不确定数据查询处理

蒋 涛[1],高云君[2],张 彬[1],周傲英[3],乐光学[1]
(1.嘉兴学院数理与信息工程学院,浙江嘉兴 314001;
2.浙江大学计算机学院,浙江杭州 310027;3.华东师范大学软件学院,上海 200062)

摘 要: 数据的不确定性在现实世界中的经济、军事、物流、金融、电信等领域普遍存在.不确定数据广泛应用于环境维护、市场分析、基于位置的服务LBS以及数量经济研究等应用.由于这些应用的重要性以及收集和累积的不确定数据数量的快速增长,查询这些数据已经成为一个重要的任务,并日益受到广大数据库研究者的关注.本文介绍了不确定数据查询的基本原理,并对不确定数据的近邻查询、逆向近邻查询、排序查询、Top-k查询以及连接查询进行了详细的讨论.同时对这些技术的优缺点进行了分析、对比.最后给出了未来的研究方向.

关键词: 不确定数据; 近邻; 逆向近邻; 连接; 查询处理
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2013)05-0966-11
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.05.021

Query Processing on Uncertain Data

JIANG Tao[1], GAO Yun-jun[2], ZHANG Bin[1], ZHOU Ao-ying[3], YUE Guang-xue[1]
(1. *College of Mathematics and Information Engineering, Jiaxing University. Jiaxing, Zhejiang* 314001, *China*;
2. *College of Computer Science and Technology, Zhejiang University. Hangzhou, Zhejiang* 310027, *China*;
3. *Shanghai Key Laboratory of Trustworthy Computing, Software Engineering Institute, East China Normal University. Shanghai* 200062, *China*)

Abstract: Data uncertainty is pervasive in various fields, for example, economy, military, logistic, finance and telecommunication, etc. Uncertain data are inherent in some important applications, such as environmental surveillance, market analysis, Location-Based Service (LBS), and quantitative economics research. Due to the importance of those applications and the rapidly increasing amount of uncertain data collected and accumulated, querying large collections of uncertain data has become an important task and has received more and more attention from the database community in recent years. This paper introduces the principle of uncertain data query, and surveys the advance of the research on uncertain data query processing, including Nearest Neighbor (NN) query, Reverse Nearest Neighbor (RNN) query, Ranking query, top-k query and join query. By a detailed comparison, the pros and cons of the techniques are discussed. In the end, the problems in current research and some future research issues are outlined.

Key words: uncertain data; nearest neighbor; reverse nearest neighbor; join; query processing

1 引言

概率数据库(即不确定数据库)自从二十世纪八十年代末期被提出,然而由于当时技术限制并未引起人们足够重视.近十年来,由于传感器网络和移动对象应用的驱动,人们重新认识到不确定数据处理的巨大价值.一方面由于不确定数据在现实的经济、军事、物流、金融以及电信等领域普遍存在[1],包括:传感器网络数据、RFID数据、基于位置数据、隐私数据、互联网数据以及金融数据等[2];另一方面因为不确定数据处理技术能够提供概率保证.自从2003年开始,学术界和工业界开始高度关注不确定数据管理技术的研究开发.目前,不确定数据查询处理技术已经成为空间和移动数据库的前沿研究领域[1~47].

不确定数据的查询方法相对于传统的确定数据查询面临更大挑战:(1)不确定数据需要使用概率查询方

收稿日期:2012-07-25;修回日期:2013-03-18
基金项目:国家自然科学基金(No.61003049);浙江省自然科学基金(No.LY12F02047, No.LY12F02019);浙江省公益性技术应用研究计划(No.2011C23130);中央高校基本科研业务费专项资金(No.2010QNA5051, No.2012QNA5018);浙江大学紫金计划重点项目;嘉兴市科技计划基金(No.2011AY1005);浙江省优秀青年教师项目(No.70611011);嘉兴学院博士启动项目(No.70510010)

法,即查询结果中需提供一个概率字段 p 说明查询结果的质量,而概率计算具有极高的计算成本;(2)不确定数据模型大多以可能世界(possible world)[49~52]理论为基础,而可能世界实例的数量与元组成指数级增长关系.这样查询空间巨大,需要高效的修剪技术,尤其对诸如传感器这种能量受限的应用环境.当前的不确定数据查询方法一般通过排序、剪枝、精化、近似取样以及索引等技术来提高查询效率[1].

2　不确定数据查询基本原理

2.1　不确定数据产生的原因

数据的不确定性在现实世界中普遍(pervasive)存在.例如:由于汽车的移动,GPS 导航系统报告的汽车位置可能只是一定范围内有效的不确定数据;搜索引擎返回给用户的页面由于未及时更新可能许多页面已经无效;广告商给出的商品优惠信息也可能因部分商品已经销售完而具有不确定性.

不确定数据产生的原因是多方面的.它可能缘于设备记录数据的精度误差、实时数据传输产生的网络延迟、数据环境的影响(例如:高压线对设备的影响)、数据丢失、数据隐私保护的需要(例如:时序隐私保护插入的扰动信息)、数据表示粒度的需要(例如:累积查询中的累加和、均值、最大值、最小值)[1].

2.2　不确定数据模型

不确定数据模型是不确定数据查询处理技术首要解决的问题.目前,已经出现了的概率数据库管理系统都通过不确定的概率模型来表示、管理数据,这些系统主要包括:美国华盛顿大学的 MystiQ[3]、斯坦福大学的 Trio[4]和 ULDBs[5]、康乃尔大学的 MayBMS[6]、普渡大学的 Orion[7]以及加拿大多伦多大学的 Conquer[8].分为基本模型和完全模型两种类型,它们大都基于可能世界语义实现.前者假定所有元组相互独立且每个以确定的概率出现,后者能够表示数据库实例任意的概率分布(例如:Orion)且元组之间可以存在任意关联.

不确定模型从可能世界的语义划分,分为属性级不确定模型(attribute-level uncertainty model)和元组级不确定模型(tuple-level uncertainty model).

(1)**属性级不确定模型**　在属性级模型中,概率数据库包含一个 n 个元组的表.每个元组有一个属性值不确定.该属性值由离散概率(或连续的概率密度函数)来描述其值的分布.考虑图 1 所示的例子:图 1(a)的表中包含 3 个元组 t_1、t_2 和 t_3,元组分数及概率显示在第 2 列,如元组 t_1 的分数可能是 100 或 70,其对应概率分别为 0.4 和 0.6.元组看作具有一定边界范围的随机变量.可能世界实例由各个元组抽取的一个可能分数值组合而成,其对应概率为各分数值的概率乘积.例如:可能世界实例 $W=\{t_1=100, t_2=92, t_3=85\}$ 的概率为 $Pr[W]=0.4\times0.6\times1$.可能世界实例的个数 $|W|$ 为各元组所包含元素个数的乘积,即 $|W|=\prod_{i=1}^{n}|t_i|$,n 表示元组个数,$|t_i|$ 表示元组 t_i 的模,即 t_i 包含的元素个数,例如:元组 t_1 的模为 2,它包含 2 个元素.

