ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.7, July 2010, pp.1550−1560
doi: 10.3724/SP.J.1001.2010.03622

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

基于容斥原理的 Skyband 基数估计方法*

赵加奎[1,2+], 杨冬青[1], 陈立军[1]

[1](高可信软件技术教育部重点实验室 北京大学 信息科学技术学院,北京 100871)

[2](中国电力科学研究院,北京 100085)

Skyband Cardinality Estimation Based on the Inclusion-Exclusion Principle

ZHAO Jia-Kui[1,2+], YANG Dong-Qing[1], CHEN Li-Jun[1]

[1](Key Laboratory of High Confidence Software Technologies of the Ministry of Education, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)

[2](China Electric Power Research Institute, Beijing 100085, China)

+ Corresponding author: E-mail: jkzhao@pku.edu.cn

Zhao JK, Yang DQ, Chen LJ. Skyband cardinality estimation based on the inclusion-exclusion principle. *Journal of Software*, 2010,21(7):1550−1560. http://www.jos.org.cn/1000-9825/3622.htm

Abstract: Skyband queries are very important for decision-making applications. To incorporate the Skyband operator into the database system, the problem of Skyband cardinality estimation must be solved, i.e., estimating the number of the Skyband elements returned by Skyband queries, which is very important for extending the query optimizer's cost model to accommodate Skyband queries. This paper proposes a space and time efficient approach to estimate the Skyband cardinality, which is based on the generalized form of the Inclusion-Exclusion Principle. Experimental results show that the proposed approach can estimate the Skyband cardinality accurately.

Key words: cardinality; Skyband query; Skyline query; database system; query optimization

摘 要: Skyband 查询是决策支持领域一类非常重要的查询.为了使数据库系统有效支持 Skyband 查询,必须解决 Skyband 基数估计的问题,即估计 Skyband 查询结果中包含的 Skyband 元素数,因为 Skyband 基数估计对于扩展数据库系统查询优化器的代价模型以便能够对 Skyband 查询进行优化非常重要.基于容斥原理的推广形式对 Skyband 基数进行理论分析并给出了时间和空间代价很小的对 Skyband 基数进行估计的算法.实验结果表明,该方法能够准确地对 Skyband 基数进行估计.

关键词: 基数;Skyband 查询;Skyline 查询;数据库系统;查询优化

中图法分类号: TP311 **文献标识码**: A

Skyline 查询[1]是决策支持领域一类非常重要的查询,因为 Skyline 查询能够返回所有不被任何其他元素支配的最优元素.但是,Skyline 查询可能忽略一些本身有价值但却被少量其他元素支配的元素,因为 Skyline 查询所涉及到的维往往无法覆盖决策支持用户所考虑的所有因素,在低维数据集中,这个问题尤为突出.因

* Supported by the National High-Tech Research and Development Plan of China under Grant Nos.2007AA01Z153, 2007AA01Z191 (国家高技术研究发展计划(863))

Received 2008-09-16; Revised 2008-11-27; Accepted 2009-03-31

此,Papadias 等人[2]将 Skyline 查询推广到 Skyband 查询并给出了利用最邻近查询[3]来处理 Skyband 查询的方法.一个 r-skyband 查询返回所有最多被其他 r 个元素支配的元素.定义 1 和定义 2 分别给出了支配和 r-skyband 元素的形式化定义.通过 r-skyband 元素的形式化定义可以看出,Skyline 查询只是 Skyband 查询的一个特例,即 0-skyband 查询.

定义 1(支配). 设 ξ_1 和 ξ_2 为 k 维空间内的两个元素,不失一般性,如果 ξ_1 在每一维上的值都不大于 ξ_2 在对应维上的值并且 ξ_1 至少在某一维上的值小于 ξ_2 在对应维上的值,则我们说 ξ_1 支配 ξ_2,记为 $\xi_1 \succ \xi_2$.

定义 2(r-skyband 元素). 如果空间内能够支配元素 ξ 的元素不超过 r 个,则 ξ 是一个 r-skyband 元素.

图 1 用一个二维决策支持问题的例子说明了 Skyline(即 0-skyband)与 1-skyband 的区别.酒店数据包含两维,即价格和到海边的距离,用户倾向于选择那些价格便宜且距海边又近的酒店.Skyline 查询返回 3 个酒店(a,d,f)供用户选择,而 1-skyband 查询返回额外 4 个酒店(b,c,e,j)供用户选择.由于酒店数据只包含两维,虽然酒店 b 比酒店 f 价格高离海边又远,但可能酒店 b 的服务好又有停车场,而酒店 f 服务相对较差又没有停车场;因此,将酒店 b 返回给用户供其选择是必要的.Buchta[4]证明了在一个包含 n 个元素的 k 维空间内包含的 Skyline 元素数的数量级为 $O((\ln n)^{k-1}/(k-1)!)$.因此,低维空间的 Skyline 查询通常返回少量元素供用户选择,而部分本身有价值但却被少量其他元素支配的元素将无法返回给用户,原因在于低维空间内的元素被其他元素支配的概率较高.Skyband 查询可以返回那些本身有价值但却被少量其他元素支配的元素,因此是决策支持领域一类非常重要的查询.

