

基于时间滑动窗口的自适应加权随机抽样算法

唐 达^{*1}, 刘 畅¹, 岳前进², 张建英¹

(1. 大连理工大学 计算机科学与技术学院, 辽宁 大连 116024;

2. 大连理工大学 工程力学系, 辽宁 大连 116024)

摘要: 为了构建传感器网络流数据的概要数据, 给出了一种基于时间滑动窗口的自适应加权随机抽样算法: AWRS/BTSW 算法. 算法根据流数据的到达时间和变化情况, 赋予流数据一定的键值, 根据流数据的键值并结合 skipping 因子生成概要数据. 在深海平台监测系统中, 流数据变化不确定, 算法可以根据数据的变化情况动态调整抽样方式, 在数据变化不稳定的情况下, 生成概要数据的准确性高; 在数据变化稳定的情况下, 生成概要数据的效率高.

关键词: 流数据; AWRS/BTSW 算法; 键值; skipping 因子

中图分类号: TP31 **文献标志码:** A

0 引言

深海平台监测系统通过分析传感器网络上采集到的数据, 来确定不同的因素对该平台的影响. 从传感器上采集到的数据是随时间不断变化的, 具有连续、无界两个特点, 即为流数据^[1]. 深海平台监测系统需要管理这类流数据, 以使用户查询. 对流数据的查询通常是连续的, 当新数据到达时增量式地返回结果. 在深海平台监测系统中, 用户查询仅仅需要一个近似结果, 并不需要获得精确的查询值. 由于存储容量的有限性, 将所有流数据都保存起来不现实, 也没有必要. 为对流数据的存储, 需要设计一个保存原始流数据的特征, 规模上远小于原流数据的概要数据结构(synopsis data structure)^[2]. 本文对此进行研究.

1 概要数据构建技术

概要数据构建技术主要有直方图(histogram)技术^[3]、小波(wavelet)技术^[4]和抽样(sampling)技术. 直方图技术只能反映数据的大致轮廓和分布特征, 小波技术是从数据集中抽取小部分数据样本, 并根据样本近似恢复数据集, 相对复杂, 两种构建技术构建的概要数据均不能满足流数据快速、连续查询的要求. 抽样技术生成概要数据的效

率高, 能满足流数据快速、连续查询的要求, 可以很好地应用到深海平台监测系统上. 现有抽样算法主要有均匀抽样(uniform sampling)、偏倚抽样(biased sampling)和权值抽样.

均匀抽样: 各元组被抽取到的概率是相等的. Vitter^[5]提出水库抽样方法, 假定数据集的总量为 A , 以 S/A 的概率抽取 S 个元组到样本集合中, 当抽取的元组数量超过 S 时, 随机删除样本集合里的一个元组, 然后将抽取出来的新元组插入到样本集合中. Gibbons 等^[6]优化了水库抽样方法, 给出了精确抽样方法, 将数据元组用 (T, C) 表示, 其中 T 表示数据元组, C 表示数据元组的个数, 在相同数据元组多的数据集合中使用该方法, 能有效地节省空间.

偏倚抽样: 各元组被抽取到的概率是不同的. Gibbons 等^[6]在精确抽样方法的基础上给出了计数抽样方法, 在抽取的元组数量超过 S 时, 将出现次数少的元素从样品集合中替换掉, 可以很方便地获得一个集合中的常用数据元组.

权值抽样: 各元组根据一定的权值进行抽样. Efraimidis 等^[7]认为均匀抽样没有区分数据元组的重要性. 其给出了加权抽样算法, 赋予数据元组相应的权值, 并根据权值进行抽样. Zhang 等^[8]认

为抽样要考虑数据的时间因素,其在加权抽样算法的基础上,综合计算数据元组的到达时间和权值,作为该数据元组的优先数,再根据优先数抽样,形成优先数随机抽样算法(PRS),解决了数据元组过期问题. Hou 等^[9]为解决多个流数据之间的多次连接操作效率低下的问题,提出一种针对多个相关流数据的概要数据生成算法.

上述抽样算法均未能考虑到流数据变化的特点,但加权抽样算法能根据用户赋予数据的权值,来衡量数据的重要性,并根据重要性进行抽样. 在深海平台监测系统中,用户事先并不知道何时到达的数据重要,只能根据经验,因此,抽取的概要数据的准确性依赖于用户的经验. 为此本文给出一种基于时间滑动窗口的自适应加权随机抽样算法: AWRS/BTSW (adaptive weighted random sampling based on time sliding window)算法,该算法通过计算流数据的平均变化率,来确定一个数据元组的权值以及 skipping 因子的值,结合 skipping 因子、权值和数据元组的到达时间来对数据进行抽样,解决了现有抽样算法生成的概要数据与原始数据的误差不确定以及数据过期问题.