Tuples	score
t_1	{(100, 0.4), (70, 0.6)}
t_2	{(92, 0.6), (80, 0.4)}
t_3	{(85, 1)}

(a)

world W	$Pr[W]$
$\{t_1=100, t_2=92, t_3=85\}$	0.4×0.6×1 = 0.24
$\{t_1=100, t_2=80, t_3=85\}$	0.4×0.4×1 = 0.16
$\{t_1=70, t_2=92, t_3=85\}$	0.6×0.6×1= 0.36
$\{t_1=70, t_2=80, t_3=85\}$	0.6×0.4×1= 0.24

(b)

图1　基于属性不确定性可能世界示例

(2)**元组级不确定模型**　在元组级模型中,每个元组的属性固定,但整个元组可能出现或不出现.其简单实现方法假定每个元组 t 以概率 $p(t)$ 独立出现.其复杂实现方法需考虑元组间存在的依赖关系,并通过生成规则(generation rules)来确定,因而更具普遍意义.图 2(a)给出了 4 个元组 t_1、t_2、t_3 和 t_4 的分数和其对应出现概率 $p(t)$,例如:元组 t_1 的分数为 100,其出现概率为 0.4.图 2(b)给出了元组之间的关联表,同一规则的元组间互斥,不同规则的元组间相互独立.例如:规则 r_2 表示元组 t_2 和 t_4 互斥,即 t_2 和 t_4 不能同时出现;规则 r_1 和 r_3 中的元组 t_1 和 t_3 相互独立.图 2(c)是元组的可能世界实例表,它包含了可能世界实例和其对应概率,例如:$W_1=\{t_1, t_2, t_3\}$ 的概率为 0.2.可能世界 W 是不确定关系表的子集.包含 n 个元组的关系表中,其可能世界实例数为 2^n 个,某个可能世界的出现概率 $Pr[W]$ 为各个生成规则 r_i 的概率 $pw(r_i)$ 的乘积,即 $Pr[W]=\prod_i^M pw(r_i)$.生成规则 r_i 的概率 $pw(r_i)$ 包括两种情形:当 $r_i \frown W=\{t\}$ 时,则 $pw(r_i)=p(t)$;当 $r_i \frown W=\varnothing$ 时,则 $pw(r_i)=1-\sum_{t_j\in r_i}p(t_j)$.

tuples	score	$p(t)$
t_1	100	0.4
t_2	92	0.5
t_3	80	1
t_4	70	0.5

(a)

ID	rules
r_1	$\{t_1\}$
r_2	$\{t_2, t_4\}$
r_3	$\{t_3\}$

(b)

world W	$Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3)=0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4)=0.2$
$\{t_2, t_3\}$	$(1-p(t_1))p(t_2)p(t_3)=0.3$
$\{t_3, t_4\}$	$(1-p(t_1))p(t_3)p(t_4)=0.3$

(c)

图2　基于元组不确定性可能世界示例

3 不确定数据索引技术

3.1 索引的基本原理

在传感器网络应用中,可能需要查询哪些传感器记录的不确定数据处于一个间隔(例如:$[a,b]$)内且其概率大于给定概率门限值[9].在移动对象数据库应用中,可能需要查询哪些移动对象处于一个给定的区域(例如:多维最小边界矩阵)内且其概率大于给定的门限值[10].那么如何处理这类查询呢? 显然,一种简单方法是先判断那些不确定对象的间隔(或区域)与给定的间隔(或区域)重叠,然后计算这些对象的概率.然而,这种方法效率低下,因为一般重叠的对象中仅有很少的对象满足概率门限值约束条件.事实上,在概率密度函数和查询间隔(或区域)已知的情况下,可以通过预计算的方法计算出给定概率门限值下间隔(或区域)的左右边界(或每维左右边界),并将不确定数据的概率密度以及概率门限值等信息保存在索引(例如:R-tree)中,这即不确定数据的索引技术.

(1)**一维索引方法** 一维索引方法通过概率密度函数积分,将所有不确定对象在指定间隔内的概率事先计算好,然后将间隔和概率值存储在索引当中.建立索引的方法类似于 R-tree,每个中间节点由子节点的边界矩阵和它们的指针组成,每个子节点另外还包括了一个由多个间隔值以及对应间隔的概率门限值组成的表;叶子节点中为不确定数据间隔和其概率密度函数.该索引技术典型的方法如 PTI(probabilistic threshold indexing)[9],它利用 x-bound 技术解决了概率门限值查询 PTQ(probabilistic threshold query).PTQ 查询返回以不少于门限值 p 的概率处于间隔范围内的所有不确定对象.x-bound[9]本质上是一种概率预计算方法,它指间隔左右两边的概率都不超过 x,即 $\int_{L_i}^{M_i.lb(x)} f_i(y) \le x$ 且 $\int_{R_i}^{M_i.rb(x)} f_i(y) \le x$,其中,$M_i$ 指第 i 个不确定对象的最小边界矩阵,L_i 和 R_i 为 M_i 的左右边界,$M_i.lb(x)$ 和 $M_i.rb(x)$ 为 M_i 的概率门限值 x 的左右间隔值.

(2)**多维索引方法** PTI 索引仅能处理一维不确定数据,事实上可以扩展 PTI 索引到多维情形,以处理给定区域的不确定数据查询.假定对象 o 的不确定区域为 $o.ur$,其概率约束区域 PCR(probabilistic constrained region)为由四条直线 l_{1-}、l_{1+}、l_{2-}、l_{2+} 围成,且满足在 l_{1-} 的左边、l_{1+} 的右边、l_{2-} 的下面、l_{2+} 的区域,且 PCR 的概率积分为给定的门限值 p,例如:对于 l_{1+} 则有 $\int_{l_{1+}}^{\infty}\int_{-\infty}^{\infty} f(x,y)\mathrm{d}y\mathrm{d}x = p$,$f(x,y)$ 为阴影区域概率密度.这样,通过判断给定不确定对象 o 与查询对象 q 的位置关系,即可判断 o 是否是查询结果而不必计算 o 与查询区域相交的概率,例如:o 与 q 的矩形区域 r_q 仅在 l_{1+} 的右边重叠,故可判断 o 出现在 r_q 中的概率小于 p.这即是支持任意概率密度的不确定数据索引 U-tree[10].该索引不限制不确定数据的分布函数.

(3)**倒排索引方法** 不确定数据的倒排索引方法源于信息检索中的倒排索引,其基本思想是维护一组列表中的列表.从倒排索引结构来看,一般外部列表在列表较大时采用随机访问方法,否则采用顺序访问;内部列表按照概率递减排序并组织成动态的结构(例如:B-Tree)以方便索引插入和删除操作,采用 B-tree 搜索或二分查找方法.扩展倒排索引来处理不确定的分类数据,这即概率倒排索引(probabilistic inverted index)[11].索引的外部列表对应每个元组,其内部列表为元组所属的分类域信息,由对象标识(tuple-id)和对象取得某值的概率组成.搜索过程首先通过匹配查询目标来决定应该搜索各列表中的那些不确定对象,然后从这些候选对象中去掉不满足门限值条件的对象;为了方便处理 PETQ(probabilistic equality threshold query)查询,文献[11]开发了行修剪、列修剪以及具有最高概率的列表优先搜索等方法.

(4)**其他索引方法** 高斯索引 Gauss-Tree[12]使用概率特征向量 pfv(probabilistic feature vectors)表示对象的不确定性.其中,每个特征值为其概率分布的方差.APLA-tree[13]通过自适应点对线性近似 APLAs 技术生成概率直方图,以表示对象任意的概率分布.同时,它通过计算查询对象 q 的期望近邻 ENN(expected nearest neighbors)来定义近邻关系,以避免高成本的概率计算.Kanagal 等人[14]则将不确定索引技术扩展到相关的不确定数据中.他们使用连接树(junction tree)来表示概率数据库的关系,利用树分区(tree partitioning)技术来创建索引.UV-Diagram(uncertain voronoi diagram)索引[15]则通过扩展凸多边形图(voronoi diagram)来支持不确定数据的近邻查询.最近,Angiulli 等人[16]则利用三角不等式、基于枢轴(pivot)选择和近似技术,提出了度量空间中的不确定对象范围查询索引 UP-index[16].

