文章编号: 1001 - 9081(2008)11 - 2729 - 03

基于不确定数据的查询处理综述

崔 斌,卢 阳
(北京大学 信息科学技术学院,北京 100871)
(bin. cui@pku. edu. cn)

摘 要:不确定数据在一些重要应用领域中是固有存在的,如传感器网络和移动物体追踪。在不确定数据上使用传统的查询方法会使查询结果出现偏差,不能满足用户的需求。因此,基于不确定数据的查询处理受到了越来越多的关注。与在确定数据上查询不同,不确定数据上的研究工作将概率引入到数据模型中来衡量不确定对象成为结果集中元素的可能性。由于问题定义和数据模型的不同,不确定数据上的查询类型也多种多样。从问题定义、数据模型、剪枝策略和算法等角度,对基于不确定数据的范围查询、top-k查询以及 skyline查询进行了介绍。

关键词:不确定数据;范围查询; top-k查询; skyline查询
中图分类号: TP18; TP311. 12 **文献标志码**: A

Survey on query processing based on uncertain data

CUI Bin, LU Yang
(*School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871, *China*)

Abstract: Uncertain data is inherent in some important application fields, such as sensor networks and mobile object tracking. Using traditional querying methods on uncertain data will bias the answer set, and hence cannot satisfy users' needs. Therefore, query processing based on uncertain data has attracted more and more attention. Different from queries on certain data, research work on uncertain data introduce probability into data model to measure the likeness of an uncertain object as one element of the answer set. Due to different problem definitions and data models, query types differentiate from each other greatly. This survey introduced range queries, top-k queries and skyline queries based on uncertain data from the views of problem definitions, data models, pruning strategies and algorithms.

Key words: uncertain data; range queries; top-k queries; skyline queries

不确定数据广泛存在于许多应用领域,如传感器网络[1]和移动物体查询[2-5]。由于不确定数据本身的特性,传统的查询处理方法已经不能高效地解决或无法解决不确定数据上的查询。因此,近些年来不确定数据上的查询处理成为研究的热点,并在 SIGMOD、VLDB和 ICDE等国际顶级会议上涌现出一批相关的研究。本文从问题定义、数据模型、剪枝策略和算法等角度对不确定数据上的范围查询[6-7]、top-k查询[8-9]以及 skyline查询[10]进行了介绍。

1 范围查询

1.1 一维空间上的范围查询

文献[6]最先提出了概率范围查询,但是范围只限于一维空间。他们使用两种辅助的索引结构来高效地访问不确定区间。

数据库中包含不确定对象 $T_i(i = 1, \cdots, n)$,每个对象包含若干个属性,其中某个属性 c上的值是不确定的。$T_i.c$的取值分布在不确定区间 $[L_i, R_i]$内,概率密度函数(Probability Density Function, PDF)为 $f_i(x)$。

文献[6]主要解决概率范围查询中的一种概率极限查询(Probabilistic Threshold Queries, PTQ):对于一个给定的区间 $[a, b]$, PTQ的查询结果是 T_i的集合, T_i满足其属性 c的取值在区间 $[a, b]$内的概率 p_i大于等于 p_q, p_q是人为设定的一个极限值。p_i的计算方法如式(1)所示,其中 R表示 $[L_i, R_i]$与 $[a, b]$交叠的区域。

$$p_i = \int_R f(x)\,\mathrm{d}x \tag{1}$$

解决 PTQ最直接的办法就是查看数据库中所有的不确定对象;如果对象与不确定区间与查询的区间有交叠,则按照式(1)计算该对象的概率;如果概率大于给定的极限值,那么这个对象应放到结果集中。但这个方法的 I/O代价很高,因为要遍历所有的不确定对象,而且计算式(1)的代价也很高。

