软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2012,23(6):1542−1560 [doi: 10.3724/SP.J.1001.2012.04200]

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

# 不确定性 Top-*K* 查询处理*

李文凤[1], 彭智勇[2+], 李德毅[3]

[1](武汉大学 软件工程国家重点实验室,湖北 武汉 430072)

[2](武汉大学 计算机学院,湖北 武汉 430072)

[3](中国电子系统工程研究所,北京 100840)

## Top-*K* Query Processing Techniques on Uncertain Data

LI Wen-Feng[1], PENG Zhi-Yong[2+], LI De-Yi[3]

[1](State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072, China)

[2](Computer School, Wuhan University, Wuhan 430072, China)

[3](Institute of Electronic System Engineering of China, Beijing 100840, China)

+ Corresponding author: E-mail: peng@whu.edu.cn

**Li WF, Peng ZY, Li DY. Top-*K* query processing techniques on uncertain data. 2012,23(6):1542−1560.** http://www.jos.org.cn/1000-9825/4200.htm

**Abstract**: Efficient processing of Top-*K* queries has always been a significant technique in the interactive environment involving massive amounts of data. With the emerging of imprecise data, the management of them has gradually raised people's attention. In contrast with traditional Top-*K* query, Top-*K* query on uncertain data presents different features both in semantics and computation. On the basis of prevailing uncertain data model and possible world semantic model, researchers have already studied multiple sound semantics and efficient approaches. This survey describes and classifies Top-*K* processing techniques on uncertain data including semantics, rank criteria, algorithms and implementation levels, and so on. Finally, the challenges and future research trends in processing of Top-*k* queries on uncertain data are predicated.

**Key words**: semantic of Top-*K* queries; processing of Top-*K* queries; rank criterion; uncertain data; possible world

**摘 要**: 高效 Top-*K* 查询处理在涉及大量数据交互的应用中是一项重要技术,随着应用中不确定性数据的大量涌现,不确定性数据的管理逐渐引起人们的重视.不确定性数据上 Top-*K* 查询从语义和处理上都呈现出与传统 Top-*K* 查询不同的特点.在主流不确定性数据模型和可能世界语义模型下,学者们已经提出了多种不确定性 Top-*K* 查询的语义和处理方法.介绍了当前不确定性 Top-*K* 查询的研究工作,并对其进行分类,讨论包括语义、排序标准、算法以及应用等方面的技术.最后提出不确定性 Top-*K* 查询面临的挑战和下一步的发展方向.

**关键词**: Top-*K* 查询语义;Top-*K* 查询处理;排序标准;不确定性数据;可能世界

**中图法分类号**: TP311 **文献标识码**: A

* 基金项目: 国家自然科学基金(61070011); 湖北省自然科学基金国际合作重点项目; 武汉市学科带头人计划(201150530139)
收稿时间: 2011-08-08; 修改时间: 2011-11-02; 定稿时间: 2012-02-15; jos 在线出版时间: 2012-03-26
CNKI 网络优先出版: 2012-03-26 13:47, http://www.cnki.net/kcms/detail/11.2560.TP.20120326.1347.001.html

## 1 引 言

随着数据采集技术的进步和网络的快速发展,人们可获取的数据量越来越大.如何从大量数据中选择最符合查询条件的信息,一直是数据管理和信息检索的重要课题.而高效 Top-*K* 查询处理在涉及大量数据交互的应用中逐渐成为一项重要技术,在数据库、网络、分布式系统等领域被广泛研究[1−3].同时,随着自动生成数据、推断数据以及大众数据的大量产生,数据往往存在大量的噪声、丢失值、错误以及不一致,不确定性数据的管理正随着现实应用逐渐被人们重视[4−7].这些应用主要包括:大规模传感器网络系统、信息抽取和数据整合系统、科学数据管理系统以及社会网络.下面是一个不确定性数据库实例.

例 1:多点温度监测可用于如火灾前期温度报警、空调环境监测、工业温度探测、粮仓、土壤、温室、养殖场、农场、冰窟热窟、矿业等多种应用,为提高监测精确程度,常常设置多个温度监测器,由于物理误差和多源监测,监测到的数据常常含有不确定性.表 1 是某时刻一个粮仓的多点温度监测数据库简表,每个温度监测器在特定的时间返回一个温度值,同一个位置有可能放置多个温度监测器,但同一时间只可能使用其中一个监测器的数据.每个监测器有一个可信度属性,反映该监测器的可靠程度.

**Table 1** A uncertain database

**表 1** 不确定性数据库实例

| | 时间 | 检测器 | 位置标号 | 温度 | 可信度 |
|---|---|---|---|---|---|
| $t_1$ | 11:45 | $M_1$ | W-101 | 50°C | 0.4 |
| $t_2$ | 11:45 | $M_2$ | W-218 | 46°C | 1 |
| $t_3$ | 11:45 | $M_3$ | E-012 | 45°C | 0.4 |
| $t_4$ | 11:45 | $M_4$ | E-012 | 44°C | 0.5 |
| $t_5$ | 11:45 | $M_5$ | S-411 | 15°C | 0.7 |
| $t_6$ | 11:45 | $M_6$ | S-411 | 10°C | 0.3 |

Rules: $(t_3 \oplus t_4)$, $(t_5 \oplus t_6)$

当考虑数据的不确定性时,Top-*K* 查询从查询语义到处理技术都面临着巨大挑战.考虑例 1 数据库上一个 Top-1 查询:

- 在 11:45 温度最高的位置

根据表 1,记录 $t_1$ 温度最高,但可靠程度只有 0.4;记录 $t_2$ 虽然温度略低,但可靠程度为 1.这种情况究竟应该返回哪条记录?记录 $t_3$ 和 $t_4$ 位置区域同为 E-012 但可靠程度不同,位置 E-012 温度值该取何值?在不确定世界里,仅仅依靠温度这一个属性决定返回哪条记录,显得不再合理.

由于不确定性的存在,Top-*K* 查询变得不再清晰和不易操作,不能再像传统 Top-*K* 查询那样,仅仅基于某分值函数返回具有最大分值的对象.幸运的是,随着不确定性数据库研究的进展和对不确定性 Top-*K* 查询关注程度加强,针对不确定性 Top-*K* 查询处理的研究工作在近年来取得了较大进展[8−16].本文将基于典型的不确定性数据模型以及可能世界语义[17−22]模型,讨论近期出现的主流不确定性 Top-*K* 查询处理技术.主要涉及的研究方向(如图 1 所示)包括:

(1) 语义合理性和数学性质研究.由于不确定性的引入,“不确定性 Top-*K* 查询究竟要返回哪些记录?”这个问题的答案不再清晰,对这个问题的不同回答形成了不确定性 Top-*K* 查询的不同语义形式.本文首先介绍学术界广泛研究的几种主要语义形式及其满足的数学性质,通过对比,对它们的合理性和满足的数学性质进行分析;

(2) 排序标准的研究.确定性 Top-*K* 对记录进行排序的标准是记录在某分值函数作用下的得分.因此,分值函数成为排序的唯一标准.在不确定性 Top-*K* 查询中,所依赖的排序标准如何变化?本文针对分值与概率的平衡问题、分值的连续与离散问题,对目前不确定性 Top-*K* 排序标准进行分析.另外,鉴于排序标准的变化导致排序结果的差异巨大,本文还将介绍统一化排序方法的新进展;

(3) 查询处理算法的研究.基于可能世界语义进行不确定性数据处理,最大的问题就是可能世界实例爆炸问题,在不确定性 Top-*K* 查询处理中,该问题仍然是我们面临的主要问题.具体而言,确定性 Top-*K*

的高效查询处理就是研究如何在最少的记录读取量和最小的实例空间内完成查询.本文将从确定性方法和近似方法两个方面展开讨论;

(4) 应用层面研究.随着不确定性 Top-*K* 查询研究的展开,学者们展开了诸如高维数据库、分布式系统、数据流等应用中不确定性 Top-*K* 的研究.