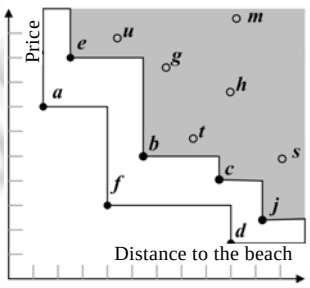


Fig.1 Skyline vs. 1-skyband

图 1 Skyline 与 1-skyband 的比较

为了使数据库系统能够有效地支持 Skyband 查询,必须解决 Skyband 基数估计的问题,即估计 Skyband 查询结果中包含的 Skyband 元素数,因为查询结果集大小的估计对于扩展数据库系统查询优化器的代价模型以便能够对包含 Skyband 的查询进行优化非常重要.查询优化的目的是为了得到一个良好的查询计划,而查询结果集大小的估计是查询优化经常遇到的一类问题.查询计划通常是一棵带标记的树,树的每个节点是一个操作(选择,投影,连接等),而标记是确定用什么样的方法来执行这个操作(例如,可以用嵌套循环连接或归并连接来执行连接操作).为了得到良好的查询计划,查询优化器通常要估计每个操作返回的结果集大小.例如,查询中的复杂谓词通常被转换成彼此之间为“与”关系的多个独立谓词,每个独立谓词对应一个选择操作,顺序执行每个选择操作得到最终结果.为了降低中间结果缓存的空间开销以及谓词判断的时间开销,查询优化器通常将返回结果集小的选择操作放在查询计划的前端.因此,查询优化器需要估计每个选择操作返回结果集的大小,这个任务通常通过估计选择操作对应谓词的选择度来实现[5].对于一个复杂查询,Skyband 查询往往只是其查询计划中的一个操作,为了降低整个查询执行过程中的中间结果缓存的空间开销以及求解各个操作的时间开销,从而得到一个良好的查询计划,查询优化器需要对 Skyband 查询结果集的大小进行估计.对于独立的 Skyband 查询,Skyband 查询结果集大小的估计对于查询优化器分配缓存存储查询结果也非常重要.本文在各维之间是无关的以及每一维上都不存在重复值的假设下,利用容斥原理的推广形式[6]对 Skyband 基数进行理论分析,并给出时间和空间代价很小的对 Skyband 基数进行估计的动态规划算法.实验结果表明,我们的算法能够对 Skyband 基数进行准确估计.各维之间是无关的这个假设被查询优化器广泛应用于对高维查询进行优化,因为多维之间

的相关性很难用数学方法来精确建模.每一维上都不存在重复值这个假设可以简化 Skyband 基数的理论分析,但本文给出的 Skyband 基数分析方法是通用的,即可以不基于无重复值假设而对 Skyband 基数进行分析,只是推导过程稍显复杂.

近几年,Skyline 查询备受关注.在数据库中求解 Skyline 的算法可以被分成非基于索引的算法和基于索引的算法.非基于索引的算法包括 D&C 算法[1],BNL 算法[1],SFS 算法[7]以及 LESS 算法[8,9]等.基于索引的算法包括基于位图索引的 Bitmap 算法[10,11]、基于 B+树索引的 Index 算法[10,11]以及基于 R 树的 NN 算法[12]和 BBS 算法[13]等.更多关于 Skyline 查询处理的研究成果可参见综述文献[14].通常情况下,基于索引的算法优于非基于索引的算法.

如前所述,Skyline 查询可能忽略一些本身有价值但却被少量其他元素支配的元素,因为其所涉及到的维往往无法覆盖决策支持用户所考虑的所有因素,在低维数据集中,这个问题尤为突出.因此,Papadias 等人[2]将 Skyline 查询推广到 Skyband 查询,并给出了利用最邻近查询[3]来处理 Skyband 查询的方法.与本文相关度较高的是 Skyline 基数估计.在各维之间是无关的以及每一维上都不存在重复值的假设下,Bentley 等人[15]和 Godfrey[16]利用归纳法推导出递推方程 $\Psi(n,k)=\Psi(n-1,k)+\Psi(n,k-1)/n$,递推方程的初始条件为 $\Psi(n,1)=1(n\geq 1)$和 $\Psi(1,k)=1(k\geq 1)$.可以利用这个递推方程来估计包含 n 个元素的 k 维空间内包含的 Skyline 元素数.但无法用这个递推方程来估计 Skyband 基数;并且,在 $n>k$ 的假设下,直接利用这个递推方程估计 Skyline 基数的时间代价为 $O(2^n)$,这个时间代价在 n 较大的情况下是无法接受的.

本文以容斥原理的推广形式为理论基础,利用概率方法对 Skyband 基数直接进行理论分析,并给出了时间和空间代价很小的对 Skyband 基数进行估计的第 1 种算法.由于 Skyline 是 Skyband 的一个特例,本文的成果同样适用于对 Skyline 基数进行估计.另外,虽然本文也是基于各维之间是无关的以及每一维上都不存在重复值的假设,但我们给出的分析方法适用于维上存在重复值的情况.

本文第 1 节介绍对 Skyband 基数进行理论分析所需的预备知识.第 2 节利用第 1 节介绍的预备知识对 Skyband 基数进行理论分析,并给出时间和空间代价很小的对 Skyband 基数进行估计的动态规划算法.第 3 节用实验证明我们给出的 Skyband 基数估计算法的准确性.第 4 节总结全文.

1 预备知识

本文对 Skyband 基数的理论分析是基于容斥原理的推广形式的.为了确保本文是自包含的,我们在定理 1 中给出了容斥原理的推广形式,具体的证明参见文献[6].为了使容斥原理能够更加方便地应用于对 Skyband 基数进行理论分析,我们给出并证明了定理 2.定理 1 给出了一个有穷集合 S 中恰好满足 m 条性质中的 r 条性质的元素数 $\Delta(m,r)$,而定理 2 给出了一个有穷集合 S 中满足 m 条性质中的最多 r 条性质的元素数 $\Gamma(m,r)$.

定理 1(容斥原理的推广形式). 设 S 是一个有穷集合,有 m 条性质,$S_1S_2...S_m$ 为 S 的 m 个子集,S_i 由 S 中满足第 i 条性质的元素构成;设 S 中恰好满足 m 条性质中的 r 条性质的元素数为 $\Delta(m,r)$,则有

$$\Delta(m,r)=\sum_{i=r}^{m}(-1)^{i-r}\binom{i}{r}T(i),$$

其中,$T(i)$由下面的等式确定:

$$T(0)=|S|,$$

$$T(1)=\sum_{i=1}^{m}|S_i|,$$

$$T(2)=\sum_{1\leq i_1<i_2\leq m}|S_{i_1}\cap S_{i_2}|,$$

$$...,$$

$$T(m)=|S_1\cap S_2\cap...\cap S_m|.$$

定理 2. 设 S 是一个有穷集合,有 m 条性质,$S_1S_2...S_m$ 为 S 的 m 个子集,S_i 由 S 中满足第 i 条性质的元素构成;设 S 中满足 m 条性质中的最多 r 条性质的元素数为 $\Gamma(m,r)$,则有

$$\Gamma(m,r)=T(0)+\sum_{i=r+1}^{m}(-1)^{i-r}\binom{i-1}{r}T(i).$$

证明:根据定理 1,$\Gamma(m,r)$可以被如下推导:

$$\Gamma(m,r)=\sum_{j=0}^{r}\Delta(m,j)=\sum_{i=0}^{r}\sum_{j=0}^{i}(-1)^{i-j}\binom{i}{j}T(i)+\sum_{i=r+1}^{m}\sum_{j=0}^{r}(-1)^{i-j}\binom{i}{j}T(i)=T(0)+\sum_{i=r+1}^{m}(-1)^{i-r}\binom{i-1}{r}T(i).$$ □