为了提高查询的效率,文献[6]中提出了概率极限索引(Probability Threshold Indexing, PTI)。PTI修改了一维的 R-tree[11]结构,将概率信息加到中间节点以加速剪枝。为 R-tree的第 j个节点的最小包含矩形(Minimal Bounding Rectangle, MBR) M_j定义左 x边界 $M_j.lb(x)$和右 x边界 $M_j.lb(x)$,使得该节点 MBR中的每个区间 $[L_i, R_i]$至多有概率 x位于左 x边界的左边和右 x边界的右边。如果下面查询区间 $[a, b]$与 M_j不相交并且极限值 p_q大于等于 x, M_j就可以被剪枝。

但是 PTI不能避免 R-tree自身的问题,一个范围查询可能会遍历有很多大区间的 MBR。为此,文献[6]中给出了另一种方法。每个区间 $[x, y]$都可以映射到二维平面中的一个点 (x, y),在 pdf是均匀分布的假设下, PTQ可以转变成三边梯形查询,使用 2D R-tree[12]加速查询。对于任意的分布,将

收稿日期: 2008 - 07 - 08;**修回日期**: 2008 - 08 - 07。
基金项目:国家自然科学基金资助项目(60603045);国家 863计划项目(2007AA01Z153)。
作者简介:崔斌(1975 -),男,浙江宁波人,研究员,主要研究方向:数据库; 卢阳(1982 -),男,吉林临江人,硕士研究生,主要研究方向:数据库。

具有相似均值和标准差的区间放到一个 MBR里,在查询的过程中可以利用均值和标准差加速剪枝。

1.2 多维空间上的范围查询

文献[7]研究了多维空间中有任意概率密度函数物体的范围查询问题。他们提出了几种剪枝策略和一种新的存取方法 U-tree来优化 I/O代价和 CPU时间。

不同于文献[6]的问题模型,文献[7]的不确定对象 T_i 包含两部分:1) 概率密度函数 $T_i.pdf(x)$,这里的 x是 d维空间中的一个点;2) d维的一个不确定区域 $T_i.ur$。概率范围查询给定一个超矩形 r_q和极限值 p_q。不确定对象 T_i的概率计算方法如式(2)。如果 p_i大于等于 p_q,将 T_i添加到结果集中。

$$p_i = \int_{T_i.ur \cap r_q} T_i.pdf(x)\,\mathrm{d}x \tag{2}$$

一个不确定对象的概率受限区域(Probabilistically Constrained Regions, PCR)的大小由参数 p确定,p的取值范围为[0, 0.5]。$T_i.pcr(p)$由四条直线 l_{1+}, l_{1-}, l_{2+}和 l_{2-}围成。l_{1+}将 $T_i.ur$分成两部分,T_i出现在 l_{1+}右侧的概率是 p;同理,T_i出现在 l_{1-}左侧的概率等于 p。类似于 l_{1+}和 l_{1-}, l_{2+}和 l_{2-}从水平方向划分 $T_i.ur$。这样,可以用下面的策略来判断不确定对象是否满足查询,减少式(2)的计算次数,只有当下述策略都不能确定 T_i是否属于结果集时,才使用式(2)。

剪枝策略:1) 当 $p_q > 0.5$时,如果 r_q不完全包含 $T_i.pcr(1-p_q)$,则 T_i不在结果集中。2) 当 $p_q \le 0.5$时,剪枝条件是 r_q不与 $T_i.pcr(p_q)$相交。

验证策略:在 d维空间中,$T_i.pcr(p)$是由二维向量 $\{T_i.pcr_{1-}(p), T_i.pcr_{1+}(p), \cdots, T_i.pcr_{d-}(p), T_i.pcr_{d+}(p)\}$确定的超矩形。1) 对于任意 p_q,如果存在 $j \in [1, d]$使得 r_q完全覆盖了 T_i的最小限定矩形在 $T_i.pcr_{j-}((1-p_q)/2)$和 $T_i.pcr_{j+}((1-p_q)/2)$之间的部分,那么 T_i满足查询。2) 当 $p_q > 0.5$时,如果存在 $j \in [1, d]$使得 r_q完全覆盖了 T_i的最小限定矩形在 $T_i.pcr_{j-}(1-p_q)$右侧的部分或 $T_i.pcr_{j+}(1-p_q)$左侧的部分,那么 T_i满足查询。3) 当 $p_q \le 0.5$时,如果存在 $j \in [1, d]$使得 r_q完全覆盖了 T_i的最小限定矩形在 $T_i.pcr_{j-}(p_q)$左侧的部分或 $T_i.pcr_{j+}(p_q)$右侧的部分,那么 T_i满足查询。

