ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.2, February 2007, pp.303−310
DOI: 10.1360/jos180303

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

Web 仓储中视图变化频率的自适应估测*

张 岩[1+], 唐世渭[1], 杨冬青[2], 李晓明[2]

[1](视觉与听觉信息处理国家重点实验室(北京大学),北京 100871)

[2](北京大学 计算机科学技术系,北京 100871)

Self-Adaptive Estimation of View Change Frequency in Web Warehouses

ZHANG Yan[1+], TANG Shi-Wei[1], YANG Dong-Qing[2], LI Xiao-Ming[2]

[1](State Key Laboratory on Machine Perception (Peking University), Beijing 100871, China)

[2](Department of Computer Science and Technology, Peking University, Beijing 100871, China)

+ Corresponding author: Phn: +86-10-62755592, E-mail: zhy@cis.pku.edu.cn, http://www.cis.pku.edu.cn/teacher/system/zhangyan/

Zhang Y, Tang SW, Yang DQ, Li XM. Self-Adaptive estimation of view change frequency in Web warehouses. *Journal of Software*, 2007,18(2):303−310. http://www.jos.org.cn/1000-9825/18/303.htm

Abstract: Refreshing materialized views is a main task of Web warehouse maintenance. As the refreshing scheme depends heavily on the base data change frequency, researchers have presented many corresponding algorithms and frequency estimators for it. Although these estimators really work, however, all of them have limitations. The bias that an estimator introduces will increase significantly when the estimated value is out of its applicable range. In this paper, a self-adaptive algorithm is presented based on Poisson process analysis, which can adjust the revisiting pattern and revisiting frequency according to the estimated change frequency. This algorithm can also tune the parameters so that the estimated value will fall into the best applicable range of the estimator. According to the experimental results, the proposed estimator is more accurate than the ones in the previous work.

Key words: Web warehouse; Webview; base data; change frequency; Poisson process

摘 要: 物化视图的刷新是 Web 仓储进行系统维护的一项主要任务,而基础数据变化频率则是刷新方案中的重要因素.在已有文献中,研究者已经给出一些关于基础数据变化规律的算法和估测器.虽然这些估测器取得了不错的效果,然而他们却忽略了这些估测器都有一定的适用范围,超出这个范围则效果急剧下降.在此,基于泊松过程进行分析,对估测器的适用范围进行了讨论,根据估测结果的偏离值和有效性对估测公式进行参数调整,同时根据估测值的大小不断调整数据源的访问频率和次数,从而使数据源访问模式和估测器互相适应,使估测器在最佳估测范围内获得估测值.实验结果表明,与已有文献中的方法相比,新提出的自适应估测算法能够取得更好的效果.

关键词: Web 仓储;Web 视图;基础数据;变化频率;泊松过程

中图法分类号: TP311 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant Nos.60573166, 60673129 (国家自然科学基金)
Received 2004-04-23; Accepted 2006-03-31

1 Web 数据变化模型

WWW 的迅猛发展,使其成为全球信息传递与共享日益重要和最具潜力的资源.如何有效地利用这个巨大的信息资源,已经成为众多研究者所面临的新课题.Web 仓储(Web warehouse)使用数据仓库方法,极大地提高了系统查询速度,为很多领域的应用打开了方便之门,尤其适合联机分析处理、决策分析等应用.然而,物化视图(materialized view)的维护成为系统的一项重要工作.在 Web 环境下,用户访问非常频繁,数据变化频率也非常高,那么,在 Web 仓储中如何提高数据的时新性(freshness),使集成数据和基础数据保持同步,这对于提高集成数据的质量、增加用户的满意度,有着重要的意义.一般情况下,数据源不会把自己的变化通知我们,我们只能通过探测的方法来监视 Web 数据的变化[1].如果能够得到 Web 数据变化的合理数学模型,那么,对于我们的研究将是巨大的支持并具有重要的意义.

在本文中,我们将假设 Web 数据的变化遵循 Poisson 过程.这是有理论依据的.Poisson 过程经常被用来描述一个随机事件序列,这些事件以固定频率重复独立发生.例如,一个城市车祸发生的情况、大型超市顾客到来的情况以及热线中心电话的多少等等,都可以用 Poisson 过程来描述.对于 Web 页面的变化来说,从一个较长的观察期来看,页面数据的变化对于 Web 仓储系统来说是随机事件,并且不断重复发生.如果变化频率在一定的时间段内基本上是固定的,那么,我们就可以应用 Poisson 过程来表示[2].

