挖掘数据流频繁模式的相关技术和算法研究综述

唐懿芳[1,2],穆志纯[1],张师超[3,4],钟达夫[2]
TANG Yi-fang[1,2],MU Zhi-chun[1],ZHANG Shi-chao[3,4],ZHONG Da-fu[2]

1.北京科技大学 信息工程学院,北京 100083
2.广东科技干部管理学院 计算机工程技术学院,广东 珠海 519090
3.广西师范大学 计算机科学与信息工程学院,广西 桂林 541004
4.悉尼理工大学 信息技术学院,澳大利亚 悉尼
1.School of Information Engineering,University of Science and Technology Beijing,Beijing 100083,China
2.Computer Engineering Technical College,Guangdong Institute of Science and Technology,Zhuhai,Guangdong 519090,China
3.College of Computer Science and Information Technology,Guangxi Normal University,Guilin,Guangxi 541004,China
4.Faculty of Information Technology,Sudney University of Technology,Sydney,Australia
E-mail:yifangt@163.com

TANG Yi-fang,MU Zhi-chun,ZHANG Shi-chao,et al.Research overview of related techniques and algorithms on frequent pattern mining in data stream.Computer Engineering and Applications,2009,45(26):121-125.

Abstract: Some characters of Data stream make that static mining method can't meet the requirements of nowadays mining application.Many new techniques and methods on frequent pattern mining in data stream have been proposed.In this paper,we give an overview of these algorithms.Firstly,the concept and characters of data stream are introduced.Then related research work about data streams are introduced at home and abroad.The characters of mining frequent pattern in data stream are analyzed,and the common techniques and the representative algorithms of mining are listed.At last,future directions in data stream mining research are discussed.

Key words: data stream;frequent pattern;synopsis data structure;decay factor;tilted time window

摘 要:数据流本身的特点使得静态挖掘方法不再满足要求。国内外学者已提出许多新的挖掘数据流频繁模式的方法和技术。对这些技术和算法进行了综述。首先介绍数据流的概念和特点,分析国内外的研究现状,总结了数据流中挖掘频繁模式的特点,并列出挖掘方法的常用技术和基于这些技术的代表性算法,最后讨论了将来的研究方向。

关键词:数据流;频繁模式;概要数据结构;衰减因子;倾斜时间窗口
DOI:10.3778/j.issn.1002-8331.2009.26.036 **文章编号:**1002-8331(2009)26-0121-05 **文献标识码:**A **中图分类号:**TP311

1 引言

近年来,数据流(Data stream)已经成为国内外研究热点。它是一种实时、连续、有序的数据序列[1-2]。数据流的大量和潜在无限的数据是由网络监控、入侵检测、情报分析、金融服务、股票交易、电子商务、电信、卫星遥感(气象、环境资源监控等)、Web 页面访问和其他动态环境产生的。如此多的应用领域都产生数据流,所以在数据流中挖掘知识显得尤为重要。

2 数据流的概念和特点

(1)概念:数据流(Data stream)[3]是连续的、无限的、快速的、随时间变化的、有序的且快速流动的数据元素组成的无限序列。这些数据元素只能被读取一次。可表示为:数据流 $DS=\{T_1,T_2,\cdots,T_N,\cdots\}$,其中 N 表示当前数据流中事务 T 的数量,每个事务 T_i 由属性集 $I=\{i_1,i_2,\cdots,i_m,\cdots\}$ 的项集构成,即 $T_i\subseteq I$。

(2)特点:①有序性、连续性、实时性,数据有序地、连续地到达并实时地变化。②无限性,大数据量,甚至是无限的数据量,存储所有的数据是不可能的。③单遍性,由于数据元素只能被读取一次,只能对数据流进行单遍扫描。④概要性,处理数据流数据时,由于内存的限制,要求构造概要数据结构。⑤低层次性和多维性,数据流的原始细节数据的概念层次较低且具有多维的特点。⑥近似性,由于采用的是概要数据结构,数据流查询以及挖掘处理得到的结果是近似的。⑦实时性,用户通常要求得到即时的处理结果。

