

多维连续不确定 XML 数据查询处理算法

张晓琳, 霍 伟, 刘立新, 崔光月
(内蒙古科技大学 信息工程学院, 内蒙古 包头 014010)

摘 要: 提出一种基于蒙特卡洛思想的数据查询处理算法 QueryMC. 在查询计算过程中, 根据小枝查询模式确定待处理多维随机变量联合概率密度函数及查询区域, 通过构造相同区域上的均匀分布随机变量将查询问题建模成相应复合函数的期望, 利用算法同时产生的随机样本集估计该期望的取值作为问题的解, 避免了传统的降维操作, 有效地减少了处理时间. 实验结果表明, 在取得理想精度的同时, 算法具有高效性.

关键词: 连续不确定 XML; 多维随机变量; 小枝查询模式; 联合概率密度; 蒙特卡洛

中图分类号: TP392

文献标识码: A

文章编号: 1000-7180(2013)06-0030-04

Queryprocessing Algorithm for Multidimensional Continuous Uncertain XML Data

ZHANG Xiao-lin, HUO Wei, LIU Li-xin, CUI Guang-yue

(School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China)

Abstract: An effective algorithm QueryMC based on Monte-Carlo method is proposed. According to the twig query pattern, the joint probability density function and the region of query are identified. Furthermore the problem of query in QueryMC is modelled into expectation of composite function by structuring random variables of uniform distribution of the same region. It could be used to avoid the traditional dimensionality reduction operation and to reduce the processing time by estimating the expectation with the random sample set. Meanwhile, the results show that the algorithm is highly efficient with ideal precision.

Key words: continuous uncertain XML; multidimensional random variables; twig query pattern; joint probabilitydensity; Monte-Carlo

1 引言

近些年来,随着多维连续不确定 XML 数据的应用愈加广泛,特别是多维连续不确定 XML 数据的存储及查询处理,例如金融领域中涉及多个随机变量的投资组合问题以及环境领域中多种生物污染指数的监测处理问题等. 对此研究人员已经提出了多种相应的数据模型及有关查询问题的处理方法^[1-2]. 对于连续不确定 XML 数据的查询处理,文献[3]中提出了积分法(Integral Method, IM)与矩形近似(Rectangles Approximation, RA)方法,前者对于随机变量概率密度函数的原函数易求得的

XML 数据有着较高效的查询响应,如均匀分布随机变量,但应用范围较窄;矩形近似方法能够处理含复杂分布类型随机变量的 XML 数据的查询,但是随着待处理连续不确定 XML 数据维数的增多,矩形近似方法将会有很高的处理时间代价. 为解决上述问题,本文提出一种基于蒙特卡洛思想的多维连续不确定 XML 数据查询处理算法 QueryMC.

2 相关定义

2.1 小枝查询模式

图 1 中给出一个连续不确定 XML 文档 (PrXML)_{BN}, 称为 P 文档树^[4-5], 记录的是某一区

收稿日期: 2012-08-01; 修回日期: 2012-09-18

基金项目: 国家自然科学基金资助项目(61163015); 内蒙古自然科学基金重点资助项目(20080404Zd21)

域内两个传感器所监测的各自子区域的多个污染生物属性相关信息. *cont* 节点存储各生物属性随机变量所满足的多维连续分布联合概率密度函数,如 $f_1(B_1, B_2, \dots, B_i)$ 等, *mux* 节点表示后代节点的互斥关系.

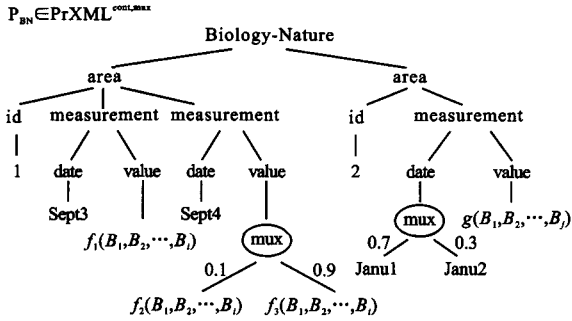


图 1 P 文档树

根据此文档树可以查询某一子区域内传感器所监测的各个生物属性随机变量在一定范围内取值的联合概率,将其与某一概率阈值进行比较,决策层据此可以判断是否需要采取措施来改善此区域的生态环境. 在此问题中,采用的是小枝 (twig) 查询模式^[6-7]. 一棵查询树,如图 2 所示,通过自身结构与谓词表达节点之间的相互关系以及相关查询信息,其中 $\text{twig}(t)$ 的叶子节点 $e_i \in E(e_1, e_2, e_3)$. 查询树小枝概率的定义如下所示.