下面我们看一下 Poisson 过程的精确描述.令 $X(t)$表示在时间段$(0,t)$内一个特别事件的发生次数,那么,如果该事件随机地(randomly)、独立地(independently)以一个固定的平均频率(a fixed average rate)λ发生,它就被称为一个频率为λ的 Poisson 过程.在 Poisson 过程中,随机变量 $X(t)$有如下特性[3]:

1. 对于任意时间点 $t_0=0<t_1<t_2<\ldots<t_n$,过程增量(发生在某个时间间隔内的事件数)$X(t_1)-X(t_0)$, $X(t_2)-X(t_1),\ldots,X(t_n)-X(t_{n-1})$是独立的随机变量.
2. 对于任意的 $s\geq 0$ 和 $t>0$,随机变量 $X(s+t)-X(t)$具有如下 Poisson 概率分布:

$$\Pr\{X(s+t)-X(s)=k\}=\frac{(\lambda t)^k}{k!}e^{-\lambda t},\quad k=0,1,\ldots$$

3. $X(0)=0$.

我们可以如下验证单位时间内事件发生的次数:

$$E[X(t+1)-X(t)]=\sum_{k=0}^{\infty}k\cdot\Pr\{X(t+1)-X(t)=k\}=\lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=\lambda e^{-\lambda}e^{\lambda}=\lambda.$$

这说明,尽管一个事件可以随机地在任何时刻发生,但对于一个 Poisson 过程来说,其平均频率固定为λ.

2 Web 页面变化频率的估测

2.1 估测器的概念

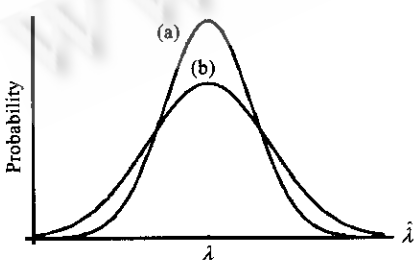

Fig.1 Two possible distribution for the estimator $\hat{\lambda}$

图 1 估测器 $\hat{\lambda}$ 的两种可能分布

在设定 Web 数据的变化模型为 Poisson 过程后,对其变化的估测任务也就是通过对基础数据的反复访问,得到其变化的平均频率λ.比如,我们在 T 时间段内观测到 X 次变化,那么,我们就可以把 X/T 作为变化频率的估测值,记作 $\hat{\lambda}=X/T$.我们把 $\hat{\lambda}$ 称为频率λ的估测器.由于 X 是一个随机变量,所以估测器 $\hat{\lambda}$ 也是一个随机变量,并遵从一定的概率分布.图 1 是 $\hat{\lambda}$ 的两种可能分布.

在本文中,我们沿用 Junghoo Cho 在文献[2]中提出的关于 $\hat{\lambda}$ 的 3 个评价指标,分别是无偏性(unbiasedness)、有效性(efficiency)和连贯性(consistency).

1. 无偏性:直观的感觉是希望 $\hat{\lambda}$ 的分布以λ的实际值为中心,数学描述是 $\hat{\lambda}$ 的期望值 $E[\hat{\lambda}]=\lambda$.
2. 有效性:即使 $E[\hat{\lambda}]=\lambda$,$\hat{\lambda}$ 的值也还是有可能偏离λ.我们当然希望这种偏离越小越好.偏离值越小,我们就认为这种分布越有效.图 1 中曲线(a)的分布比曲线(b)更有效.
3. 连贯性:直观地,我们希望当加大统计(观察)样本时,$\hat{\lambda}$ 的值应该逼近λ.其数学描述如下:

令 $\hat{\lambda}_n$ 为样本大小为 n 时的估测器,那么,当且仅当下式满足时,我们称 $\hat{\lambda}_n$ 是连贯的:

$$\lim_{n\to\infty}\Pr\{|\hat{\lambda}_n-\lambda|\le\varepsilon\}=1,\quad \text{对于任何正数}\varepsilon.$$

2.2 具有一般性的估测器

一般情况下,Web 站点只告诉我们其目前状态,而不透露任何历史记录,这时对 Web 页面变化频率进行估测是一件比较困难的事情.文献[2]中作了比较细致的讨论,然而,其中的一些结果不够深入.本文将在这些探讨的基础上走得更远一些.