基金项目:北京市教委重点学科共建项目资助。
作者简介:唐懿芳(1976-),女,博士研究生,副教授;穆志纯(1952-),男,博士生导师,教授;张师超(1962-),男,澳大利亚悉尼理工大学博士生导师,广西师范大学教授。
收稿日期:2008-05-16 **修回日期:**2008-08-04

3　挖掘数据流的研究现状

Henzinger等人于1998年在论文"Computing on Data Stream"中首次将数据流作为一种数据处理模型提出来[4]。从2000年开始,数据流作为一个热点研究方向出现在数据挖掘与数据库领域的几大顶极会议中，如VLDB、SIGMOD、SIGKDD、ICDE、ICDM等会议每年都有多篇有关数据流处理的文章。目前,国外在数据流挖掘方面有两个比较有影响的研究小组:

(1)一个是Stanford大学的R.Motwani教授领导的研究小组,他们的研究侧重在数据流管理、数据流的连续查询和数据流的聚类方面[1,5-7],提出了不同于传统DBMS的DSMS(Data Stream Management System)概念,他们的研究得到了美国国家自然科学基金的资助,开发出一个现在已经公开的数据流关系的原型系统:STREAM。STREAM是一个通用的数据流管理原型系统，该系统提供了一种连续查询语言CQL(Continuous Query Language)，它既可以处理数据流又可以处理关系型数据。文[8]提供了如图1的STREAM的结构图,以便于理解该数据流管理系统。

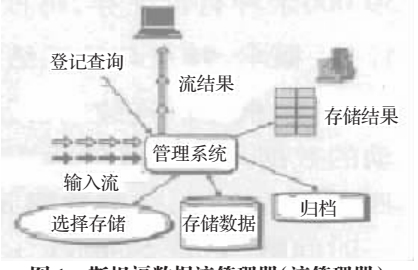

图1　斯坦福数据流管理器(流管理器)

(2)另一个是UIUC的C.Aggarwal和J.Han教授领导的研究小组。该研究小组侧重在数据流分析方面,对于数据流的在线分析,从聚类、分类、频繁项集挖掘以及可视化等角度做了大量研究工作,提出了倾斜时间窗口(tilted-time window)策略[9-12],采用不同时间粒度保存数据流的信息,他们的研究得到了美国军方和国家自然科学基金的资助。

目前国内在数据流挖掘方面比较有影响的主要有以下几个团队:(1)东北大学的于戈教授、王大玲教授等组成的团队,文[13]提出了数据流中一种快速启发式频繁模式挖掘方法。(2)中国人民大学的孟小峰教授在文[14]指出了数据库发展趋势的泛数据时代已经来临,并提出了对流数据当前存在问题的解决办法,并翻译了J.Han教授的《数据挖掘概念与技术》。(3)上海交通大学的谢康林教授,他们的主要研究方向是数据流的频繁集的挖掘[15]。(4)华中科技大学的李庆华教授,他们的研究小组对数据流的挖掘进行了相关的研究,提出了一些解决相关问题的算法[16]。(5)浙江大学的潘云鹤教授,从网络TCP流量的模拟对数据流相关问题进行研究[17]。除此之外,还有许多发表的相关文章,限于篇幅,不能一起列出。但总结起来,大部分的数据流处理模型都可以归结为图2的结构[18]:

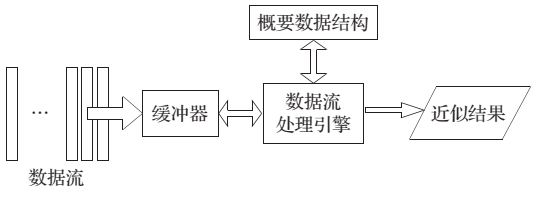

图2　数据流处理模型

该模型通过处理引擎(Data stream processing engine)对数据流进行快速处理并实时更新保存在内存中的概要数据结构,缓冲器(Buffer)的设置便于数据的批量处理,概要数据结构存储了数据流的统计特征,是对数据流的一种压缩总结,借助于它就不需要对数据流进行多遍扫描即可获得误差可控的近似结果。该模型中的数据流处理引擎和概要数据结构的设计成为影响整个过程的关键因素,其性能的优劣直接影响处理算法的时间和空间复杂度。在一般情况下,这两个组件都是算法设计的重要内容,但在数据流处理中,由于存储空间的限制,算法的空间复杂度尤为重要,它是评价其性能优劣的首要指标[1]。由此可知，数据流处理的重点就是设计有效的概要数据结构,使其能够满足数据的近似处理要求,得到误差可控的结果。