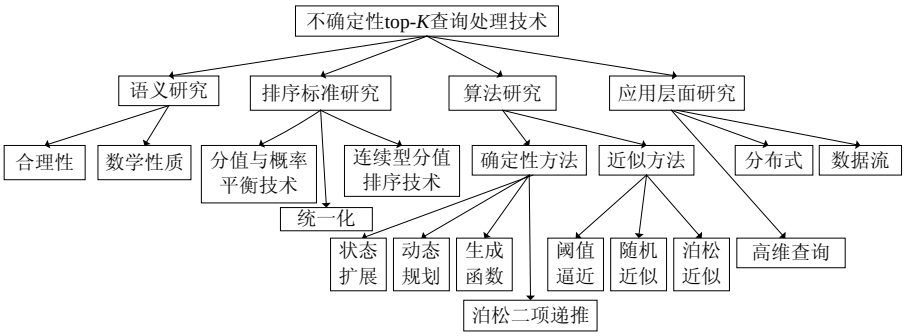



Fig.1 Classification of Top-k query processing techniques on uncertain data

图 1 不确定性 Top-*K* 查询处理技术分类

### 1.1 标 号

本文描述的技术中将使用统一的符号标记,表 2 中将列出本文频繁使用的标号及含义.

**Table 2** Frequently used notations

**表 2** 正文标号表

| 符号 | 含义 |
|---|---|
| $D$ | 不确定性数据库 |
| $w\in W$ | $w$:一个可能世界实例,$W$:所有可能世界形成空间 |
| $\|w\|$ | 可能世界实例的规模(可能世界实例中记录数) |
| $\|W\|$ | 可能世界空间 $W$ 的规模(可能世界实例数) |
| $p(w)$ | 可能世界 $w$ 的概率 |
| $t_i$ | 不确定性数据记录 |
| $p(t_i)$ | 不确定性记录的存在概率 |
| $\tau_j$ | 生成规则,存在约束,依赖规则 |
| $p(\tau_j)$ | 生成规则的概率(所有生成规则内记录概率和) |
| $\|\tau_j\|$ | 生成规则的规模(生成规则中记录数) |
| $f(t_i)=f_i$ | 分值函数 $f$ 作用下,$t_i$ 的得分为 $f_i$ |
| $R_k$ | 不确定性 Top-*K* 查询结果集 |

### 1.2 大 纲

本文第 2 节介绍主流的不确定性数据模型和可能世界语义.第 3 节将介绍目前广泛使用的不确定性 Top-*K* 查询语义,并分析其合理性和满足的数学性质.第 4 节将从不确定性数据排序标准的角度阐述目前不确定性 Top-*K* 的研究进展.第 5 节从确定性方法和近似方法讨论不确定性 Top-*K* 查询处理的算法.第 6 节介绍各应用层面的不确定性 Top-*K* 研究技术.第 7 节总结各技术并给出未来的研究方向.

## 2 基础知识

从 20 世纪 80 年代初就已经开始了针对不确定性数据的研究[17,19,20],为了将不确定性引入数据模型,出现了很多用来描述不确定性数据的模型[19,20,22−24].文献[25]中,从模型特点和表达能力比较了各数据模型,提出了下层完备、上层不完备的两层数据模型.而文献[26]则针对关系型、半结构化、数据流、高维数据中主要的不确定性数据模型进行了比较分析.在不确定性 Top-*K* 查询处理的研究中,大部分的研究工作都不是基于完备的不

确定数据模型展开的,主要原因是完备数据模型上的查询难于推理和展开处理.作为基础背景,本文将首先介绍几种主流不确定性数据模型[17,20,25,27]以及被广泛应用的可能世界语义模型[17,18,21].

### 2.1 不确定性数据模型

不确定性数据模型中,第 1 个提出的完备模型是 *c*-table[17],*c*-table 由 *c*-tuples 组成,*c*-tuples 具备以下特点:

(1) 某些属性值自由变量代替;

(2) 每条记录都有一个属性 condition,定义了该记录中自由变量满足的关系范式;

(3) 整个 *c*-table 可能还会有一个全局约束条件.

对全部的变量进行赋值,每次能满足所有约束条件的赋值会形成一个可能表实例.表 3 和表 4 分别是一个 *c*-table 及其一个可能表实例的例子.

**Table 3** An instance of *c*-table model

**表 3** *c*-table 实例

| ID | Attr.1 | Attr.2 | con |
|---|---|---|---|
| 001 | 5 | z | |
| 002 | *x* | 4 | $x\neq z\wedge x\neq 1$ |
| 003 | 7 | *y* | $x=y\vee z=y$ |

**Table 4** A possible state of the *c*-table instance

**表 4** *c*-table 的可能表实例

| ID | Attr.1 | Attr.2 |
|---|---|---|
| 001 | 5 | 3 |
| 002 | 6 | 4 |
| 003 | 7 | 6 |

*c*-table 之所以完备,是允许变量的约束任意.但在实际的应用中,定义变量的任意约束,通常意味着读取和推理的代价很高.事实上,很多不确定性数据处理,包括不确定性 Top-*k* 查询处理,关注的不确定性主要包括以下两个方面:

(1) 属性级不确定性.在一个不确定性数据库 *D* 中,如果有一个或多个属性是不确定性的:

a. 属性值是一个离散值的集合,每个离散值关联一个出现概率;

b. 属性值是一个可能值的连续分布,关联一个概率密度函数,则此数据库含属性级不确定性.

实例化该数据库时,每条记录在其不确定性属性值或分布中抽取一个可能值,形成一个实例表.很多实际应用如传感器、电子标签、GPS 值形成的数据记录含属性级不确定性,文献[27]中描述了针对关系模型进行属性级不确定性扩展的 probability ?-table 模型;

(2) 记录级不确定性.在一个不确定性数据库 *D* 中,假如记录不含不确定性属性,但整个数据库中的每条记录都以一定概率出现,则此数据库含记录级不确定性.更复杂的记录级不确定性还含有一组生成规则,每个生成规则含有一组记录,规定该组记录满足的约束条件.通常,生成规则有两种:

a. 互斥规则,规定该组记录只能有一个出现,不能同时出现;

b. 共存规则,规定该组记录必须同时存在.

probability or-set table[20]是针对记录级不确定性对关系模型的一种重要扩展.

同时含有属性级和记录级不确定性的 probability or-?-set table[22,25]又被称为 *x*-relation[20].但在目前已有的不确定性 Top-*K* 查询技术中,同时涉及两种不确定性的并不常见,经常是仅仅关注其中一方面或只处理其简单情况.例如,属性级不确定性只处理离散值不确定性,又或是记录级不确定性只处理不含生成规则的情况.

### 2.2 可能世界模型

目前研究的主流不确定性数据库为概率数据库,它建立在可能世界模型的基础上,可能世界空间由一系列可能世界实例组成,即 $W=\{w_1,w_2,\ldots,w_n\}$,$P:W\to[0,1]$是其上一个概率分布,且 $\sum_{j=1,\ldots,n} p(w_j)=1, p(w_j)>0$.每一个可能世界实例对应一个确定性数据库,其中,那些非确定性属性是满足约束条件的确定值.可能世界语义是不确定性查询处理技术的出发点和基础.