假设集成系统 WR 在一段时间 T 之内访问了数据源中某元素 n 次,其中探测到该元素变化了 X 次,没有变化的次数为X次.最直观地,如果不考虑我们漏掉的(即未探测到的)那些元素变化情况,我们可以把X/T作为一个该元素变化频率λ的估测器.然而,该估测器的无偏性和有效性都很差,并且不是连贯的.也就是说,即使我们增大对该元素的访问次数,仍然无助于估测结果的提高.因此,我们就必须寻找新的估测器.

一般地,我们记 $r=\lambda/f$,则有 $\hat{\lambda}=\hat{r}\cdot f$,这样我们就可以直接对估测器 $\hat{r}$ 进行讨论,使论述更加方便.令 f 为该元素的探测频率,那么在时间段 $I(=1/f)$内,该元素不会发生变化的概率为

$$q=\Pr\{X(t+i)-X(t)=0\}=\frac{\lambda^0 e^{-\lambda I}}{0!}=e^{-\lambda I}=e^{-\lambda/f}=e^{-r}.$$

因此,$\hat{r}=-\log q=-\log\left(\frac{\overline{X}}{n}\right)=-\log\left(\frac{n-X}{n}\right)$,这样就有估测器 $\hat{r}=-\log\left(\frac{n-X}{n}\right)$.

现在我们来看估测器 $\hat{r}=-\log\left(\frac{n-X}{n}\right)$.注意到,当 $n=X$ 时,$-\log 0=\infty$,为了避免该奇异值的出现,我们对上述估测器进行修正.假设估测器具有这样的形式:$\hat{r}=-\log\left(\frac{n-X+a}{n+b}\right)$,其中,$a,b>0$,那么,显然可以避免当 $n=X$ 时出现奇异值.后面我们会通过实验表明这种假设是有道理的.我们先来推算一下 a 和 b 的值.

我们知道 X 是一个随机变量,由 $q=e^{-r}$ 可得:

$$\Pr\{X=n-i\}=\binom{n}{i}(1-q)^{n-i}q^i=\binom{n}{i}(1-e^{-r})^{n-i}(e^{-r})^i.$$

进而,
$$E[\hat{r}]=E\left[-\log\left(\frac{n-X+a}{n+b}\right)\right]=-\sum_{i=0}^{n}\log\left(\frac{i+a}{n+b}\right)\binom{n}{i}(1-e^{-r})^{n-i}(e^{-r})^i.$$

根据 Taylor 展开[4],有

$$e^{-r}=\sum_{k=0}^{\infty}\frac{1}{k!}(-r)^k=1-r+\frac{r^2}{2!}-\frac{r^3}{3!}+\frac{r^4}{4!}-\cdots,$$

$$1-e^{-r}=1-\sum_{k=0}^{\infty}\frac{1}{k!}(-r)^k=r-\frac{r^2}{2!}+\frac{r^3}{3!}-\frac{r^4}{4!}+\cdots$$

这样,$E[\hat{r}]$ 的上述表示就可以展开为对 r 的无穷级数,

$$E[\hat{r}]=-\sum_{i=0}^{n}\log\left(\frac{i+a}{n+b}\right)\binom{n}{i}(1-e^{-r})^{n-i}(e^{-r})^i=\sum_{k=0}^{\infty}c_k r^k.$$

显然,$E[\hat{r}]$ 的值并不等于 r.然而,为了获得更好的无偏性(unbiasedness),我们希望这个值能够和 r 尽量接近,并且当 $n\to\infty$时,能够有 $E[\hat{r}]=r$(以满足连贯性).这就要求 $c_0=0$,且 $\lim_{n\to\infty}c_1=1$.

