大数据环境下的多源数据演化更新研究

余 放[1] 陈盛双[1] 李石君[2] 余 伟[2]
(武汉理工大学理学院 武汉 430070)[1] (武汉大学计算机学院 武汉 430072)[2]

摘 要 大数据环境下的多源数据呈现出数据量大、数据种类多、数据变化快的特点,这些特点对数据更新提出了新的挑战。通过分析大数据下多源数据的特点,定义了演化数据的概念,基于此建立了大数据的动态变频遍历更新模型。首先通过抽象数据的演化方式,建立了演化数据的势与稳定性概念,从而推导出更一般的代数意义上的演化运算工具;其次通过将运算工具导入大数据数据更新的实际应用中,推导出基于概率的变频遍历与动态权值模型;最后通过实验验证了在大数据环境下动态变频遍历模型(Dynamic Frequency Conversion Traversal,DFCT)对多源数据具有较高的更新效率。

关键词 大数据,演化数据,DFCT模型,数据更新

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.034

Research on Evolution and Updating among Multi-source Data Based on Big Data

YU Fang[1] CHEN Sheng-shuang[1] LI Shi-jun[2] YU Wei[2]
(Department of Science, Wuhan University of Technology, Wuhan 430070, China)[1]
(Computer School, Wuhan University, Wuhan 430072, China)[2]

Abstract Multi-source data based on big data presents the characteristics of a large amount of data, a great variety of data and data changing quickly. These characteristics put forward a new challenge to data updating. The concept of evolutionary data was defined by the analysis of the characteristics among multi-source data based on the big data. Based on this, a dynamic frequency conversion traversal data updating model was created. Firstly, abstracting the data evolutionary way and establishing the concept of evolutionary potential and stability of data, a more general evolutionary computing tools in algebra sense was derived. Secondly, frequency conversion traversal and dynamic weighting model based on probability was deduced by deriving a more general evolutionary computing tools in algebra sense. Finally, by importing tools into the practical application of data updating, dynamic frequency traversal model of multi-source data is verified by experiment with high updated efficiency on big data.

Keywords Big data, Evolutionary data, DFCT model, Data updating

1 引言

大数据时代,数据呈现出新的特点,即需要处理的数据量更大,需要处理的数据种类更多,数据产生和变化的速度更快[1]。这些特点对大数据环境下的数据更新又提出了新的要求,即能否设计一种更快速更智能的数据更新方式以提高数据更新的效率。目前数据更新的方法可分为全量更新与增量更新。在大数据环境中,数据量大且数据变化频繁,采用全量更新是不适合的。增量更新则是检测某个时间段内发生变化的数据,并且只更新此类局部数据,更新效率相对有所提高。

数据增量更新方式主要有3种:触发器法、日志分析法、快照差分法[3]。触发器对数据库进行增量检测,是一种简单的定位数据变动的方式,但是编写触发器是需要成本的,同时由于每次更新都需要启动,这使得触发器本身对系统的性能造成了一定影响。日志分析法更新有很好的效率,对系统性能会影响较小,然而此方法只适用于带有日志管理系统的数据库,同时还必须具有分析日志文件的工具。快照差分法通过比较不同时点上的快照文件,发现增量数据并将其更新,应用范围较广。但是获取快照本身的查询操作对系统性能会造成影响,同时每次更新操作都要对比快照文件,从而进一步对系统性能造成负担。

数据对现实世界实体进行抽象描述与度量,现实世界的实体的各项属性记录随时间动态变化。同时,世上没有绝对的准确的认识,人类对于实体的认识只是在无限接近于客观。

到稿日期:2015-11-18 返修日期:2016-02-26 本文受国家自然科学基金项目(61502350),湖北省自然科学基金项目(2014CFB289)资助。
余 放(1988—),男,硕士生,主要研究方向为数据挖掘、机器学习,E-mail:yufang9977@163.com;陈盛双(1964—),男,硕士,教授,主要研究方向为金融数学与数据挖掘,E-mail:chenshsh@whut.edu.cn;李石君(1964—),男,博士,教授,CCF会员,主要研究方向为大数据、互联网搜索与挖掘等,E-mail:shjli@whu.edu.cn;余 伟(1987—),男,博士,讲师,主要研究方向为数据质量评估、数据抽取与数据融合,E-mail:yuwei@whu.edu.cn。

基于以上两点，可以将数据的更新过程理解为数据朝着无限接近于客观的方向动态演化的过程。本文试图在抽象层面探讨大数据下的数据动态演化的过程，建立一种更为一般的具有较广应用性的数据更新模型。

