

一种最大向量平均个数的估计方法

杨永滔, 王意洁

(国防科技大学计算机学院并行与分布处理国家重点实验室, 长沙 410073)

摘 要: 提出一种估计 n 个 d 维向量中最大向量平均个数的方法。该方法通过分析单个向量与其他向量子集的支配关系, 求出最大向量平均个数的解析式。证明解析式满足已知的递归关系, 得到最大向量平均个数的近似估计。与已有方法相比, 该方法可应用到估计 k 个其他向量支配的平均个数问题。

关键词: 最大向量; 平均个数; 支配; 近似估计上界; 复杂性计算; skyline 查询

Estimating Method of Maxima Vector Average Number

YANG Yong-tao, WANG Yi-jie

(National Key Laboratory for Parallel and Distributed Processing, School of Computer,
National University of Defense Technology, Changsha 410073, China)

【Abstract】 A new method is proposed for estimating the average number of maxima in a set of n vectors in d -dimensional space. The new method reveals the probability that a vector is dominated by a set of other vectors, and obtains the analytic form of the average number of the maxima. After the fact that the analytic form satisfies a known recurrence is proved, the estimation of the average number can be derived naturally. Compared with the existed methods, the new method can be extended to the general problem of estimating the average number of the vectors that are only dominated by other k vectors.

【Key words】 maxima vector; average number; dominance; asymptotic estimation upper bound; complexity computation; skyline queries

DOI: 10.3969/j.issn.1000-3428.2011.19.017

1 概述

最大向量问题早期在计算几何领域得到广泛研究, 如凸包的计算等^[1-2]。目前, 文献[3]提出的 skyline 查询实际上是最大向量问题在数据处理领域的应用和扩展。随着 skyline 查询得到广泛研究, 最大向量问题也再次得到重视和进一步的研究, 如利用最大向量平均个数的近似估计分析各类 skyline 查询算法的复杂性^[4]。估计最大向量平均个数的方法主要有文献[5-6]的方法; 文献[7]讨论了上述方法给出的近似估计结果在 skyline 查询背景下的应用。文献[5]方法主要基于分析向量支配关系关于维数和向量个数的递归性质, 而文献[6]方法基于分析某个向量成为最大向量的概率。基于以上方法的研究, 本文提出一种估计 n 个 d 维向量中最大向量平均个数的方法。

2 基本概念

2.1 支配关系和最大向量

给定 d 维空间 2 个向量 $a=(a_1, a_2, \dots, a_d)$ 和 $b=(b_1, b_2, \dots, b_d)$, a 支配 b (记作 $a < b$), 当且仅当 $a \neq b$ 且 $a_i \geq b_i (1 \leq i \leq d)$ 。令 D_n 为 d 维空间中 n 个向量的集合, 则最大向量是 D_n 中所有不被其他向量支配的向量。令 $D_{\max}$ 表示所有最大向量的集合。 D_n 是 d 维空间中的任意 n 个向量的集合, 因此, 当 D_n 中包含不同向量时, $D_{\max}$ 的大小一般也会变化。本文研究的问题就是估计 $D_{\max}$ 的平均大小。

2.2 问题假设

估计最大向量平均个数通常做出便于数学处理但比较合理的假设^[8-9]。本文假定各个向量之间是相互独立的; 向量中

各维度的取值也是独立的, 而且在同一维度上, 各向量的取值都不相同。

3 相关工作

目前, 最大向量平均个数的近似估计得到了广泛研究, 但是大部分研究集中于在不同背景下应用关于最大向量平均个数近似估计的结果, 估计的基本方法仍然是文献[5-6]的方法, 本节主要介绍这 2 种方法。

3.1 文献[5]方法

文献[5]将 n 个 d 维向量视作 $n \times d$ 矩阵, 并通过对该取值矩阵的递归分析得到关于最大向量平均个数的递归公式, 最后解递归公式得到最终的近似估计。

根据同一维度上各向量取值都不相同的假设, 某一维度上各向量的 n 个取值可以排序, 并且根据排序结果可以按照相对顺序给这 n 个值分别赋予从 1 到 n 的值, 从而每个维度上的 n 个分量都可以映射为 $(1, 2, \dots, n)$ 上某个置换。因此, n 个 d 维向量实际上对应一个 $n \times d$ 矩阵(也称为值矩阵), 其每列都是 $(1, 2, \dots, n)$ 的一个置换。令 S 表示所有 $(n)^d$ 个值矩阵的集合, $A(n, d)$ 表示最大向量的平均个数。对 S 中每个 $n \times d$