通过对 $E[\hat{r}]$ 展开式的分析,有

$$c_0=-\log\left(\frac{n+a}{n+b}\right)\binom{n}{n}=-\log\left(\frac{n+a}{n+b}\right),$$

$$c_1=-\log\left(\frac{n+a}{n+b}\right)\cdot\binom{n}{n}\cdot(-n)-\log\left(\frac{n-1+a}{n+b}\right)\cdot\binom{n}{n-1}\cdot(1)=\log\left(\frac{n+a}{n+b}\right)\cdot n-\log\left(\frac{n-1+a}{n+b}\right)\cdot n=n\cdot\log\left(\frac{n+a}{n-1+a}\right).$$

这样,我们就要使 $-\log\left(\frac{n+a}{n+b}\right)=0$ 以及 $n\cdot\log\left(\frac{n+a}{n-1+a}\right)=1$ 都能尽量满足.

由前式,我们得到 $a=b$;对于后式,我们知道 $\lim_{n\to\infty}\log\left(\frac{n+a}{n-1+a}\right)^n=1$.这样,无论 a 的值如何,当 $n\to\infty$时,后式均可满足.在这种情况下,我们希望找出一个 a 的最佳值,使后式在 n 较小时就能逼近 1.

经过变换,有

$$\frac{n+a}{n-1+a}=e^{\frac{1}{n}},$$

进而可解出

$$a=\frac{n-(n-1)\cdot e^{\frac{1}{n}}}{e^{\frac{1}{n}}-1}.$$

根据 Taylor 展开[4],

$$e^{\frac{1}{n}}=\sum_{k=0}^{\infty}\frac{1}{k!}\left(\frac{1}{n}\right)^k=1+\frac{(1/n)^1}{1!}+\frac{(1/n)^2}{2!}+\frac{(1/n)^3}{3!}+\cdots,$$

所以,

$$a=\frac{e^{\frac{1}{n}}\cdot 0.5-n\cdot e^{\frac{1}{n}}+0.5+n}{e^{\frac{1}{n}}-1}+0.5=0.5+\frac{\sum_{k=0}^{\infty}\frac{(1/n)^k}{(k+2)!}\cdot\left(\frac{1}{2}-\frac{1}{k+3}\right)}{n+\sum_{k=0}^{\infty}\frac{(1/n)^k}{(k+2)!}}.$$

显然,当 $n\to\infty$时,$a\to 0.5$.这样就有 $b=a=0.5$,从而得到估测器:

$$\hat{r}=-\log\left(\frac{n-X+0.5}{n+0.5}\right),$$

或记作

$$\hat{r}=-\log\left(\frac{\overline{X}+0.5}{n+0.5}\right).$$

其中,$\overline{X}=n-X$.这与文献[2]中得到的结果相同,但文献[2]的推导过程存在着明显的不足:它是通过观察图中曲线的收敛趋势而得到 $a=0.5$,并没有严格的推导.

下面,我们从观察这个估测器的偏离性开始我们的分析.图 2 是文献[2]中的效果图,从图中可以看出,$\frac{E[\hat{r}]}{r}$ 随着 r 的增大而迅速减小,当 n 比较小时(如 $n=3$),这种偏离尤为明显.作为估测器,应该尽量保持较小的偏离性,$\frac{E[\hat{r}]}{r}$ 的值最好在 1 的周围.为了获得更好的效果,我们考虑对 $\hat{r}=-\log\left(\frac{\overline{X}+0.5}{n+0.5}\right)$ 作进一步的调整.

我们仍然从估测器的公式 $\hat{r}=-\log\left(\frac{n-X+a}{n+b}\right)$ 开始分析.一方面我们知道,

$$E[\hat{r}]=\sum_{i=0}^{n}\log\left(\frac{n+b}{i+a}\right)\binom{n}{i}(1-e^{-r})^{n-i}(e^{-r})^i,$$

当 n 和 r 的值都保持一定时,增大 $\log\left(\frac{n+b}{i+a}\right)$ 的值可以使 $E[\hat{r}]$ 增大,从而使 $\frac{E[\hat{r}]}{r}$ 值增大,降低估测器的偏离性;

另一方面,我们又想使 $-\log\left(\frac{n+a}{n+b}\right)=0$ 能够尽量得到满足,这样,最好还是保持 $a=b$.于是,希望通过降低 a 的值来达到我们的目的:当 a 减小时,由于 $0\le i\le n$,且 $\frac{n+a}{i+a}=1+\frac{n-i}{i+a}$,所以,$\frac{n+a}{i+a}$ 值增大,从而 $\log\left(\frac{n+a}{i+a}\right)$ 也增大.于是,$\sum_{i=0}^{n}\log\left(\frac{n+b}{i+a}\right)\binom{n}{i}(1-e^{-r})^{n-i}(e^{-r})^{i}$ 值增大.