2 等值计数划分与计数向量

为使模型具有普遍性，首先定义等值计数划分与计数向量。

2.1 等值计数划分与划分映射

定义 1 $\alpha_1,\alpha_2,\cdots,\alpha_n$ 为关系模式 R 下某一实体 E 具有的 n 条记录，其下标 i 代表遍历序数。设 A 为记录集合，对于 $A=\{\alpha_1,\alpha_2,\cdots,\alpha_n\}$，若存在 $A_i,A_j\subset A$，使得 $A_i\cap A_j=\varnothing$ $(i,j=1,2,\cdots,\lambda)$，$\{\lambda\in Z\mid 1\leqslant\lambda\leqslant n\}$ 且 $A_1\cup A_2\cup\cdots A_\lambda=A$，则记 $A_1,A_2,\cdots,A_\lambda$ 为 A 的一个划分。若划分满足 $\forall\alpha_{i_1},\alpha_{i_2}\in A_i,\alpha_{j_1},\alpha_{j_2}\in A_j$，有 $value(\alpha_{i_1})=value(\alpha_{i_2}),value(\alpha_{j_1})=value(\alpha_{j_2})$（$value(\ )$表示属性的记录值）恒成立，则称此划分为等值计数划分，记为 A_c，并称 $A_c=(A_1,A_2,\cdots,A_\lambda)^{\mathrm{T}}$ 为 A 以 λ 为维度的等值计数划分（下文简称计数划分）。$\forall A=\{\alpha_1,\alpha_2,\cdots,\alpha_n\}$，若相应的计数划分 $A_c=(A_1,A_2,\cdots,A_\lambda)^{\mathrm{T}}$ 存在，则称从 A 到 A_c 的映射 φ_c 为划分映射。

2.2 计数向量及其映射

定义 2 $C(A)=\varphi_d(A_c)=(\dim(A_1),\dim(A_2),\cdots,\dim(A_\lambda))$ 为 A 的计数向量。并称从 A 到 $C(A)$ 的映射为一级计数映射，从 A_c 到 $\varphi_d(A_c)$ 的映射为二级计数映射。

$$A\xrightarrow{\varphi_c}A_c\xrightarrow{\varphi_d}C(A)=\varphi_d(A_c) \tag{1}$$

定理 1 $\forall A=\{\alpha_1,\alpha_2,\cdots,\alpha_n\}$，有 $\dim[C(A)]\leqslant\dim(A)$。

证明：取任意整数 $l,m\in[1,n]$，当 $value(\alpha_l)\neq value(\alpha_m)$ 时，根据计数划分，定义 1 满足：$value(\alpha_{i_1})=value(\alpha_{i_2}),value(\alpha_{j_1})=value(\alpha_{j_2})$，则必有 $\dim(A_l)=\dim(A_m)=1$，所以 $\dim[C(A)]=n$，则 $\dim[C(A)]=\dim(A)$。

当存在 $value(\alpha_i)=value(\alpha_j)$ 时，则至少存在一子划分 $A_r\subset A_c$，且包含至少 2 个相同元素。

所以 $\dim[C(A-A_r)]<n-1,\dim[C(A-A_r)]+\dim[C(A_r)]<n$，所以 $\dim[C(A)]<\dim(A)$ □

定理 1 说明计数映射是一个降维映射。

3 多源数据演化的代数描述

3.1 多源数据下的理性数据采纳者偏好

实体 E 按关系模式 R 被多个信息源描述，则每一描述构成一条记录，记录包含相应实体的各属性具体值。数据源本身是具有权重的，例如对现实中某一品牌具体型号的彩电，不同家电网站介绍以及电商相应的售卖页面简介就构成了对这款彩电的不同的信息源。对于理性消费者（本质是消费领域的理性数据采纳者），当两个网站针对同一商品的介绍参数出现不一致时，他们更愿意相信相对大型网站的简介信息。这种网站积累的信用构成衡量信息源的权重的一个方面。而当数据不一致地出现在两家相同信用的网站时，理性消费者会参考第三方或更多网站信息，以“少数服从的多数”的原则去遴选出准确的信息。这是一致信息的网站数量构成度量信息重要性的另一方面。

3.2 多源属性的降维数值化

设 $A=\{\alpha_1,\alpha_2,\cdots,\alpha_n\}$ 为针对某一实体 E 被 k 个数据源记录的属性集合，$\alpha_i\in A$ 为第 i 个数据源生成的记录向量，则记录向量的维度对应 E 的实体型中属性的个数，默认空值属性的值为 $\varnothing$，A 中元素的个数即 A 的维度 n 代表信息源的数量。由定义 1 与定义 2 可知，一级计数映射或二级计数映射在映射原理上并不涉及记录向量 α_i 的具体内容（理论上其记录内容可以是多种数据格式），而只关注不同数据源的记录一致性以及数据源本身的权重。根据式(1)，E 的属性集 A 可通过计数映射对其进行降维数值化生成一个计数向量 $C(A)$。

3.3 EVO 元素与进化势

定义 3 $dia(M)=(a_{11},a_{22},\cdots,a_m)^{\mathrm{T}}$，其中矩阵 $M\in\mathscr{R}^{n\times n}$，$a_{ij}$ 为 M 第 i 行第 j 列的元素，$i,j=1,2,\cdots,n$。

定义 4 EVO 运算 $\odot$：$\forall\eta,\xi\in\mathscr{R}^n$，

$$\eta^{\mathrm{T}}\odot\xi=\max[dia(\xi\cdot\eta^{\mathrm{T}})]-\eta_{-p_1}^{\mathrm{T}}\cdot\xi_{-p_1} \tag{2}$$