数据流挖掘的对象可以是多条数据流,也可以是单条数据流。挖掘多条数据流的主要目的是分析多条并行到达的数据流之间的关联[19-24]。对单数据流的挖掘则涵盖了分类、频繁模式挖掘、聚类等多项传统数据挖掘中的主要任务。挖掘变化的数据流是一项特殊的任务。主要对单数据流频繁项集和频繁模式的挖掘研究现状进行总结,并对存在的问题和未来的研究方向提出一些观点。

4　数据流中频繁模式的挖掘

频繁模式挖掘定义为发现数据集中频繁出现的模式集,如果一个模式的计数满足最小支持度，则这个模式是频繁的,可描述如下:数据流 $DS=\{T_1,T_2,\cdots,T_N,\cdots\}$,其中 N 表示当前数据流中事务 T 的数量,每个事务 T_i 由属性集 $I=\{i_1,i_2,\cdots,i_m,\cdots\}$ 的项集构成。即 $T_i\subseteq I$。如果一个模式 $P(P\subseteq I)$,在数据流 DS 中的频率由所有包含模式 P 的事务数组成,记为 $f(p)$,当其支持度 $sup(p)/N\geqslant S_{min}$ 时,称 P 为频繁模式,其中用户定的最小支持度阈值 $S_{min}\in(0,1)$。

在静态数据集中,可伸缩的频繁模式挖掘方法得到了广泛的研究,最有代表性的就是Apriori算法[25]。然而,数据流实时、连续、有序、快速到达的特点以及在线分析的应用需求,对数据流挖掘算法提出了诸多挑战。总结起来,根据数据流本身的特点,数据流挖掘算法需要具有以下特点:

(1)单次线性扫描。即算法只能按数据的流入顺序依次读取数据一次。

(2)低时间复杂度。为了跟上数据流的流速,处理每个数据项的时间不能太长。

(3)低空间复杂度。数据流的挖掘算法是在主存中进行,算法的空间复杂度不能随数据量无限增长。

(4)结果的近似性。数据流无法存储和再次扫描,只能得到近似的处理结果。

(5)实时响应性。算法必须能响应用户在线提出的任意时间段内的挖掘请求。

(6)自适应性。能适应动态变化的数据与流速。

许多现有的频繁模式挖掘算法需要系统多次扫描整个数据集[25],很多数据挖掘的前期步骤还有一个预处理的过程[26],但对无限的数据流来说是不现实的。一个现在的不频繁项集将来可能变成频繁项集,因此不能将其忽略。一个频繁项集也可能变成不频繁项集,非频繁项集呈指数级增长,不可能全部记录。如何对数据流中的频繁项集进行更新，对不频繁项集进行剪枝,频繁项集如何组合成频繁模式。

有两种可能的办法来解决以上问题。第一种是仅保持一个预先确定的项或项集的有限集。该方法用途不大,因为它需要系统将考察范围限制在预定义的项集上,这不符合数据流无限且实时的要求。第二种方法是推导出答案的近似解,实际上近似结果常常足以满足需要。后面将详细介绍采用此思想的算法。总结起来,数据流频繁模式挖掘的任务就是在有限的存储空间下,通过近似算法对模式的频率进行估计,并尽可能减少其相对误差,从而获得满足最小支持度要求的频繁模式。正因为是近似解,结果就可能会存在漏报(False negative,即没有获得所有正确的频繁模式)或误报(False positive,即获得了非频繁模式)的可能,且这两种误差相互制约。而挖掘数据流频繁模式的算法就是依具体应用情况而采取不同的折中,使误差尽可能减少。