设 S 是一个有穷集合,有 3 条性质,$S_1S_2S_3$ 为 S 的 3 个子集,S_i 由 S 中满足第 i 条性质的元素构成;根据定理 1,恰好满足 3 条性质中的两条性质的元素数:

$$\Delta(3,2)=T(2)-3T(3)=|S_1\cap S_2|+|S_1\cap S_3|+|S_2\cap S_3|-3|S_1\cap S_2\cap S_3|.$$

根据定理 2,满足 3 条性质中的最多一条性质的元素数:

$$\Gamma(3,1)=T(0)-T(2)+2T(3)=|S|-|S_1\cap S_2|-|S_1\cap S_3|-|S_2\cap S_3|+2|S_1\cap S_2\cap S_3|.$$

设 S 中一共有 20 个元素.其中,满足性质 1 的元素有 10 个,满足性质 2 的元素有 9 个,满足性质 3 的元素有 8 个,同时满足性质 1 和性质 2 的元素有 3 个,同时满足性质 2 和性质 3 的元素有 3 个,同时满足性质 1 和性质 3 的元素有 3 个,同时满足 3 条性质的元素有 1 个,则恰好满足 3 条性质中的两条性质的元素数为

$$T(2)-3T(3)=3+3+3-3\times1=6,$$

满足 3 条性质中的最多一条性质的元素数为

$$T(0)-T(2)+2T(3)=20-(3+3+3)+2\times1=13.$$

2 基数估计

对 Skyband 基数进行估计的本质问题是要求得 m 个其他元素中最多有 r 个元素能够支配某一元素的概率,求得了这个概率即可求得某一元素成为 Skyband 元素的概率,亦即可以对 Skyband 基数进行估计.以容斥原理的推广形式为基础的定理 2 给出了求解一个有穷集合中满足 m 条性质中的最多 r 条性质的元素数的方法.将其中的求解满足 m 条性质中的最多 r 条性质的元素数的问题映射到求解 m 个元素中最多 r 个元素能够支配某一元素的概率的问题,即可求得某一元素成为 Skyband 元素的概率.因此,容斥原理恰好可以用来对 Skyband 基数进行估计.本节在各维之间是无关的以及每一维上都不存在重复值的假设下对 Skyband 基数进行理论分析,并给出时间和空间代价很小的对 Skyband 基数进行估计的算法.通过将定理 2 中的求解满足 m 条性质中的最多 r 条性质的元素数的问题映射到求解 m 个元素中最多 r 个元素能够支配某一元素的概率的问题,引理 1 给出并证明了一个 k 维空间内的 m 个元素中最多 r 个元素能够支配某一元素的概率 $\mathrm{P}\{D_r(m,k)\}$.从引理 1 的证明可以看出,只要给出各维数据的分布函数,即可求得 $T'(i)$.因此,本文给出的分析方法适用于分析维上存在重复值的情况.

引理 1. 设$\xi_0\xi_1\xi_2\ldots\xi_m$ 为 k 维空间内的 m+1 个元素,空间的各维之间是无关的并且每一维上都不存在重复值;设 $\mathrm{P}\{D_r(m,k)\}$为 k 维空间内的 m 个元素$\xi_1\xi_2\ldots\xi_m$中最多 r 个元素能够支配ξ_0的概率,则有:

$$\mathrm{P}\{D_r(m,k)\}=1+\sum_{i=r+1}^{m}\frac{(-1)^{i-r}}{(i+1)^k}\binom{i-1}{r}\binom{m}{i}.$$

证明:把定理 2 中的有穷集合 S,第 i 条性质以及第 i 个子集 S_i 分别映射到整个概率空间,ξ_i 支配ξ_0 以及ξ_i 支配ξ_0的概率,则定理 2 中的 $T(i)$可以被映射到 $T'(i)$并重新定义如下:

$$T'(0)=1,$$

$$T'(1)=\sum_{i=1}^{m}\mathrm{P}\{\xi_i\succ\xi_0\},$$

$$T'(2)=\sum_{1\le i_1<i_2\le m}\mathrm{P}\{\xi_{i_1}\succ\xi_0\wedge\xi_{i_2}\succ\xi_0\},$$

$$\ldots,$$

$$T'(m)=\mathrm{P}\{\xi_1\succ\xi_0\wedge\xi_2\succ\xi_0\wedge\ldots\wedge\xi_m\succ\xi_0\}.$$

一个元素与其他 i 个元素一共 i+1 个元素,由于每一维上都不存在重复值,则 i+1 个元素中的任一元素在某一维

上成为 i+1 个元素中最差元素的概率为 1/(i+1),即任一元素在某一维上被所有其他 i 个元素支配的概率为 1/(i+1).由于空间的各维之间是无关的,则一个元素在 k 维上被所有其他 i 个元素支配的概率为 $1/(i+1)^k$,则有:

$$T'(i)=\binom{m}{i}\frac{1}{(i+1)^k}.$$

根据定理 2,P{$D_r(m,k)$}可以被如下推导:

$$\mathrm{P}\{D_r(m,k)\}=T'(0)+\sum_{i=r+1}^{m}(-1)^{i-r}\binom{i-1}{r}T'(i)=1+\sum_{i=r+1}^{m}\frac{(-1)^{i-r}}{(i+1)^k}\binom{i-1}{r}\binom{m}{i}.$$ □

根据引理 1,k 维空间内的 3 个元素$\xi_1\xi_2\xi_3$中最多有一个元素能支配ξ_0的概率 P{$D_1(3,k)$}可以表示如下:

$$\begin{aligned}\mathrm{P}\{D_1(3,k)\}&=T'(0)+\sum_{i=2}^{3}(-1)^{i-1}\binom{i-1}{1}T'(i)=T'(0)-(T'(2)-2(T'(3))\\&=1-(\mathrm{P}\{\xi_1\succ\xi_0\wedge\xi_2\succ\xi_0\}+\mathrm{P}\{\xi_1\succ\xi_0\wedge\xi_3\succ\xi_0\}+\\&\quad\mathrm{P}\{\xi_2\succ\xi_0\wedge\xi_3\succ\xi_0\}-2\mathrm{P}\{\xi_1\succ\xi_0\wedge\xi_2\succ\xi_0\wedge\xi_3\succ\xi_0\}).\end{aligned}$$