基金项目: 国家“973”计划基金资助项目(2011CB302601); 国家自然科学基金资助项目(60873215); 湖南省自然科学基金杰出青年基金资助项目(S2010J5050); 高等学校博士学科点专项科研基金资助项目(200899980003)

作者简介: 杨永滔(1981—), 男, 博士研究生、CCF 会员, 主研方向: 不确定数据管理, 数据流查询; 王意洁, 教授、博士生导师

收稿日期: 2011-05-20 **E-mail:** scholarnudt@163.com

矩阵 r , 令 $M(r)$ 表示 r 中最大向量个数。则根据 $A(n, d)$ 的定义, 有 $A(n, d) = (\sum_{r \in S} M(r)) / (n!)^d$ 。

文献[5]得到的关于 $A(n, d)$ 的递归公式^[5]如下:

$$A(n, d) = A(n-1, d) + A(n, d-1)/n \quad (1)$$

其中, $n, d \geq 2$ 。式(1)所示递归关系的初值显然为 $A(1, d) = 1 (d \geq 1)$ 和 $A(n, 1) = 1 (n \geq 1)$ 。解出式(1)便得到 $A(n, d)$ 的近似估计为 $O((\ln n)^{d-1})$ (文献[5]中定理 2)。

3.2 文献[6]方法

文献[6]的方法将最大向量问题离散化从而可以使用组合数学技巧来分析。与此不同, Butcha 的方法直接使用初等微分技巧。

文献[6]的方法假定各向量是独立同分布的, 并且各向量 x 的分量分别服从相同的连续分布 $F_i(x_i) (i=1, 2, \dots, d)$ 。文献[6]的第 2 个假设意味着同一维度上各分量取值不同的概率为 1。文献[6]的 2 个假设结合起来便满足第 2.2 节的假设。在文献[6]假设下, 向量 $(x_1, x_2, \dots, x_d)$ 不被某个随机选择的向量支配的概率是 $1 - \prod_{i=1}^d (1 - F_i(x_i))$, 因此它不被其他 $n-1$ 个独立且随机的向量支配的概率是 $(1 - \prod_{i=1}^d (1 - F_i(x_i)))^{n-1}$ 。该概率关于 $x_i (i=1, 2, \dots, d)$ 的积分的便是某个随机向量在 n 个向量集合中成为最大向量的概率; 因为所有向量是同分布的, 则:

$$A(n, d) = n \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left(1 - \prod_{i=1}^d (1 - F_i(x_i)) \right)^{n-1} \prod_{j=1}^d dF_j(x_j) =$$

$$n \int_0^1 \dots \int_0^1 (1 - x_1 \dots x_d)^{n-1} dx_1 dx_2 \dots dx_d$$

最后可得:

$$A(n, d) = \frac{1}{(d-1)!} (\ln n)^{d-1} + \frac{\gamma}{(d-2)!} (\ln n)^{d-2} + O((\ln n)^{d-3})$$

由上式可知, 文献[6]方法得到的 $A(n, d)$ 的近似估计也是 $O((\ln n)^{d-1})$, 但更精确。

4 最大向量平均个数的估计方法

本文提出的新估计方法与文献[5]方法一样, 也采取离散化的方法。与已有的文献[5-6]方法不同的是, 本文方法将最大向量平均个数的估计问题置于一个一般性框架下考虑, 因此, 本文方法除了给出最大向量平均个数的近似估计, 还可以给出一般的仅被 k 个其他向量支配的向量平均个数的近似估计 ($k=0$ 时即最大向量)。

令 $D_n = \{a^1, a^2, \dots, a^n\}$ 表示 n 个向量的集合, a^i 表示向量 a^i 的第 l 个分量。以 D_n 中每个向量作为一行, 即形成文献[5]方法中的一个值矩阵 r 。同样令 S 表示所有值矩阵的集合。在给定值矩阵 r 下可以定义 M_i^r 为 $\{j | a^j \in (D_n - \{a^i\}) \wedge a^j < a^i\}$, 即 M_i^r 是给定值矩阵 r 下 D_n 中所有支配 a^i 的向量的索引序号的集合。令 $I_n = \{1, 2, \dots, n\}$, 则 M_i^r 是 I_n 的子集。