3.2 索引方法性能比较

当对象不确定区域很小时,基于不确定间隔索引[9]的查询效率很高.但是,由于不确定区域被当作索引原子单位,如果没有进一步搜索不确定区域,算法不能判断与查询区域层叠的对象是否是查询结果.这样,如果对象的 MBR 太大,使得基于 PCRs 的索引 U-tree[10]在范围查询中显著优于基于不确定间隔的索引.然而,U-tree 的缺点在于不能很好支持非矩阵区域或不沿轴

向排列的矩阵区域的不确定对象的范围查询.倒排索引针对存在大量低概率的分类时有较高的效率.常用的启发式能够较好优化查询性能,但各种策略的效力依赖于数据特性.UI-tree[17]结合了基于 R-tree 的索引[10]和倒排索引方法[11].基于 Voronoi Diagram 的索引不能直接利用已有的地理信息计算方法,同时它的分区数目与不确定对象的数目成指数级增长关系,因而具有极高成本.但是,该索引在概率近邻查询方面具有天生的优越性.高维空间不确定数据索引方法具有更高复杂度.一般可采用子空间聚类[18]和网格分区等方法来实现.综合对比分析,可以得出下列结论:

(1)不确定数据索引方法与数据概率模型紧密相关.为了简化计算,连续模型可以抽取分布特征来表示 pdf[12]或预计算概率边界[10]等方法来实现;

(2)不确定数据的索引方法依赖于数据分布特征(例如:不确定区域形状和概率密度函数)以及支持的概率查询类型;

(3)概率近邻查询索引相对于范围查询具有更高的难度.因为概率近邻查询与其它对象有关,例如:Xie 等人[15]提出的 UV-diagram 索引.

(4)不确定字符串、基于度量函数的空间应用等领域的不确定数据查询则需要借助于不确定数据的度量空间索引方法,如:UP-index[16].

为了便于对比分析,表 1 总结了几种典型的概率索引方法.

表 1　各种不确定数据索引对比

索引名称	概率模型	支持的概率查询	索引特性
PTI[9]	连续模型	基于间隔的门限值查询(PTQ)	较依赖不确定区域大小
U-tree[10]	连续模型	任意概率密度范围查询	依赖不确定区域形状和门限值
Inverted index[11]	离散模型	概率相等门限值查询(PETQ)	主要应用于空间数据库的关键词搜索
PDR-tree[11]	离散模型	基于分类不确定数据的门限值和 Top-k 查询	简单灵活,效率较高
UV-Diagram[15]	连续模型	近邻查询	概率近邻查询效率较高
UP-index[16]	连续模型	多值概率密度函数的范围查询	用于度量空间,主要技术也可建立向量空间索引
UI-tree[17]	混合模型	范围查询、Top-k 范围查询、相似连接查询	支持混合概率模型并支持多种范围查询

4　不确定数据的近邻查询

4.1　概率近邻查询的定义和分类

传统近邻定义为距离查询对象 q 最近的对象.由于不确定对象包含多个数据实例,且每个实例以一定概率出现并满足约束条件:所有对象出现的概率之和为 1.因而,它的近邻指那些可能成为 q 近邻的数据对象.其近邻查询结果为一个数据集合 S,集合中的每个对象 $o\in S$ 有大于零的概率成为 q 的近邻.这种空间上的不确定性为位置不确定性(也可理解为不确定数据值的不确定性).另一种不确定性为存在不确定性[19],它假定数据库中的每个对象 O 都存在,但是以一定的概率 $p(O)(0<p(O)<1)$存在,且不确定对象仅包含一个数据实例.

从类型方面来划分,不确定数据的近邻查询分为:基本的不确定数据近邻查询[20~23]和扩展的不确定数据近邻查询[24~29].前者从基本近邻查询定义出发,研究不确定数据的范围查询[20]、最近邻查询[21,22]以及累积查询[23].后者研究各种查询变体,例如:概率约束近邻查询[24]、概率组近邻查询[26]、概率反向近邻查询[27~29]等.

4.2　基本的不确定数据近邻查询

4.2.1 查询定义及分类

基本的不确定数据近邻查询包括:范围查询[20]、最近邻查询[21,22]、累积查询[23]等.

不确定数据范围查询检索出与查询区域相交且概率大于零($p>0$)的所有不确定对象,不像精确数据范围查询仅仅返回处于查询区域中的精确对象,它需要根据不确定对象的概率密度 $f(o)$ 计算其概率 $p(o)=\int_{o\sim R} f(o)$,保留那些概率大于查询概率门限值的数据对象.

不确定数据最近邻查询则指检索出与查询对象 q 最近且 $p>0$ 的不确定对象,一般查询结果为多个对象(不像精确对象最近邻查询仅仅返回一个对象).其特点在于:近邻概率计算取决于所有对象之间的相对位置关系,因而比范围查询具有更高复杂度.

不确定数据累积查询检索出满足累积条件(例如:最大值或累加值)且概率 $p>0$ 的不确定对象.由于需考虑不同对象间的相互影响,累积查询比范围和近邻查询具有更高难度.

Cheng 等人[20]最先对不确定数据查询进行分类,将其分为基于值和基于实体的概率查询两种类型.例如:查询传感器 S_{18} 记录的风速是一种概率单值查询.这两种查询进一步可以根据它们是否包含累积特性,划分为累积查询和非累积查询.基于值的概率查询返回一

个满足约束条件的间隔值和其对应概率或对象和其概率;基于实体的概率查询检索出满足查询条件的一组对象及其概率.这些查询类型包括:基于实体的概率范围查询 ERQ[20]、基于实体的概率近邻查询 ENNQ[20]、基于值的最小值查询 VMinQ[20]等.

4.2.2 不确定数据近邻查询处理方法

不确定数据近邻查询(例如:ENNQ)过程一般可分为映射(projection)、修剪(pruning)、边界(bounding)和评估(evaluation)四个阶段[20].

(1)映射阶段:根据应用的不确定模型为每个对象计算不确定区域.例如:移动对象的不确定区域形状由不确定模型、对象上次记录的位置、经过的时间和最大速度决定;

(2)修剪阶段:修剪掉那些成为查询对象 q 的近邻概率为零的对象,以减少昂贵的近邻概率计算成本.例如:如果对象 A 与 q 的最近距离比对象 B 对应的最远距离大,则 A 可以修剪掉.因此,算法关键在于计算每个不确定对象 o_i 与 q 的最远距离 $fDist_i$,然后取它们的最近距离 $nDist = \min_{i=1,\cdots,n}(fDist_i)$作为修剪边界($n$ 为不确定对象数目),以修剪与 q 的最近距离比 $nDist$ 大的对象;

(3)边界阶段:为了进一步提高效率,需要修剪掉那些与查询区域相交且处于查询区域之外的不确定对象区域.查询区域为以 q 为中心,$nDist$ 为半径的超球体.

(4)评估阶段:计算每个对象 o_i 的概率,基本原理是转化 o_i 近邻概率计算为其概率密度函数 pdf 和累积密度函数 cdf 计算.具体方法为:在距离区间[$nDist_i$, $fDist_i$]范围内进行概率积分,其中,$nDist_i$ 表示对象 o_i 距离 q 的最近距离.基本原理为:一个对象 o_i 以距离 r 成为查询对象 q 的近邻的概率 p_i 等于对象 o_i 和 q 的距离为 r 的概率 $P_1 = p_i(r)$乘以每个其它对象 o_k 距离 q 的距离大于或等于 r 的概率 $P_2 = \prod_{k=1\wedge k\neq i}^{n}(1 - P_k(r))$之积;其中,$n$ 为不确定对象数目,$P_k(r)$为不确定对象 o_k 与 q 距离为 r 的概率.进一步地,当按最近距离升序排列且下标 i 表示第 i 个最近距离对象时,o_i 的概率 P_2 可简化为 $\prod_{k=i+1}^{n}(1 - P_k(r))$.

其它不确定数据近邻查询处理方法还包括:聚类[22]、取样[22]以及索引[19]等方法.

4.3 扩展的不确定数据近邻查询

扩展不确定数据近邻查询,一方面可以改变查询条件,例如:约束查询结果的约束近邻查询[24]和 k 近邻查询[25]、将单查询对象 q 改变为多查询对象 $Q = \{q_1, q_2, \cdots, q_n\}$的组近邻查询[26];另一方面可以改变查询语义,例如:从逆向角度思考的概率反向近邻查询[27~29]、寻找最好近邻集合的替代近邻查询[30]、以及从多目标决策角度研究的 Skyline 查询[31~34]等.常见处理技术包括:基于空间关系的修剪方法[26~27]、概率预计算的修剪方法[24,25]以及概率边界修剪方法[25~29].下面总结处理这类查询的基本原理、原则和方法.