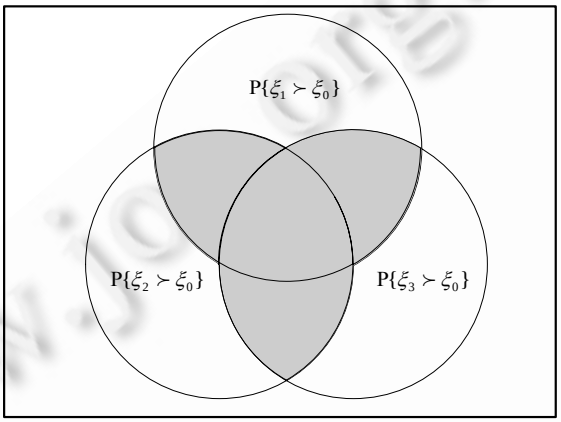


Fig.2 Probability that ξ_0 is dominated by at most one of $\xi_1\xi_2\xi_3$

图 2 ξ_0 至多被$\xi_1\xi_2\xi_3$中的一个元素支配的概率

图 2 给出了 P{$D_1(3,k)$}的 Venn 图.整个矩形的面积对应完整的概率空间,即 $T'(0)$,其值等于 1.3 个圆形的面积分别对应ξ_1支配ξ_0的概率,ξ_2支配ξ_0的概率以及ξ_3支配ξ_0的概率.灰色区域的面积对应于 3 个元素$\xi_1\xi_2\xi_3$中至少两个元素能够支配ξ_0的概率,其值等于 $T'(2)-2T'(3)$.因此,

$$\mathrm{P}\{D_1(3,k)\}=T'(0)-(T'(2)-2T'(3)).$$

设 k=2,即空间为二维的,根据引理 1 的证明可知,从$\xi_1\xi_2\xi_3$中选两个元素,这两个元素都能支配 ξ_0 的概率为 $1/3^2=1/9$,$\xi_1\xi_2\xi_3$这 3 个元素都能支配ξ_0的概率为 $1/4^2=1/16$.因此,3 个二维元素$\xi_1\xi_2\xi_3$中最多有一个元素能支配ξ_0的概率为

$$\mathrm{P}\{D_1(3,2)\}=1-(\mathrm{P}\{\xi_1\succ\xi_0\wedge\xi_2\succ\xi_0\}+\mathrm{P}\{\xi_1\succ\xi_0\wedge\xi_3\succ\xi_0\}+\mathrm{P}\{\xi_2\succ\xi_0\wedge\xi_3\succ\xi_0\}-2\mathrm{P}\{\xi_1\succ\xi_0\wedge\xi_2\succ\xi_0\wedge\xi_3\succ\xi_0\})=19/24.$$

基于引理 1,定理 3 给出了在一个包含 n 个元素的 k 维空间内包含的 r-skyband 元素数的期望值 $\Psi_r(n,k)$.由定理 3 可知,在一个包含 4 个元素的 2 维空间内包含的 1-skyband 元素数的期望值 $\Psi_1(4,2)=19/6=4\mathrm{P}\{D_1(3,2)\}$.但是,定理 3 给出的求解 $\Psi_r(n,k)$的公式中包含组合数,这使得 $\Psi_r(n,k)$的值非常难以计算.例如,从 100 个不同元素中选取 50 个元素的方法数超过了一个 64 位整数的存储范围.鉴于此,定理 4 给出了一种递归求解 $\Psi_r(n,k)$的方法,这种方法不会导致整数溢出.但是,在 $n>k$ 的合理假设下,直接利用这个递推方程估计 Skyband 基数的时间代价为 $O(2^{n-r})$,而这个时间代价在 $n-r$ 较大的情况下是无法接受的.图 3 给出了直接利用定理 4 给出的递推方程求解 $\Psi_1(4,3)$的例子,从这个例子可以看出,直接利用定理 4 给出的递推方程求解 $\Psi_1(4,3)$包含大量重复计算,即图中用虚线画出的部分.如果能够去除重复计算,利用定理 4 给出的递推方程求解 $\Psi_r(n,k)$的时间代价将会显著降低.

定理 3. 设 k 维空间内有 n 个元素,空间的各维之间是无关的并且每一维上都不存在重复值;设 $\Psi_r(n,k)$为空间内 r-skyband 元素数的期望值,则有:

$$\Psi_r(n,k)=n+\sum_{i=r+1}^{n-1}\frac{(-1)^{i-r}}{(i+1)^{k-1}}\binom{i-1}{r}\binom{n}{i+1}.$$

证明:根据引理 1, $\Psi_r(n,k)$可以被如下推导:

$$\Psi_r(n,k)=n\cdot \mathrm{P}\{D_r(n-1,k)\}=n+\sum_{i=r+1}^{n-1}\frac{(-1)^{i-r}}{(i+1)^{k-1}}\binom{i-1}{r}\binom{n}{i+1}.$$ □

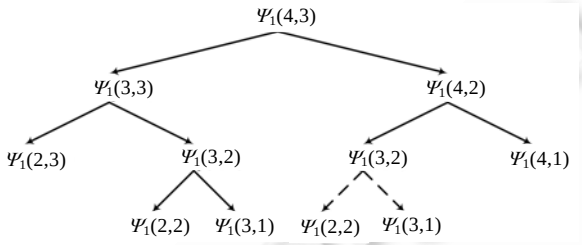


Fig.3 Evaluating $\Psi_1(4,3)$ using the recursive algorithm

图 3 利用递归算法求解 $\Psi_1(4,3)$

定理 4. $\Psi_r(n,k)$可以递归表示如下:

$$\Psi_r(n,k)=\Psi_r(n-1,k)+\frac{\Psi_r(n,k-1)}{n},$$

该递推方程的初始条件为 $\Psi_r(n,1)=r+1(n\geqslant r+1)$和 $\Psi_r(r+1,k)=r+1(k\geqslant 1)$.