5 挖掘数据流频繁模式的相关技术和算法

在动态数据集上挖掘频繁项是一项困难的任务[27]。数据流的单次线性扫描进一步增加了这项任务的难度。针对数据流的特点,学者们提出了许多相关的技术。

5.1 概要数据结构和相应的近似算法

在数据流处理系统中,由于数据量远大于可用内存,系统无法在内存中保存所有扫描过的数据,而数据流挖掘经常要求读取这些数据,为了避免代价昂贵的磁盘存取,必须要构造一种新的数据结构,以保留扫描过的信息,通常称为概要数据结构。近年来国内外学者对这种数据结构进行了深入研究,提出了很多有效的构造方法,如采样技术[28-29]、小波变换[30-32]、直方图技术[33-35]、哈希方法[36-37]、梗概技术[38-40],字典树结构[13]等各种高效的树型存储结构等。他们已广泛应用于数据流的处理和挖掘中,相关文献[41-42]已经进行了较为全面的论述,在此略过。

基于概要数据结构的算法都是近似算法。这是因为在构建概要数据结构时,不可避免地存在着信息的损失。概要数据结构只能近似还原原有数据。但通常这种近似算法对数据流挖掘已经够用,它的误差 ε 可以控制在一定的范围内。比较有代表性的算法有基于 Count Sketch 数据结构的 Count Sketch 算法[43],此算法利用有限内存空间通过一趟扫描来估计数据流中最大频繁项集,利用 Count Sketch 结构可在数据流中可靠地估计频繁项集的频率,算法强调估计频繁项集的方法,但没有量化误差的范围。另外,文[44]提出了一个近似的数据流频繁项挖掘算法 Lossy Counting 算法。算法保存多个三元组记录:(e,f,Δ),其中,e 是数据流中的数据项,f 为 e 的估计频率,Δ是 e 的最大可能误差,即若记 e 的真实频率为 f_e,则 $f_e \leqslant f+\Delta$。对于选定的参数 ε,每当算法遇到一个没有被记录的新元素 e'时,就生成一个新元组$(e',1,\lfloor \varepsilon N \rfloor)$,每当算法读取$\lceil 1/\varepsilon \rceil$个数据项时,就删除所有 $f+\Delta \leqslant \varepsilon N$ 的元组。其中 N 为目前已读取的总的数据项数目。通过处理,算法保证所有 $f_e \geqslant \lfloor \varepsilon N \rfloor$的元素都被记录,且 $f \leqslant f_e \leqslant f+\varepsilon N$。对于任意的频繁度阈值 $s>\varepsilon$,输出满足条件 $f \geqslant (s-\varepsilon)N$ 的项就保证所有 $f_e \geqslant sN$ 的项都被输出,且所有输出项都满足条件 $f_e \geqslant (s-\varepsilon)N$。在实际应用中,大部分数据的出现频率都较低,通过采用上述方法,算法不需要记录出现频率较低的数据,从而既节省了计算空间,同时又保证了输出的质量。

5.2 滑动窗口技术和基于滑动窗口技术的算法

使用滑动窗口减少了算法需要处理的数据量,并对挖掘变化的数据流提供支持,另一方面,有些应用只对最近的数据感兴趣,要求算法对以当前时间为终点的某个滑动窗口内的数据进行处理。

SWF 算法[45]是最先提出采用滑动窗口技术的流挖掘算法。SWF 算法使用一个滑动窗口包括最近几个固定数量的事务,在滑动窗口内部计算频繁项集,即只挖掘最近几个事务,找到候选集。当数据流往前流动,滑动窗口更新,原来旧的事务去掉,新事务进入滑动窗口,但如何完全去除旧事务的影响,SWF 算法并没有解决。Geoff Hulten 等人提出的 CVFDT[46]这种方法可以精确地对滑动窗口内的计算结果进行增量式地更新且每个样例的复杂度仅为 O(1)。但是如果要保存滑动窗口内的所有数据,有时需要进行磁盘存取,这是不适合数据流特点的。为减少滑动窗口内数据所占用的空间,另一种方法以降低滑动窗口上计算的精度为代价,使用小于滑动窗口内数据体积的空间,支持滑动窗口上计算的增量式更新。算法 StaStream[47]将数据流划分为小的固定长度的段(bucket),对每个段仅保存段内数据的概要信息。滑动窗口在这些段上滑动,当流入的数据积累成一段时,抽取这一段的概要信息,将其加入滑动窗口,并从滑动窗口中删除最早的段。内存只需保存滑动窗口多个段的概要信息,此时滑动窗口的增量式更新粒度由一个数据项增大为一个数据段。这种方法通常只支持大小为段整数倍的滑动窗口上的计算[22]。