定义 1 查询树小枝概率 $\text{twigProb}(t_i)$

$$\text{twigProb}(t_i) =$$

$$\frac{\text{twigProb}(t_{i-1}) * \text{pathProb}(e_i) * \text{predicatesProb}(e_i)}{\text{pathProb}(\text{common}(t_{i-1}, e_i))}$$

式中, $\text{pathProb}(e_i)$ 为 e_i 节点所在单枝路径概率, $\text{pathProb}(\text{common}(t_{i-1}, e_i))$ 为 e_i 节点所在单枝与后续递归部分的公共路径概率, $\text{predicatesProb}(e_i)$ 表示当 e_i 节点的取值满足谓词中相应的查询条件时的概率值.

例如,对于图 2 所示的查询树 $\text{twig}(t)$,根据其谓词查询信息可以给出该查询树的小枝概率 $\text{twigProb}(t) = 0.7 * \int_{a_1}^{b_1} \dots \int_{a_j}^{b_j} g(B_1, B_2, \dots, B_j) dB_1 \dots dB_j$.

显然,对于小枝模式查询问题,在确定了相关节点,多维连续分布随机变量联合概率密度函数以及查询区域之后,转化成了求多维积分的问题.

2.2 多维连续分布

以 n 维正态分布为例,给出如下定义.

定义 2 称随机变量 $X = (X_1, X_2, \dots, X_n)$ 服从参数为 $N(a, B)$ 的 n 维正态分布,记为 $X \sim N(a, B)$,其联合概率密度函数为

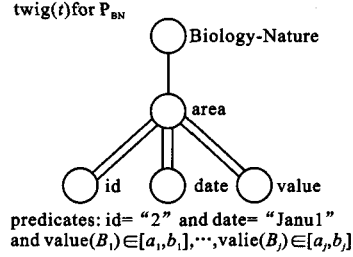


图 2 twig 查询树

$$f(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{n/2} (\det B)^{1/2}} \exp \left\{ -\frac{1}{2} \right.$$

$$\left. (x-a)^T B^{-1} (x-a) \right\},$$

式中, $B = [b_{ij}]_n$ 是 $n \times n$ 阶协方差矩阵,当 $i=j$ 时,记录各随机变量的方差, $i \neq j$ 时,记录各随机变量之间的相关系数与相应标准差的乘积. 显然, $b_{ij} = b_{ji}$, B^{-1} 是其逆矩阵, a 为实值向量,记录各随机变量均值. 特别地,当 $X_1, X_2, \dots, X_n$ 相互独立时,联合概率密度函数为

$$f(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{n/2} b_{11} b_{22} \dots b_{nn}} \exp \left\{ -\frac{1}{2} \left[\frac{(x_1 - a_1)^2}{b_{11}^2} + \dots + \frac{(x_n - a_n)^2}{b_{nn}^2} \right] \right\}.$$

3 QueryMC 查询处理算法

3.1 蒙特卡洛积分思想

蒙特卡洛积分是一种数值积分方法,使用满足独立性和均匀性统计检验的多维随机样本集估计积分 $I_n = \int_D \dots \int f(x_1, \dots, x_n) dx_1 \dots dx_n$ 的取值,其中 D 是 n 维积分区域,为一个 n 维超立方体,在此描述 n 个连续分布随机变量的取值范围,即 $D = [a_1, b_1] * [a_2, b_2] * \dots * [a_n, b_n]$,则 D 的体积为 $V_D = (b_1 - a_1)(b_2 - a_2) \dots (b_n - a_n)$. 函数 $f(x_1, \dots, x_n)$ 描述 n 个随机变量满足的某 n 维连续分布,据此和独立采样的 N 个 n 维样本 $(r_{11}, \dots, r_{1n}), \dots, (r_{N1}, \dots, r_{Nn})$,利用连续分布随机变量数字特征性质及样本平均值原理可将积分 I_n 转化为

$$I_n = \int_D \dots \int f(x_1, \dots, x_n) dx_1 \dots dx_n = E \left[\frac{f(x_1, \dots, x_n)}{g(x_1, \dots, x_n)} \right] \quad (1)$$

式中, $g(x_1, \dots, x_n)$ 为构造的 D 区域上均匀分布的 n 维随机变量联合概率密度函数:

$$g(x_1, \dots, x_n) = \frac{1}{V_D} \quad (2)$$

进而可以将原积分式处理为式(3)所示:

$$I_n = \frac{V_D}{N} \sum_{i=1}^N f(r_{i1}, \dots, r_{in}) \quad (3)$$

根据式(3)可知,在足够样本数目下可以给出理想精度的无偏估计结果.