其中，$\max[dia(\xi\cdot\eta^{\mathrm{T}})]$ 为向量 $dia(\xi\cdot\eta^{\mathrm{T}})$ 中最大值元素，并称这个值对应的元素为演化元素，简称为 EVO 元素，其在 A_c 中所对应的数据为演化数据，记为 EVO 数据。当存在一个以上相等 EVO 元素时，规定 $\odot$ 运算选出的所有最大值 EVO 元素需合并为一个。P_k（此处 $k=1$）表示下标为 k 时，参与合并元素的个数。$\eta_{-p_1},\xi_{-p_1}\in\mathscr{R}^{n-1}$ 为向量 $dia(\xi\cdot\eta^{\mathrm{T}})$ 中去掉 p_1 个 EVO 元素后，保持余下元素次序而形成的 $n-p_1$ 维向量。EVO 数据的现实意义在于它是权重与计数整合的最大者，代表在现阶段最接近于客观的数据。EVO 运算用以计算数据的演化程度，称 $\eta^{\mathrm{T}}\odot\xi$ 为 EVO 元素的演化势。

3.4 EVO 元素稳定性

当 $\eta^{\mathrm{T}}\odot\xi>0$ 时，称 EVO 元素具备稳定性，其对应的数据趋于稳定不变。而当 $\eta^{\mathrm{T}}\odot\xi\leqslant 0$ 时，由于定义 4 规定了 $\max[dia(\xi\cdot\eta^{\mathrm{T}})]$ 项已选出所有相等 EVO 元素，合并统一后的 EVO 元素加权值较余下多源非 EVO 元素加权值演化势弱，称此时 EVO 元素是非稳定的。其对应的数据为非稳定的 EVO 数据。

当计数向量 $C(A)$ 中 EVO 元素非稳定时，其演化势受不一致数据的削弱，说明计数向量通过一级计数映射的逆映射对应记录值是倾向于演化的，更新的 EVO 数据可能蕴含在新产生的不一致数据中，只是在当次测量时间段内，这种演化虽然具备量的累积却未达到质的变化。

3.5 演化势的层级

当 $\eta^{\mathrm{T}}\odot\xi\leqslant 0$ 时，EVO 元素非稳定，运算改写为如下递推关系：

$$\begin{aligned}R_i&=\eta_{-p_i}^{\mathrm{T}}\odot\xi_{-p_i}\\&=\max[dia(\xi_{-p_i}\cdot\eta_{-p_i}^{\mathrm{T}})]-\eta_{-p_{i+1}}^{\mathrm{T}}\cdot\xi_{-p_{i+1}}\end{aligned} \tag{3}$$

其中，整数 $i\in[0,n]$，可以发现式(2)为式(3)在 $i=1$ 时的特例。R_i 表示势在递推运算中进行到的层级。每一层递推关系可导出一个 $\max[dia(\xi_{-p_i}\cdot\eta_{-p_i}^{\mathrm{T}})]$ 项，当 $P_i=0$ 时，此项即为 EVO 元素。当 $P_i\geqslant 1$ 时，称 $\max[dia(\xi_{-p_i}\cdot\eta_{-p_i}^{\mathrm{T}})]$ 为第 i 级主项，用以度量其对 EVO 元素的削弱程度，根据式(3)递推层级结构可以得到多个层级主项，其层级越大，对 EVO 数据的削弱程度越小。同时由式(3)可知，第 i 层的主项对 EVO

元素影响的削弱程度恰好被同层的进化势 R_i 度量表示出来。若 $R_{i+x}>0$(x 为正整数),由于此层级的势达到稳定,则停止递推过程,并称 $r=i+x$ 为 EVO 元素的稳定度。

4 动态变频遍历模型 DFCT

4.1 EVO 数据的定位

数据库进行更新时需对外部数据进行一次遍历,对于某一实体 E 的属性,假设遍历出的所有记录依计数划分映射为 n 类,通过二级计数映射成为一个 n 维计数向量 $C(A)$。而 $C(A)$ 内每一元素都代表具有相同记录值的属性个数,不妨设第 i 个元素所对应的数据源的平均权重为 w_i,则 n 维 $C(A)$ 对应的权值向量为 $W=(w_1,w_2,\cdots,w_n)$,规定 $w_i>0,i=1,2,\cdots,n$。至此通过式(2)运算即可得遴选 EVO 数据的一种方法,即

$$M=\max\{dia[C(A)_{-P_1}\cdot W^{\mathrm{T}}_{-P_1}]\} \tag{4}$$

其中,M 为 EVO 数据的影响因子。当 M 确定时,则唯一确定了 $dia[C(A)_{-P_k}\cdot W^{\mathrm{T}}_{-P_k}]$ 上 EVO 元素的序数,假设序数为 m($0<m\leqslant n$),则此时 $M=w_m.\dim(A_m)$,A_m 即为 EVO 元素在计数向量 A_c 中的位置。当确定 A_m 后,即可建立起与属性集 A 的索引关系。由此定位到 A 中的 EVO 数据。