国内学者董逸生教授研究小组根据数据流的特点,提出了一种发现滑动窗口中频繁闭合模式的新方法:DS-CFI[48]算法。此算法将滑动窗口分割为若干个基本窗口,以基本窗口为更新单位,利用已有的频繁闭合模式挖掘算法计算每个基本窗口的潜在频繁闭合项集,将此频繁闭合项集存储在概要数据结构 DSCFI-tree 中,从而可快速挖掘滑动窗口中的所有频繁闭合模式。

5.3 多窗口技术及相应算法

基于滑动窗口的方法一般都要求用户事先指定窗口的大小,算法在运行过程中只能给出此滑动窗口上的计算结果。而在许多应用中,用户可能在线提出某个窗口上的挖掘请求,此窗口的大小没有事先确定,而且窗口的终点也不是当前时刻。为了支持这样的应用需求,学者们提出一种多窗口方法,支持用户的在线挖掘请求。

多窗口技术在内存或磁盘中保存数据流上多个窗口内数据的概要信息。如 CluStream 算法所使用的 pyramidal 时间框架[10],但此算法每个窗口的范围都是从数据流起始点到窗口建立的时刻点,所以窗口内的数据存在重叠。

5.4 衰减因子及采用衰减因子的代表算法

数据流中实时在线的特点使所蕴涵的知识更有可能随着时间改变,一个现在的不频繁项集将来可能变成频繁项集,因此不能将其忽略。一个频繁项集也可能变成不频繁项集。引入时间衰减因子来消除历史数据对当前计算结果的影响。在这种方法中,每个数据项都被赋予一个随时间不断减小的衰减因子[49],数据项的值与衰减因子相乘后再参与计算。因此,数据项对计算结果的影响随时间的推移逐渐减小。

estDec[50]算法就是此类算法的代表。它利用时间衰减因子 d 来消除废弃数据对现在结果的影响。其中衰减因子 d 由衰减基 b 和衰减基周期 h 计算如下:$d=b^{-(1/h)}$。衰减基 b 决定了每个衰减单元(即固定的窗口大小)衰减的幅度,衰减基 b 取值越大,说明衰减的程度越强,而衰减基周期 h 是数据项集从 1 衰

减为 1/b 所经过的衰减单元的数量,由 b 和 h 就可决定 d。estDec 算法维护着一个监控树,监控树把频繁项集加入树中,最后把保留的从根到叶结点作为频繁模式输出。整个算法分为 4 个阶段:(1)项集计数更新阶段;(2)项集插入阶段;(3)频繁项集选择阶段;(4)剪枝阶段。在项集计数更新阶段和项集插入阶段都利用了时间衰减因子 d 来消除废弃数据对现在结果的影响。即随着时间的变化,原本是频繁项可能会变成非频繁项。引用衰减因子 d 消除了频繁项误报的可能。这种方法实现简单,但计算结果的意义不是非常明确。在使用滑动窗口的算法中,用户明确地知道是对哪些数据进行处理,而在使用衰减因子的方法中,每项数据都只是部分地参与了运算,用户无法确定计算结果由哪些数据得到。此外,所有衰减因子的算法都只考虑了以时间为轴心的衰减,衰减因子并不能根据应用领域灵活设定,将来的工作可考虑以其他参数为轴心的衰减因子。