文献[7]中针对空间开销问题,提出了保守功能盒(Conservative Functional Boxes, CFB)的概念来限定 PCR,并给出了计算方法,对上述策略也进行了相应修改。文献[7]进而提出了 U-tree索引结构来加快剪枝和验证过程。U-tree的结构类似于 R^*-tree[13],在叶节点存储不确定对象,U-tree的中间节点用 CFB代替 MBR。

2 Top-k查询

文献[8]基于可能世界语义[14-17]提出了解决 top-k查询的不确定数据模型。在该模型中,每个元组属于数据库的概率被称为置信度,产生规则是任意的逻辑公式用来确定有效的世界,每个可能世界是元组的联合。通过假设世界中存在的元组可以根据元组的置信度以及产生规则计算世界的概率。

文献[8]中将 top-k查询分为两种,即 U-Topk和 U-kRanks:

假设 D是不确定数据库,可能世界空间 $PW = \{PW^1, \cdots, PW^n\}$。$T = \{T^1, \cdots, T^n\}$是长度为 k的元组向量的集合。对于每一个 T^i: T^i的元组根据得分函数 F排序; T^i是非空世界集合 $PW(T^i) \subseteq PW$的 top-k结果。基于 F的 U-Topk查询返回 $T^* \in T$, T^*是在所有可能世界上成为 top-k最大聚集概率的元组向量,计算式如下:

$$T^* = \arg\max_{T^i \in T}\left(\sum_{w \in PW(T^i)} (Pr(w))\right) \tag{3}$$

假设 D是不确定数据库,可能世界空间 $PW = \{PW^1, \cdots, PW^n\}$。对于 $i = 1, \cdots, k$, $\{x_i^1, \cdots, x_i^m\}$是元组的集合。每个元组 x_i^j根据得分函数 F在非空集合 $PW(x_i^j) \subseteq PW$中出现在排名为 i的位置。基于 F的 U-kRnaks返回 $\{x_i^*; i = 1, \cdots, k\}$, x_i^*的计算式如下:

$$x_i^* = \arg\max_{x_i^j}\left(\sum_{w \in PW(x_i^j)} (Pr(w))\right) \tag{4}$$

为了说明算法,引入了搜索状态 s_l表示一个长度为 l的元组向量,根据得分函数在一个或多个可能世界中成为 top-l结果。假设 d是目前从数据库中访问的元组数, s_l的概率 $P(s_l) = Pr(s_l \wedge \neg I(s_l, d))$,其中 $I(s_l, d)$表示在已访问的元组中而不在 s_l中的元组。将 s_l转变为 $s_{l,i}$,下标 i表示该向量最后的可见元组所在的位置。对于 U-Topk查询,使用 OPTU-Topk算法,该算法的基本思想:1) 设置一个以搜索状态概率优先级排序的队列 Q,初始化时插入 $s_{0,0}$,概率 $P(s_{0,0}) = 1$。2) 当 Q不为空时不断执行下面操作:从 Q中弹出 $s_{l,i}$;如果 $l = k$时返回 $s_{l,i}$,否则根据 i和 d的比较情况选择下一个要访问的元组 t,分别对 t可以加入和不加入 $s_{l,i}$两种情况,将 $s_{l+1,i+1}$和 $s_{l,i+1}$插回 Q。

对于 U-kRanks查询使用 OPTU-kRanks算法,该算法的基本思想是:在计算排名 i时,对于一个新来的元组 t,计算其在所有可能世界在排名 i上出现的概率 $P_{t,i}$;如果 $P_{t,i}$比目前答案的概率大,并且比将未见元组考虑进来时的概率也大,那么 t是结果集中排名为 i的元组。