(1)**概率约束近邻查询和概率 k 近邻查询** 约束的概率近邻算法[24]通过增加容忍度(tolerance)参数约束条件,以修剪掉概率较小的大部分查询对象.其一般思路是依据距离信息和概率密度变化点位置信息划分距离 $R_i = |o_i - q|$成多个距离子分区 $S = \{S_1, S_2, \cdots, S_M\}$($M$ 为子分区数目),然后预计算 R_i 在每个分区 S_j 中的概率和累积概率,最后基于这些预计算的概率信息提出评估近邻概率的上边界和下边界计算公式.概率门限值 k 近邻查询 T-k-PNN[25]返回包括 k 个近邻对象组成的集合 R 且满足集合概率 $p(R)$大于给定的门限值 T,能够应用于基于位置的服务、传感器监控以及生物管理系统.T-k-PNN 算法首先通过 k-边界过滤方法(k-bound filtering)移除不可能成为 q 的 k 近邻的对象,它定义 k-边界为升序排列的第 k 个最远距离 f_k;然后通过引理 $p(R) \leq \prod_{oi\in S'} Pr(r_i \leq f_k)$计算 R 的概率上边界,进一步约简搜索空间.其中 $\forall R' \subseteq R$, r_i,表示对象 o_i 与 q 之间的距离;最后在验证阶段利用对象的累积概率信息[24]提出了集合概率 $p(R)$的上边界和下边界函数,更进一步减少 $p(R)$的计算成本.

(2)**概率组近邻查询和概率替代近邻查询** 概率组近邻 PGNN(probabilistic group nearest neighbor)[26]查询返回一个不确定对象集合,以使得每个对象到 Q 的距离最小,同时其概率大于一个门限值.PGNN 在森林扑火(起火点模型为不确定对象,多个扑火者为查询对象)、多图像特征的近邻搜索等领域具有重要应用价值.给定不确定对象 o 和 p,如果 p 与 Q 的上边界累积距离 $UB_adist(p, Q)$小于等于 o 与 Q 的下边界累积距离 $LB\_adist(o, Q)$,则 p 可以被修剪掉.另一方面,当将不确定对象的概率分布已知时,则可以预计算出不确定对象基于某种区域(例如:球形区域)的概率上下边界,从而避免不确定数据对象的概率计算.在概率修剪中,如果 $UB\_adist(p_{1-\beta}, Q) < LB\_adist(o, Q)$,则仍然可以修剪掉 p.其中,$p_{1-\beta}$为处于 p 的$(1-\beta)$-超球面区域中的任意对象,$\beta \in [0, \alpha]$

概率替代近邻查询[30]是另一种重要的扩展不确定数据近邻查询.给定查询对象 q,近邻候选对象(成为 q 的近邻的概率大于零)o_1 和 o_2,如果 o_1 比 o_2 更可能靠近 q,则说 o_1"替代"(supersede)o_2.如果某不确定对象替代所有其它近邻候选对象,则说该对象是 q 的替代近邻(superseding nearest neighbor,SNN).当没有一个对象

能够替代其它所有近邻候选集对象时,SNN 查询则返回一个最小的近邻集合(称为 SNN 核),集合中的每个对象能够替代集合外的每个近邻候选对象.SNN 核可以当作是不确定对象中的最好近邻对象集合.

(3)**概率逆向近邻查询**　概率逆向近邻查询算法 PRNN(probabilistic reverse nearest neighbor)[27],返回那些成为查询对象 q 的逆向近邻(RNN)概率大于给定的门限值的不确定对象.从可能世界语义角度处理 PRNN 查询是不同的研究思路[28].在金融或图像数据分析、传感器数据监控、多目标决策以及商业计划等领域中,常常需要识别一个给定对象(例如:q)在其它对象中的重要性(例如:排序或优先级),这即 PIR(probabilistic inverse ranking)查询[29].概率逆向查询[27~29]能够反映查询对象 q 对其它不确定对象的影响,在决策支持、环境信息处理、资源分配、个性化市场策略、游戏开发以及军事策略计划等方面具有广阔的应用价值.主要处技术包括:基于垂直平分线的空间修剪[27,28]和概率上下边界修剪方法[29]等.

(4)**概率 Skyline 查询**　不确定数据的 Skyline 查询本质上是一种扩展概率近邻查询.它能够应用到市场分析[31]、多目标决策、数量经济学、环境维护[32]等领域.概率 Skyline 查询模型通常假定不确定对象包含多个相互独立的实例,且它们以相同的概率出现.不确定对象 U 处于 Skyline 中的概率为每个实例 $u\in U$ 的概率密度与 u 不被其它所有不确定对象 V 控制的概率乘积在整个 U 的数据空间中的积分[31].概率 Skyline 查询检索出所有概率大于给定门限值的不确定对象.一般处理技术包括:概率上下边界修剪[31~34]、层次分区方法[31]、与近邻紧密相关的空间修剪方法[32]、基于概率分布特性的概率预计算方法[32]、数据流环境下 Skyline 集合的动态维护技术[33]以及概率 Top-k 控制查询 PTD(probabilistic top-k dominating)中的近似处理技术[34]等.

5　不确定数据的排序查询

排序是数据分析和决策支持的基础操作.传统排序查询返回参考函数值最高的前 k 个排序对象.不确定数据排序查询需在传统排序查询的基础上再考虑概率的语义.因为,确定数据库中的元组排序值为一个确定值,而不确定数据元组排序值则是一个服从某种分布的随机变量.严格的不确定数据排序查询将给出每个不确定对象基于某种语义的确切排序值.而不确定数据的 Top-k 查询则是一种非严格的排序查询,仅仅给出不确定对象的排序值范围,例如:排序值$r\in[1,k]$.

5.1　概率排序查询语义

为了适应不同应用需求,通常不同概率排序方法的语义并不保持一致,例如:有的 Top-k 查询返回 k 个查询结果,而另一些则并不是精确的 k 个.文献[36,37]总结提出了概率排序查询应满足的五个基本属性,即 exact-k、包容性(containment)、唯一排序(unique ranking)、排序不变性(value invariance)和排序稳定性(stability).文献[38]在此基础上补充了公平性(fairfulness)属性的语义要求.这些属性含义分别如下:

(1)exact-k 属性指出 Top-k 查询结果 R_k 应该为精确的 k 个,即$|R_k|=k$;(2)containment 属性表明 $R_k\subset R_{k+1}$,当使用$\subseteq$代替$\subset$时,该属性则变成弱包含性;(3)unique ranking 属性要求 $R_k\subset R_{k+1}$,当使用$\forall i\neq r_k(i)\neq r_k(j)$,$r_k(i)$表示排序值为 i 的输入元组标识;(4)value invariance 属性要求对于任意 k 值,改变对象排序分数的值而不改变排序分数的相对位置时,排序结果不变.例如:对于分数$70\leq 80\leq 92\leq 100$,当替换它们成$1\leq 2\leq 3\leq 1000$时,排序结果仍然相同;(5)stability 属性指出 Top-k 查询不会拒绝更大概率的元组作为查询结果,同时它也表明比 Top-k 概率更低的元组不会成为 Top-k 查询的结果;(6)fairfulness 属性表明如果对象 X 比对象 Y 更概率靠近,那么当 Y 属于 Top-k 查询时,X 一定属于 Top-k 查询.事实上,概率排序方法与排序中使用的不确定分数函数紧密相关.最近,Soliman M A 等人[39]系统地分析了不确定分数函数的排序语义和计算结果敏感性.