4.2 数据演化

数据演化遍历的过程可以分解成两个方面:1)多源数据空间中对同一实体的记录的数据不一定相同,即数据不一致现象,很多学者在此问题上做过研究,然而更为关键的问题是当出现有新数据与原数据不一致时,应当吸纳哪一条数据作为对实体接近真实的描述,如何设置一种标准来判断接近于真实的数据;2)多源数据空间中对实体描述的集合,其数据量是庞大的,而由于实体的属性值变动是相对平稳的小概率事件,按传统方式全部遍历则会消耗过多计算资源,因此有必要设计一种优化的遍历方式使得计算的效率增加。

4.3 更新赋值

假设实体 E 在数据库的属性记录为 α_e,遍历的过程需按式(4)计算出 EVO 元素所在位置,M 所对应的划分标记 $A_m\in A_c$ 指明了其标记的唯一记录值 $\hat{\alpha}\in A$,对比 $\hat{\alpha}$ 与 α_e,若 $\hat{\alpha}=\alpha_e$,则说明数据在当前阶段并未发生变化。若 $\hat{\alpha}\neq\alpha_e$,说明数据发生变动进行更新赋值,$\alpha_e\leftarrow\hat{\alpha}$。按式(3)计算实体 E 的多源记录集合 A,逐层求出进化势 R_i,直至 $R_{i+k}>0$ 递推停止。令第 i 层级的主项 $\max[dia(C(A)_{-P_i}\cdot W^{\mathrm{T}}_{-P_i})]$ 对应的计数向量分量为 $\dim(A_k)$,则存在两种情况,即 $\dim(A_k)=1$ 或 $\dim(A_k)>1$,前者的意义在于,P_i 层级中,削弱上一层级的主项的新记录仅来自于一个信息源,说明 A_k 对应的平均权值相对同层级其他划分权值要大。$\dim[C(A)]>1$ 时的情况则是下文 DFCT 模型关注的内容。

4.4 重复非 EVO 元素出现的概率密度

正确客观的数据具有唯一性,而错误数据具有多样性,可能出现多种不一致情况。实体被充分多的数据源进行描述后,若多次遍历 E 的记录依概率 P 收敛于正确属性值 α_e,则不一致现象出现的概率为 $P(\bar{\alpha}_e)=1-P$。在总体出现数据不一致的条件下,任一数据源不一致记录出现的概率密度为:

$$P(\alpha_i|\bar{\alpha}_e)=1/\prod_{j=1}^{\dim(E)}[domian(\alpha_{ij})] \tag{5}$$

其中,$\alpha_i\in A$,$P(\alpha_i|\bar{\alpha}_e)$ 表示在不一致数据条件下,α_i 记录出现某一具体值的概率,$domian(a_{ij})$ 表示第 i 行记录第 j 列属性的域包含的可能取值的个数(在连续型随机变量中为分段区间的个数),$\dim(E)$ 表示 E 的实体型所包含的属性个数。

式(5)表明实体型中包含的属性越多,其取到特定记录值的概率越小,同时由于数据不能超出其属性域(事实上,超出属性域的记录不满足关系模式 R 中的函数依赖关系,根据模型设计更新程序时,可将权重赋 0 值,排除其影响),当属性个数既定时,属性域含有的元素量愈多,概率也会愈小。基于定义 1,令计数划分 A_c 的子集 A_i 是一个由等值非 EVO 数据组成的集合,对于 A_i 中的相等记录是依概率 $P(\alpha|\bar{\alpha}_e)$ 收敛于特定记录 α 且独立重复 $\dim(A_i)$ 次后形成。则任意非 EVO 元素对应的计数划分 A_i 形成的概率为:

$$P(A_i)=C_n^{\dim(A_i)}P^{\dim(A_i)}(\alpha|\bar{\alpha}_e)[1-P(\alpha|\bar{\alpha}_e)]^{[n-\dim(A_i)]} \tag{6}$$

其中,n 表示遍历的外部数据源数量,在大数据环境中,遍历数据源数量较大,而真实世界实体的属性数量也是较大的,通常导致 $P(\alpha|\bar{\alpha}_e)$ 是一个很小的概率。此时可对式(6)作如下推导。

$$\begin{aligned}
P(A_i)&=\lim_{\substack{n\to\infty\\P(\alpha|\bar{\alpha}_e)\to 0}} C_n^{\dim(A_i)}P^{\dim(A_i)}(\alpha|\bar{\alpha}_e)[1-P(\alpha|\bar{\alpha}_e)]^{[n-\dim(A_i)]}\\
&=\lim_{\substack{n\to\infty\\P(\alpha|\bar{\alpha}_e)\to 0}} \frac{n(n-1)\cdots[n-\dim(A_i)+1]}{\dim(A_i)!}P^{\dim(A_i)}(\alpha|\bar{\alpha}_e)[1-P(\alpha|\bar{\alpha}_e)]^{\left[\frac{n\cdot P(\alpha|\bar{\alpha}_e)}{P(\alpha|\bar{\alpha}_e)}-\dim(A_i)\right]}\\
&=\lim_{\substack{n\to\infty\\P(\alpha|\bar{\alpha}_e)\to 0}} \frac{\dfrac{n^{\dim(A_i)}\cdot P^{\dim(A_i)}(\alpha|\bar{\alpha}_e)}{\dim(A_i)!}\cdot[1-P(\alpha|\bar{\alpha}_e)]^{\left[-\frac{1}{P(\alpha|\bar{\alpha}_e)}\right]\cdot[-nP(\alpha|\bar{\alpha}_e)]}}{[1-P(\alpha|\bar{\alpha}_e)]^{\dim(A_i)}}\\
&=\lim_{\substack{n\to\infty\\P(\alpha|\bar{\alpha}_e)\to 0}} \frac{[nP(\alpha|\bar{\alpha}_e)]^{\dim(A_i)}}{\dim(A_i)!}\cdot e^{-nP(\alpha|\bar{\alpha}_e)}
\end{aligned}$$