证明:根据定理 3 以及组合公式$\binom{n}{i+1}=\binom{n-1}{i+1}+\binom{n-1}{i}$, $\Psi_r(n,k)$可以被如下推导:

$$\begin{aligned}\Psi_r(n,k)&=n+\sum_{i=r+1}^{n-1}\frac{(-1)^{i-r}}{(i+1)^{k-1}}\binom{i-1}{r}\binom{n}{i+1}\\&=n+\sum_{i=r+1}^{n-2}\frac{(-1)^{i-r}}{(i+1)^{k-1}}\binom{i-1}{r}\binom{n-1}{i+1}+\sum_{i=r+1}^{n-1}\frac{(-1)^{i-r}}{(i+1)^{k-1}}\binom{i-1}{r}\binom{n-1}{i}\\&=\Psi_r(n-1,k)+\frac{\Psi_r(n,k-1)}{n},\end{aligned}$$

该递推方程的初始条件为

$$\Psi_r(n,1)=n+\sum_{i=r+1}^{n-1}(-1)^{i-r}\binom{i-1}{r}\binom{n}{i+1}=r+1(n\geqslant r+1),$$

$$\Psi_r(r+1,k)=r+1+\sum_{i=r+1}^{r}\frac{(-1)^{i-r}}{(i+1)^{k-1}}\binom{i-1}{r}\binom{r+1}{i+1}=r+1(k\geqslant 1).$$ □

通过去除重复计算,算法 1 利用定理 4 给出的递推方程给出了一种求解 $\Psi_r(n,k)$的非递归算法.这种算法的时间代价和空间代价非常小,分别为 $O((n-r)k)$和 $O(k)$.因此,可以将其嵌入数据库系统的查询优化器,以用于对 Skyband 基数进行估计.图 4 给出了利用算法 1 求解 $\Psi_r(n,k)$的过程.算法的空间开销主要是两个长度为 k 的数组 α与 β,因此空间代价为 $O(k)$.首先将 $\alpha[1],\alpha[2],\ldots,\alpha[k]$的值初始化为 $\Psi_r(r+1,1),\Psi_r(r+1,2),\ldots,\Psi_r(r+1,k)$的值.根据递推方程的初始条件,

$$\Psi_r(r+1,1)=\Psi_r(r+1,2)=\ldots=\Psi_r(r+1,k)=r+1.$$

接下来,计算 $\Psi_r(r+2,1),\Psi_r(r+2,2),\ldots,\Psi_r(r+2,k)$的值,并将其保存在 $\beta[1],\beta[2],\ldots,\beta[k]$中.根据递推方程的初始条件, $\Psi_r(r+2,1)=r+1$,因此 $\beta[1]$的值被始化为 $r+1$.根据递推方程有

$$\Psi_r(r+2,2)=\Psi_r(r+1,2)+\Psi_r(r+2,1)/(r+2),$$

即$\beta[2]=\alpha[2]+\beta[1]/(r+2)$,因此可计算$\Psi_r(r+2,2)$的值并将其保存在$\beta[2]$中.同理,可依次计算$\Psi_r(r+2,3)$,$\Psi_r(r+2,4)$,…,$\Psi_r(r+2,k)$的值并将其保存在$\beta[3]$,$\beta[4]$,…,$\beta[k]$中.接下来,通过同样的方法利用保存在$\beta[1]$,$\beta[2]$,…,$\beta[k]$中的$\Psi_r(r+2,1)$,$\Psi_r(r+2,2)$,…,$\Psi_r(r+2,k)$的值来计算$\Psi_r(r+3,1)$,$\Psi_r(r+3,2)$,…,$\Psi_r(r+3,k)$的值,并将其保存在$\alpha[1]$,$\alpha[2]$,…,$\alpha[k]$中.这个过程持续下去,直到计算$\Psi_r(n,1)$,$\Psi_r(n,2)$,…,$\Psi_r(n,k)$的值将其保存在$\alpha[1]$,$\alpha[2]$,…,$\alpha[k]$中或$\beta[1]$,$\beta[2]$,…,$\beta[k]$中,并将$\alpha[k]$或$\beta[k]$保存的值作为$\Psi_r(n,k)$的值返回为止.显然,算法 1 的时间复杂度为 $O((n-r)k)$.从严格意义上来讲,算法 1 是一种动态规划算法[17],因为算法 1 是基于定理 4 给出的递推方程及其初始条件设计的一种非递归算法,并且算法的每一步都给出了对应子问题的精确解.

算法 1. 估计 Skyband 基数$\Psi_r(n,k)$.

输入: (1) n:空间包含的元素数;(2) k:空间的维数;(3) r:r-skyband.

输出: Skyband 基数的期望值.

```
01.  if n≤r+1 then return n;
02.  if k=1 then return r+1;
03.  for i=1 to k do α[i]←r+1;
04.  φ←0;
05.  for i=r+2 to n do
06.      φ←(φ+1) mod 2;
07.      if φ=1 then
08.          β[1]←r+1;
09.          for j=2 to k do β[j]←α[j]+β[j−1]/i;
10.      else //φ=0
11.          α[1]←r+1;
12.          for j=2 to k do α[j]←β[j]+α[j−1]/i;
13.      end
14.  end
15.  if φ=1 then return α[k] else return β[k];
```

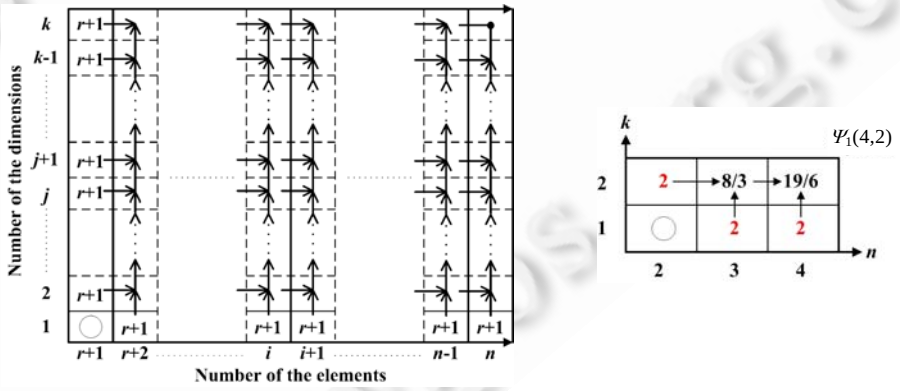


Fig.4 Evaluating $\Psi_r(n,k)$ and $\Psi_1(4,2)$ using the non-recursive algorithm

图 4 利用非递归算法求解$\Psi_r(n,k)$和$\Psi_1(4,2)$

图 4 还给出了利用算法 1 求解$\Psi_1(4,2)$的示例过程,由递推方程的初始条件可知$\Psi_1(2,2)=2$和$\Psi_1(3,1)=2$,然后可由递推方程$\Psi_1(3,2)=\Psi_1(2,2)+\Psi_1(3,1)/3$ 计算得到$\Psi_1(3,2)=8/3$.由递推方程的初始条件还可知$\Psi_1(4,1)=2$,接下来,即可由递推方程$\Psi_1(4,2)=\Psi_1(3,2)+\Psi_1(4,1)/4$ 计算得到$\Psi_1(4,2)=19/6$.