文献[9]提出了另一种 top-k查询——概率极限值 top-k查询(PT-k查询)。采用的不确定数据模型也是基于可能世界语义模型。不确定表 T包含许多元组,每个元组 t属于 T的概率是 $Pr(t)$。产生规则 R规定了不能同时存在的元组, R的概率 $Pr(R)$为 R中所有元组的概率和。T上所有的产生规则构成了产生规则集合 R_T。一个可能世界 W是 T的子集。W的存在概率计算公式如式(5)所示。PW是所有可能世界的集合。

$$Pr(W) = \prod_{R \in R_T, |R \cap W| = 1} Pr(R \cap W) \prod_{R \in R_T, |R \cap W| = \varnothing} (1 - Pr(R)) \tag{5}$$

一个传统的 top-k查询 Q可以直接应用到可能世界 W上,用 $Q^k(W)$表示结果集中的 k个元组。一个元组 t的 top-k概率计算公式如式(6)所示,在没有歧义的情况下我们简写为 $Pr^k(t)$。则 PT-k查询结果集中所包含元组的 top-k概率至少是 p, p是给定的极限值。

$$P_{Q,T}^k(t) = \sum_{W \in PW,\ t \in Q^k(W)} Pr(W) \tag{6}$$

处理查询时,可以将处理空间从表 T降低为满足 top-k查询谓词的元组集合 $P(T)$,相应的从产生规则中去掉不在 $P(T)$中的元组。对于 $P(T)$中的一个元组 t,它是否属于 $Q^k(W)$取决于 $P(T)$中统治它的元组,所有统治它的元组构成了统治集 S_t。假设各元组相互独立,即不考虑产生规则, $P(T)$中的元组 $t_1, \cdots, t_n$按照排名递降的顺序排列, $Pr(S_t, j)$表示 S_t中 j个元组出现在可能世界的概率。t_i的 top-k概率计算式(7)可以根据文献[18]中的泊松二项递归公式算出。

$$Pr^k(t_i) = Pr(t_i) \sum_{j=1}^{k} Pr(S_{t_{i-1}}, j-1) \tag{7}$$

考虑产生规则时,如果能够消除统治集中元组之间的依

赖关系,就能用式(7)计算元组的 top-k概率。作者对元组 t和产生规则 R的元组之间的关系进行了讨论,并给出两个规则:1)如果 R中的任何一个元组的排名都比 t高,将 R中的所有元组压缩成一个元组放到表 T中。2)R的元组数大于 1,如果 R中的元组 t满足谓词,计算 t的 top-k概率时所用的表为统治集 S_t除去 R中的元组后与 $\{t\}$的并集。为了加速查询,文献[9]给出了四条剪枝规则以及采样算法和泊松近似算法。

3　Skyline查询

Skyline查询在数据挖掘领域以及多规则决策应用领域发挥着重要作用。自从文献[19]将 skyline算子引入到关系数据库系统,skyline查询就受到了数据库研究者广泛的关注。近年来,随着传感器和移动物体数据库的发展,不确定数据日益增多,基于不确定数据的 skyline查询成为研究热点。

文献[10]在离散分布的不确定对象上提出了 p-skyline查询,并给出了自底向上和自上向下两种算法。离散分布的不确定对象是指每个对象存在多个实例。为了简化模型易于分析,作者假设不确定对象之间是相互独立的并且每个实例出现的概率相同。

不确定对象 $U=\{u_1,\cdots,u_m\}$有 m个实例。实例 u成为 skyline的概率以及不确定对象 U成为 skyline的概率分别由式(8)和(9)计算得到。其中符号"$\rightarrow$"表示统治关系,V表示其他任意的不确定对象。p-skyline查询是对于给定的极限值 p,查找成为 skyline的概率至少为 p的不确定对象。

$$Pr(u)=\prod_{V\neq U}\left(1-\frac{|\{v\in V\mid v\rightarrow u\}|}{|V|}\right) \tag{8}$$