令 $\lim\limits_{\substack{n\to\infty\\P(\alpha|\bar{\alpha}_e)\to 0}}[nP(\alpha|\bar{\alpha}_e)]=\theta$

从推导结果可以发现,大数据环境下的小概率等值非 EVO 记录的出现次数服从一个参数为 θ 的泊松分布。由于其密度函数是凸函数,不难推导出其密度在 $\dim(A_i)=[\theta]$ 或 $\dim(A_i)=[\theta]+1$(此处[ ]为取整符号)时达到最大,此后随 $\dim(A_i)$ 的积累单调减小趋于 0^+。θ 作为常数均值,针对具体实体可在历史遍历中通过参数估计获得。该公式的实际意义在于对同一实体的不一致描述是存在的,并且发生相同值的不一致错误概率在遍历初期可能是增长的。例如对于测试

成本很高的实体，当某一数据源对其测试后并公布各测试出的属性记录时，其他数据源为减少成本，可能采取复制或转发方式构成相同的新记录，此时 θ 便会偏大，然而当 $\dim(A_i)$ 累积到超过极值区域后，根据泊松分布特点，其概率密度则会迅速衰减。当重复遍历出此类等值小概率事件时，下一遍历期 EVO 记录可能蕴藏在对应的划分中，应提高对此划分的遍历频率。

至此得到了变频遍历的理论基础，即可以设定一个与泊松分布的概率密度成反比的遍历频率映射来调节遍历频率。

$$P(\alpha|\bar{\alpha}_e)\rightarrow f[P(A_i)] \tag{7}$$

式(7)的现实意义在于，由于不一致记录无统一性，不一致数据出现多个相同记录值的概率是相对较小的，当实际遍历出现此类小概率事件时，应当对其权值进行调整，并提高对其遍历的频率，因为此类记录存在着相对更高的演变为 EVO 数据的可能性。直观理解便是真相只有一个，但错误可能有多种，当多个数据源的错误都一样时，便需要怀疑这种相同的错误是否真的为错。

针对具体实体的特性，上式的负相关函数表达式可以设定多种形式，在此不妨假设最简单的反比形式用以讨论其性质。

4.5 泊松变频函数

定义 5

$$f_{t_{i+1}}[P(A_i)]=\frac{[\theta^{\dim(A_i)}/\dim(A_i)!]\cdot\upsilon}{(t_{i+1}-t_i)\cdot e^{\theta}} \tag{8}$$

其中，$f[P(A_i)]$为泊松概率变频函数，υ 为调节系数，用来调节负相关的程度，$t_{i+1}-t_i$ 为一个单位时间间隔。分析以上定义可以推导出动态变频函数的性质，即本期划分 A_i 的维度与下期对其遍历的频率成反比，实际上当 A_i 本期出现的相同非 EVO 记录概率很小时，程序则按式(8)自动加强了对其的“关注”，在下期采用更高的频率对其检查；相反，当 A_i 中非 EVO 相同数据出现次数变小时，程序则在下期自动降低检查频率。由此，在大数据多源数据环境下进行数据更新，不同划分遍历频率是不同的，在规定时间段内，相对于所有数据实施全局遍历，动态变频遍历效率有所提高。

4.6 泊松动态权值函数

现分析式(4)中的平均权重项 $W^{T}_{\underline{P}_1}$，作为遴选 EVO 元素的依据之一，其分量权值的赋予来源于两方面：1)对应数据源稳定的内在信用因素；2)数据源所在计数划分 A_i 形成的概率 $P(A_i)$。

定义 6

$$w_{(\tau_i,P(A_i),\sigma_\tau,\sigma_\rho)}=\frac{\log_{\sigma_\tau}\tau_i-\log_{\sigma_\rho}^{P(A_i)}}{\dim[C(A)]} \tag{9}$$

其中，w 为计数划分 A_i 的泊松动态权重，$\tau_i\in(0,1)$定义为 A_i 的信用评级，参数 $P(A_i)$表示数据源所在计数划分 A_i 形成的概率。参数 $\sigma_\tau,\sigma_\rho(\sigma_\tau,\sigma_\rho>1)$分别表示 τ_i 与 $P(A_i)$的平衡参数，针对具体更新项目调节量纲并控制 w 为正数。规定下列两种情况直接对 w 赋值，无需经过式(9)计算。1)当 A_i 违反关系模式 R 中的函数依赖或完整性约束等必要条件时，则对 w 赋 0 值。2)当 A_i 的记录中包含绝对权威数据源的数据时，直接对 w 赋 1 值。其中绝对权威数据源指在特定情况下，依照法律或制度等保证其数据记录是唯一可信的信息发布机构。

