，流数据模型可分为三种类型[5]：

(1)快照模型(snapshot model)。

处理数据的范围限制在两个预定义的时间戳之间。

(2)界标模型(landermark model)。

处理数据的范围从某个已知的初始时间点到当前时间点为止。界标窗口可以表示为 $W=(T_1,T_2,\cdots,T_t)$，其中 T_1 为初始时间点，T_t 为当前时间点。

(3)滑动窗口模型(sliding window model)。

处理数据的范围由某个固定大小的滑动窗口确定。滑动窗口可以表示为 $W=(T_{t-w+1},\cdots,T_t)$，其中 W 为指定的时间点，T_t 为当前时间点。

后两种模型采用较多。界标模型通常将流数据起始点 T_1 作为数据处理的初始时间点，因此，算法是对内存中的所有流数据进行处理，新数据和旧数据的影响力被同等对待。滑动窗口模型一般都要求用户事先指定窗口大小，算法在运行过程中只能给出此滑动窗口上的计算结果，非常适合对最近时间段内的数据进行处理。

1.1.2 查询语言

结构化查询语言(SQL)由于其简洁明了、方便实用和功能齐全的特点，倍受人们的欢迎。目前，绝大多数关系型 DBMS 都支持 SQL。传统 SQL 是对存储在硬盘等外部介质上的静态数据进行操作，这使它无法胜任对流数据的查询操作[2]。

为了适应流数据快速、实时、连续到达的特点，连续查询的概念在流数据管理系统 Tapestry[6]中被首次提出。流数据查询语言可分成三种类型[2]：

(1)基于关系的流查询语言。

基于关系的 DSMS 使用类 SQL 语言，通过流－关系算子将连续的流数据转化为关系模型进行处理，得到的关系型结果可以直接输出，查询过程支持窗口和排序。例如 CQL[7]是基于关系的流查询语言。

(2)基于对象的流查询语言。

基于对象的 DSMS 也使用类 SQL 语言，它通过对流数据相关元素的提取，按照层次类型进行处理。这种 DSMS 在设计上采用层次型查询框架，把查询任务分成两个阶段：前期对流数据进行简单分类，后期对分类后的流数据进行具体操作，这种分阶段查询使得系统的整体性能得到优化。例如 Tribeca[8]使用的就是基于对象的流查询语言。

(3)基于过程的流查询语言。

基于过程的 DSMS 通过数据流程的各种算子构造查询，为用户提供了一种自定义查询方案的机制，用户可以通过组织算子组合出查询框架，从而控制流数据的流向并获得自己需要的各种信息。例如 Aurora 系统[9]，为用户提供了一个基于 Java 的图形接口，用户可随时通过它定义和修改自己的查询计划。

1.1.3 系统框架

DBMS 按功能大致可划分为模式翻译、应用程序编译、交互式查询、数据的组织与存取、事务运行管理和数据库维护等组成部分。

在挖掘流数据的过程中，会出现数据流速较快、数据量庞大、数据仅通过内存一次、挖掘结果短时间响应、内存消耗快和 CPU 处理效率低等问题，为解决上述问题，提出了一个 DSMS 的系统框架，如图 1 所示。