3 实验与分析

本节通过实验验证上一节中给出的对 Skyband 基数进行估计的方法的准确性.我们选择在一个较低维空间(三维空间)和一个较高维空间(六维空间)内进行验证.我们利用 GSL(GNU Scientific Library: http://www.gnu.org/software/gsl/)来生成各维数据.根据概率论,如果某一维上的数据是连续分布的,则在这一维上存在重复值的概率为 0.因此,我们把用 GSL 生成的正态连续分布的数据作为各维数据.在维数确定的情况下,测试不同大小的数据集对应的 Skyband 基数.数据集包含的元素数从 100 增加到 1 000,每步增加 100 个元素.对于每一步,我们用 GSL 生成 1 000 个数据集并测试这 1 000 个数据集对应的 Skyband 基数的最大值、最小值和平均值,并把测试结果与理论结果进行比较.图 5 和图 6 分别给出了在三维空间和在六维空间内进行验证的结果.可以看出,实验测得的 Skyband 基数的平均值与利用我们给出的对 Skyband 基数进行估计的方法计算出的理论值基本相同,并且实验测得的 Skyband 基数的最大值和最小值与理论值的偏差很小.从而验证了我们给出的对 Skyband 基数进行估计的方法是正确的,也是准确的.由于本文的工作首次给出了对 Skyband 基数进行估计的方法,因此本实验部分没有也无法与 Skyband 基数估计的其他相关工作进行比较.

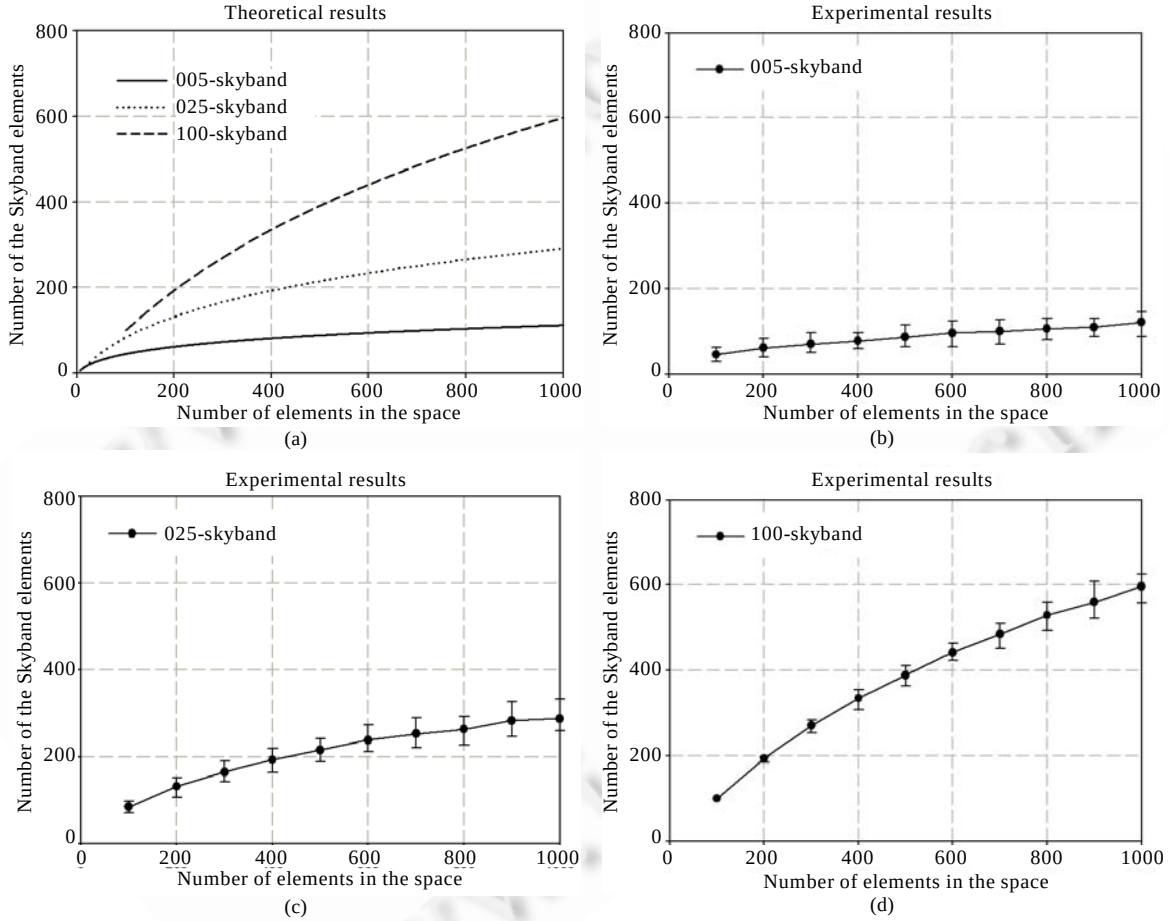


Fig.5 Skyband cardinality in the 3-dimensional space

图 5 三维空间内的 Skyband 基数

通过对三维空间和六维空间的实验结果进行比较可以看出,在空间内包含的数据元素数和 r 值确定的情况下,较高维空间内包含的 r-skyband 元素多于较低维空间内包含的 r-skyband 元素,原因在于,较高维空间内的一个元素被其他元素支配的概率较小.设 $\xi_1\xi_2...\xi_n$ 为 k 维空间内的 n 个元素,将这 n 个元素扩展到 $k+k'$ 维,$\xi_1'\xi_2'...\xi_n'$