2 AWRS/BTSW 算法

2.1 相关定义

定义 1 数据平均变化率 假设在时刻 i 的数据为 t_i , 则时刻 i 的数据变化率 Δi 为 $\left| \frac{t_i - t_{i-1}}{t_{i-1}} \right|$, 则从时刻 m 到 n 之间的时间段 Δt 的数据平均变化率为 $\overline{\Delta i} = \frac{\sum_{i=1}^{\Delta t} \Delta i}{\Delta t}$.

定义 2 skipping 因子 若时间段 Δt 内的数据平均变化率小于阈值 ξ , 该时间段的所有数据元组的 skipping 因子的值为 true, 否则为 false. skipping(Δt) 函数定义如下:

$$\text{skipping}(\Delta t) = \begin{cases} \text{true}; & \overline{\Delta i} < \xi \\ \text{false}; & \overline{\Delta i} \geq \xi \end{cases}$$

定义 3 数据集的稳定度 将数据集分成若干个数据区间, d_s 表示数据平均变化率不超过阈值的数据区间的个数, d_a 表示总数据区间的个数, 则该数据集的稳定度为 $s = d_s / d_a$.

定义 4 相对误差 从原始数据里查询的结果定义为 Q_r , 从概要数据里查询的结果定义为 Q_s , 则相对误差为 $e = Q_s / Q_r$.

2.2 数据项的键值

从流数据变化特点出发, 根据流数据的平均变化率赋予数据项相应的权值, 令 $w(x)$ 为单调递增函数, $w(x)$ 函数取 $x^{1/\lambda}$ (λ 为正整数), 数据变化越快, 赋予的权值就越大. 权值计算公式如下:

$$w_{\Delta i} = w(\overline{\Delta i}) = \overline{\Delta i}^{1/\lambda} \quad (1)$$

结合基本窗口技术, 综合考虑权值和时间因素并将其作为数据项的键值, 解决时间滑动窗口的数据过期问题. 键值计算公式如下:

$$x_i = \alpha g(t_i) + \beta f(w_i, \mu_i) \quad (2)$$

其中 α 和 β 是权衡数据到达时间和权值的两个参数. 用 v_i 表示数据流中的第 i 个数据项, x_i 表示数据项 v_i 的键值, t_i 表示数据项 v_i 的到达时间, w_i 表示数据项 v_i 的权值, μ_i 表示数据项 v_i 生成的一个随机数. $g(x)$ 为单调递增函数, 数据到达时间越早, 它的值就越小. 对于变量 $x, y, f(x, y)$ 是一个单调递增函数, 例如 x^y .

当新数据到达时, 时间窗口向后移, 周期性地计算当前滑动窗口的数据项的键值, 可以将滑动窗口平均分为 k 个子窗口 ($s[0], s[1], \dots, s[k-1]$). 假设 v_i 在子窗口 $s[m]$, 则当前滑动窗口的各个数据项 Δt 的键值 x_i 计算如下:

$$x_i = \alpha g(k - m + 1) + \beta f(w_{\Delta i}, \mu_i) \quad (3)$$

假定 l 是一个正整数, $g(x)$ 取 $x^{-1/l}$, $f(x, y)$ 取 $y^{1/x}$, 代入式(3)可得

$$x_i = \alpha(k - m + 1)^{-1/l} + \beta \mu_i^{1/w_{\Delta i}} \quad (4)$$

由式(1)可得

$$x_i = \alpha(k - m + 1)^{-1/l} + \beta \mu_i^{1/\overline{\Delta i}^{1/\lambda}} \quad (5)$$

新到来的 Δt 时间段的数据, 将其放入 $s[k]$ 中, 将 $m = k$ 代入式(5) 计算可得

$$x_i = \alpha + \beta \mu_i^{1/\overline{\Delta i}^{1/\lambda}} \quad (6)$$

2.3 AWRS/BTSW 算法步骤

当 Δt 时间段的数据到达时, 首先计算该时间段的数据平均变化率 $\overline{\Delta i}$, 然后根据数据平均变化率计算其 skipping 因子. 如果 skipping 因子的值为真, 则滑过这些元组; 否则, 根据数据平均变化率给这段数据项赋予一定的权值, 并结合到达时间计算这段数据项的键值; 最后再根据键值进行抽样. 具体算法如下:

输入: 包含 n 个数据项的流数据 S

输出: 基于时间的滑动窗口 T 的概要数据集 R

(1) 将新到达的时间跨度为 T 的 h 个数据项加入到当前时间滑动窗口;

(2) 将当前时间滑动窗口 T 按照 Δt 的时间间隔, 平均切分为 k 个子窗口 ($s[0], s[1], \dots, s[k-1]$), $k = T/\Delta t$;

(3) 分别计算每个子窗口的平均数据变化率 $\overline{\Delta i}$ 及各个数据项的权值 w_i ;

(4) 计算各个数据项的键值 x_i ;

(5) for($i = h + 1; i \leq n; ++i$), 重复步骤 (5) ~ (13);

(6) 计算下一个 Δt 时间间隔的数据平均变化率 $\overline{\Delta i}$;

(7) if skipping(Δt) == true, 跳跃 Δt 个时间跨度段, 转步骤 (5);

(8) 将 Δt 时间跨度的数据项读取到 $s[k]$;

(9) 计算 Δt 时间跨度里的各个数据项的键值 x_i ;

(10) for($j = 1; j \leq \Delta t; ++j$), 重复步骤 (10) ~ (12);

(11) 查找当前滑动窗口中键值最小的数据项, 假设最小键值为 MIN , 并且有 $R[m] = MIN$;

(12) if $x_i > MIN - \epsilon$

用键值为 x_i 的数据项替换 $R[m]$;

(13) 窗口向前滑动, $s[i] \rightarrow s[i-1]$ ($i = 1, 2, \dots, k$).

3 算法分析与比较

AWRS/BTSW 算法首先计算平均数据变化率, 时间复杂度为 $o(n)$, 将流数据中的数据元组抽取到样本集, 时间复杂度为 $o(h \log n/h + \Delta t(h-1))$, 算法在数据稳定的情况下滑过 Δt 个时间跨度, 假设流数据中的数据项的稳定度为 s , 则该算法总时间复杂度为 $o(n + (n-h)(1-s)((h \log n/h)/\Delta t + (h-1)))$, 其中 $s \in [0, 1]$, $\Delta t \in [2, h]$.

从算法的时间复杂度计算公式可以得出, 该算法依赖数据集的稳定度 s 和时间跨度 Δt . 从深海平台监测系统取出 3 万个数据, 用该算法生成 5 000 个概要数据. 其中测试环境为操作系统 Windows XP、CPU 2.6 GHz Pentium 4、1 GB 内存.

对稳定度相同 (假设数据集的稳定度为 25%), 时间跨度 Δt 不同 (分别为 5 000、3 000、1 000、900、800、100、2 s) 的数据集进行实验, 生成概要数据所用时间如图 1 所示.

在时间跨度 Δt 相同 (时间跨度 Δt 设置为 5

s), 稳定度不同 (分别为 0、10%、25%、50%、75%) 的数据集中, 将 AWRS/BTSW 算法和优先数随机抽样 (PRS) 算法生成概要数据所用的时间和准确性进行对比, 其中 PRS 算法中的数据项的权值用随机函数生成, 分别运行两种算法并求其算术平均值. 两种算法生成概要数据所用时间和相对误差如图 2、3 所示.