针对第 2.1 节中涉及的不确定性,可以用图 2 示意其结构.一个不确定性数据库可以分别或同时含有属性级和记录级不确定性;而对于不确定性属性,其值可以离散或连续;对于以一定概率存在的记录之间,可以没有生成规则也可以有生成规则,含有生成规则时,生成规则可以是互斥、共存或其他规则[28].

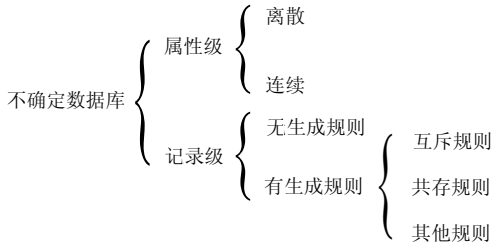



Fig.2 The uncertainty levels in a uncertain database

图 2 不确定性数据库的组成

通过表 5~表 7,我们会看到几个典型的不确定性数据库及其可能世界空间的实例,分别是:

(a) 仅含属性级不确定性,且不确定性属性值离散;

(b) 仅含属性级不确定性,且不确定性属性值连续;

(c) 仅含记录级不确定性,且有互斥生成规则.

表 5 是属性级不确定性数据库的一个实例,不确定性属性值为离散值.表 6 是不确定属性值为连续值的一个实例,其中,$N(15,5)$是期望为 15、方差为 5 的正态分布;$B(1000,0.5)$是二项分布.值得注意的是,属性值连续时,可能世界有可能不可数.

表 7 是对引言中例 1 的抽象,它是一个仅含记录级不确定性的数据库,每条记录有一个附属属性,用来表示该记录存在的概率.在规则表中,每条规则$\tau_i$中的记录相互排斥,即不出现在同一个可能世界中.

**Table 5** An instance of attribute-level uncertainty (discrete value distribution)

**表 5** 属性值离散的不确定性数据库

| 记录 | 属性值 | 可能世界 |
|---|---|---|
| $t_1$ | {(50,0.3),(15,0.7)} | $w_1$={$t_1$=50,$t_2$=40,$t_3$=35} |
| $t_2$ | {(40,0.6),(20,0.4)} | $w_2$={$t_1$=50,$t_2$=20,$t_3$=35} |
| $t_3$ | {(35,1)} | $w_3$={$t_1$=15,$t_2$=40,$t_3$=35} |
| | | $w_4$={$t_1$=15,$t_2$=20,$t_3$=35} |

**Table 6** An instance of attribute-level uncertainty (continuous value distribution)

**表 6** 属性值连续的不确定性数据库

| 记录 | 属性值区间 | 分布函数 |
|---|---|---|
| $t_1$ | [4,10] | $f$=1/6 |
| $t_2$ | [0,200] | $f$=$N$(15,5) |
| $t_3$ | [0,1000] | $f$=$B$(1000,0.5) |

**Table 7** An instance of tuple-level uncertainty with exclusion rules

**表 7** 含互斥规则的记录级不确定性数据库

| 记录 | 值 | 概率 | 互斥规则 | | 可能世界 | |
|---|---|---|---|---|---|---|
| $t_1$ | 50 | 0.4 | $\tau_1$ | {$t_3$,$t_4$} | $w_1$={$t_1$,$t_2$,$t_5$} | $w_7$={$t_2$,$t_5$} |
| $t_2$ | 46 | 1 | $\tau_2$ | {$t_5$,$t_6$} | $w_2$={$t_1$,$t_2$,$t_3$,$t_6$} | $w_8$={$t_2$,$t_3$,$t_6$} |
| $t_3$ | 45 | 0.4 | | | $w_3$={$t_1$,$t_2$,$t_4$,$t_5$} | $w_9$={$t_2$,$t_4$,$t_5$} |
| $t_4$ | 44 | 0.5 | | | $w_4$={$t_1$,$t_2$,$t_6$} | $w_{10}$={$t_2$,$t_6$} |
| $t_5$ | 15 | 0.7 | | | $w_5$={$t_1$,$t_2$,$t_3$,$t_5$} | $w_{11}$={$t_2$,$t_3$,$t_5$} |
| $t_6$ | 10 | 0.3 | | | $w_6$={$t_1$,$t_2$,$t_4$,$t_6$} | $w_{12}$={$t_2$,$t_4$,$t_6$} |

实际应用中的不确定性数据库,是以上 3 种典型数据库中的一种或者组合情况.目前,Top-$K$ 查询处理中涉及最复杂的情况就是既含有属性级不确定性,又含有记录级不确定性和生成规则,且不确性属性值连续.

### 2.3 可能世界语义下的计算问题

可能世界语义下的基本计算问题包括:(1) 可能世界空间规模和可能世界实例规模的计算;(2) 可能世界概率(分布)的计算.

可能世界空间规模即一个可能世界空间中实例的个数.对于第 2.2 节中涉及的属性为离散值的属性级不确定性数据库,假设每条不确定性属性 $t_i$ 有 $s_i$ 个可选属性,则其可能世界空间规模 $|W|=\prod_{i=1}^{n}s_i$ .例如第 2.2 节中的例(a),3 条记录的不确定属性可选值分别为 2,2,1,则可能世界空间规模为 2×2×1=4.属性为连续值时,可能世界规

模为无穷.但无论属性值连续还是离散,每个可能世界实例的规模|*w*|=*n*.

对于含生成规则的记录级不确定性数据库,可能世界规模与生成规则有关.由于组成一个可能世界实例需要在概率为 1 的生成规则中取一个记录,而在概率小于 1 的生成规则中取 0 个或 1 个记录,根据组合数学的计算方法可知,可能世界空间规模 $|W|=\prod_{p(\tau_j)=1}|\tau_j|\prod_{p(\tau_j)<1}(|\tau_j|+1)$;如果概率为 1 的生成规则有 *m* 条,则可能世界实例的规模|*w*|∈[*m*,*n*].如第 2.2 节中的例(c),生成规则 $\tau_1$ 概率小于 1,$\tau_2$ 概率等于 1,则可能世界空间规模为 2×1×(2+1)×2=12.

每个可能世界实例都有一定的存在概率,而可能世界空间中所有可能世界实例的概率形成了可能世界概率分布.对于属性级不确定性数据库,每个可能世界实例的概率为所选属性值概率积,即 $p(w_k)=\prod_{t_i\in w_k}p(t_i)$.例如第 2.2 节中的例(a),$p(w_1)$=0.3×0.6×1=0.18.对于记录级不确定性数据库,每个可能世界实例的概率也与生成规则有关,计算公式为 $p(w_k)=\prod_{t_i\in w_k}p(t_i)\prod_{\tau_j\cap w_k=\Phi}(1-p(\tau_j))$,它是所有出现记录的概率积与未出现任何记录的生成规则的不出现概率的乘积.例如第 2.2 节中的例(c), $p(w_1)$=0.4×1×(1–0.9)×0.7=0.028.不论是属性级不确定性数据库还是记录级不确定性数据库,及有无生成规则,可能世界空间的概率和为 1,即 $\sum_{w\in W}p(w)=1$.

## 3 不确定性 Top-*K* 查询的语义研究

目前,不确定性数据库的研究虽然涉及关系型数据、半结构化数据、本体数据[20,26,29]等,由于关系数据模型的应用最为广泛,不确定性数据库的研究焦点仍然在关系型数据上.因此,大部分的不确定性 Top-*k* 研究也都建立在不确定性关系型数据库上.确定性关系数据库的 Top-*K* 查询的语义非常清晰,就是基于某个分值函数计算每个记录的分值,返回数据库中具有最大分值的前 *k* 个记录.分值函数的计算是基于记录的属性值的,不确定数据模型中,属性值可能有多重选择,此时分值该如何计算?如何比较?不确定数据模型中,记录可能有一定的存在概率,此时分值该如何计算?如何比较?这些问题使究竟该返回哪些记录这个问题变得不再清晰.