3.2 QueryMC 查询处理算法

算法 1 是 QueryMC 的主算法. 其中, getRandom(r)函数的功能是利用线性同余法产生独立均匀分布的 0~1 间随机数. 算法根据各随机变量的积分上下限 ul[], ll[] 以及维数 dim 调用该函数产生特定区间内的多维随机数样本, 样本数为 number (MC). 之后利用样本平均值原理, 将其代入被积函数, 获得统计平均值近似代替待求积分值, 作为结果返回.

算法 1 QueryMC

输入: 含积分区间的多维连续分布随机变量的联合概率密度函数

输出: 表示随机变量在其各自特定区间内取值联合概率的积分值

```
(1)begin
(2)getRandom(r) //产生 0-1 间均匀分布的随机数
(3)for i=1 to number(MC) //采集相应数量的样本集
(4) $r_i = (u * r_{i-1} + v) \bmod m$  //线性同余法
(5)return  $r_i$ 
(6)getf(x[]) //获取被积函数
(7) $z = f(x_1, x_2, \dots, x_n)$ 
(8)return  $z$ 
(9)getIntegral(dim, ul[], ll[]) //蒙特卡洛积分
(10)for m=1 to number(MC)
(11)for j=1 to dim
(12) $x[j] = ll[j] + (ul[j] - ll[j]) * \text{getRandom}(r_i)$ 
(13)integral = integral + getf(x[j])/number(MC)
(14)for k=1 to dim
(15)integral = integral * (ul[k] - ll[k])
(16)return integral
(17)end
```

对于一般积分操作,随着维数的增加,计算量将会显著的增加,同时误差也将会随之增大. 而 QueryMC 算法的本质在于随机模拟,收敛速度不依赖于问题的维数,因此用于多维积分计算时,误差与维数和积分区域无关,而且与传统处理方式矩形近似不同,QueryMC 算法没有使用大量循环操作来进

行降维处理,而是直接将多维连续分布随机变量联合概率密度函数作为处理对象,这种方式无疑提高了查询计算的效率.

4 实验结果和分析

实验是在 Intel Core2 CPU(2.5 GHz), 2 GB RAM, 160 GB DISK 的 PC 机上运行, 操作系统是 Windows XP Professional. 工具为 Visual C++ 6.0, 实验测试所用数据集是在 Xmark^[8] 的基础上进行合成, 利用随机化算法产生 4 组不同维数的随机变量联合概率密度函数节点, 每组节点个数为 100, 数据分布类型为多维正态分布.

实验从计算精度与处理时间两个角度对 QueryMC 与矩形近似算法进行比较. 以二维正态分布随机变量节点为例 $X \sim N\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{bmatrix} 4, & 0.8\sigma_1\sigma_2 \\ 0.8\sigma_1\sigma_2, & 16 \end{bmatrix}\right)$, 图 3 中给出了矩形近似算法的步长(Delta)与精度的关系, 图 4 是其相应的处理时间, 可以看出, 随着精度要求的提高, 步长减小, 分段数目增加, 误差将逐渐减小, 而相应地, 处理时间将呈指数增长趋势, 因而对于此类查询问题, 矩形近似算法具有指数时间代价.

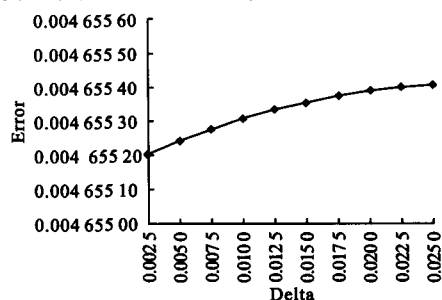


图3 矩形近似算法精度

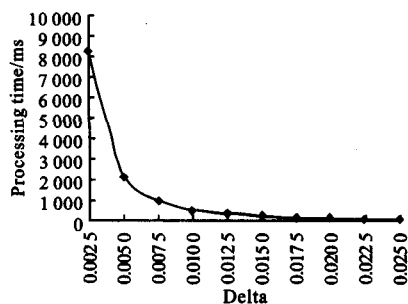


图4 矩形近似算法执行时间

图 5 和图 6 中给出了相同情况下查询在 QueryMC 算法策略下的计算精度与处理时间. 可以看出, 随着精度要求的提高, 样本数目增多, 同时, 处理时间呈线性增长趋势.