5.5 自适应地启发式算法

由于数据流是动态变化的,处理数据流的算法必须能够根据数据分布的变化以及数据流流速的变化自动调节算法的处理策略。在数据流频繁模式挖掘中,批处理方法平均处理时间短,但需要积攒足够的数据,实时性差且查询粒度粗,而启发式方法可以直接处理数据流,但处理速度慢。鉴于此原因,文[13]提出一种自适应地启发式算法 FPIL-Stream,它基于一种改进的字典树结构 IL-tree,在更新模式和生成新模式的过程中,可以快速定位历史模式。此算法结合了倾斜窗口策略,可以详细记录历史信息,并且提供了更细的查询粒度。

5.6 倾斜时间窗口

在数据流分析中,人们通常对细尺度上的当前变化感兴趣,但在粗尺度上对长期变化感兴趣。例如,在股票交易中,用户感兴趣的股市波动规律通常是最近几天,甚至是最近几分钟的规律,而对较早的交易数据不会特别重视。所以可在不同粒度层上记录时间,最近的时间在最细的粒度上记录,较远的时间在较粗的粒度上记录,粗略程度取决于应用,这样的时间维模型叫做倾斜时间框架[51](tilted time frame)。这种模型对许多分析任务来说是足够的,也能保证驻留在内存中的数据最小。倾斜时间框架主要有三种模型:(1)自然倾斜时间框架模型;(2)对数倾斜时间框架模型;(3)渐进对数倾斜时间框架模型。

基于 FP-tree 模型的算法 FP-stream[9]充分结合了倾斜时间窗口和 FP-Growth 的特点,利用自然倾斜时间窗口改进而成的对数倾斜时间窗口维护频繁模式以解决时效问题。此算法研究在数据流中构造、维护和更新 FP-stream 结构的有效算法,提出了计算和维护所有频率模式的方法,并动态更新它们。建立一个框架来挖掘带近似支持度的时间敏感模式,为每个模式在多时间粒度上增量维护一个倾斜时间窗口,在这种框架下可以构建和回答感兴趣的查询。

6 结语

数据流频繁模式挖掘是当前一个非常重要的研究方向,在许多领域有着广泛的应用,该文总结了最近几年国内外在该领域的研究成果,介绍了数据流及其频繁模式挖掘的相关特点,以及针对这些特点所提出的相关技术。基于目前挖掘数据流频繁模式的研究现状,以下方面的研究将得到更多的关注:(1)探索改进的概要数据结构;(2)对趋势和变化的实时检测[12,52-53];(3)对树、图等更加复杂模式的挖掘[54];(4)高效的异常挖掘算法;(5)数据流选择因子的估算方法 ε 和 δ 等。

参考文献:

[1] Babcock B,Babu S,Datar M,et al.Models and issues in data stream systems[C]//Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART-SIGART Symposium on Principles of Database Systems.Madison,USA:ACM Press,2002:1-16.

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

[3] Mouratdis K I.Data stream processing:An overview of recent research[D].Hong Kong:Hong Kong University of Science and Technology,2003.

[4] Henzinger M R,Raghavan P,Rajagopalan S.Computing on data streams,SRC Technical Note 1998-011[R].Digital Systems Research Center:Palo Alto,California,1998.

[5] O'Callaghan L,Mishra N,Meyerson A,et al.Streaming-data algorithms for high-quality clustering[C]//Proc of IEEE International Conference on Data Engineering,2002.

[6] Guha S,Mishra N,Motwani R,et al.Clustering data streams[C]//Proc of IEEE Symposium on Foundations of Computer Science(FOCS'00),2000:71-80.

[7] Guha S,Meyerson A,Mishra N,et al.Clustering data streams:Theory and practice[J].IEEE Transactions on Knowledge and Data Engineering,2003,15(3):515-528.

[8] The Stream Group.Stanford data stream management system(Lastest Overview Paper)[EB/OL].(2005).http://db.stanford.edu.

[9] Giannella C,Han Jia-wei,Jian Pei,et al.Mining frequent patterns in data streams at multiple time granularities[C]//Proc of the NSF Workshop on Next Generation Data Mining,2002.