针对不确定性 Top-*K* 查询,学者们从不同侧面和不同应用的需要给出了不同的查询语义.最优先考虑的是查询语义的合理性,即该语义和返回结果能合理解释并满足实际的查询需求.因此,学者们根据实际的查询需求定义了各式各样的不确定性 Top-*K* 查询语义.这些不确定性 Top-*K* 查询语义看似都满足了一定的应用需求,但实际上无论从形式化定义满足的数学性质还是从返回结果上,都存在巨大差异,因此出现了对不确定性 Top-*K* 查询语义满足的数学性质的研究.

第 3.1 节首先介绍目前出现的比较有影响的不确定性 Top-*K* 查询语义[8−13,15,30,31],并对其进行合理性分析.第 3.2 节将对不确定性 Top-*K* 查询语义满足的数学性质进行介绍和分析[10,15,30].

### 3.1 不确定性Top-*K*查询语义的合理性

目前,不确定性 Top-*K* 查询语义的研究有多种定义,比较有影响的包括 U-Top*K*[9,32],U-*k*Ranks[9,16],PT-*k*[11−13],Global-Top*K*[15],Expected Rank[10,30],E-Score Rank[10,30],*c*-typical-Top*K*[31]等.它们分别适应不同的应用场景,下面我们将给出各语义的形式化定义及简要解释.

**定义 1(U-Top*K*)**. 设 *D* 是不确定性数据库,其可能世界空间为 $W=\{w_1,w_2,\ldots,w_n\}$,$T=\{T_1,T_2,\ldots,T_m\}$,$T_i$ 是长度为 *k* 的记录向量,如果 $T_i$ 中 *k* 个记录是根据某分值函数 *f* 计算的分值排序序列且对应某些可能世界的前 *k* 个记录,则基于 *f* 的 U-Top*K* 查询返回 $T^*\in T$ 满足 $T^*=\arg\max_{T_i\in T}\left(\sum_{w\in w(T_i)}p(w)\right)$.即将 $T_i$ 对应的那些可能世界概率求和,具有最大概率和的 $T_i$ 是 U-Top*K* 查询的返回结果.

**定义 2(U-*k*Ranks)**. 设 *D* 是不确定性数据库,其可能世界空间为 $W=\{w_1,w_2,\ldots,w_n\}$.对于排序位置 *i*=1,…,*k*,分别对应一组记录 $t_i^1,t_i^2,\ldots,t_i^m$,它们是在某可能世界中,根据分值函数 *f* 排序后出现在位置 *i* 的记录.基于 *f* 的 U-*k*Ranks,返回 *k* 个记录$\{t_i^*\mid i=1,\ldots,k\}$,满足 $t_i^*=\arg\max_{t_i^j}\left(\sum_{w\in w(t_i^j)}p(w)\right)$.也就是将出现在位置 *i* 的可能世界概率求和,取概率最大者作为返回结果的第 *i* 个记录.

**定义 3(PT-*K*)**. 设 $D$ 是不确定性数据库,其可能世界空间为 $W=\{w_1,w_2,\ldots,w_n\}$,$Q=\{q_1,q_2,\ldots,q_m\}$,$q_i$ 是记录集合,分别对应每个可能世界按某分值函数 $f$ 排序的前 $k$ 个记录.在 $Q$ 的基础上计算每个记录位于 Top-$k$ 的概率 $p(t)=\sum_{t\in q_i} p(w_i)$,设定概率阈值 $p(0<p\leqslant 1)$,PT-$K$ 查询返回 $T=\{t_i|p(t_i)\geqslant p\}$.即对每个记录出现在 Top-$K$ 位置的可能世界概率求和,取大于等于 $p$ 的记录作为 PT-$K$ 的结果返回(不限于 $k$ 个).

**定义 4(global-Top*K*)**. 设 $D$ 是不确定性数据库,其可能世界空间为 $W=\{w_1,w_2,\ldots,w_n\}$,$Q=\{q_1,q_2,\ldots,q_m\}$,$q_i$ 是记录集合,分别对应每个可能世界按某分值函数 $f$ 排序的前 $k$ 个记录.在 $Q$ 的基础上计算每个记录 $t$ 位于 Top-$k$ 的概率 $p(t)=\sum_{t\in q_i} p(w_i)$,称为该记录的 global-Top$K$ 概率,global-Top$K$ 查询返回具有最大 global-Top$K$ 概率的 $k$ 个记录.

**定义 5(expected rank)**. 设 $D$ 是不确定性数据库,其可能世界空间为 $W=\{w_1,w_2,\ldots,w_n\}$.假设根据分值函数 $f$,可能世界 $w_i$ 中在记录 $t$ 前的记录数记为 $rank_{w_i}(t)$,则 $t$ 的 Expected Rank 分值为

$$r(t)=\sum_{w_i\in W,t\in w_i} p(w)\cdot rank_{w_i}(t).$$

Expected Rank 按 Expected Rank 分值排序记录($t$ 在未出现它的可能世界中 $rank_{w_i}(t)=|w_i|$).

**定义 6(E-score rank)**. 设 $D$ 是不确定性数据库,其可能世界空间为 $W=\{w_1,w_2,\ldots,w_n\}$.假设根据分值函数 $f$,在可能世界 $w_i$ 中计算 $t$ 的分值为 $score_{w_i}(t)$,则 $t$ 的 E-score 分值定义为 $e(t)=\sum_{w_i\in W,t\in w_i} p(w)\cdot score_{w_i}(t)$,E-Score Rank 按 E-Score 分值排序记录.

**定义 7($c$-typical-Top$K$)**. 设 $D$ 是不确定性数据库,其可能世界空间为 $W=\{w_1,w_2,\ldots,w_n\}$,$T=\{T_1,T_2,\ldots,T_m\}$,$T_i$ 是长度为 $k$ 的记录向量.如果 $T_i$ 中 $k$ 个记录是根据某分值函数 $f$ 计算的分值排序序列且对应某些可能世界的前 $k$ 个记录,对这些可能世界概率求和 $p(T_i)=\sum_{w\in w(T_i)} p(w)$,并对 $T_i$ 中所有记录求总 $f$ 分值 $s(T_i)=\sum_{t\in T_i} f(t)$ 形成分布 $S$,在 $S$ 中寻找最典型的 $c$ 个分值 $\{s_1,\ldots,s_c\}=\arg\min_{\{s_1,\ldots,s_c\}} E[\min_{\{s_1,\ldots,s_c\}}|S-s_i|]$,$c$-typical-Top$K$ 返回典型分值中具有最大概率的记录向量 $T_i=\arg\max_{s(T_i)=s_i} p(T_i)$,$1\leqslant i\leqslant c$ (分值典型性).

从以上定义中可以看出,不确定性 Top-$K$ 查询中涉及的一个很重要的计算问题是可能世界概率,假设例 1 中属性值一列为根据某分值函数 $f$ 计算的得分(此处分值直接为温度值),我们在表 8 中给出每个可能世界的概率,并根据以上定义求各查询 Top-2 结果,见表 9.