令 $P(M < i)$ 表示当且仅当 D_n 中下标在 M 中的向量都支配 a^i 的概率。因为 a^i 不可能支配自身, 所以如果 $i \in M$, 则 $P(M < i) = 0$; 因此, 只需考虑 $M \subseteq I_n - \{i\}$ 的情况。此外, 因为各值矩阵都是等概率出现的, 所以 $P(M < i) = \sum_{r \in S, M_i^r = M} 1/|S|$ 。

4.1 最大向量的平均个数估计

引理 1 令 $|M|$ 表示集合 M 的基数, 则对于 I_n 的 2 个子集 M_1 和 M_2 , 如果 $|M_1| = |M_2|$, 则 $P(M_1 < i) = P(M_2 < i)$ 。

证明: 令 $S_1 = \{r | r \in S \wedge M_i^r = M_1\}$, $S_2 = \{r | r \in S \wedge M_i^r = M_2\}$ 。因为 $P(M < i) = \sum_{r \in S, M_i^r = M} 1/|S|$, 所以只需证明 $|S_1| = |S_2|$ 即可。

任给 $r \in S_1$, 因为 $|M_1| = |M_2|$, 所以可对 r 做行变换得到 r' , 使得 $|M_2| = |M_i^{r'}|$, 则 $r' \in S_2$, 因此存在 S_1 到 S_2 的映射。同理可证存在 S_2 到 S_1 的映射, 因此 S_1 和 S_2 之间存在一一映射, 则 $|S_1| = |S_2|$ 。由此可知引理成立。

引理 2 任给 $M \subseteq I_n - \{i\}$, 且 $|M| = k$, 则:

$$\sum_{M' \supseteq M} P(M' < i) = 1/(k+1)^d \quad (2)$$

其中, M' 是 $I_n - \{i\}$ 的子集; d 是向量的维度。

分以下 2 种情况证明:

(1) 当 $k=0$ 时, $M = \emptyset$, 则 $I_n - \{i\}$ 的任意子集都包含 M , 而任给 $r \in S$, M_i^r 都是 $I_n - \{i\}$ 的子集; 因此 $\sum_{M' \subseteq I_n - \{i\}} P(M' < i)$ 涉及了 S 中所有的 r , 因此它的值为 1, 即 $1/(0+1)^d$ 。

(2) 当 $k \neq 0$ 时, 令 $V_l = \{a^j | j \in M\} \cup \{a^i\}$, 则 a^i 是 V_l 中最小值的概率是 $1/(k+1)$ 。“任何维度 l 上 a^i 都是 V_l 中最小值”等价于“集合 $\{a^j | j \in M\}$ 中的向量都支配 a^i ”, 并且因为各维之间独立, 所以该事件发生的概率是 $1/(k+1)^d$ 。令 S_1 表示所有满足该事实的值矩阵集合, 则 $|S_1|/|S| = 1/(k+1)^d$ 。令 S_2 表示所有满足 $M_i^r \supseteq M$ 的值矩阵 r 的集合。显然, $S_1 = S_2$ 。 $P(M < i) = \sum_{r \in S, M_i^r = M} 1/|S|$, 则 $|S_2|/|S| = \sum_{M \subseteq M', M' \subseteq I_n - \{i\}} P(M' < i)$, 则引理成立。

以 $k=1, n=2, d=2$ 的简单情况为例。 D_n 有 4 种不同取值 (即值矩阵 r), 分别是 $\{(1,1), (2,2)\}$, $\{(1,2), (2,1)\}$, $\{(2,1), (1,2)\}$ 和 $\{(2,2), (1,1)\}$ 。因为只有当 r 为 $\{(2,2), (1,1)\}$ 时, $M_2^r = \{1\}$, 所以 $P(\{1\} < 2) = 1/4$ 。而根据引理 2, 当 $M = \{1\}$ 时, $I_2 - \{2\}$ 的子集中惟有其本身包含 $\{1\}$, 因此, $P(\{1\} < 2) = 1/(1+1)^2 = 1/4$, 与根据前面 $P(M < i)$ 的定义得出的结果是一样的。

根据引理 1, 可以将满足 $|M| = k$ 的概率 $P(M < i)$ 统一记作 m_k 。关于 m_k 有如下定理 (C_n^k 表示 n 个对象中选 k 个的组合数)。

定理 任给:

$$k (0 \leq k \leq n-1), \quad P(M < i) = m_k = \sum_{j=0}^{n-1-k} \frac{(-1)^j C_{n-1-k}^j}{(k+j+1)^d}$$