$$Pr(U)=\frac{1}{m}\sum_{u\in U}Pr(u) \tag{9}$$

自底向上的方法引入了两个特殊实例 $U_{\min}$和 $U_{\max}$,它们分别表示取 U的所有实例在各维上的最小值和最大值所组成的实例。我们可以限定 U的概率为 $Pr(U_{\min})\geq Pr(U)\geq Pr(U_{\max})$。使用下面四种剪枝策略可以加快查询:1)如果 $Pr(U_{\min})<p$,那么 U不是 p-skyline元素;如果 $Pr(U_{\max})\geq p$,那么 U在 p-skyline中;2)假设对于 U中任意实例 u,$Pr^+(u)$和 $Pr^-(u)$分别是 u的概率上限和概率下限,如果 $Pr^+(u)$的平均值小于 p,那么 U不在 p-skyline中;如果 $Pr^-(u)$的平均值大于等于 p,则 U是 p-skyline的元素。3)假设 U和 V是两个不同的对象,如果 U的一个实例 u满足 $V_{\max}\rightarrow u$,那么 $Pr(u)=0$。4)假设 U和 V是两个不同的对象,U'是 U的一个子集并且 $U_{\max}\rightarrow V_{\min}$,如果 $(|U-U'|/|U|)\times\min_{u\in U'}\{Pr(u)\}$小于 p,则 V不在 p-skyline中。为了减少每个不确定对象实例概率的计算,可以将对象的实例分层,概率按层次递减。基于上述策略,自底向上的方法使用堆和 R-tree等结构组织算法。

自顶向下的方法为每个不确定对象建一棵划分树。划分树是二叉树,建造过程类似于 kd-tree[20]。页节点包含实例及最小限界矩形,中间节点维护后代的最小限界矩形。对于划分树的一个节点 N,其最小限定矩形的左下角和右上角的顶点分别用 $N_{\min}$和 $N_{\max}$表示。N中任意一个实例 u的概率可以限定为 $Pr(N_{\min})\geq Pr(u)\geq Pr(N_{\max})$。进一步如果不确定对象 U的划分树有 n个叶节点 $N_1,\cdots,N_n$,那么 U的概率可用不等式(10)限定。剪枝策略:1)假设 N是不确定对象 U的划分树的一个节点,如果存在一个对象 V使得 $V_{\max}\rightarrow N_{\min}$,那么 N可以被剪枝;2)假设 N是不确定对象 U的划分树的一个节点,如果 $Pr(N_{\min})=Pr(N_{\max})$,那么 N可以被剪枝;3)假定 p是给定的极限值,如果不确定对象 U的划分树有 n个叶节点 $N_1,\cdots,N_n$,如果式(10)中最左侧表达式的值大于等于 p,那么 U是 p-skyline元素,如果最右侧表达式的值小于 p,那么 U不在 p-skyline中。

$$\frac{1}{|U|}\sum_{i=1}^{n}|N_i|\,Pr(N_{i,\max})\leq Pr(U)\leq\frac{1}{|U|}\sum_{i=1}^{n}|N_i|\,Pr(N_{i,\min}) \tag{10}$$

4　结语

由于真实世界应用中数据固有的不确定性,以及相关应用系统需求的增加,针对不确定数据管理的研究已经成为数据库技术研究中一个新的热点。本文综述了基于不确定数据的查询处理研究。与传统的查询处理相比,不确定数据上的数据模型引入了概率,结果集多以概率为衡量指标。为了减少查询代价,多种优化技术已经应用到不确定数据查询当中,比如剪枝策略和索引。目前,不确定数据的管理、模型以及查询处理的研究受到了广泛关注,但是还有很多研究问题亟待解决,比如不确定数据建模、不确定数据整合、不确定数据管理的评价指标和不确定数据挖掘等。

参考文献:

[1] SILBERSTEIN A, BRAYNARD R, ELLIS C, *et al*. A sampling-based approach to optimizing top-k queries in sensor networks[C]// Proceedings of the 22nd International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2006: 68.