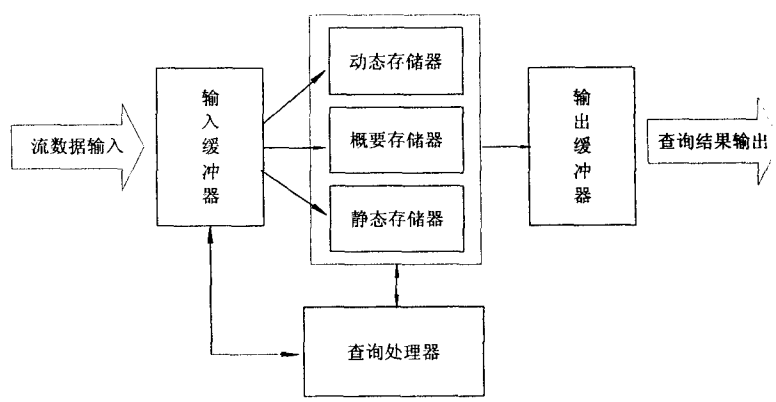

图 1 DSMS 的系统框架

上图中，输入缓冲器的设置便于批量处理流数据，动态存储器存储窗口查询操作所需要的数据，概要存储器保存流数据的统计信息，静态存储器根据特定的需求对某阶段的原始流数据进行保存，查询处理器响应用户提出的实时查询请求并对查询进行优化，最后

由输出缓冲器输出查询结果。

1.2 存储方式

目前大部分 DBMS 是关系型的,这类系统一般都提供多种存取数据的方法,最常用的存取方法有三类[10]:第一类是索引方法,主要是 B+ 树索引方法;第二类是聚簇方法;第三类是 HASH 方法。

流数据由于规模庞大、流速很快,因此在辅存上保存全部流数据是不现实的,DSMS 一般只保存最近的历史信息,对于比较重要的流数据,则保存其概要数据信息。由于流数据查询过程基本上都是连续查询,所查询的数据只是全部流数据的一部分,因此 DSMS 采用滑动窗口的机制处理查询,下面的三种方法都是基于滑动窗口的。

1.2.1 线性回归

Teng 等人提出一种 ATF 方法[11]压缩存储最近一段时间窗口内的信息。通过线性回归分析,可以用以前的信息预测模式的频率变化情况。文中定义模式在 T 时刻的支持度为该时刻包含该模式的客户数目与客户总数之比,模式的频率计数 f 为窗口中该模式在 t 时刻之前的支持度的平均值。AFT 用来记录模式频率计数变化,由四元组 $(t_s, \sum t_f, \sum f, \sum f^2)$ 表示。其中 t_s 表示开始记录的时间,$\sum t_f$ 是时间与相应频率计数乘积的和,$\sum f$ 与 $\sum f^2$ 分别是频率计数的和与平方和。

算法最终为模式维护一个 ATF 列表,分段记录整个数据流上模式的频繁情况。给定任意时间段 $t = [t_1, t_2]$,由 $f(t_1)$ 和 $f(t_2)$ 可以计算出时间段 t 内该模式的平均支持度。ATF 方法记录信息量小,只能以估计的方式回答模式的平均频繁状况,查询精度有限,适合于对时间精确度要求比较高的应用领域。

1.2.2 滑动窗口抽样

滑动窗口抽样根据处理的角度不同,可分为基于计数的和基于时间的滑动窗口抽样。基于计数的滑动窗口抽样是指大小固定的滑动窗口,所处理的数据以某个正整数 N 作为基本单位,只有当输入的数据达到一定的数目之后才做相应的处理。正整数 N 可以自己设定,在 N 已知的情况下从中抽取 n 个数据就是基于计数的滑动窗口抽样,典型的代表算法有 Chain-sample 抽样算法[12]。

基于时间的滑动窗口抽样是指滑动窗口的大小 N 可以变化,窗口模型以固定的时间间隔 T 作为参数,文献[13]给出了一种基于时间的滑动窗口抽样算法:

定义两个时间戳 t_{i-1} 和 t_i,则 $[t_{i-1}, t_i]$ 构成一个时间段 T。t_{i-1} 和 t_i 分别称为 T 的下界和上界,$t_i - t_{i-1}$ 称为 T 的跨度,记为 ΔT。

① t_i 时刻起,读入最先到达的 n 个元组,分别存储在数组 array[0],…,array[$n-1$] 中作为候选样品。

② 顺序读入剩余的元组,处理机制如下:

(a) 判断是否处理完所有窗口中的元组,完成则转到 ③;否则继续(b)。

(b) $t = n$。

(c) 读入第 $t+1$ 个元组,计算 R = TRUNC(($t+1$) * RANDOM( ))。其中 $0 \leqslant R \leqslant t$。若 $R < n$,则第 $t+1$ 个元组以 $n/(t+1)$ 的概率替换候选样品 array[R];若 $R \geqslant n$,重复(c)。

(d) 直到读完窗口中的 N 个元组。

③ 输出 array[0],…,array[$n-1$],即为此滑动窗口的抽样。

1.2.3 倾斜时间窗口

Giannella 等人提出使用倾斜时间窗口模型挖掘频繁项集的近似集[14]。具体做法是,对于最近项集的频率以一个较好的时间粒度保存,对于历史项集的频率用粗糙的时间粒度保存。为了减少倾斜时间窗口中项集 X 的频率记录的数目,需要进行剪枝。假设 $\text{freq}_j(X)$ 是项集 X 在时间单元 T_j 上的估计频率,N_j 是在 T_j 内所接收事务的数量,其中 $1 \leqslant j \leqslant T$。对某个参数 m,$1 \leqslant m \leqslant T$,频率记录 $\text{freq}_1(X), \cdots, \text{freq}_m(X)$ 当公式(1)和(2)成立时被剪枝:

$$\exists n \leqslant T, \forall i, 1 \leqslant i \leqslant n, \text{freq}_i(X) < \sigma N_i \tag{1}$$

$$\forall l, 1 \leqslant l \leqslant m \leqslant n, \sum_{j=1}^{l} \text{freq}_j(X) < \varepsilon \sum_{j=1}^{l} N_j \tag{2}$$

剪枝确保了如果在时间间隔 T 内返回所有估计频率超过 $(\sigma - \varepsilon)N$ 的项集,则不会在 T 上丢失任何频繁项集,所返回项集的估计频率至多比它们的实际频率少 εN。

2 结束语

在管理系统和存储方式两个方面对流数据和传统数据进行了比较,着重讨论了流数据的特点、挖掘过程中出现的问题以及解决方法,指出了流数据越来越多地被人们使用,揭示了流数据正成为将来的一个热点研究领域。

参考文献:

[1] Golab L, Ozsu M T. Issues in data stream management[J]. ACM SIGMOD Record, 2003, 32(2):5-14.

[2] Henzinger M, Raghavan P, Rajagopalan S. Computing on data streams[R]. Palo Alto, California: Digital Systems Research

Center,1998.
[3] 周明中,龚 俭.数据流管理系统综述[J].计算机工程,2006,32(2):10-12.
[4] Post G V. Database Management Systems: Designing and Building Business Applications[M]. 3rd Edition. 冯建华,刘旭辉,周维续译.北京,机械工业出版社,2006.
[5] 金澈清,钱卫宁,周傲英.流数据分析与管理综述[J].软件学报,2004,15(8):1172-1181.
[6] Terry D,Goldberg D,Nichols D,et al. Continuous queries over append-only databases[C]//In Proc. of the 1992 ACM SIGMOD Intl. Conf. on Management of Data. San Diego, California:[s.n.],1992:321-330.
[7] Arasu A, Babu S, Widom J. An Abstract Sematics and Concrete Language for Continuous Queries over Streams and Relations[R]. US:Stanford University,2002.
[8] Sullivan M,Heybey A. Tribeca: A System for Managing Large Databases of Network Traffic[C]//In Proc. USENIX Annual Technical Conf. New Orleans, Louisiana:[s.n.],1998.
[9] Carney D, Cetinternel U, Cherniack M, et al. Monitoring streams - A New Class of Data Management Applications [C]//In Proc. Int. Conf. on Very Large Data Bases. Hong Kong:[s.n.],2002:215-226.
[10] 萨师煊,王 珊.数据库系统概论[M].北京,高等教育出版社,2000.
[11] Golab L,Bijay K. On Concurrency Control in Sliding Window Queries over Data Streams[C]//In Proceeding 10th International Conference on Extending Database Technology. Munich,Germany:[s.n.],2006:608-626.
[12] Babcock B, Data M, Motwan I R. Sampling from a moving window over streaming data[C]//In ACM-SIAM Symposium on Discrete Algorithms. San Francisco, CA, USA:[s.n.],2002.
[13] 葛君伟,公丕强,刘兆宏.一种存储和索引历史数据流数据的方法[J].计算机应用研究,2007,24(6):104-106.
[14] Giannella C,Han J,Pei J,et al. Mining Frequent Patterns in Data Streams at Multiple Time Granularities[C]//In Kargupta et al. Data Mining:Next Generation Challenges and Future Directions.[s.l.]:MIT/AAAI Press,2004.