为 $k+k'$维空间内对应的 n 个元素.对于 k 维空间内的一个元素ξ_i,如果存在另一元素ξ_j支配它,则在扩展后的 $k+k'$维空间内,ξ_j'可能支配ξ_i'也可能不支配ξ_i',具体情况视新增维上的值来确定.对于 k 维空间内的一个元素ξ_i,如果另一元素ξ_j 不能支配它,则无论新增维上为何值,在扩展后的 $k+k'$维空间内,ξ_j'肯定不能支配ξ_i'.根据上述分析可知,在空间内包含的数据元素数和 r 值确定的情况下,一个元素被其他元素支配的概率随着维数的增大而减小,因此 Skyband 元素数随着维数的增大而增大.从实验结果还可以看出,在空间内包含的数据元素数和维数确定的情况下,r 值较大的 Skyband 包含的 r-skyband 元素较多.设$\xi_1\xi_2...\xi_n$为 k 维空间内的 n 个元素,在求解 r-skyband 和[$r+r'$]-skyband(r'>0)的情况下,根据定义 2,如果元素 ξ_i 是一个 r-skyband 元素,则它肯定是一个[$r+r'$]-skyband 元素;如果ξ_j 是一个[$r+r'$]-skyband 元素,则它不一定是一个 r-skyband 元素.根据上述分析可知,在空间内包含的数据元素数和维数确定的情况下,r-skyband 元素数随着 r 值的增大而增大.

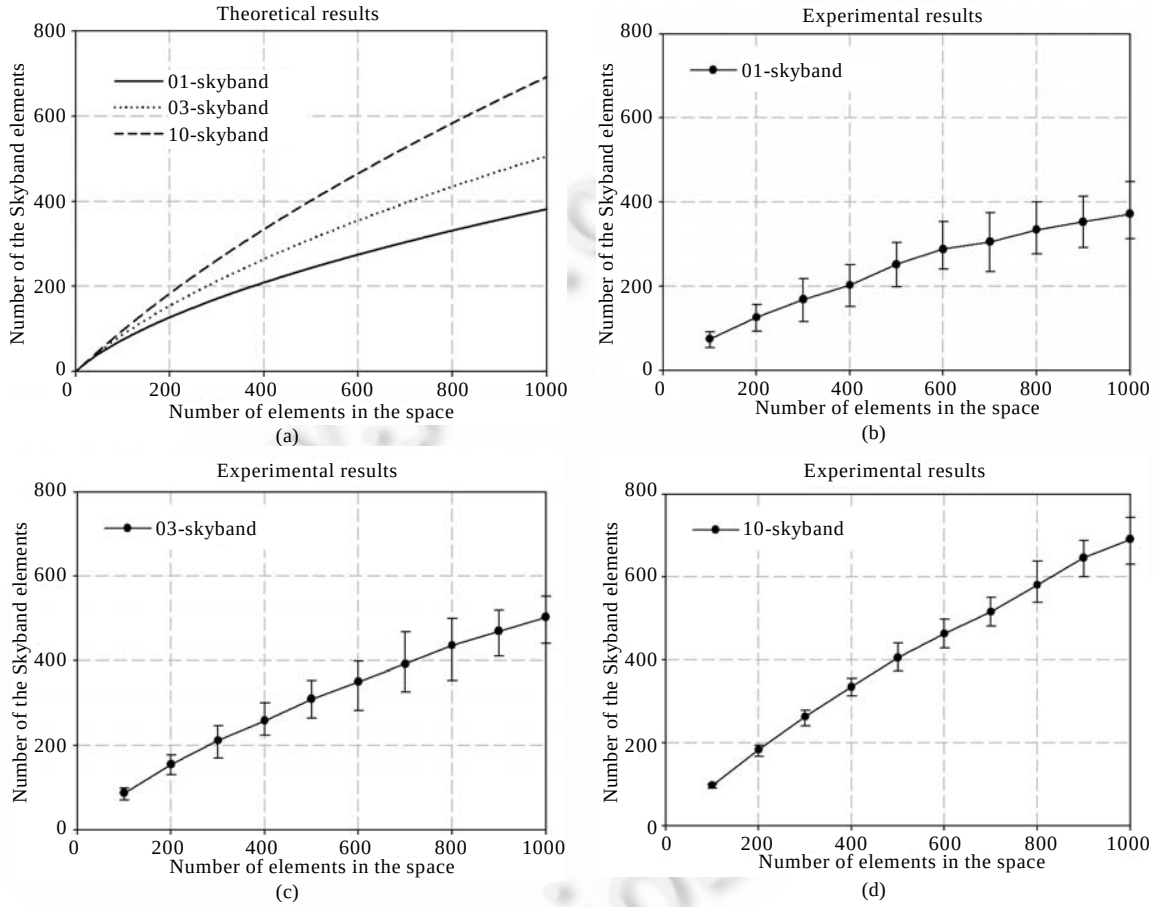


Fig.6 Skyband cardinality in the 6-dimensional space

图 6 六维空间内的 Skyband 基数

另外,从实验结果还可以看出,在维数和 r 值确定的情况下,当空间内包含的数据元素较多时,r-skyband 元素也较多.为了对这个性质进行解释,我们首先给出后缀 r-skyband 元素的概念.设$\xi_1\xi_2...\xi_n$为 k 维空间内长度为 n 的元素序列,对于其中的任一元素ξ_i,如果$\xi_{i+1}\xi_{i+2}...\xi_n$内能够支配$\xi_i$的元素不超过 r 个,则ξ_i是一个后缀 r-skyband 元素.设$\xi_1\xi_2...\xi_n$和$\xi_0\xi_1\xi_2...\xi_n$分别为 k 维空间内长度为 n 和 n+1 的元素序列,则ξ_i(i>0)在两个元素序列中成为后缀 r-skyband 元素的概率相同,因为后续元素都是相同的.ξ_0可能为后缀 r-skyband 元素,但这个概率随着后续元素数的增大而减小.因此,在维数和 r 值确定的情况下,后缀 r-skyband 元素数随着序列内包含的元素数的增大而增大,而且增幅越来越小.下面证明 k–1 维空间内包含的后缀 r-skyband 元素数等于 k 维空间内包含的 r-skyband

元素数.设$\xi_1'\xi_2'...\xi_n'$为 k–1 维空间内长度为 n 的元素序列,将其扩展到 k 维,$\xi_1\xi_2...\xi_n$ 为 k 维空间内对应的元素序列,ξ_i 在新增维上的值为 i.设新增维上的支配关系为大于关系,则 k–1 维空间内的后缀 r-skyband 元素与 k 维空间内的 r-skyband 元素一一对应,即 k–1 维空间内包含的后缀 r-skyband 元素数等于 k 维空间内包含的 r-skyband 元素数.因此,r-skyband 元素数与后缀 r-skyband 元素数随数据元素数增长的趋势是相同的.根据上述分析可知,在维数和 r 值确定的情况下,r-skyband 元素数随空间内包含的元素数的增大而增大,且增幅越来越小.