**Table 8** The possible worlds for the uncertain database in example 1

**表 8** 例 1 可能世界概率表

| 可能世界 | 概率 | 可能世界 | 概率 |
|---|---|---|---|
| $w_1=\{t_1,t_2,t_5\}$ | $p_1$=0.028 | $w_7=\{t_2,t_5\}$ | $p_7$=0.042 |
| $w_2=\{t_1,t_2,t_3,t_6\}$ | $p_2$=0.048 | $w_8=\{t_2,t_3,t_6\}$ | $p_8$=0.072 |
| $w_3=\{t_1,t_2,t_4,t_5\}$ | $p_3$=0.14 | $w_9=\{t_2,t_4,t_5\}$ | $p_9$=0.21 |
| $w_4=\{t_1,t_2,t_6\}$ | $p_4$=0.012 | $w_{10}=\{t_2,t_6\}$ | $p_{10}$=0.018 |
| $w_5=\{t_1,t_2,t_3,t_5\}$ | $p_5$=0.112 | $w_{11}=\{t_2,t_3,t_5\}$ | $p_{11}$=0.168 |
| $w_6=\{t_1,t_2,t_4,t_6\}$ | $p_6$=0.06 | $w_{12}=\{t_2,t_4,t_6\}$ | $p_{12}$=0.09 |

**Table 9** The result sets of Top-2 by different semantics on the uncertain database in example 1

**表 9** 例 1 各查询 Top-2 结果表

| 查询类型 | Top-2 | Prob. (score) | 可能世界空间 |
|---|---|---|---|
| U-Top2 | $t_1t_2$ | 0.4 | $\{w_{1\sim6}\}$ |
| U-2Ranks | $t_2,t_2$ | 0.6,0.4 | $\{w_{7\sim12}\}$, $\{w_{1\sim6}\}$ |
| PT-2 ($p$=0.3) | $t_1,t_2,t_4$ | 0.4,1,0.3 | $\{w_{1\sim6}\},\{w_{1\sim12}\},\{w_9,w_{12}\}$ |
| Global-Top2 | $t_2,t_1$ | 1,0.4 | $\{w_{1\sim12}\},\{w_{1\sim6}\}$ |
| Expected rank-2 | $t_2,t_1$ | 0.4,0.63 | — |
| E-score-2 | $t_3,t_2$ | 46,22 | — |
| 1-typical-Top2 | $t_2t_4$ | 0.3 | — |

从表 9 的查询结果看,同样是 Top-2 查询,语义不同则查询结果迥然,这是因为每个查询语义的定义都是情景相关的.

下面从各查询语义在可能世界中的兼容性、对 *f* 分值的依赖性及有序性这 3 个方面进行对比,见表 10.

我们发现:(1) 大部分不确定性 Top-*K* 查询都考虑了查询结果在可能世界中的兼容性问题.这个结论很显然也很合理,用户总是希望得到的结果集或结果序列是能够在同一可能世界共存的.而 U-*k*Ranks 更多地关注排在某位置的高概率记录,更多地用于单一记录查询而不是一组记录查询;(2) 大部分不确定性 Top-*K* 查询都不依赖排序函数 *f* 的分值.这说明不确定性 Top-*K* 查询从根本上关注的仍然是记录间的相对位置,除非分值的大小程度与记录重要性直接相关;(3) 大部分不确定性 Top-*K* 查询结果都是有序的.PT-*K* 虽然无序且结果集也未必是 *k* 个元素,但是 Global-Top*K* 在它的基础上以概率排序,仍然可以得到有序结果.U-*k*Ranks 的无序性更进一步说明它更关注单一记录查询,实际上在第 5 节中我们可以看到,经过排序整合 U-*k*Ranks 也可以有序.

**Table 10** Rank criterion comparison of various Top-k semantics on uncertain database

**表 10** 不确定性 Top-*k* 查询语义对比表

| 查询方式 | 兼容性 | 分值依赖性 | 有序性 |
|---|---|---|---|
| U-Top*k* | Y | N | Y |
| U-*k*Ranks | N | N | Y |
| PT-*k* (*p*=0.3) | Y | N | N |
| Global-Top*k* | Y | N | Y |
| Expected Rank | Y | N | Y |
| E-score | Y | Y | Y |
| 1-typical-Top*k* | Y | Y | Y |

总之我们可以看到,不确定性 Top-*K* 查询总体上仍然关注记录的相对位置,并以有序的方式呈现查询结果.不同语义的差别关键点在概率和排序分值的平衡方式上.

### 3.2 不确定性Top-*K*查询语义的数学性质

从第 3.1 节可以看到,尽管大部分不确定性 Top-*K* 查询语义遵循了传统确定性 Top-*K* 的合理解释,但它们在形式化定义满足的数学性质上还存在很大差异,因此,Zhang 和 Cormode 等人提出了一系列不确定性 Top-*K* 查询应该满足的数学性质[10,15,30,33],并分析证明了部分不确定性 Top-*K* 查询语义分别能满足的性质[33].

**性质 1(exact *K*)**. 设 $R_k$ 是某不确定性 Top-*K* 查询的返回结果集,$|D|\geqslant k$ 时,则$|R_k|=k$.

**性质 2(faithfulness)**. $t_1,t_2\in D$,如果 $t_1$ 的分值和概率都大于 $t_2$ 且 $t_2\in R_k$,则 $t_1\in R_k$.

**性质 3(containment)**. 对于任何正整数 $k$,$R_k\subset R_{k+1}$.

**性质 4(unique ranking)**. 设 $r(i)$是在结果集中位于 $i$ 位置的记录 ID,对于任何排序位置 $i,j,i\neq j$,则 $r(i)\neq r(j)$.

**性质 5(value invariance)**. 排序分值 $v_1\leqslant v_2\leqslant\ldots\leqslant v_k$,对应不确定性 Top-*K* 结果序列,在不改变分值相对位置的前提下,排序分值的大小不影响查询结果.

**性质 6(stability)**. $t_i\in R_k$ 时,增大 $t_i$ 的排序分值或概率不会使 $t_i\notin R_k$;$t_i\notin R_k$ 时,减小 $t_i$ 的排序分值或概率不会使 $t_i\in R_k$.

**Table 11** Property comparison of various Top-*K* semantics on uncertain database

**表 11** 不确定性 Top-*K* 查询语义满足的性质对比表

| Semantics | Exact *k* | Faithfulness | Containment | Uni.Ranking | Val.invariance | Stability |
|---|---|---|---|---|---|---|
| U-Top*k* | N | Weak | N | Y | Y | Y |
| U-*k*Ranks | Y | N | Y | N | Y | N |
| PT-*k* (*p*=0.3) | N | Weak | Weak | Y | Y | Y |
| Global-Top*k* | Y | Y | N | Y | Y | Y |
| Exp. Rank | Y | Weak | Y | Y | Y | Y |
| E-score | Y | Y | Y | Y | N | Y |
| 1-typical-Top*k* | Y | N | N | Y | N | N |

不确定性 Top-*K* 查询究竟应该满足哪些基本性质,目前来说是一个相对开放的问题,而不确定性 Top-*K* 查询语义的应用相关性更增加了该问题的难度,例如某些应用更关心排序分值的典型性[31],而某些应用更关心记录的位置概率等[9].总体上看,研究不确定性 Top-*K* 查询语义满足的数学性质作用主要有:

1. 进一步规范不确定性 Top-*K* 语义的定义.在此项研究开展前,各学者基本上都是从各自面临的应用场

景出发定义需要的语义,很少考虑该语义满足那些性质、需要满足那些性质.语义性质的研究,可以指导更为科学、规范、合理的不确定性 Top-*K* 语义定义;

2. 由于不确定性 Top-*K* 查询语义满足的性质反映了结果集的数学性质,对指导查询处理意义重大.例如在文献[34]中,由于 Expected Rank 具有包含性和稳定性,在处理滑动窗口的 Top-*K* 解集时,可以不保存整个窗口的记录,而仅仅通过维护含 Top-*K* 解集的最小子集——紧致集,实现连续的 Top-*K* 回答.因此,结合语义满足的数学性质,更有可能设计出高效的查询算法.