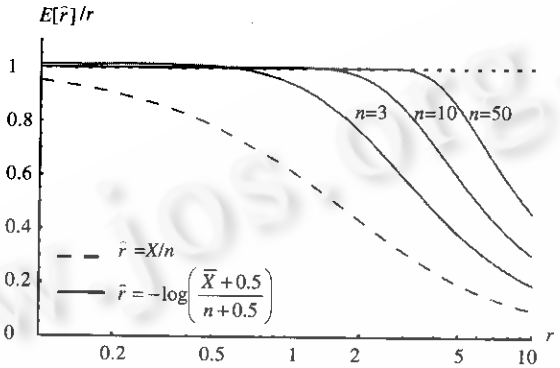

Fig.2 The bias for the estimator $\hat{r}=-\log\left(\frac{\overline{X}+0.5}{n+0.5}\right)$

图 2 估测器 $\hat{r}=-\log\left(\frac{\overline{X}+0.5}{n+0.5}\right)$ 的偏离情况

通过分析我们发现,当 a 取(0,1)之间的任意一个值时,都可以在一定的 r 值范围内取得较好的效果.这样,实际上是得到了一组估测器 $\hat{r}=-\log\left(\frac{\overline{X}+a}{n+a}\right)$,并且,可以根据文献[5]的二项式理论,使用与文献[2,6]中相同的方法证明这组估测器都是连贯的.其中,a 的取值影响着估测器的数学期望和方差,不存在一个值,让这两项指标同时取得最好的效果.如果进行综合比较,那么,a=0.4 可以取得更好的效果.

我们考察 $\hat{r}=-\log\left(\frac{\overline{X}+0.4}{n+0.4}\right)$,把它与 $\hat{r}=-\log\left(\frac{\overline{X}+0.5}{n+0.5}\right)$ 进行比较,结果见表 1.

从表 1 可以看出,$\hat{r}=-\log\left(\frac{\overline{X}+0.4}{n+0.4}\right)$ 确实取得了较好的结果.当 r 增大时,其偏离情况有着明显的改善.

Table 1 Comparison of $\frac{E[\hat{r}]}{r}$ for different values of a when n=3

表 1 当 n=3 时,a 取不同值情况下的 $\frac{E[\hat{r}]}{r}$ 比较

	r=0.1	0.5	1.0	1.5	1.8	2.0
a=0.5	1.011 5	1.006 7	0.951 5	0.862 5	0.804 1	0.765 6
a=0.4	1.050 7	1.063 1	1.021 2	0.934 9	0.875 0	0.834 7

3 自适应估测算法

从上面的分析可以看出,对源数据的访问次数 n 是估算公式中的重要因子.不同的 Web 页面,其源数据的更新频率不同,所以该因子也应该不同.如何根据估测情况合理调整该因子,是本文要考虑的一项工作.

以估测器 $\hat{r}=-\log\left(\frac{n-X+a}{n+a}\right)$ 为例,当 a 取(0,1)中的任意值时,估测器都有自己的适用范围.当 a 减小时,估测器的偏离值减小,但统计方差 σ/r 增大.因为该估测器具有连贯性,理论上我们可以通过提高样本数量 n 来提高估

测的准确性.但是,n 过大显然让我们无法接受.所以,对于不同的 r 值,可以选用不同的 a 值进行估测,从而在保证估测效果的前提下,尽量减少访问次数 n.我们进行估测的指导思想是:在估测过程中,通过对估测公式中参数 a 的调整,同时根据变化频率 λ 值的大小不断调整数据源的访问频率和次数,从而使数据源访问模式和估测器互相适应,使估测器在最佳估测范围内获得 λ 值.下面我们对估测器的最佳估测范围进行分析.