4.7 DFCT 模型整体框架与算法

DFCT 模型的整体框架如图 1 所示。

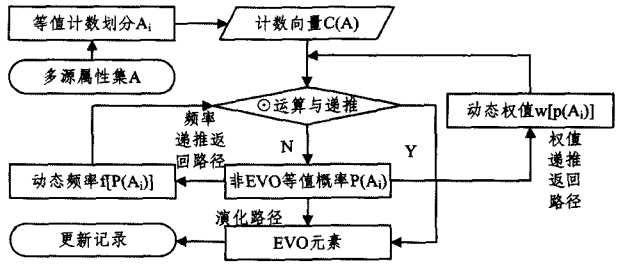

图 1 DFCT 模型整体框架

DFCT 的计算框架如算法 1 所示。

算法 1 DFCT 框架

输入：抽样实体属性对应的计数向量 C(A)

输出：EVO 元素对应的划分

1. $\{A\}\leftarrow RAND_{\omega}(B_1\cap B_2\cap\cdots\cap B_n)$
2. $\{C(A)\}\leftarrow\{A\}$
3. $M_{k+1}\leftarrow\max\{dia[C(A^k)_{-P_1}\cdot W^{kT}_{-P_1}]\}$
4. if $M_{k+1}=M_k$ then
5. $k+1=k$
6. else
7. for each EVO$\in A^k$ do
8. EVO$\leftarrow M_{k+1}$
9. for each $A_i^{k+1}\subset A^{k+1}$ do
10. $P(A_i)\leftarrow\dim(A_i)$
11. for each $f_{t_{k+1}}[P(A_i)]\in F$ do
12. $f_{t_{k+2}}[P(A_i^{k+1})]=f_{t_{k+1}}[P(A_i^k)]$
13. for each $w_{(\tau_i,P(A_i),\sigma_\tau,\sigma_\rho)}\in W$ do
14. $w_{(\tau_i^{k+1},P(A_i^{k+1}),\sigma_\tau,\sigma_\rho)}=w_{(\tau_i^{k},P(A_i^{k}),\sigma_\tau,\sigma_\rho)}$
15. end for
16. $W^{(k+1)T}_{-P_1}=W^{kT}_{-P_1}$
17. end for
18. end for
10. end for
20. end if

其中，k 表示遍历的序数，F 表示泊松动态频率，W 表示动态权重所需满足的域。

5 实验分析

5.1 实验平台与数据

本实验的目的是通过具体实例应用，测试大数据环境下多源数据动态变频遍历更新在规定时间段内的更新性能。实验系统环境为 R 语言 version 3.2.2，运行于 Windows7 64 位操作系统，硬件环境为 AMD FX-8350 4.0GHz CUP，DDR3 1600 8GB RAM。为客观分析 DFCT 过程的性质，数据采用 B2C 电商平台数据集，以十大门类商品为实体，自定义网络爬虫，并采集 2014 年 1 月至 2015 年 8 月时间段内的共 997631 件商品实体。各平台下实体门类如表 1 所列。

表 1　数据集各分类的分布总体数量

Type	JD	YIXUN	YHD	AMAZON
Mobile	13140	12757	10611	12791
Digital	44324	69786	49108	59240
Computer	14102	19683	9799	13725
Appliance	96800	105379	66881	60497
Houseware	22284	35436	40482	25152
Jewelry	82452	68328	78481	62113
Cosmetics	25215	21326	15205	21742
Sports	5483	5567	2492	4532
Food	4164	6060	6285	3894
Dailyuse	17236	26409	11567	20686

5.2　实验评价标准

数据更新性能分两个评价指标：1)数据更新准确率；2)更新所需时间。定义在规定时间段上的数据更新准确率为在这个时间段内遍历正确更新的记录数量与 B2C 平台实际发生变动记录数量的比值。

EVO 数据被定义为最进化数据，在多次遍历后其值最接近正确值。本次实验中设置参照，实体所在的官网属性记录不参与遍历，而设为已知的参照组数据。正确更新的记录指参照组实际发生变动的记录数量与遍历数据源错误更新的 EVO 记录数量之差。错误的更新定义为 3 种情况：1)B2C 平台实体记录发生变动，且参照数据发生相同变动但未被 DFCT 模型检测出；2)B2C 平台实体记录发生变动，且参照数据发生相同变动，但 DFCT 模型检测出的 EVO 数据与前两者不符；3)B2C 平台实体记录与参照数据均未改变，而 DFCT 模型检测与两者不符的 EVO 数据。则准确率计算公式如式(10)所示：

$$E(t)=\frac{N_t(\Delta A)-N_t(\Delta A')-N_t(\Delta \tilde{A})}{N_t(\Delta A)} \tag{10}$$