[10] Aggarwal C,Han J,Wang J,et al.A framework for clustering evolving data streams[C]//Proc of Int Conf on Very Large Data Bases(VLDB'03),Berlin,Germany,2003:81-92.

[11] Dora Cai Y,Clutter D,Pape G,et al.MAIDS mining alarming incidents from data streams[C]//Proc of the 23rd ACM SIGMOD,Paris,France,2004.

[12] Dong G,Han J,Lakshmanan L V S,et al.Online mining of changes from data streams:Research problems and preliminary result[C]//Proc of ACM SIGMOD Workshop on Management and Processing of Data Streams,2003.

[13] 张昕,李晓光,王大玲,等.数据流中一种快速启发式频繁模式挖掘方法[J].软件学报,2005,16(12):2099-2105.

[14] 孟小峰,周龙骧,王珊.数据库技术发展趋势[J].软件学报,2004,15(12):1822-1836.

[15] 徐利军,谢康林,徐虹.基于数据流的频繁集挖掘[J].上海交通大学学报,2006,40(3):502-506.

[16] 赵峰,李庆华,金莉.多维流序列并行算法研究[J].小型微型计算机系统,2007,28(2):333-336.

[17] 董亚波,陈宇峰,鲁东明,等.面向大规模网络的聚集 TCP 流量模拟方法研究[C]//全国网络与信息安全技术研讨会,2005:136-142.

[18] 潘云鹤,王金龙,徐从富.数据流频繁模式挖掘研究进展[J].自动化学报,2006,32(4):594-602.

[19] Papadimitriou S,Sun J,Faloutsos C.Streaming pattern discovery in multiple time-series[C]//Proc of the 31st VLDB Conf,2005:697-708.

[20] Sakurai Y,Papadimitriou S,Faloutsos C.BRAID:Stream mining through group lag correlations[C]//Proc of the 2005 ACM SIGMOD

Intl Conf on Management of Data,2005:599-610.

[21] Zhu Y,Shasha D.StatStream:Statistical monitoring of thousands of data streams in real time[C]//Proc of the 28th VLDB Conf,2002:358-369.

[22] Guha S,Gunopulos D,Koudas N.Correlating synchronous and asynchronous data streams[C]//Proc of The 9th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining,2003:529-534.

[23] Yang J.Dynamic clustering of evolving streams with a single pass[C]//Proc of the 19th IEEE Intl Conf on Data Engineering(ICDE'03),2003:695-697.

[24] Dai B R.Clustering on demand for multiple data streams[C]//Proc of the Fourth IEEE Intl Conf on Data Mining(ICDM'04),2004:367-370.

[25] Agrawal R,Srikant R.Fast algorithms for mining association rules[C]//Proceedings of the 1994 Very Large Data Bases.Santiago de,Chile:Morgan Kanfmann,1994:487-499.

[26] 唐懿芳,钟达夫,严小卫.基于聚类模式的数据清洗技术[J].计算机应用,2004,24(5):116-119.

[27] Cormode G,Muthukrishnan S.What's hot and what's not:Tracking nost frequent items dynamically[J].ACM Trans on Database Systems,2005,30(1):249-278.

[28] Vitter J S.Random sampling with a reservoir[J].ACM Transactions on Mathematical Software(TOMS),1985,1(11):37-57.

[29] Gibbons P B,Matias Y.New sampling-based summary statistical for improving approximate query answers[C]//Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data.Washington D C,USA:ACM Press,1998:331-342.

[30] Garofalakis M,Gibbons P B.Wavelet synopses with error guarantees[C]//Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data.Madison,USA:ACM Press,2002:476-487.

[31] Chakrabarti K,Garofalakis M,Rastogi R,et al.Approximate query processing using wavelet[C]//Proceedings of the 2000 International Conference on Very Large Data Bases.Cairo,Egypt:Morgan Kaufmann,2000:111-122.

[32] Gilbert A C,KOtidis Y,Muthukrishnan S,et al.Surfing wavelets on streams:One pass summaries for approximate aggregate queries[C]//Proceedings of 2001 International Conference on Very Large Data Bases.Roma,Italy:Morgan Kanfmann,2001:79-88.