5.2　Ranking 排序查询处理方法

概率排序查询时间复杂度远高于概率近邻查询,它可以根据不同应用背景从不同侧面给出查询定义,例如:基本的概率排序查询 PIR[29]和 PRank[40]、分布式环境下的概率排序查询[41]、偏序域中概率排序查询[42]以及移动轨迹中的概率查询[43]等.从算法设计思路看,首先要区分查询概率模型:离散的或连续的概率模型.因为不同模型处理方法差异巨大:离散概率模型一般可以借助精确处理方法,例如:利用概率的上下边界修剪计算空间[40];而连续概率模型则常需要借助于近似求解方法[38,42].其次要充分考虑不同概率排序查询的特性,例如:基本概率查询[40]需要研究如何减少物化具有指数级的可能世界,分布式概率查询[41]需要考虑如何合并查询结果并尽量减少通讯成本.再次需区分不同类型的不确定数据源:关系型、半结构型等,例如:概率混排 MashRank[44]算法能处理多个不确定数据源.最后要考虑排序分数是全序关系[29]还是偏序关系[42].

从处理方法看,不同概率排序查询处理方法差异巨大.这些方法主要包括:概率上下边界修剪[40]、基于垂直平分线划分平面的空间修剪[29]、动态规划的迭代计算方法[40]、单一站点的局部排序[41]、近似处理方法[42,44,45](例如:蒙特卡罗取样[44],泊松近似[45])、利用

参数平衡排序分数和概率的排序方法[46-47].概率排序查询一般需要集成多种类型的处理方法.下面归纳总结几种常见的处理技术.

5.2.1 概率上下边界修剪

该方法的基本思路是首先根据不确定对象的分数上下边界确定可能的候选对象.例如:2-PRank 概率排序查询[40]中包含 6 个不确定对象 A、B、C、D、E、F,如果第 2 个下边界 $LB_f(D)$大于对象 A、B、C 的上边界,因而对象 A、B、C 不可能成为查询结果.PIR[29]利用查询对象 q 的上边界 $UB\_f(q)$(下边界 $LB\_f(q)$)分数确定 q 的排序范围$[R^q_{\min}, R^q_{\max}]$,$R^q_{\min}$和 $R^q_{\max}$分别表示 q 在不确定数据中的最小和最大排序位置.这里的排序分数一般为指定的单调函数,也可为自定义函数,例如:Zhang 等人[38]提出的期望排序(expected rank)和中值排序(median rank)语义的分数函数.

然后根据候选集合中不确定对象的概率上下边界修剪,返回最终的查询结果,例如:k-PRank 查询检索出 k 个不确定对象 $o_1, o_2, \cdots, o_m, \cdots, o_k$,且每个对象 o_m 排序在第 m 位的概率最高.PRank 算法中,假定 LB_Pr_1, $LB\_Pr_2, \cdots, LB\_Pr_k$ 分别表示概率 $Pr_1(o_1)$, $Pr_2(o_2)$, $\cdots$, $Pr_k(o_k)$的低边界,其中,$Pr_m(o_m)$表示对象 o_m 有第 m 大分数的最大概率,那么如果对象 o 有第 m 大排序分数的概率上边界 $UB\_Pr_m(o) \le LB\_Pr_m(P_m)$, $m \in [1,k]$,对象 o 可以修剪掉.这里的概率边界计算通常采用动态规则方法或预计算方式,以大规模减少概率计算成本.

5.2.2 动态规划的迭代计算方法

动态规划算法满足最优化原理,能够将求解的问题分解为若干子问题(阶段),且下一个子阶段的求解依赖于上一个子阶段的解,最终根据最后子阶段的求解来依次求得其它阶段的解.假定对象 o_m 有第 m 大分数的最大概率为 $Pr(o_m)$,有第 i 大分数的概率为 $Pr(o_m, i)$,而对象 o_{m-1}有第 $m-1$ 大分数的最大概率和 $i-1$ 大分数的概率分别为 $Pr(o_{m-1})$和 $Pr(o_{m-1}, i-1)$,对象 o_{m-2}有 $i-1$ 大分数的概率为 $Pr(o_{m-2}, i-1)$,则可以得出概率排序查询的迭代计算公式:$Pr(o_m, i) = Pr(o_m) \times Pr(o_{m-1}, i-1) + (1 - Pr(o_{m-1})) \times Pr(o_{m-2}, i-1)$.这样,不确定对象的概率排序问题转化成了动态规划问题.

事实上,PIR 查询[29]和 k-PRank 查询[40]都运用动态规划算法的思想.例如:k-PRank 查询中假定对象 o 的当前分数值 s 要排序在第 m 位置的概率 $S(N,m)$表示任意选取的 $m-1$ 个对象 $o_1, o_2, \cdots, o_{m-1}$的分数值大于 s 的概率乘积值 P_1 和其它的对象 $D\{o, o_1, o_2, \cdots, o_{m-1}\}$小于 s 的概率乘积值 P_2 之积 $P_1 * P_2$ 的累加和,概率排序查询迭代形式为 $S(N,m) = S(N-1,m) * (1 - G(o_N)) + S(N-1, m-1) * G(o_N)$.其中,$G(o_N)$表示第 N 个对象 o_N 的分数值大于等于 s 的概率.

5.2.3 近似处理方法

近似处理技术能够在保证一定精度的前提条件下极大地提高算法效率,它是连续模型概率排序查询的关键技术.分为两种类型:近似取样技术和近似计算技术.前者采用随机采样技术(也称为蒙特卡罗—Monte Carlo 方法[48]),保证了在具有巨大实例数目的可能世界中随机选取部分实例样本且样本达到一定数目的条件下,计算的概率能够比较接近于真实结果.独立样本使用基本蒙特卡罗方法产生,连续模型中的相关样本则需要使用马尔可夫链的蒙特卡罗方法(Markov Chain Monte Carlo)[42].近似计算技术则能够较大程度提高概率计算的效率,其前提条件是需要确定不确定数据的概率分布函数.例如:包含 n 个不确定对象 $X_1, X_2, \cdots, X_n$ 的泊松二项分布中,不确定对象 X_i 值为 1 的概率 $Pr(X_i = 1) = p_i$,且令 $X = \sum_{i=1}^{n} \{X_i\}$, $u = [X] \sum_{i=0}^{n} \{p_i\}$,那么可以使用累积分布函数$(\int_{\mu}^{\infty} t^{\lfloor k+1-1 \rfloor} e^{-t} \mathrm{d}t) / \lfloor k \rfloor !$计算概率 $Pr(X < k)$的近似值.

5.3 Top-k 排序查询处理方法

确定数据库 Top-k 查询假定元组分数 s 为元组的排序维度,然后返回 k 个分数最大对象.然而概率数据库多了一个表示元组出现概率的维度 p.因而,不确定数据的 Top-k 排序查询中需要综合考"元组分数"和"元组出现概率"两个因素,同时考虑不同 Top-k 查询语义.单纯从某一个维度排序都是毫无意义的,例如:仅对元组分数排序,那么可能返回概率为 0 的元组作为结果.Top-k 查询广泛应用于数据探索、决策支持以及数据清洗等场景.

5.3.1 Top-k 查询算法设计思路

由于需要考虑排序并累积所有可能世界中的概率值,不确定数据的 Top-k 查询面临极大挑战.除了考虑排序操作中分数和概率的平衡外,Top-k 查询还需考虑几个重要设计因素:(1)区分确定性和近似算法设的特性;(2)如何通过查询语义特点避免确定性算法展开整个可能世界空间.例如:基于概率上边界的状态空间扩展方法[49]、基于门限值修剪的动态规划方法[37]等;(3) 如何通过近似计算来大幅度提高查询效率,例如:U-Top$k$ 查询[49]的动态马尔可夫取样方法、MS-Top$k$ 查询[50]中的蒙特卡洛(monte-carlo)取样方法、PT-k 查询[45]中的蒙特卡洛积分方法、泊松分布近似计算方法[45]等;(4)如何对连续分布的属性值进行离散取样,例如:分数区域等距离取样方法、按概率相等的方法划分区间的等深离散取样

方法、样条取样方法(将任意分布的密度函数采用样条逼近的方式近似为线性函数);(5)如何在特定场景下设计 Top-k 查询,例如:针对数据流环境下的 Pk-top$k$ 查询[51]、分布式环境下的 Top-k 查询[41].