其中，$N_t(\Delta A)$为在 t 单位时间段内 DFCT 遍历出的发生变动的 EVO 记录总数，$N_t(\Delta A')$表示在 t 单位时间段内第一、二两类错误记录的数目。$N_t(\Delta \tilde{A})$表示第三类错误记录产生的数量。由式(10)可知，当第三类错误累积到一定程度时准确率可以为负值。更新过程时间由耗时比来度量，即在同平台同环境下动态变频遍历耗时与全局遍历耗时之比。当耗时比越小时，说明计算效率越高。

每组分 DFCT 遍历和全局遍历两次实验进行，其中后者为对照组，预设规定时间段基期标准遍历频率为 10 次，而后程序每次遍历频率依概率结果浮动调节，将数据集按 2014 年 1 月至 2015 年 8 月的时间顺序平均分成 20 个时间段。

5.3　结果分析

数据集各分类的分布总体数量如表 2 所列。

表 2　数据集各分类的分布总体数量

Type	准确率(%)	耗时比(%)	备注
Mobile	86.70	32.10	
Digital	92.90	27.20	
Computer	88.30	20.50	
Appliance	81.10	37.20	
Houseware	59.40	29.80	
Jewelry	91.70	12.00	单属性变化为主
Cosmetics	80.20	46.70	
Sports	74.40	23.20	
Food	69.50	28.30	
Dailyuse	64.60	16.40	

由表 2 可以看出，更新的准确率与耗时比不存在明显相关关系。但针对不同的实体门类，其指标具备一定特点。即更新换代较快且规格属性统一的实体门类更新准确率较高，均达到 80%以上，甚至超过 90%，例如智能手机、数码产品、电脑。同时它们的耗时比相对较小，例如电脑类的耗时不到全局遍历的 1/4，为 20.5%。更新较慢且规格属性统一度不高的实体门类对应的准确率较低，且耗时比较大。例如家居用品、生活用品与食物类。珠宝类的准确率超过 90%且耗时比最低为 12.0%，原因在于虽然珠宝存在多属性，但同型号实体的多种属性变化很小，更多的则是价格这一单一属性在频繁变动，由此带来的遍历次数更少。同时由于价格这一单一属性的频繁变动，遍历结果能更准确地依概率收敛于 EVO 元素的对应值。通过图 2 能够清晰看出 DFCT 对不同门类商品的更新性能变化趋势。

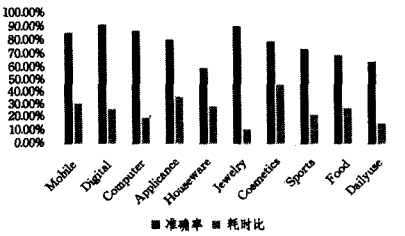

图 2　更新性能

5.4　实验结论

DFCT 模型适用于属性记录变化较为频繁的多元实体，其更新效率比传统的全局遍历更新有明显提高。但是在数据源较少时准确率只能趋近于 100%，而无法达到完全准确。这意味着现阶段该模型不适合要求精确更新的数据实体。然而对于诸如 B2C 信息集成平台或者比价助手等信息辅助平台，其数据变动频繁和精度要求相对宽松，采用 DFCT 更新方法相比传统数据更新具有一定优势。

结束语　DFCT 模型是从数据演化的角度建立起的动态变频遍历模型，遍历频率与遴选 EVO 元素的权值依赖于等值非 EVO 元素出现的概率。演化运算还构建了演化势的概念，如何将演化势与遍历频率关联起来是值得今后研究的。同时根据实验发现虽然 DFCT 模型的耗时比较低，但更新精度仍有待提高。其中参数的调节和 EVO 元素稳定性的关系是值得进一步分析的。

参 考 文 献

[1] Chen Shi-min. Big Data Analysis and Data Velocity[J]. Journal of Computer Research and Development, 2015, 52(2): 3333-3342(in Chinese)
陈世敏. 大数据分析与高速数据更新[J]. 计算机研究与发展, 2015, 52(2): 3333-3342

[2] Li Jian-zhong, Liu Xian-min. An Important Aspect of Big Data: Data Usability[J]. Journal of Computer Research and Development, 2013, 50(6): 1147-1162(in Chinese)
李建中, 刘显敏. 大数据的一个重要方面：数据可用性[J]. 计算机研究与发展, 2013, 50(6): 1147-1162

[3] Tian J, Guo H, Hu H, et al. OFDM Signal Sensing over Doubly-Selective Fading Channels[C]//2010 IEEE Global Telecommu-

nications Conference (GLOBECOM 2010). IEEE, 2010: 1-5

[4] Cheng Xue-qi, Jin Xiao-long, Wang Yuan-zhuo, et al. Survey on Big Data System and Analytic Technology[J]. Journal of Software, 2014(9): 1889-1908(in Chinese)
程学旗,靳小龙,王元卓,等. 大数据系统和分析技术综述[J]. 软件学报,2014(9):1889-1908

[5] Meng Xiao-feng, Li Yong, Zhu Jian-hua, et al. Social Computing in the Era of Big Data: Opportunities and Challenges[J]. Journal of Computer Research and Development, 2013, 50(12): 2483-2491(in Chinese)
孟小峰,李勇,祝建华,等. 社会计算:大数据时代的机遇与挑战[J]. 计算机研究与发展,2013,50(12):2483-2491