[33] Ioannidis Y E,Poosala V.Histogram-based approximation of set-valued query-answers[C]//Proceedings of the 1999 International Conference on Very Large Data Bases.Edinburgh,UK:Morgan Kaufmann,1999:174-185.

[34] Guha S,Koudas N,Shim K.Data-streams and histograms[C]//Proceedings of the 33rd Annual ACM Symposium on Theory of Computing.Hersonissos,Greece:ACM Press,2001:471-475.

[35] Guha S,Koudas N.Approximating a data stream for querying and estimation:Algorithms and performance evaluation[C]//Proceedings of the 18th International Conference on Data Engineering.San Jose,USA:IEEE Press,2002:567-576.

[36] Bloom B.Space/time tradeoffs in hash coding with allowable errors[J].Communications of the ACM,1970,13(7):422-426.

[37] Estan C,Varghese G.New directions in traffic measurement and accounting[C]//Proceedings of the First ACM SIGCOMM Workshop on Internet Measurement.San Francisco,USA:ACM Press,2001:75-80.

[38] Cormode G,Muthukrishnan S.An improved data stream summary:The count-min sketch and its applications[J].Journal of Algorithms,2005,55(1):58-75.

[39] Thorup M,Zhang Y.Tabulation based 4-universal hashing with applications to second moment estimation[C]//Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms.New Orleans,USA:ACM Press,2004:615-624.

[40] Cormode G,Muthukrishnan S.Summarizing and mining skewed data streams[C]//Proceedings of the Fifth SIAM International Conference on Data Mining.Newport Beach,USA:Society for Industrial and Applied,2005.

[41] Gaber M M,Aaslavsky A,Krishnaswamy S.Mining data streams:A review[J].ACM SIGMOD Record,2005,34(2):18-26.

[42] Jin C,Qian W,Zhou A.Analysis and management of streamlining data:A survey[J].Journal of Software,2004,15(8):1172-1181.

[43] Charikar M,Chen K,Farach-Colton M.Finding frequent items in data streams[J].Theoretical Computer Science,2004,312:3-15.

[44] Manku G S,Motwani R.Approximate frequency counts over data streams[C]//Proc of the 28th VLDB Conf,2002:346-357.

[45] Lee C H,Lin C R,Chen M S.Sliding-window filtering:An efficient method for incremental mining on a time variant database[J].Inform System,2005,30(3):227-244.

[46] Hulten G,Spencer L,Domingos P.Mining time-changing data streams[C]//Proc of the 7th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining,2001:97-106.

[47] Zhu Y,Shasha D.StatStream:Statistical monitoring of thousands of data streams in real time[C]//Proc of the 28th VLDB Conf,2002:358-369.

[48] 刘学军,徐宏炳,董逸生,等.基于滑动窗口的数据流闭合频繁模式的挖掘[J].计算机研究与发展,2006,43(10):1738-1743.

[49] Aggarwal C C.A framework for projected clustering of high dimensional data streams[C]//Proc of the 30th VLDB Conf,2004:852-863.

[50] Chang J H,Lee W S.Finding recent frequent itemsets adaptively over online data streams[C]//Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.Washington,USA:ACM Press,2003:487-492.

[51] Chen Y,Dong G,Han Jia-wei,et al.Multi-dimensional regression analysis of time-series data streams[C]//Proc 2002 Int Conf Very Large Data Bases(VLDB'02),2002:323-334.

[52] Aggarwal C C.A framework for diagnosing changes in evolving data streams[C]//Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data.San Diego,USA:ACM Press,2003:575-596.

[53] Ben-David S,Gehrke J,Kifer D.Detecting change in data streams[C]//Proceedings of the 30th International Conference on Very Large Data Bases.Toronto,Canada:Morgan Kaufmann,2004:180-191.

[54] Chen G,Wu X,Zhu X.Sequential pattern mining in multiple streams[C]//Proceedings of the Fifth IEEE International Conference on Data Mining.Houston,USA:IEEE Press,2005:585-588.