证明: 如果 $|M| = k (0 \leq k \leq n-1)$, 则 $I_n - \{i\}$ 中满足 $M' \supseteq M$ 且 $|M'| = k+i (0 \leq i \leq n-1-k)$ 的 M' 的个数是 C_{n-1-k}^i 。式(2)可以写作 $\sum_{i=0}^{n-1-k} C_{n-1-k}^i m_{k+i} = 1/(k+1)^d$, 则可得到如下方程组:

$$\begin{aligned} m_{n-1} &= 1/n^d \\ m_{n-2} + m_{n-1} &= 1/(n-1)^d \\ &\vdots \\ C_{n-1-k}^0 m_k + C_{n-1-k}^1 m_{k+1} + \dots + C_{n-1-k}^{n-2-k} m_{n-2} + C_{n-1-k}^{n-1-k} m_{n-1} &= 1/(k+1)^d \\ &\vdots \\ C_{n-1}^0 m_0 + C_{n-1}^1 m_1 + \dots + C_{n-1}^{n-2} m_{n-2} + C_{n-1}^{n-1} m_{n-1} &= 1 \end{aligned}$$

解出方程组便得到定理的结论。

定理表明当 $k=0$ 时, m_0 即是 a_i 为最大向量的概率, 并且 m_0 与 i 无关, 则:

$$A(n, d) = \sum_{i=1}^n P(\emptyset < i) = n \cdot m_0 = n \sum_{j=0}^{n-1} \frac{(-1)^j C_{n-1}^j}{(j+1)^d} \quad (3)$$

容易验证式(3)满足文献[5]方法中的递归关系 $A(n, d) = A(n-1, d) + A(n, d-1)/n$ 及其初始条件 $A(1, d) = 1$ 和 $A(n, 1) = 1$ 。因此, 便得到 $A(n, d)$ 的近似估计为 $O((\ln n)^{d-1})$ 。

4.2 k 个向量的支配情况

文献[5-6]方法以及式(3)中讨论的 $A(n, d)$ 只是定理在

$k=0$ 时的特例的应用, 实际上定理的结论可以推广到估计一般地被 k 个其他向量支配的向量平均个数的问题。

令:

$$S(n, d) = nC_{n-1}^k m_k = nC_{n-1}^k \sum_{j=0}^{n-1-k} \frac{(-1)^j C_{n-1-k}^j}{(k+j+1)^d}$$

其中, $n \geq k+1$; $d \geq 1$; $k \geq 0$ 。因为 $I_n - \{i\}$ 中基数为 k 的不同子集个数为 C_{n-1}^k , 所以 $S(n, d)$ 即为仅被 k 个其他向量支配的向量的平均个数。本文将文献[5]方法求解递归关系的过程推广到一般情况得出 $S(n, d)$ 的上界估计。

首先, 容易验证下面的递归关系:

$$S(n, d) = S(n-1, d) + S(n, d-1)/n \quad (4)$$

其中, $n \geq k+2$; $d \geq 2$ 。得到递归关系式(4)的初始条件为:

$$(1) S(n, 1) = 1, n \geq k+1。$$

$$(2) S(k+1, d) = 1/(k+1)^{d-1}, d \geq 1。$$

建立生成函数 $G_n(z) = \sum_{d=0}^{\infty} S(n, d+1)z^d (n \geq k+1)$, 则由式(4)可得:

$$G_n(z) = G_{n-1}(z)/(1-z/n) (n \geq k+2) \quad (5)$$

根据第(2)个初始条件可得:

$$G_{k+1}(z) = 1/(1-z/(k+1))$$

结合式(5), 得:

$$G_n(z) = \prod_{i=k+1}^n 1/(1-z/i)$$

令 $S_j = \sum_{i=k+1}^n 1/i^j (j \geq 1)$, 则根据文献[8]中式(33)~式(38)

的分析, 可知 $G_n(z)$ 中的 z^d 的系数, 即 $S(n, d+1)$, 满足:

$$S_1^d/d! \leq S(n, d+1) \leq S_1^d (k \geq 0) \quad (6)$$

当 $k=0$ 时, 根据文献[5]的方法, $S(n, d+1) = A(n, d+1) = O((\ln n)^d)$; 当 $k \geq 1$ 时, 根据调和数的近似公式 $\sum_{i=1}^n 1/i = O(\ln n)$, 可知 $S_1 = O(\ln n - \ln k)$, 则由式(6)可得: $S(n, d+1) = O((\ln n - \ln k)^d)$ 。最后可得:

$$S(n, d) = \begin{cases} O((\ln n)^{d-1}) & k=0 \\ O((\ln n - \ln k)^{d-1}) & k \geq 1 \end{cases}$$