## 4 不确定性 Top-*K* 查询的排序标准研究

从第 3 节中可以看到,研究者在不确定性数据库上定义了许多新的 Top-*K* 查询语义,例如 U-Top*K*,返回具有最大概率的 Top-*K* 记录向量;U-*k*Ranks,返回每个位置最大概率出现的那个记录;PT-*K* 返回以概率 *p* 以上出现在 Top-*K* 的记录集,它们基本上都有特定的应用场景,语义的提出具有合理性.但它们返回的结果集又有什么特点呢?Li 采用规范 Kendall 距离[14]对比同一数据集上 5 种不确定性 Top-*K* 返回的结果,发现各语义返回结果序列差异显著,有些结果甚至完全相反.最根本的原因是采用的排序标准不同,而在不确定性 Top-*K* 查询中,影响排序的因素主要有两个:排序函数分值和记录的概率.

不确定性 Top-*K* 查询处理中,并没有研究者专门对排序标准进行研究,多是在讨论某查询处理方法时进行一些扩展和思考.本节从排序分值与概率的平衡技术、连续型分值排序技术以及统一化排序方法 3 个方面来探讨不确定性 Top-*K* 的排序标准.

### 4.1 分值与概率的平衡技术

按照对分值排序和概率的处理先后顺序不同,可以将分值与概率平衡方式划分为 3 类:第 1 类是先排序再求取概率,第 2 类是先求取概率再排序,第 3 类是同时综合考虑排序和概率.

U-Top*K* 在平衡分值与概率时采取的就是第 1 类方式.它首先将每个可能世界空间记录按排序分值排序,截取每个可能世界空间的前 *k* 个记录形成一个 *k* 长度排序序列,这些 *k* 长度排序序列与其所在的可能世界一样,是有一定存在概率的,找到拥有最大存在概率的 *k* 长度排序序列就找到了 U-Top*K* 的解.事实上,这种平衡的方式有很多弊端.例如文献[31]提出,U-Top*K* 求得的 Top-*K* 序列往往概率很小.这点显而易见,所有可能世界概率和为 1,*k* 越大,Top-*K* 序列可能情况就越多,每个可能的 Top-*K* 序列概率就越小.只按这种微小的概率差别来区分优劣,很多时候并不客观.因此,文献[31]中提出,在求得所有 Top-*K* 序列之后,概率并不应该成为唯一的衡量标准,分值应该重新纳入考虑范围,比如考虑 Top-*K* 向量总分值的典型性程度.

第 2 类方式是先考虑记录在某位置的位置概率,再按概率在各位置最优分配的方式形成全排序(Top-*K* 排序).文献[32]在文献[9]中位置概率 U-*i*Ranks 的基础上,提出依据 Top-*K* 位置上记录的位置概率总和最大的优化目标,用二分图匹配的方式找出最优全序.这种平衡方式也并不是没有缺点,比如:得到的序列有可能在任何可能世界都不存在;概率总和最大化的目标也许并不适用所有场景等.

第 3 类方式是同时考虑排序和概率,比较典型的是 Expected Rank 和 global Top*K* 的处理方式.Expected Rank 是按记录在所有可能世界的期望排位来排序;而 global-Top*K* 则将分值序列和记录在 Top-*K* 的概率排序序列看成是两个待合并的排序列表,用传统合并排序的方式形成全序.

不管是在形式化语义中还是在查询处理中,分值与概率的平衡都是不确定性 Top-*K* 查询的焦点问题.因此,合理适用的分值概率平衡方式在不确定性 Top-*K* 查询中至关重要.

### 4.2 连续分值排序技术

属性级不确定性数据库中可以存在连续型属性,当此属性是分值函数的参考依据时,就出现了排序分值是连续分布的情况.连续分值的存在直接挑战传统排序标准,因为作为传统排序标准的分值函数总能根据记录的唯一得分形成记录的全序,而连续分值形成的却是记录偏序[32].

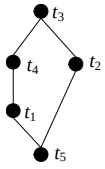
目前,解决偏序问题基本上都是采用 Soliman 在文献[32]中提到的概率偏序模型.如图 3 所示.

概率偏序模型将连续分值的定义域按端点排序,每个记录 $t$ 有 $low$,$up$ 两端点.如果某记录 $t_i$ 的下限 $low_i$ 大于另一记录 $t_j$ 的上限 $up_j$,则 $t_i$ 对 $t_j$ 全支配;如果两记录定义域相交,则存在概率支配关系,可以通过联合密度函数在相交区域的二重积分 $p(t_i>t_j)=\int_{low_i}^{up_i}\int_{low_j}^{y} f_i(x)f_j(y)\mathrm{d}y\mathrm{d}x$, $p(t_j>t_i)=1-p(t_i>t_j)$ 来计算两记录之间的相互支配概率.在文献[35]中,通过使用生成函数将计算位置概率的问题转化为一维积分.虽然如此,连续分值条件下,概率的计算始终是一个积分问题.而积分本身就是一个很复杂的运算,Soliman 采用蒙特卡洛积分法[32]计算分值为均匀分布时的支配概率,但没有给出对于任意密度函数的分值分布该如何计算.

**Table 12** An database with uncertain continuous scores

**表 12** 连续分值实例

| 记录 | 分值区间 | 分值密度 |
|---|---|---|
| $t_1$ | [5,7] | $f_1$ |
| $t_2$ | [6,8] | $f_2$ |
| $t_3$ | [9,9] | $f_3$ |
| $t_4$ | [7,8] | $f_4$ |
| $t_5$ | [1,1] | $f_5$ |




$$\begin{cases} p(t_2>t_4)=\int_6^8\int_7^y f_2(x)f_4(y)\mathrm{d}y\mathrm{d}x \\ p(t_1>t_2)=\int_5^7\int_6^y f_1(x)f_2(y)\mathrm{d}y\mathrm{d}x \end{cases}$$

Fig.3 Probabilistic partial order model

图 3 概率偏序模型

为了能处理任意密度函数的分值分布,有 3 种方法值得考虑.最简单直接的方法是将连续分值离散化.如图 4 所示,将每个记录的分值区域划分为等距离的区间段,每个区间段看成均匀分布.该区间段概率按照均匀分布积分密度函数获得,而分值则取该区间段中点值.这样,任意密度分布的分值都可以转换成以一定概率存在的离散属性值,再以离散属性级不确定性的方式进行后续处理.可以预见,离散的粒度直接决定处理的复杂度和精确度,通常很难平衡.第 2 种方法是随机近似积分的方法.可类似于计算积分的 MC(蒙特卡洛随机近似)方法,在相交区间随机生成各记录的一个分值,将这些随机分值排序,形成一个样本.多次取样后,根据样本中各记录在每个位置的出现频度估算位置概率.因为文献[35]中对记录区间整体都随机取样,所以只能处理均匀分布的分值.但只要我们对每记录生成分值时采用文献[36]中按概率相等的方法划分区间,如图 5 所示,划分为 $k$ 个区间段,每个区间段积分概率为 $1/k$,随机在各区间段生成样本分值,可保证高概率区间段取样频率相对较高,同样可以处理任意分布的分值.第 3 种方法是在文献[35]中提出的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(或简单函数),再按线性函数求积分即可,如图 6 所示,是用三次样条逼近将高斯分布函数近似为 4 段多项式函数.