4 结束语

Skyband 查询是决策支持领域一类非常重要的查询,为了使数据库系统能够支持 Skyband 查询,必须解决 Skyband 基数估计的问题,因为这对于查询优化器对 Skyband 查询进行优化非常重要.本文在各维之间是无关的以及每一维上都不存在重复值的假设下提出了基于容斥原理的推广形式对 Skyband 基数进行理论分析的方法,并给出了时间和空间代价很小的对 Skyband 基数进行估计的动态规划算法.由于 Skyline 查询只是 Skyband 查询的一个特例,我们给出的算法可以用于对 Skyline 基数进行估计.另外,只要给出各维数据的分布函数,那么,我们给出的基于容斥原理的推广形式对 Skyband 基数进行理论分析的方法同样适用于维上存在重复值的情况.实验结果表明,我们给出的算法能够准确地对 Skyband 基数进行估计,因此可以将其嵌入数据库的查询优化器用于对 Skyband 基数进行估计.

致谢 在此,我们向对本文的工作给予支持和建议的北京大学信息学院屈婉玲教授和张立昂教授以及澳大利亚新南威尔士大学计算机系林学民教授表示感谢.

References:

[1] Börzsönyi S, Kossmann D, Stocker K. The Skyline operator. In: Proc. of the 17th Int'l Conf. on Data Engineering. Heidelberg: IEEE Computer Society, 2001. 421−430. http://www.informatik.uni-trier.de/~ley/db/conf/icde/icde2001.html

[2] Papadias D, Tao YF, Fu G, Seeger B. Progressive Skyline computation in database systems. ACM Trans. on Database Systems, 2005,30(1):41−82. [doi: 10.1145/1061318.1061320]

[3] Hjaltason GR, Samet H. Distance browsing in spatial databases. ACM Trans. on Database Systems, 1999,24(2):265−318. [doi: 10.1145/320248.320255]

[4] Buchta C. On the average number of maxima in a set of vectors. Information Processing Letters, 1989,33(2):63−65. [doi: 10.1016/0020-0190(89)90156-7]

[5] Markl V, Megiddo N, Kutsch M, Tran TM, Haas PJ, Srivastava U. Consistently estimating the selectivity of conjuncts of predicates. In: Proc. of the 31st Int'l Conf. on Very Large Data Bases. Trondheim: ACM, 2005. 373−384.

[6] Rosen KH. Discrete Mathematics and Its Applications. 5th ed., Boston: McGraw-Hill, 2002.

[7] Chomicki J, Godfrey P, Gryz J, Liang D. Skyline with presorting. In: Dayal U, Ramamritham K, Vijayaraman TM, eds. Proc. of the 19th Int'l Conf. on Data Engineering. Bangalore: IEEE Computer Society, 2003. 717−816.

[8] Godfrey P, Shipley R, Gryz J. Maximal vector computation in large data sets. In: Böhm K, Jensen CS, Haas LM, Kersten ML, Larson P, Ooi BC, eds. Proc. of the 31st Int'l Conf. on Very Large Data Bases. Trondheim: ACM, 2005. 229−240.

[9] Godfrey P, Shipley R, Gryz J. Algorithms and analyses for maximal vector computation. The VLDB Journal, 2007,16(1):5−28.

[10] Tan KL, Eng PK, Ooi BC. Efficient progressive Skyline computation. In: Proc. of the 27th Int'l Conf. on Very Large Data Bases. Roma: Morgan Kaufmann Publishers, 2001. 301−310.

[11] Eng PK, Ooi BC, Tan KL. Indexing for progressive skyline computation. Data & Knowledge Engineering, 2003,46(2):169−201. [doi: 10.1016/S0169-023X(02)00208-2]

[12] Kossmann D, Ramsak F, Rost S. Shooting stars in the sky: An online algorithm for skyline queries. In: Bressan S, Chaudhri AB, Lee ML, Yu JX, Lacroix Z, eds. Proc. of the 28th Int'l Conf. on Very Large Data Bases. Hong Kong: Morgan Kaufmann Publishers, 2002. 275−286.

[13] Papadias D, Tao YF, Fu G, Seeger B. An optimal and progressive algorithm for Skyline queries. In: Halevy Y, Ives ZG, Doan AH, eds. Proc. of the 2003 ACM SIGMOD Int'l Conf. on Management of Data. San Diego: ACM, 2003. 467−478.

[14] Wei XJ, Yang J, Li CP, Chen H. Skyline query processing. Journal of Software, 2008,19(6):1386−1400 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/1386.htm [doi: 10.3724/SP.J.1001.2008.01386]

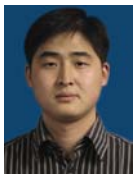
[15] Bentley JL, Kung HT, Schkolnick M, Thompson CD. On the average number of maxima in a set of vectors and applications. Journal of the ACM, 1978,25(4):536−543. [doi: 10.1145/322092.322095]

[16] Godfrey P. Skyline cardinality for relational processing. In: Seipel D, Torres JMT, eds. Proc. of the 3rd Int'l Symp. on Foundations of Information and Knowledge Systems. Wilhelminenburg Castle: Springer-Verlag, 2004. 78−97.

[17] Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms. 2nd ed., Cambridge: MIT Press, 2001.

附中文参考文献:

[14] 魏小娟,杨婧,李翠平,陈红.Skyline 查询处理.软件学报,2008,19(6):1386−1400. http://www.jos.org.cn/1000-9825/19/1386.htm [doi: 10.3724/SP.J.1001.2008.01386]



赵加奎(1979－),男,辽宁铁岭人,博士,主要研究领域为数据库,决策支持,Web 信息处理,数据挖掘.

陈立军(1968－),男,博士,副教授,主要研究领域为数据库,数据流,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,数据仓库,Web 数据集成,移动数据挖掘.