[2] CHENG R, PRABHAKAR S, KALASHNIKOV D V. Querying imprecise data in moving object environments[C]// Proceedings of the 19th International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2003: 723 - 725.

[3] CHENG R, KALASHNIKOV D, PRABHAKAR S. Querying imprecise data in moving object environments[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(9): 1112 - 1127.

[4] CHEN L, ÖZSU M T, ORIA V. Robust and fast similarity search for moving object trajectories[C]// Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2005: 491 - 502.

[5] LJOSA V, SINGH A K APLA: Indexing arbitrary probability distributions[C]// Proceedings of the 23rd International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2007: 946 - 955.

[6] CHENG R, XIA Y, PRABHAKAR S, *et al*. Efficient indexing methods for probabilistic threshold queries over uncertain data[C]// Proceedings of the 30th International Conference on Very Large Data Bases. San Francisco: Morgan Kaufmann, 2004: 876 - 887.

[7] TAO YU-FEI, CHENG R, XIAO XIAO-KUI, *et al*. Indexing multi-dimensional uncertain data with arbitrary probability density functions[C]// Proceedings of the 31st International Conference on Very Large Data Bases. New York: ACM Press, 2005: 922 - 933.

[8] SOLMAN M A, ILYAS I F, CHANG K C C. Top-k query processing in uncertain databases[C]// Proceedings of the 23rd International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2007: 896 - 905.

[9] HUA M, PEI J, ZHANG W, *et al*. Ranking queries on uncertain data: A probabilistic threshold approach[C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2008: 673 - 686.

[10] PEI J, JIANG B, LIN X, *et al*. Probabilistic skylines on uncertain data[C]// Proceedings of the 33rd International Conference on Very Large Data Bases. New York: ACM Press, 2007: 15 - 26.

(下转第 2744页)

已有的隐私数据模型基本都是通过使用多个视图来表现同一隐私数据的不同侧面,一些具有隐私保护功能的数据库原型系统的访问控制模块直接通过限定用户/角色能访问的视图范围来实现对隐私数据的保护。基于视图的访问控制机制已经非常成熟,这里不再详述。

目的的概念最早出现在文献[8]中提出的 Hippocratic隐私数据库原型系统中,用来表示被收集数据的用途,只有查询的访问目的与被访问元组或属性的预期目的一致时,系统才会执行查询。在基于目的的访问控制机制中,所有目的被组织为一个层次结构,对隐私数据的访问目的必须被限定在数据提供者所定义的预期目的之中。文献[7]中提出了一种新的基于目的的访问控制策略。该策略与以往基于目的的访问控制策略的区别在于它能处理复杂的、具有层次结构的数据模型,如 XML和对象—关系数据模型。其次,该策略通过扩展 RBAC来判断用户是否能访问特定目的的数据。

3.2 基于隐私保护的分发控制

基于隐私保护的分发控制技术通过提高查询结果集中隐私数据的不确定性和不可区分性,能在一定程度上防止数据窥探者利用数据挖掘技术来推理隐私信息。

不可区分性是指数据窥探者不能将现实中的个体与数据库中存储的特定记录相对应。目前实现不可区分性的方法主要包括 k-anonymous模型和 ℓ-diversity模型。k-anonymous通过在准标识属性上使用数据泛化技术,使至少 k个数据元组在准标识属性上具有相同的属性值,隐私窥探者因此很难将某个元组与这个元组所代表真实个体相联系。ℓ-diversity模型建立在 k-anonymous的基础之上,其基本思想是确保在每个经过泛化的数据分块中,敏感属性上有至少 $\ell(\ell \geq 2)$ 个不同的值,并且这 ℓ个值有大致相同的出现概率,从而保证敏感属性取值的多样性,防止同质化攻击的产生。然而,这两种方法都无法确保隐私窥探者一定不能获取个体的敏感信息。

不确定性是指数据窥探者不能确切地知道一个特定个体的隐私数据值。目前实现不确定性的方法主要包括在数据项上增加数据噪声,用数据泛化和数据压缩方法对敏感数据进行变换等,这些方法在提高不确定性的同时也会造成隐私数据质量一定程度的下降。