这 3 种方法考虑的都是在分值的分布函数已知的情况下去求解,其实,现实中数据采集得到的多是分布未知的值.研究分布未知,只能得到大量离散值情况下的不确定性 Top-$K$ 查询也许更具有现实意义.

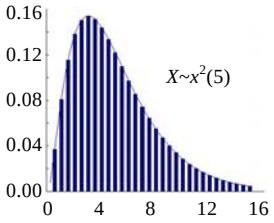



Fig.4 Discretization of a distribution into uniform intervals

图 4 分值等距离散

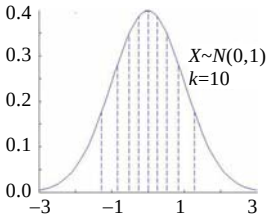



Fig.5 Discretization of a distribution into equidepth intervals

图 5 等深离散取样

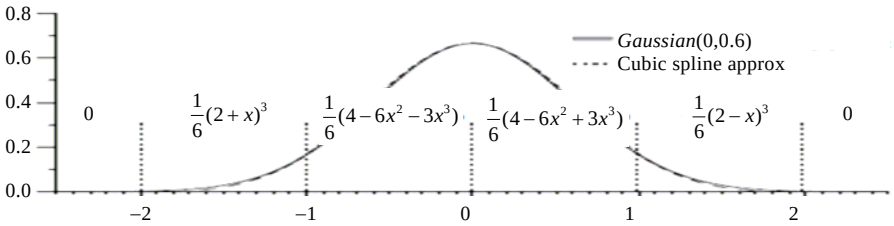



Fig.6 Approximating a distribution using a Cubic Spline

图 6 样条逼近

### 4.3 统一化的排序方法

从各不确定性 Top-*K* 查询语义定义过程来看,具体应用的需要在概率和分值平衡方式设计中起了主导作用,这就导致各不确定性 Top-*K* 查询语义下返回结果差异巨大.在例 1 中,假如应用只关心温度,不关心测温器的可靠性,进行排序的结果为$\{t_1,t_2,t_3,t_5\}$;假如应用只关心测温器的可靠性,进行排序的结果为$\{t_2,t_5,t_4,t_1\}$.这只是两种极端情况,事实上,各不确定性 Top-*K* 语义在平衡概率和分值时,根据不同应用需要,分别有各自权衡的标准.这些权衡标准虽然不同,但都满足了部分应用的排序要求.

Normalized Kendall Distance 是众多比较 Top-*K* 序列的准则中有相当多优势的一种[37],它通过比较两序列中记录对的逆序数和跨度差别衡量两序列的差异.例如,序列$\{t_1,t_2,t_3,t_5\}$和$\{t_2,t_5,t_4,t_1\}$中,同时出现的 3 个记录对$\{t_1,t_2\},\{t_2,t_5\},\{t_1,t_5\}$有 1 对逆序,而跨度差别为 2,−1,−1.逆序对和跨度差别数越小,两序列越接近.Li 在文献[14]中,通过求解同一不确定性数据库数据上的各不确定性 Top-*K*,使用 Normalized Kendall Distance 度量结果序列的距离,发现它们不仅不具有相似性,差别还相当大.这样的比较结果并不意外,主要是因为平衡分值和概率的方式很多,而究竟如何平衡则取决于应用所看重的属性.这样看来,现存的不确定性 Top-*K* 语义也许没有任何一个算得上是完全“正确”的定义[14].Li 等人就统一化排序做出尝试[14,35],提出既然排序结果与应用偏好直接相关,可以设计一个带参数的排序函数,参数的选取直接关系分值和概率的平衡,而参数的设置可以通过应用需要进行学习.这样,不确定性数据库上进行排序实际上变成了一个多准则优化的问题,不同参数的排序函数生成不同的排序结果.

## 5 不确定性 Top-*K* 查询的算法研究

从第 3 节各不确定性 Top-*K* 的语义可以看出,得到查询结果的原始方法是展开不确定性数据库的整个可能世界空间,根据各不确定性 Top-*K* 语义在每个可能世界求解,最后整合指定答案涉及的可能世界的概率.展开整个可能世界空间是 NP 问题,极其耗时,而生成规则更增加了展开的难度.其实,根据某不确定性 Top-*K* 查询的特点,有时候求解并不需要展开整个可能世界空间.对不确定性 Top-*K* 查询处理的确定性算法技术的研究焦点,就

是如何利用查询语义的特点避免展开整个可能世界空间[9−12,15,30−32],从而提高查询效率.这方面的内容在第 5.1 节中重点讨论.第 5.2 节讨论不确定性算法技术.通常,在用户不需要完全精确答案的前提下,为了进一步提高查询效率,很多学者都采用近似的方法来处理不确定性 Top-*K* 查询,以牺牲少量精度的方式换取更高的处理效率[11,12,14,30,32,35,36].第 5.3 节中,将对各算法进行复杂性分析.生成规则的存在增加了问题的复杂程度,不失一般性,不确定性 Top-*K* 查询对生成规则进行处理的技术在第 5.4 节中加以介绍.

### 5.1 确定性算法技术

在文献[9]中,Soliman 证明了按分值排序读取记录可以使 U-Top*K* 和 U-*k*Ranks 读取最少记录完成查询.事实上,在不确定性 Top-*K* 查询处理的算法中,记录的读取顺序几乎都是采用按分值排序.不确定 Top-*K* 查询最基本的处理方法是将之看成一个状态空间搜索问题,研究转化为如何实例化最少的状态.在文献[9]中的 U-Top*K* 和 U-*k*Ranks 查询、文献[31]中 *c*-typical-Top*K* 查询以及文献[32]中分值连续时不确定性 Top-*K* 查询中,都曾采用此思想处理.当所求 Top-*K* 具有最优子结构性质,还可以采用动态规划的方法.在文献[9]中求解的 U-*k*Ranks、文献[31]中求解的 *c*-typical-Top*K*、文献[15]求解的 global Top*K* 都有此性质.根据记录独立时概率运算的基本规则,在文献[11−13,38]中处理 PT-*k*,U-*k*Ranks 以及 U-Top*k* 采用了泊松二项递推的方法进行求解.基于生成函数的方法首先将记录关系模型转化为规则树,每个树节点对应一个生成函数,根据生成函数的系数来辅助求解 Top-*K*[14].以下介绍 4 种方法的思想.

#### 5.1.1 状态空间扩展的方法

状态空间扩展设定每个记录 $t$ 有两个状态 $t$ 和$\neg t$,概率分别是 $p(t)$和 $1-p(t)$.设 $s_l$ 是长度为 $l$ 的状态,如果已经扫描 $m$ 个记录,状态 $s_l$ 的概率记为 $p(s_l)=\prod_{t_i\in s_l,i\leqslant m}p(t_i)\prod_{t_j\in s_l,j\leqslant m}(1-p(t_j))$,那么对第 $m$+1 个记录 $t_{m+1}$ 分别有状态 $t_{m+1}$和$\neg t_{m+1}$加在 $s_l$ 后,形成长度为($l$+1)的新状态 $s_lt_{m+1}$和长度为 $l$ 的新状态 $s_l\neg t_{m+1}$,其概率分别为 $p(s_l)p(t_{m+1})$和 $p(s_l)p(1-(t_{m+1}))$.由于记录的概率都小于 1,因此通过添加新记录的方式得到的新状态,概率小于等于扩展前概率.由于概率递减性质的保证,在求 Top-*K* 时可以方便地使用队列或者设置概率上限,有效提高查询效率.如果配合有效的剪枝策略则会加速算法,例如文献[9]中求取 U-Top*K* 时可以将状态按长度分成等价类,每个等价类取概率最大者扩展.文献[32]中分值是连续的,状态扩展时是根据分值定义域端点的支配关系决定每一步可扩展的状态数,概率计算支配记录的联合概率.