对于 a=0.4 时的估测器,如果希望 $\left|\frac{E[\hat{r}]-r}{r}\right|<5\%$,那么,当 $r\le 1$ 时,n=3 即可;当 $1\le r\le 2$ 时,n 最好大于 7;而当 $2\le r\le 4$ 时,n 最好为 35 以上.也就是说,当 n 一定时,r 值较小时无偏性较好.而从有效性考虑,则会得到不同的结论:统计方差 σ/r 的值与 r,n 都有关系,当 r 值较小时,σ/r 的值比较大.所以,当 r 值较大时有效性较好.

综合考虑,当 a=0.4 时,如果设定 n=7,那么 r 保持在(1,2)内时,估测器的偏离和统计方差都可以保持在所要求的范围内.我们选择访问次数 n=7 是因为这个访问次数可操作性比较强.若 r 值大于 2 那么,这时估测器的统计方差较小但偏离较大,我们可以通过减小 a 值来提高估测器的无偏性.若 r 值小于 1,那么,这时估测器的统计方差较大,我们可以通过适当增大 a 值来提高估测器的有效性.当 r<0.4 时,增大 a 值会明显增大估测器的偏离,但对于减小估测器的统计方差却没有明显效果.这就说明,在访问次数一定的前提下,对基础数据过于频繁地监测访问不仅给系统资源带来极大的压力,而且未必能够得到好的估测效果.此时,我们或者通过提高访问次数 n 的值,或者通过减小访问频率来提高 r 值,然后再寻找合适的 a 值.

对于 n=7,在要求估测器的偏离<5%,同时 σ/r<60%的情况下,我们可以通过计算给出 r 值变化时符合要求的 a 值,见表 2.

Table 2 The best values for a when n=7

表 2 当 n=7 时,a 的最适合值

The range of r	(0.5,0.9)	(0.9,2)	(2,2.7)	(2.7,3)	(3,3.3)	(3.3,3.7)
The best values for a	0.7	0.4	0.3	0.25	0.2	0.15

由于我们无法得到确切的 r 值,因此只能由 $\hat{r}$ 来代替.实际上,当 n=7 时,只有在 X 也等于 7 时,$\hat{r}$=2.92,才越过了 a=0.4 的最佳估测范围的上限.根据 $\hat{r}$ 值,此时 a 应该取为 0.2.但因为每次探测都发现了变化,r 的实际值此时可能比 $\hat{r}$ 要大.为了减小误差,此时应该加大访问频率或增加访问次数.

另外,由于 λ 的动态性,我们应该对 λ 的历史值进行比较和分析,以判断其是否发生明显变化,变化规律如何,然后根据其变化规律,不断调整数据源访问方案和 λ 估测方案.下面,我们给出 λ 值的自适应估测算法 SAA(self-adaptive algorithm),其中提到的估测公式是指 $\hat{r}=-\log\left(\frac{n-X+a}{n+a}\right)$.

算法. SAA.

Step 1. 对于一个元素 e(它可能是 Web 页面,也可能是 Web 数据元组),可以先根据经验粗略估计出其变化频率 λ_0,然后使用频率 $f_0=\lambda_0$ 来访问它.访问 7 次以后,令 a=0.4,按照频率估测公式求出 $\hat{r}_0$ 值.

Step 2. 对于 $\hat{r}_i$ 的值,按照下面几种情况分别处理:

(1) 若 $0.5\le\hat{r}_i\le 0.9$,则不再继续访问;令 a=0.7,求出 $\hat{r}_{i+1}$ 值,再计算 $\hat{\lambda}=\lambda_{i+1}=\hat{r}_{i+1}\cdot f_i$.

(2) 若 $0.8<\hat{r}_i\le 2$,则不再继续访问;令 a=0.4,求出 $\hat{r}_{i+1}$ 值,再计算 $\hat{\lambda}=\lambda_{i+1}=\hat{r}_{i+1}\cdot f_i$ 即可.

(3) 如果 $\hat{r}_i>2$,当 f_i 不太大时,令 $f_{i+1}=2\cdot f_i$,然后对数据源进行 7 次访问,再按照 a=0.4 计算 $\hat{r}_{i+1}$;当 f_i 已经比较大时,保持 $f_{i+1}=f_i$ 不变,再继续访问 8 次,连同原来的 7 次共 15 次,按照 a=0.4 计算 $\hat{r}_{i+1}$.如果 $\hat{r}_{i+1}<2$,则可以结束操作,计算 $\hat{\lambda}=\lambda_{i+1}=\hat{r}_{i+1}\cdot f_{i+1}$;否则,需要继续提高访问频率并增加访问次数.