基于上述设计思想的不确定数据 Top-k 查询方法主要有:U-Top$k$ 查询[49]、U-kRanks 查询[49]、MS-Top$k$[50]、PT-k 查询[45]、Pk-top$k$ 查询[51]、基于 x-relation 模型的 Top-k 查询[52]、k-Selection 查询[53]、Top$k$-PNN[54]、c-Typical-Top$k$[55]、Top-k 控制查询[34,56]、PTkS[57]查询等.

5.3.2 Top-k 排序查询总结和讨论

不确定数据 Top-k 查询可以从不同查询需求来考虑不同查询语义,同时考虑不确定数据的分数和概率排序.

(1)针对累积概率统计对象,存在单个元组和元组集的累积概的区别.例如:U-kRanks 查询、PT-k 查询和 Pk-top$k$ 查询针对单个元组统计在不同可能世界中的累积概率,而 U-Top$k$ 和 k-Selection 查询则统计元组集合的累积概率.

(2)针对单个元组的分数排序前提条件,存在元组在某个排序情况下(如:排第 1 位)和在 Top-k 范围排序(例如:排在前 k 位)情况下的区别.前者如 U-kRanks 查询,后者如 PT-k 查询.

(3)针对累积概率的查询条件,可以分为概率范围查询和概率 k 近邻(kNN)查询.例如:PT-k 为概率范围查询(即概率门限值查询),而 Pk-topk 为概率 kNN 查询.当 $k=1$ 时,即概率最大查询,如:U-Top$k$ 查询、U-kRanks 查询和 k-Selection 查询等.

(4)针对元组的概率含义,计算前 k 个数据对象(或元组)的概率,可以是依据可能世界语义计算的元组存在最大累积概率[49~52]或最好有效存在概率[53],也可为根据控制关系计算的控制最多其它对象的累积概率[34~56],或是根据空间位置关系计算的最靠近指定对象 q 的累积概率[57].

6 不确定数据的连接查询

连接查询是一种非常重要的数据库操作,它基于某些查询谓词将两个数据集合并成一个由数据对象对组成的单一集合.不确定数据的连接查询在移动和基于位置的服务、图像和多媒体数据库、人脸识别与指纹分析系统等应用中存在广泛应用价值,例如:移动和位置服务中,移动对象的位置信息由于连续移动而无法获取其精确位置,不确定数据的连接查询能够通知移动用户是否他的一个朋友进入他们的邻近区域.

6.1 不确定数据连接查询的定义和分类

给定两个由不确定对象组成的关系 R 和 S,不确定数据连接查询 $J(R,S)$是一个笛卡尔积集合 $J(R,S)=(a,b,s(a,b))\in R\times S\times[0\cdots1]$,其中 $s(a,b)$表示根据给定的比较操作和连接谓词计算出的两个不确定对象 a 和 b 之间的分数.表 2 总结了目前出现的概率连接查询.

表 2　概率连接查询分类

参考文献	概率查询类型	连接谓词	分数类型
[58]	PJQ,PTJQ,PTop$k$JQ	ε-Range	Similarity Distance
[59]	PJQ,PTJQ	ε-Range	Similarity Distance
[60]	PTJQ,PTop$k$JQ	ε-Range	Spatial Distance
[64]	PTJQ,PTop$k$JQ,Other	Any	Any

不确定数据连接查询可以根据数据表示形式、查询类型、连接谓词以及分数函数等标准进行分类.不确定数据的数据表示形式主要包括连续概率模型、离散概率模型和空间概率模型.不确定数据连接查询类型包括:概率连接查询 PJQ(probabilistic join query)、概率门限值连接查询(probabilistic threshold join query)和概率 Top-k 连接查询 PTop$k$JQ(probabilistic Top-k join query)等.不确定数据连接查询从连接谓词角度划分,则有 ε-范围和 k 近邻(k-NN)连接查询.从连接分数来划分,则包括布尔值比较操作、相似距离查询和空间距离查询等.

6.2 连接查询算法设计思路

传统的基于精确数据的连接操作一般使用循环比较、块循环比较、排序合并、哈希合并以及索引等方法实现.目前,不确定数据连接查询一方面继承前述处理精确数据的方法,另一方面还需对已有方法进行适当修改或开发新的算法,以考虑概率维度.连接查询主要在不确定数据表示、距离度量类型、连接查询类型和查询谓词以及结果的表示方面有所区别.

(1)**基于置信度的连接查询处理**　基于置信度的连接查询算法根据候选元组的置信度来约简搜索空间,它不考虑连接相关的对象属性和连接谓词.一般处理方法是将两个关系 R 和 S 中的元组对象的置信度按照降序排列处理[64].假定对象之间相互独立,那么两个对象连接对的可能性最大为它们置信度之积.这种方法可大大加速 PTJQ、PTop$k$JQ 等查询执行效率,并且使用简单块循环比较方法即可实现.

(2)**概率相似连接查询处理**　概率相似连接查询要考虑对象间的置信度和相似关系,它一般仅有非常少比例的候选对象满足连接谓词.因而,高效的修剪方法能够较大程度提高其查询效率.对于连续不确定模型,体现两个不确定对象间的相似分数需要连续的概率密度函数表示,它使得表示两个对象间的相似概率也是一种连续的 pdf.

离散不确定模型概率相似查询处理技术主要有聚类[58]、取样[61]、概率预计算[61]以及概率上边界修剪[62,63]等方法.文献[58]利用 k-means(k 均值)聚类算法和取样方法将每个对象的取样点分成 k 个组,每个聚类通过一个最小边界超矩阵近似表示,大大约简了连接处理的计算复杂度;同时,通过过滤-精炼的多步查询范例进一步提高修剪效率.Jestes 等人[62]研究概率字符串连接问题,则使用期望编辑距离 EED(expected edit distance)作为相似度量,并开发了基于 q-gram[62]的低边界过滤和上边界连接对选取等高效修剪方法.Lian 等人[63]则利用 Jaccard 距离[63]修剪、概率上边界修剪以及累积修剪方法,讨论了基于可能世界语义的概率集合相似连接 PS^2J(probabilistic set similarity join).

连续模型概率相似连接[60]将每个不确定对象当作一个随机变量,每个数据项通过可能值的间隔范围和概率分布来表示.文献[59]提出了数据项层次、页面层次以及索引层次修剪方法.数据项层次修剪通过建立连接谓词过滤概率的上下边界从而避免对象间昂贵的概率评估过程;页面层次修剪为每个节点增加 x-bound[9]以避免页面访问;索引层次修剪进一步将页面组织成树结构,从而可以提高连接查询的 I/O 处理性能.

(3)**概率空间连接查询处理** 概率空间连接查询[60]要考虑不确定对象间的概率属性和空间关系,使用空间谓词(例如:相交和层叠等)和距离谓词(比如:距离范围和近邻等)处理查询.常见处理方法利用空间索引(例如:R-tree)以及空间位置关系(例如:空间中的对象间的最大距离和最小距离等),如应用于地理图形和生物医学图像数据处理的不确定数据流连接查询 USJ(join on uncertain data streams)[61].

7 研究展望

从已有研究成果看,不确定数据查询主要集中在近邻查询、索引、排序查询和连接查询.尽管研究成果已经相当丰富,但仍值得深入,仍有许多符合现实应用需求的新的查询方法没有被提出来,例如:度量空间中的概率查询.

综合国内外最新研究成果,我们认为扩展已有著名的精确数据查询算法到不确定数据中、探索新的满足查询需求的以及开发高效的查询处理算法将是未来的重要方向.另外,基于不确定图数据的查询处理技术[65,66]、分布式环境下不确定数据查询算法[41,67]、基于数据流环境下不确定数据查询算法[51,61,68]、基于安全需求的不确定数据查询算法以及基于多数据源的不确定数据查询算法[44]也可能成为未来持续研究的热点.