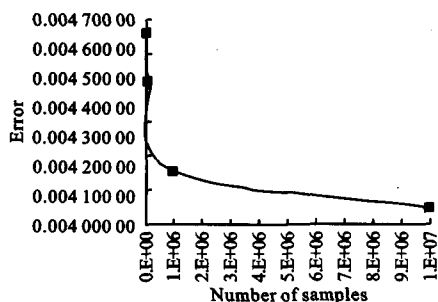


图5 QueryMC 算法精度

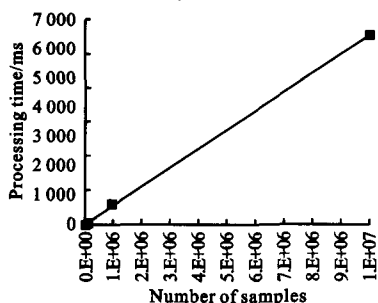


图6 QueryMC 算法执行时间

图7所示为处理数据集中不同维数(dimensionality, dim)节点时的情况,即通过控制矩形近似算法的步长和 QueryMC 算法的样本集大小,使结果在相似的理想精度范围时两种算法的平均处理时间。可以看出,QueryMC 算法在处理时间方面逐渐体现出优势,而矩形近似算法则有相对较高的处理时间代价,并且随着待处理连续不确定 XML 数据维数的增多,这种趋势将越加明显。

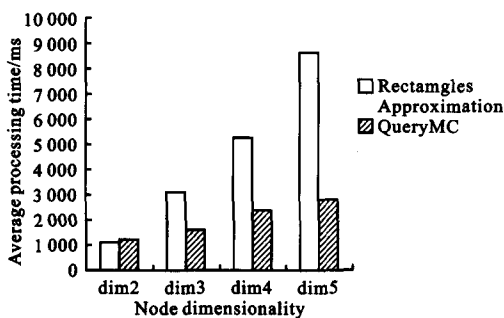


图7 Xmark 数据集上两种算法平均处理时间

5 结束语

在处理多维连续不确定 XML 数据的矩形近似算法的基础上,本文利用小枝模式查询信息,根据多维连续分布概率密度函数自身特点,提出了基于蒙特卡洛思想的 QueryMC 数据处理算法,避免了时间代价较高的降维操作。实验测试结果表明,算法在计算精度,处理时间方面有较好的性能。本文主要考

虑多维连续不确定 XML 数据的普通查询模式,今后的工作将考虑聚集查询,knn, top-k 等更为复杂的查询处理模式。

参考文献:

- [1] Kimelfell B, Sagiv Y. Matching twigs in probabilistic XML[C]//Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, 2007: 27—38.
- [2] Thiagarajan A, Miadden S. Querying continuous functions in a database system[C]//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2008: 791—804.
- [3] Scholte T. Managing continuous uncertain data by a probabilistic XML database management system[R]. Netherlands: University of Twente, 2008.
- [4] Abiteboul S, Kimelfeli B, Sagiv Y, et al. On the expressiveness of probabilistic XML models[J]. The VLDB Journal, 2009, 18(5): 1041—1064.
- [5] Abiteboul S, Chan H, Kharlamov E, et al. Aggregate queries for discrete and continuous probabilistic XML[C]//proceedings of the 13th International Conference on Database Theory. New York: ACM Press, 2010: 50—61.
- [6] Kimelfeld B, Sagiv Y. Modeling and querying probabilistic XML data[J]. The ACM SIGMOD Record, 2008, 37(4): 69—77.
- [7] Kimelfeld B, Kosharovskiy Y, Sagiv Y. Query efficiency in probabilistic XML models[C]//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2008: 701—714.
- [8] Busse R, Carey M, Florescu D, et al. Xmark an XML benchmark project[EB/OL]. [2012-08-10]. <http://monetdb.cwi.nl/xml/index.html>.

作者简介:

张晓琳 女, (1966—), 博士, 教授。研究方向为数据库理论与技术。

霍伟 男, (1986—), 硕士研究生。研究方向为 XML 数据查询技术。

刘立新 女, (1983—), 硕士, 助教。研究方向为 XML 数据管理、数据挖掘。

崔光月 女, (1989—), 硕士研究生。研究方向为 XML 数据模型。