(上接第 97 页)

参考文献:

[1] 梁路洪,艾海舟,徐光祐,等.人脸检测研究综述[J].计算机学报,2002,25(5):449-458.
[2] Hafed Z M, Levine M D. Face recognition using the discrete cosine transform[J]. International Journal of Computer Vision,2001,43(3):167-188.
[3] 于威威,滕晓龙,刘重庆.复杂背景下人眼定位及人脸检测[J].计算机仿真,2004,21(12):185-188.
[4] 陶 亮,庄镇泉.复杂背景下人眼自动定位[J].计算机辅助设计与图形学学报,2003,15(1):38-42.
[5] 冯建强,刘文波,于盛林.基于灰度积分投影的人眼定位[J].计算机仿真,2005,22(4):75-77.
[6] 吕东辉,王 滨.YCbCr 空间中一种基于贝叶斯判决的肤色检测方法[J].中国图象图形学报,2006,11(1):47-52.
[7] Tao Liang, Zhuang Zhen-quan. An effective approach for frontal face verification[J]. Journal of Image and Graphics, 2003,8(8):860-865.

(上接第 100 页)

署的兴趣,可以使用防御模型作为增值服务来为用户提供更好的服务,从而增加收入项目,因此 ISP 也就有相当大的兴趣来部署。另外,所有必要的防御措施都由最后的 ISP 来管理,它是防御模型的受益者。

在以后的工作中,比率控制方法仍然是一个重点,在网络拥塞的节点可以考虑对流量进行一定的疏导,保证正常流量尽可能不受到攻击的干扰。另外在 IPv6 的情况下,20 位标记位性能的提高也有待验证。

参考文献:

[1] Schneier B. Secrets and Lies: Digital Security in a Networked World[M]. New York:John Wiley & Sons,2000.
[2] Belenky A,Ansari N. On deterministic packet marking[J]. Computer Networks,2007,51:2677-2700.
[3] Burch H,Cheswick H. Tracing anonymous packets to their approximate source[C]//Proc. USENIX LISA Conf. New Orleans,LA:[s.n.],2000:319-327.
[4] Stoica I,Zhang H. Providing Guaranteed Services Without Per Flow Management[C]//Proc. the 1999 ACM SIGCOMM Conf. Boston,MA:[s.n.],1999:81-94.
[5] Yaar A,PerrigA,Song D. Pi:a path identification mechanism to defend against DDoS attacks[C]//Proceedings of the IEEE Symposium on Security and Privacy. Berkeley: IEEE Press, 2003:93-107.
[6] Yaar A, Perrig A, Song Dawn. StackPi: new packet marking and filtering mechanism for DDoS and IP spoofing defense [R]. US:Carnegie Mellon University,2003.
[7] Rivest R L. The MD5 message digest algorithm[S]. RFC 1321,Internet Activities Board, Internet Privacy Task Force, 1992.
[8] 孙知信,李清东.基于源目的 IP 地址对数据库的防范 DDos 攻击策略[J]. Journal of Software, 2007, 18(10): 2613-2623.