综上所述,数据库隐私保护机制的研究已经取得了一些进展,但在支持复合隐私数据保护的数据模型、精确表达隐私规格的策略定义语言、平衡数据隐私性和可用性的数据泛化技术等方面,还需要进行更加深入的研究和探讨。

4 结语

数据库是当今信息社会中数据存储和处理的核心,它的可信性必须得到保证。在相关领域研究人员的不懈努力下,一系列针对数据库的安全机制有了很大的发展。但到目前为止,在数据库作为服务提供的开放式环境下,尚没有将访问控制、密文查询和隐私保护机制有效融合在一起的可信数据库系统出现,在多级密钥的管理与分发策略、高查准率的密文查询方法、基于推理控制的隐私数据保护、具备分布式特征的可信前端模型等方面,我们还需要付出进一步的努力。

参考文献:

[1] BELL D E, LAPADULA L J. Secure computer systems: Mathematical foundations, ESD-TR-73-278[R]. Bedford: The Mitre Corporation, 1973.

[2] SANDHU R, COYNE E J, FEINSTEIN H L, *et al*. Role-based access control models[J]. IEEE Computer, 1996, 29(2): 38 - 47.

[3] PARK J, SANDHU R. The $UCON_{ABC}$ usage control model[J]. ACM Transaction on Information and System Security, 2004, 7(1): 128 - 174.

[4] HACIGUMUS H, IYER B, MEHROTRA S. Providing database as a service[C]// Proceedings of the 18th International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2002: 29 - 38.

[5] FERRER J D. A new privacy homomorphism and applications[J]. Information Processing Letters, 1996, 60(5): 277 - 282.

[6] HACIGUMUS H, IYER B, LI C, *et al*. Executing SQL over encrypted data in the database-service-provider model[C]// Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2002: 216 - 227.

[7] BYUN J W, BERTINO E, LIN. Purpose based access control of complex data for privacy protection[C]// Proceedings of the 10th ACM Symposium on Access Control Models and Technologies. New York: ACM Press, 2005: 102 - 110.

[8] AGRAWAL R, KIERNAN J, SRIKANT R, *et al*. Hippocratic databases[C]// Proceedings of the 28th International Conference on Very Large DataBases: VLDB 2002. San Francisco: Morgan Kaufmann, 2002: 143 - 154.

(上接第 2731页)

[11] GUTTMAN A. R-trees: A dynamic index structure for spatial structure for spatial searching[C]// Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 1984: 47 - 57.

[12] GOLDSTEIN J, RAMAKRISHNAN R, SHAFT U, *et al*. Processing queries by linear constraints[C]// Proceedings of the 16th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. New York: ACM Press, 1997: 257 - 267.

[13] BECKMANN N, KRIEGEL H P, SCHNEIDER R, *et al*. The R*-tree: An efficient and robust access method for points and rectangles[C]// Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 1990: 322 - 331.

[14] MIELINSKI T, LIPSKI J W. Incomplete information in relational databases[J]. Journal of the ACM, 1984, 31(4): 761 - 791.

[15] ABITEBOUL S, KANELLAKIS P, GRAHNE G. On the representation and querying of sets of possible worlds[C]// Proceedings of the 1987 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 1987: 34 - 48.

[16] SARMA A D, BENJELLOUN O, HALEVY A, *et al*. Working models for uncertain data[C]// Proceedings of the 22nd International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2006: 7.

[17] DALVI N, SUCIU D. Management of probabilistic data: foundations and challenges[C]// Proceedings of the 26th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. New York: ACM Press, 2007: 1 - 12.

[18] LANGE K. Numerical analysis for statisticians[M]. Berlin: Springer-Verlag, 1999.

[19] BÖRZSÖNYI S, KOSSMANN D, STOCKER K. The skyline operator[C]// Proceedings of the 17th International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2001: 421 - 230.

[20] BENTLEY J L. Multidimensional binary search trees used for associative searching[J]. Communications of the ACM, 1975, 18(9): 509 - 517.