(4) 若 $\hat{r}_i<0.5$,则 λ 必然较小,在短期内的访问可能无法探测到 e 的变化.这时,可令 $f_{i+1}=f_i/3$,然后再访问 4 次;把前面的 7 次访问转换为 3 次访问(取第 1,4,7 次的结果)后,加上这 4 次,令 a=0.4 再计算 $\hat{r}_{i+1}$.如果 $\hat{r}_{i+1}>0.5$,则可以结束操作,计算 $\hat{\lambda}=\lambda_{i+1}=\hat{r}_{i+1}\cdot f_{i+1}$;否则,需要继续减小访问频率并延长访问周期.

Step 3. (1) 比较 $\hat{\lambda}$ 的历史值,找出其剧烈变化点:当 $\left|\hat{\lambda}_i-\hat{\lambda}_{i-1}\right|>k\cdot\bar{\hat{\lambda}}$ 时,我们认为 $\hat{\lambda}_i$ 是一个剧烈变化点,两次

剧烈变化点之间为一个变化周期 T_λ.其中,k 为根据经验决定的预先设定值,一般在(0.1,0.3)之间;$\overline{\hat{\lambda}}$ 为 $\hat{\lambda}$ 的算术平均值.

(2) 对 Web 页面数据的访问在变化周期 T_λ 的前期完成,使用估测的 $\hat{\lambda}$ 值指导在变化周期 T_λ 剩余时间的工作(如系统刷新等).

END 算法 SAA.

4 视图变化频率估测

Web 页面数据的变化遵循 Poisson 过程,而对于 Web 仓储中的物化视图来说,它可能来源于多处 Web 页面数据.一般情况下,我们不考虑数据源的相互关联性,而认为各处数据的变化都是独立发生的,这样,基础数据项的变化都是独立的 Poisson 过程.根据文献[3],多个 Poisson 过程的合成仍然是一个 Poisson 过程,其变化频率为各变化频率之和.我们来看一个例子,如图 3 所示,视图 V_1,V_2,V_3,V_4 的构成中使用到基础数据项 R,S,T,U,Q.

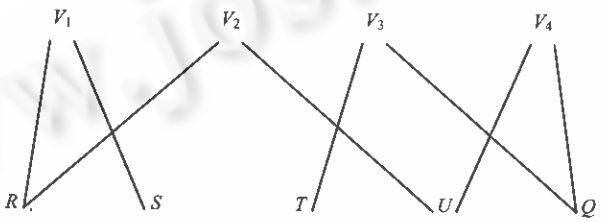

Fig.3 Views and their underneath base data

图 3 视图和它们的基础数据

我们仍然通过对基础数据的监测来分析物化视图内容是否发生变化.视图 V_1 的构成用到基础数据项 R 和 S,那么,其变化频率等于 R 和 S 的变化频率之和.而对 V_1 的监测实际上应该是监测 R 和 S,这两项中有一项发生变化,就会导致 V_1 的变化.这样,V_1 就应该进行更新维护了.

5 实验结果

我们在一些具有代表意义的.com 网站(包括国内和国外的站点)上选取 400 个适合的页面,进行了为期 1 个月的追踪.我们每天访问每个页面两次,并粗略地认为每次变化都被我们探测到了(变化频率明显过快的页面将被我们替换成变化适中的页面),这样我们就可以得到页面"实际的"变化频率.然后,从这些页面中找出那些变化了 10 次以下的(这些页面平均变化时间间隔为 3 天以上),共有 185 个页面.由于这 185 个页面平均变化频率不是很快,这样,我们的"粗略认为"(即我们探测到了页面的每次变化)就更为准确.

我们对这些页面进行估测,使用的仍然是我们的访问记录,只不过作了些调整,采用它的一个子集:假设监测访问无法每天两次这样频繁地进行,而是每 5 天一次,也就是只取 6 次监测记录进行估测.我们把估测结果和页面"实际"变化频率作了一个对比,如图 4 所示.