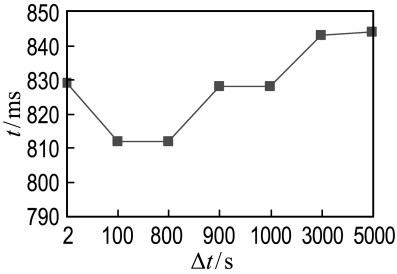


图 1 AWRS/BTSW 算法所用时间

Fig. 1 AWRS/BTSW algorithm efficiency

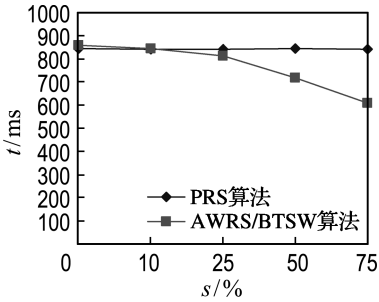


图 2 算法所用时间的比较

Fig. 2 Comparison of algorithm efficiencies

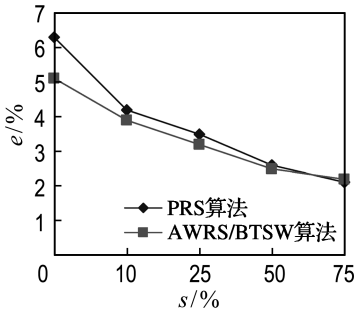


图 3 算法相对误差的比较

Fig. 3 Comparison of accuracy of algorithm

从图 1、2 可以看出, 稳定度相同, AWRS/BTSW 算法所用时间在 2~100 s 时随时间跨度 Δt 的增大而减小, 在 800~5 000 s 时随时间跨度 Δt 的增大而增大. 当时间跨度 Δt 相同时, 数据集的稳定度越高, 该算法生成概要数据的效率越高.

从图 2、3 可以看出, 数据集的稳定度越高, AWRS/BTSW 算法生成概要数据的效率越高;

数据集的稳定度越低,与 PRS 算法相比该算法的准确性越高.当数据集的稳定度在 10% 以上时,该算法的效率和准确性都比 PRS 算法高;当数据集的稳定度在 10% 以下时,该算法的准确性比 PRS 算法高,效率略低.

4 结 论

本文总结和分析了概要数据构建的几种抽样方法,给出了一种基于时间滑动窗口的自适应加权随机抽样算法:AWRS/BTSW 算法,解决了目前抽样算法在深海平台检测系统等数据变化不确定的应用中,抽取出的概要数据的准确性与原始数据之间的误差不确定的问题.与其他的抽样算法相比,该算法能根据数据变化的情况,动态生成概要数据,在数据变化稳定的情况下生成概要数据的效率高,在数据变化剧烈的情况下生成概要数据的准确性高,相对误差较小.

参考文献：

[1] Babcock B, Babu S, Datar M, *et al.* Models and issues in data streams [C] // **Proceedings of the Twenty-First ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems**. New York:ACM, 2002:1-16.
[2] Gibbons P B, Matias Y. Synopsis data structures for massive data sets [C] // **Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms**.

Philadelphia:SIAM, 1999.
[3] Gibbons P B, Matias Y, Poosala V. Fast incremental maintenance of approximate histograms [J]. **ACM Transactions on Database Systems**, 2002, **27**(3):261-298.
[4] Matias Y, Vitter J S, Wang M. Wavelet-based histograms for selectivity estimation [J]. **ACM SIGMOD Record**, 1998, **27**(2):448-459.
[5] Vitter J S. Random sampling with a reservoir [J]. **ACM Transactions on Mathematical Software**, 1985, **11**(1):37-57.
[6] Gibbons P B, Matias Y. New sampling-based summary statistics for improving approximate query answers [C] // **Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data**. New York:ACM, 1998:331-342.
[7] Efraimidis P S, Spirakis P G. Weighted random sampling with a reservoir [J]. **Information Processing Letters**, 2006, **97**(5):181-185.
[8] ZHANG Long-bo, LI Zhang-huai, ZHAO Yi-qiang, *et al.* A priority random sampling algorithm for time-based sliding windows over weighted streaming data [C] // **Proceedings of the 2007 ACM Symposium on Applied Computing**. New York: ACM, 2007: 453-456.
[9] HOU Wei, YANG Bing-ru, WU Chen-shen, *et al.* RedTrees: A relational decision tree algorithm in streams [J]. **Expert Systems with Applications**, 2010, **37**(9):6265-6269.

Adaptive weighted random sampling algorithm based on time sliding window

TANG Da^{*1}, LIU Chang¹, YUE Qian-jin², ZHANG Jian-ying¹

(1. School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China;
2. Department of Engineering Mechanics, Dalian University of Technology, Dalian 116024, China)

Abstract: To obtain the synopsis data of the stream data from the sensor networks, an adaptive weighted random sampling algorithm, AWRS/BTSW algorithm, is provided based on time sliding window. Firstly, the algorithm assigns a key value to a stream data according to its arrival time and changes; then, based on the key value and skipping factor, the algorithm generates synopsis data. In the deep sea platform monitoring system, the change of stream data is unknown. The algorithm can dynamically adjust sampling according to data variation. If the data is stable, it generates synopsis data efficiently. Even if the data is unstable, it also generates accurate synopsis data.

Key words: stream data; AWRS/BTSW algorithm; key value; skipping factor