(上接第 55 页)

了优越性。

(4)返回 path, 算法结束。

综上所述, 本文的演化方法同时处理多个演化需求, 分析演化需求之间的关系, 从局部范围内全局地选择演化路径, 减少了演化策略选择及回溯次数, 降低了演化代价。文献[6]的 MINI 算法及文献[7]的 COST 算法只是本文算法的一个特例, 它们在这样的情形下与本文算法的效果是一致的: 上下文窗口内的所有划分都只含一个单位演化需求即所有演化需求都是 disrelated 关系, 因此本文的算法时间和空间复杂度始终小于或等于上述结果, 表明本文算法更优。

4 结束语

本文提出基于上下文窗口的演化代价约束的个体演化方法, 实验表明其相对于传统方法降低了演化代价。本文算法的优势在于分析了演化需求之间的关系, 利用上下文窗口内信息进行路径选择决策。

由于潜在代价需要领域专家的参与, 进一步的研究工作是寻找减少人工参与的个体演化方法, 以及个体演化后如何与客观实际保持一致, 例如文中例子将 worker 设为概念 animal 的实例。

5 结束语

本文研究一组向量中最大向量的平均个数的问题, 提出了一种新的估计方法。该方法分析了单个向量被其他向量或向量集合支配的情况, 得出向量被某个基数为 k 的向量支配的概率。以此为基础, 不仅可以证明使用已有方法得出的结论, 还可以推广到估计一般的被 k 个其他向量支配的向量的平均个数的问题。

参考文献

- [1] Preparata F P. Computational Geometry[M]. New York, USA: Springer, 1985.
- [2] 张忠武, 吴信才. 平面海量散乱点集凸壳算法[J]. 计算机工程, 2009, 35(9): 43-45.
- [3] Borzsonyi S, Kossmann D, Stocker K. The Skyline Operator[C]// Proceedings of the 17th International Conference on Data Engineering. Washington D. C., USA: IEEE Computer Society, 2001: 421-430.
- [4] Godfrey P, Shipley R, Gryz J. Algorithms and Analyses for Maximal Vector Computation[J]. The VLDB Journal, 2007, 16(1): 5-28.
- [5] Bentley J L, Kung H T, Schkolnick M, et al. On the Average Number of Maxima in a Set of Vectors and Applications[J]. Journal of the Association for Computing Machinery, 1978, 25(4): 536-543.
- [6] Buchta C. On the Average Number of Maxima in a Set of Vectors[J]. Information Processing Letters, 1989, 33(2): 63-65.
- [7] Godfrey P. Skyline Cardinality for Relational Processing[C]// Proceedings of FoIKS'04. Berlin, Germany: Springer-Verlag, 2004: 78-97.
- [8] Knuth D E. 计算机程序设计艺术[M]. 苏运霖, 译. 北京: 国防工业出版社, 2007.

编辑 索书志

参考文献

- [1] Bernes L T, Handler J, Lassila O. The Semantic Web[EB/OL]. (2001-05-17). <http://www.scientificamerican.com/article.cfm?id=the-semantic-web>.
- [2] Stojanovic L, Maedche A, Motik B, et al. User Driven Ontology Evolution Management[C]//Proc. of the 13th Int'l Conf. on Knowledge Engineering and Knowledge Management. [S. l.]: IEEE Press, 2002.
- [3] Ljiljana S, Alexander M, Nenad S, et al. Ontology Evolution as Reconfiguration Design Problem Solving[C]//Proc. of the 2nd Int'l Conf. on Knowledge Capture. New York, USA: ACM Press, 2003.
- [4] 宋免寰, 张志祥. 基于概念代数的本体演化[J]. 计算机工程, 2009, 35(13): 23-25.
- [5] 金龙飞, 刘 磊. 一种本体演化波及效应分析方法[J]. 电子学报, 2006, 34(8): 1469-1474.
- [6] 刘 晨, 韩燕波, 陈旺虎, 等. MINI——一种可减小变更影响范围的本体演化算法[J]. 计算机学报, 2008, 31(5): 711-720.
- [7] 周 栩, 罗景文, 周 桐, 等. 一种基于演化代价约束的本体演化方法[J]. 吉林大学学报, 2010, 48(4): 646-652.

编辑 陈 文