在图 4 中,r_0 是指使用估测器 X/T(X 为监测到的变化次数,T 为监测时间),r_1 为估测器 $-\log\left(\frac{\overline{X}+0.5}{n+0.5}\right)$,而 r_2 为估测器 $-\log\left(\frac{\overline{X}+0.4}{n+0.4}\right)$.

一个好的估测器,其估测结果与实际值的比率显然应该围绕 1 分布.从图 4 中可以看出:SAA 的效果显然是最好的;r_1 和 r_2 的结果也都远远好于 r_0;r_1 与 r_2 相比,后者要略强些.

6 结 论

Web 页面刷新是搜索引擎和 Web 仓库系统中最主要的维护工作,而页面变化频率则是决定刷新方案的最

重要的因素.本文主要研究了如何基于不完整的页面数据变化记录对 Web 页面变化频率λ进行估测.

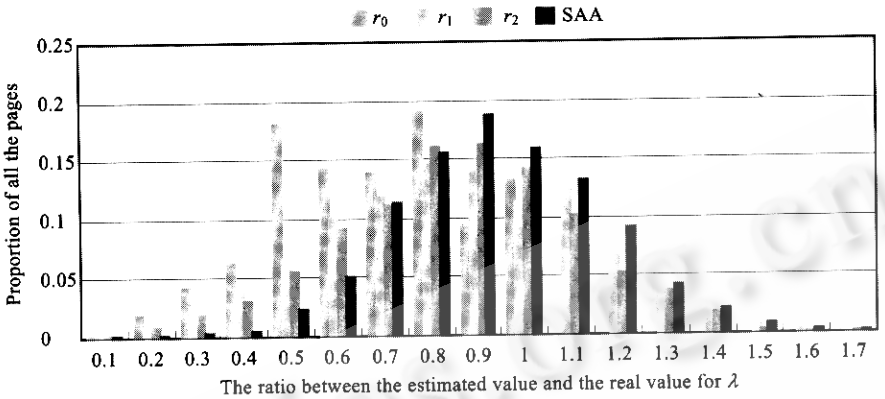

Fig.4 Comparison of effect of the estimators

图 4 估测器效果对比

事实上,文献[2]在这方面做了相当多的工作,并取得了很好的结果.然而,它缺少对估测器的适应性和综合性讨论.一般来说,估测器有 3 个评价指标,分别是无偏性(unbiasedness)、有效性(efficiency)和连贯性(consistency)[2,6].由于这 3 个指标几乎不可能达到同时最优,这就需要我们进行综合平衡.我们在对λ的估测过程中,根据这 3 个评价指标对估测器进行参数调整,同时根据λ值的大小不断调整数据源的访问频率和次数,从而使数据源访问模式和估测器互相适应,并使估测器在最佳估测范围内获得λ值.实验结果表明,我们的自适应估测算法效果优于文献[2]中的估测器.

References:

[1] Junghoo C, Hector GM. Synchronizing a database to improve freshness. In: Chen WD, Naughton JF, Bernstein PA, eds. Proc. of the 2000 ACM Int'l Conf. on Management of Data (SIGMOD). New York: ACM Press, 2000. 117−128.

[2] Junghoo C, Hector GM. Estimating frequency of change. ACM Trans. on Internet Technology, 2003,3(3):256−290.

[3] Taylor HM, Karlin S. An Introduction to Stochastic Modeling. 3rd ed., New York: Academic Press, 1998.

[4] George BT, Ross LF, Maurice DW, Frank RG. Thomas' Calculus. 10th ed., Reading: Addison-Wesley, 2002.

[5] Dennis DW, William M, Richard LS. Mathematical Statistics with Applications. 5th ed., Boston: PWS Publishing, 1997.

[6] Junghoo C. Crawling the Web: Discovery and maintenance of a large-scale Web data [Ph.D. Thesis]. Stanford: Stanford University, 2001.

张岩(1970－),男,河南南阳人,博士,副教授,主要研究领域为数据库技术,Web 信息处理.

唐世渭(1939－),男,教授,博士生导师,CCF 高级会员,主要研究领域为数据库系统,数据仓库,数据挖掘.

杨冬青(1945－),女,教授,博士生导师,CCF 高级会员,主要研究领域为数据库系统实现技术,Web 环境下的信息集成与共享,典型应用领域的数据库技术.

李晓明(1957－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机网络和海量信息.