参考文献

[1] 周傲英,金澈清,王国仁,李建中.不确定性数据管理技术综述[J].计算机学报,2009,32(1):1-16.
3Zhou Ao-ying, Jin Che-qing, Wang Guo-ren, Li Jian-zhong. A survey on the management of uncertain data[J]. Chinese Journal of Computers, 2009, 32(1): 1-16. (in Chinese)

[2] 李建中,于戈,周傲英.不确定性数据管理的要求与挑战[J].中国计算机学会通讯,2009,5(4):6-14.

[3] Boulos J, Dalvi N, Mandhani B, et al. MYSTIQ: a system for finding more answers by using probabilities[A]. Proc of ACM SIGMOD Conf[C]. Maryland, USA: ACM, 2005. 891-893.

[4] Agrawal P, Benjelloun O, Das Sarma A, et al. Trio: a system for data, uncertainty, and lineage[A]. Proc of the 32nd VLDB Conf[C]. Seoul, Korea: ACM, 2006. 1151-1154.

[5] Benjelloun O, Sarma A D, Halevy A, et al. Databases with uncertainty and lineage[J]. The VLDB Journal, 2008, 17(2): 243-264.

[6] Antova L, Koch C, Olteanu D. Fromcomplete to incomplete information and back[A]. Proc of ACM SIGMOD Conf[C]. Beijing, China: ACM, 2007. 713-724.

[7] Singh S, Mayfield C, Mittal S, et al. Orion 2.0: native support for uncertain data[A]. Proc of ACM SIGMOD Conf[C]. Vancouver, BC, Canada: ACM, 2008. 1239-1242.

[8] Fuxman A, Fazli E, Miller R J. Conquer: efficient management of Inconsistent databases[A]. Proc of ACM SIGMOD Conf[C]. Vancouver, Canada: ACM, 2008. 155-166.

[9] Cheng R, Xia Yu-ni, Prabhakar S, et al. Efficient indexing methods for probabilistic threshold queries over uncertain data[A]. Proc of VLDB Conf[C]. Toronto: ACM, 2004. 876-887.

[10] Tao Yu-fei, Cheng R, Xiao Xiao-kui, et al. Indexing multi-dimensional uncertain data with arbitrary probability density functions[A]. Proc of VLDB Conf[C]. Trondheim, Norway: ACM, 2005. 922-933.

[11] Singh S, Mayfield C, Prabhakar S, et al. Indexing uncertain categorical data[A]. Proc of IEEE ICDE Conf[C]. Istanbul: IEEE Computer Society, 2007. 616-625.

[12] Bohm C, Pryakhin A, Schubert M. Thegauss-tree: efficient object identification in databases of probabilistic feature vectors[A]. Proc of IEEE ICDE Conf[C]. Boston, MA, USA: IEEE Computer Society, 2006. doi: 10.1109/ICDE.2006.159.

[13] Ljosa V, Singh A K. APLA: indexing arbitrary probability distributions[A]. Proc of IEEE ICDE Conf[C]. Istanbul: IEEE Computer Society, 2007. 946-955.

[14] Kanagal B, Deshpande A. Indexing correlated probabilistic databases[A]. Proc of ACM SIGMOD Conf[C]. Rhode Island: ACM, 2009. 455-468.

[15] Xie Xi-ke, Cheng R, Yiu Man Lung, et al. UV-diagram: a

voronoi diagram for uncertain spatial databases[J]. The VLDB Journal, 2012, doi: 10.1007/s00778-012-0290-x.

[16] Angiulli F, Fassetti F. Indexing uncertain data in general metric spaces[J]. IEEE Trans on Knowl and Data Engineering, 2012, 24(4): 1640-1657.

[17] Zhang Ying, Lin Xue-ming, Zhang Wen-jie, et al. Effectively indexing the uncertain space[J]. IEEE Trans Knowledge and Data Eng, 2010, 22(9): 1247-1261.

[18] 庄毅,胡海洋,胡华.基于质心片的不确定高维索引研究[J].电子学报,2011,38(5):1136-1142.
21Zhuang Yi, Hu Hai-yang, Hu Hua. Centroid-slice-based uncertain high-dimensional indexing structure[J]. Acta Electronica Sinica, 2011, 38(5): 1136-1142. (in Chinese)

[19] Yiu M L, Mamoulis N, DaiXiang-yuan, et al. Efficient evaluation of probabilistic advanced spatial queries on existentially uncertain data[J]. IEEE Trans on Knowl and Data Eng, 2009, 21(1): 108-122.

[20] Cheng R, Kalashnikov D, Prabhakar S. Evaluating probabilistic queries over imprecise data[A]. Proc of ACM SIGMOD Conf[C]. San Diego, California: ACM, 2003. 551-562.

[21] Cheng R, Kalashnikov D, Prabhakar S. Querying imprecise data in moving object environments[J]. IEEE Trans on Knowl and Data Eng, 2004, 16(9): 1112-1127.

[22] Kriegel H P, Kunath P, Renz M. Probabilistic nearest-neighbor query on uncertain objects[A]. Proc of DASFAA Conf[C]. Bangkok, Thailand: Springer, 2007. 337-348.

[23] Ross R, Subrahmanian V S, Grant J. Aggregate operators in probabilistic databases[J]. Journal of the ACM, 2005, 52(1): 54-101.

[24] Cheng R, ChenJin-chuan, Mokbel M, et al. Probabilistic verifiers: evaluating constrained nearest-neighbor queries over uncertain data[A]. Proc of IEEE ICDE Conf[C]. Cancun, Mexico: IEEE Computer Society, 2008. 973-982.

[25] Cheng R, Chen Lei, Chen Jin-chuan, et al. Evaluating probability threshold k-nearest-neighbor queries over uncertain data[A]. Proc of ACM EDBT Conf[C]. Saint Petersburg, Russia: ACM, 2009. 672-683.

[26] Lian Xiang, Chen Lei. Probabilistic group nearest neighbor queries in uncertain databases[J]. IEEE Trans Knowledge and Data Eng, 2008, 20(6): 809-824.

[27] Lian Xiang, Chen Lei. Efficient processing of probabilistic reverse nearest neighbor queries over uncertain data[J]. The VLDB Journal, 2009, 18(3): 787-808.

[28] Cheema M A, Lin Xue-ming, Wang Wei, et al. Probabilistic reverse nearest neighbor queries on uncertain data[J]. IEEE Trans on Knowl and Data Eng, 2010, 22(4): 550-564.

[29] Lian Xiang, Chen Lei. Probabilistic inverse ranking queries in uncertain databases[J]. The VLDB Journal, 2011, 20(1): 107-127.

[30] Yuen S M, Tao Yu-fei, Xiao Xiao-kui, et al. Superseding nearest neighbor search on uncertain spatial databases[J]. IEEE Trans on Knowl and Data Eng, 2010, 22(7): 1041-1055.

[31] Pei Jian, Jiang Bin, Lin Xue-ming, Yuan Yi-dong. Probabilistic skylines on uncertain data[A]. Proc of VLDB Conf[C]. Vienna, Austria: ACM, 2007. 15-26.

[32] Lian Xiang, Chen Lei. Reverse skyline search in uncertain databases[J]. ACM Transactions on Database Systems, 2010, 35(1), doi: 10.1145/1670243.1670246.

[33] Zhang Wen-jie, Lin Xue-ming, Zhang Ying, et al. Probabilistic skyline operator over sliding windows[J]. Information Systems. 2012, http://dx.doi.org/10.1016/j.is.2012.03.002.

[34] Lian Xiang, Chen Lian. Top-k dominating queries in uncertain databases[J]. Information Sciences, 2013, 226: 23-46.