[6] Shi Jin-gang, Bao Yu-bin, Leng Fang-ling, et al. Study on Log-Based Change Data Capture and Handling Mechanism in Real-Time Data Warehouse[C]// Proceedings of 2008 International Conference on Computer Sience and Software Engineering. Wuhan, 2008: 478-481

[7] Li Shi-jun, Yu Jun-qing, Ou Wei-jie. Web Information Extraction Based on HTML Pattern Algebra[J]. Journal of Computer Research and Development, 2006, 43(9): 1644-1650(in Chinese)
李石君,于俊清,欧伟杰. 基于 HTML 模式代数的 Web 信息提取方法[J]. 计算机研究与发展,2006,43(9):1644-1650

[8] Building the Data Warehouse [M]. New York: John Wiley & Sons, 1996

[9] Korn F, Muthukrishnan S, Zhu Y. Checks and balances: Monitoring data quality problems in network traffic databases[C]// Proc of the 29th IntConf on Very Large Databases. San Francisco, USA, 2003: 536-547

[10] Xu K S, Kliger M, Hero A O I. Evolutionary spectral clustering with adaptive forgetting factor[C]// International Conference on Acoustics, Speech, and Signal Processing, 1988 (ICASSP-88). 2010: 2174-2177

[11] Wang Y, Liu S X, Feng J, et al. Mining Naturally Smooth Evolution of Clusters from Dynamic Data[C]// Proc. of SIAM Conf. on Data Mining. 2007: 125-134

[12] Li J, Li S. Evolutionary Hierarchical Dirichlet Process for Timeline Summarization[C]// Meeting of the Association for Computational Linguistic. 2013: 556-560

[13] Kim H D, Lee D H, Choe H, et al. The evolution of cluster network structure and firm growth: a study of industrial software clusters[J]. Scientometrics, 2014, 99(1): 77-95

[14] Hedeler C, Belhajjame K, Fernandes A A A, et al. Dimensions of Dataspaces[M]// Dataspace: The Final Frontier. Springer Berlin Heidelberg, 2009: 55-66

[15] Ci Xiang, Ma You-zhong, Meng Xiao-feng, et al. Method for Top-K Query on Big Data in Cloud[J]. Journal of Software, 2014, 25(4): 813-825(in Chinese)
慈祥,马友忠,孟小峰,等. 一种云环境下的大数据 Top-K 查询方法[J]. 软件学报,2014,25(4):813-825

[16] Peng Yuan-hao, PAN Jiu-hui. Study on Incremental Data Capturing Method Based on Log Analysis[J]. Computer Engineering, 2015, 6(6): 56-60(in Chinese)
彭远浩,潘久辉. 基于日志分析的增量数据捕获方法研究[J]. 计算机工程,2015,6(6):56-60

(上接第 182 页)

national Conference on Knowledge Discovery and Data Mining, 2001. New York, NY, USA: ACM, 2001: 97-106

[8] Rutkowski L, Jaworski M, Pietruczuk L, et al. A New Method for Data Stream Mining Based on the Misclassification Error [J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(5): 1048-1059

[9] Gama J. Learning Decision Trees from Dynamic Data Streams [J]. Journal of Universal Computer Science, 2005, 11(8): 1353-1366

[10] Mena Torres D, Aguilar Ruiz J S. A similarity-based approach for data stream classification[J]. Expert Systems with Applications, 2014, 41(9): 4224-4234

[11] Gama J, Fernandes R, Rocha R. Decision Trees for Mining Data Streams[J]. Intelligent Data Analysis, 2006, 10(1): 23-45

[12] Andromeda T, Marsono M N, Ru L H. Online Data Stream Learning and Classification with Limited Labels[C]// Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics, 2014. Yogyakarta, Indonesia: Indonesia journals, 2014: 161-164

[13] Widyantoro D H. Exploiting Unlabeled Data in Concept Drift Learning[J]. Jurnal Informatika, 2007, 8(1): 54-62

[14] Lindstrom P, Delany S J, B M Namee. Handling Concept Drift in a Text Data Stream Constrained by High Labelling Cost[C]// Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, 2010. Florida, USA: AAAI, 2010: 32-37

[15] Masud M M, Gao J, Khan L, et al. Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(6): 859-874

[16] Xiao M, Guo Y. Semi-Supervised Kernel Matching for Domain Adaptation[C]// Proceedings of the 26th AAAI Conference on Artificial Intelligence, 2012. North America: AAAI, 2012: 1183-1189

[17] Kobayashi N, Inui K, Matsumoto Y. Extracting Aspect-Evaluation and Aspect-of Relations in Opinion Mining[C]// Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007. Prague: Association for Computational Linguistics, 2007: 1065-1074

[18] Li L H, Jin X M, Long M S. Topic Correlation Analysis for Cross-Domain Text Classification[C]// Proceedings of the 26th AAAI Conference on Artificial Intelligence, 2012. North America: AAAI, 2012: 998-1004

[19] Blitzer J, McDonald R, Pereira F. Domain adaptation with structural correspondence learning[C]// Proceedings of the Conference on Empirical Methods in Natural Language, 2006. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006: 120-128