[35] Lazaridis L, Mehrotra S. Progressive approximate aggregate queries with a multi-resolution tree structure[A]. Proc of ACM SIGMOD Conf[C]. Santa Barbara: ACM, 2001. 401-412.

[36] Jestes J, Cormode G, Li Fei-fei, Yi Ke. Semanticsof ranking queries for probabilistic data[J]. IEEE Trans on Knowl and Data Eng, 2011, 23(12): 1903-1917.

[37] Zhang Xi, Chomicki J. Semantics and evaluation of top-k queries in probabilistic databases[J]. Distributed and Parallel Databases, 2009, 26(1): 67-126.

[38] Zhang Ying, Lin Xue-ming, Zhu Guo-ping, et al. Efficient rank based kNN query processing over uncertain data[A]. Proc of IEEE ICDE Conf[C]. California, USA: IEEE, 2010. 28-39.

[39] Soliman M A, IIyas I F, Martinenghi D. Rankingwith uncertain scoring functions: semantics and sensitivity measures[A]. Proc of ACM SIGMOD Conf[C]. Athens, Greece: ACM, 2011. 805-816.

[40] Lian Xiang, Chen Lei. Ranked queries processing in uncertain databases[J]. IEEE Trans on Knowl and Data Eng, 2010, 22(3): 420-436.

[41] Li Fei-fei, Yi Ke, Jestes J. Ranking distributed probabilistic data[A]. Proc of the ACM SIGMOD Conf[C]. Rhode Island, USA: ACM, 2009. 361-374.

[42] Soliman M A, Ilyas Ihab F, Ben-David S. Supporting ranking queries on uncertain and incomplete data[J]. The VLDB Journal, 2010, 19(4): 477-501.

[43] Trajcevski G, Tamassia R, Cruz I F, et al. Ranking continuous nearest neighbors for uncertain trajectories[J]. The VLDB Journal, 2011, 20(5): 767-791.

[44] Soliman M A, Ilyas Ihab F, Saleeb M. Building ranked mashups of unstructured sources with uncertain information[J]. PVLDB, 2010, 3(1): 826-837.

[45] Hua Ming, Pei Jian, Lin Xue-ming. Ranking queries on uncer-

tain data[J]. The VLDB Journal, 2011, 20(1): 129 – 153.
[46] Li Jian, Deshpande A. Ranking continuous probabilistic datasets[J]. PVLDB, 2010, 3(1): 638 – 649.
[47] Li Jian, Saha B, Deshpande A. A unified approach to ranking in probabilistic databases[J]. The VLDB Journal, 2011, 20(2): 249 – 275.
[48] Karp R, Luby M. Monte-carlo algorithms for enumeration and reliability problems[A]. Proc of the 15th Annual ACM Symposium on Theory of Computing(STOC)[C]. Boston, Massachusetts: ACM, 1983. 56 – 64.
[49] Soliman M A, Ilyas Ihab F, Chang Kevin C.-C. Probabilistic top-k and ranking-aggregate queries[J]. ACM Trans Database Syst, 2008, 33(3): doi: 10.1145/1386118.1386119.
[50] Re C, Dalvi N, Suciu D. Efficient top-k query evaluation on probabilistic data[A]. Proc of IEEE ICDE Conf[C]. Istanbul, Turkey: IEEE Computer Society, 2007. 886 – 895.
[51] Jin Che-qing, Yi Ke, Chen Lei, Yu Jeffrey Xu, Lin Xue-ming. Sliding-window top-k queries on uncertain streams[J]. The VLDB Journal, 2010, 19(3): 411 – 435.
[52] Yi Ke, Li Fei-fei, Kollios G, et al. Efficient processing of top-k queries in uncertain databases with x-relations[J]. IEEE Trans on Knowl and Data Eng, 2008, 20(12): 1669 – 1682.
[53] Liu Xing-jie, Ye Mao, Xu Jian-liang, et al. k-selection query over uncertain data[A]. Proc of DASFAA Conf[C]. Tsukuba, Japan: Springer, 2010. 444 – 459.
[54] Beskales G, Soliman M A, Ilyas I F. Efficient search for the top-k probable nearest neighbors in uncertain databases[J]. PVLDB, 2008, 1(1): 326 – 339.
[55] Ge T, Zdonik S, Madden S. Top-k queries on uncertain data: on score distribution and typical answers[A]. Proc of SIGMOD Conf RI[C]. USA: ACM. 375 – 388.
[56] Zhang Wen-jie, Lin Xue-ming, Zhang Ying, Pei Jian. Threshold-based probabilistic top-k dominating queries[J]. The VLDB Journal, 2010, 19(2): 283 305.
[57] Lian Xiang, Chen Lei. Shooting top-k stars in uncertain databases[J]. The VLDB Journal, 2011, 20(6): 819 – 840.
[58] Kriegel H P, Kunath P, Pfeifle M, et al. Probabilistic similarity join on uncertain data[A]. Proc of DASFAA Conf[C]. Singapore: Springer, 2006. 295 – 309.
[59] Cheng R, Singh S, Prabhakar S, et al. Efficient join processing over uncertain data[A]. Proc of ACM CIKM Conf[C]. New York: ACM, 2006. 738 – 747.
[60] LjosaV, Singh A K. Top-k spatial joins of probabilistic objects [A]. Proc of IEEE ICDE Conf[C]. Cancun, Mexico: IEEE Computer Society, 2008. 566 – 575.
[61] Lian Xiang, Chen Lei. Similarity join processing on uncertain data streams[J]. IEEE Trans on Konwl and Data Eng, 2011, 23(11): 1718 – 1734.
[62] Jestes Jeffrey, Li Fei-fei, Yan Zhe-peng, et al. Probabilistic string similarity joins[A]. Proc of ACM SIGMOD Conf[C]. Indianapolis, Indiana: ACM, 2010. 327 – 338.
[63] Lian Xiang, Chen Lei. Set similarity join on probabilistic data [A]. Proc of VLDB Conf[C]. Singapore: ACM, 2010. 650 – 659.
[64] Agrawal P, Widom J. Confidence-aware join algorithms[A]. Proc of IEEE ICDE Conf[C]. Shanghai, : IEEE, 2009. 628 – 639.
[65] Lian Xiang, Chen Lei. Efficient query answering in probabilistic RDF graphs[A]. Proc of ACM SIGMOD Conf[C]. Athens, Greece: ACM, 2011. 157 – 168.
[66] Li Jian-zhong, Zou Zhao-nian, Gao Hong. Mining frequent subgraphs over uncertain graph databases under probabilistic semantics[J]. The VLDB Journal, 2012, 21(6): 753 – 777.
[67] Ding Xiao-feng, Jin Hai. Efficient and progressive algorithms for distributed skyline queries over uncertain data[J]. IEEE Trans on Knowl and Data Eng, 2012, 24(8): 1448 – 1462.
[68] Ding Xiao-feng, Lian Xiang, Chen Lei, Jin Hai. Continuous monitoring of skylines over uncertain data streams[J]. Information Sciences, 2012, 184(1): 196 – 214.
[69] Dallachiesa M, Nushi B, Mirylenka K, et al. Uncertain time-series similarity: Eturn to the basics[J]. Proc of the VLDB Endowment, 2012, 5(11): 1662 – 1673.

作者简介

蒋 涛 男, 1973 年生于湖南衡阳, 博士, 讲师. 研究方向为时空数据库查询、Skyline 计算、不确定数据处理.

高云君(通信作者) 男, 1977 年生, 博士, 副研究员. 研究方向为空间数据库、GIS 系统、Skyline 查询.
E-mail: gaoyj@zju.edu.cn

张 彬 女, 1978 年生于江苏徐州, 讲师, 硕士. 研究方向为时空数据库查询、Skyline 计算、数据流数据查询.

周傲英 男, 1965 年生, 博士, 教授. 研究方向为空间数据库、数据挖掘 P2P 计算和系统、Web 搜索

乐光学 男, 1963 年生, 博士, 教授. 研究方向为 P2P 网络、传感器网络.