#### 5.1.2 动态规划的方法

动态规划的目标是发现求解 Top-*K* 问题的最优子结构性质.U-*k*Ranks 查询在位置 $i$ 处有最高概率的记录,为了得到 $t_m$ 记录在 $i$ 位置的概率 $p[t_m,i]$,只需要知道 $t_{m-1}$ 记录在 $i$−1 位置的概率 $p[t_{m-1},i-1]$和 $t_{m-2}$ 记录在 $i$−1 位置的概率 $p[t_{m-2},i-1]$,$p[t_m,i]=p(t_m)\times(p[t_{m-1},i-1]+(1-p(t_{m-1})\times p[t_{m-2},i-1]))$.又有所有记录在位置 $i$ 的概率和为 1,因此当 $t_m$ 在位置 $i$ 的概率比余下记录在位置 $i$ 的概率大时,即可停止.类似的,global-Top*K* 需要计算每个记录 $t_i$ 在位于 Top-*K* 内的概率 $p(k,t_i)$,只需要知道 $t_{i-1}$ 位于 Top-*K* 内的概率 $p(k,t_{i-1})$以及 $t_{i-1}$ 位于 Top-($K$−1)的概率 $p(k,t_{i-1})$,通过关系式 $p(k,t_i)=p(t_i)\times\left(p(k,t_{i-1})\times\frac{1-p(t_{i-1})}{p(t_{i-1})}+p(k-1,t_{i-1})\right)$求解.因为 global-Top*K* 概率计算函数具有单调性,可以设定阈值进行剪枝加速算法.*C*-typical-Top*K* 查询需要先生成 Top-*K* 向量的分值分布,在 Top-*K* 概率-分值分布图上寻找 $c$ 个最典型 Top-*K* 向量,在生成 Top-*K* 概率分值分布时,从记录 $t_i$ 开始的 Top-$j$ 的分值分布 $D_{i,j}$ 是由 $D_{i+1,j}$ 和 $D_{i+1,j-1}t_i$ 组合而成.最后 $D_{1,k}$ 为所求概率-分值分布.

#### 5.1.3 生成函数的方法

生成函数的方法求解目标是每个记录的位置概率.在按分值排序的记录序列中,每个记录不存在的概率和存在的概率分别作为 $x^0$ 和 $x^1$ 的系数,则每个记录可以表示为一元一次函数$(1-p(t))x^0+p(t)x^1$.在记录级不确定性数据库不存在生成规则时,将排在 $t_i$ 前的所有记录$\{t|t\in T_{i-1}\}$的表示函数相乘再乘以 $p(t_i)x$,形成记录 $t_i$ 的生成函数 $F^i(x)=\left(\prod_{t\in T_{i-1}}(1-p(t))x^0+p(t)x^1\right)\times(p(t)x^1)=\sum_{j\geqslant 0}c_jx^j$,$t_i$ 在位置 $j$ 的概率刚好是 $x^j$ 的系数 $c_j$.这样,可以将问题转化为多项式系数求解问题.

当存在生成规则时,将记录关系转化为规则树,规则内节点的生成函数设定相应求取规则,通过求解节点生成函数中的变量系数,同样可以得到记录的位置概率.这种方法的好处是在生成规则扩展情况下,计算方法可扩展;而生成函数间的递推关系可以使求解效率得到有效提高.

5.1.4 泊松二项递推的方法

泊松二项分布是指每次实验有两种对立结果:$n$ 次实验相互独立,每次实验发生概率是一个常数;而 $n$ 相对较大,某事件在 $n$ 次实验中发生 $x$ 次的概率.当不确定性数据库只存在记录级不确定性且没有生成规则时,可以将每条记录的出现与否看成实验的两种对立结果:$n$ 次实验代表数据库的 $n$ 条记录,数据规模越大,实验次数 $n$ 越多.PT-$k$ 和 global-Top$K$ 都需要求解每条记录出现在 Top-$K$ 的概率.如果将记录按分值排序,记录 $t$ 出现在 Top-$K$ 中的概率可以理解为事件:在排序序列中,排位在 $t$ 前的那些记录同时出现小于等于 $k$–1 个记录的概率,因此可以使用泊松二项分布的递推方法.

假设 $p^k(t_i)$表示记录 $t_i$ 在 Top-$K$ 的概率,$p(t_i,j)$表示记录 $t_i$ 在位置 $j$ 的概率, $p(s_{t_i}, j)$ 表示在 $t_i$ 前同时出现 $j$ 个记录的概率,则记录 $t_i$ 在 Top-$K$ 内的概率可以表示为记录 $t_i$ 在位置 1,…,$k$ 的概率和,即排序列表中出现在 $t_i$ 前同时出现 1,…,($k$–1)个的概率.因存在递推关系 $p(s_{t_i}, j) = p(s_{t_{i-1}}, j-1)p(t_i) + p(s_{t_{i-1}}, j)(1-p(t_i))$ ,因此可以方便地使用动态编程方法来实现.

因为记录级不确定性数据库大致满足泊松二项分布的条件,在很多 Top-$K$ 处理中都体现了泊松二项递推的思想[11−14,35,38].当存在生成规则时,只要对生成规则相应处理,就可以使问题转化成简单情况.

### 5.2 近似算法技术

从不确定性 Top-$K$ 处理的精确算法中发现,大多关键求解概率具有递减性质.再加上其他的一些限制条件,在求解一些特定不确定性 Top-$K$ 时,例如 expected Rank Top-$K$[10,30]和 global Top$K$[15,33],本来需要扫描所有记录,但根据一些已知条件会发现某位置后的记录在解集中的可能性微乎其微,因此可以利用已知结果维护一个逐渐逼近真实值的阈值,使得求解为精确度非常高的近似解集.这样的方法可以极大地缩短求解时间,有时候甚至是数量级的.近似求解不仅可以通过逼近阈值来实现,还可以通过随机取样的方法来实现.在取样条件得当和样本数足量的前提下,很多时候可以更有效率地得到很高精度的解集.随机近似在 PT-$K$[11,12]的求解中、在分值连续情况下[32]各不确定性 Top-$K$ 求解中以及 $c$-typical-Top$K$[31]分值合并时均有使用.在第 5.1 节中看到,泊松二项递推在精确求解时非常有效,如果数据量非常大,即使记录概率不等,通过求取期望,求解目标也可以近似看成泊松二项分布,而采用泊松近似得到解集.下面分别介绍 3 种近似方法.

5.2.1 阈值逼近

在 Expected Rank 求解中,需要求解每个记录在所有可能世界的期望位置 $r(t) = \sum_{w_i \in W, t \in w_i} p(w) \cdot rank_{w_i}(t)$ ,再求取其中最好的 $k$ 个.如果记录按分值的期望顺序排列,每个记录的排位在各可能世界中变动范围是有限的.通过这个事实,按分值期望顺序扫描记录,不断更新已扫描记录的期望排位上限 $r^+(t_i)$和重新计算尾部记录排位下限 $r^-$,在马尔可夫不等式性质的保证下,扫描记录越多,$r^+(t_i)$和 $r^-$越接近真实值.当有 $k$ 个记录的期望排位上限大于尾部记录排位下限时算法结束,得到的结果具有非常高的近似程度.

Global-Top$K$[15,33]需要两个列表:一个按记录分值排序 $L_1$,一个按记录概率排序 $L_2$.算法仍然按照 $L_1$ 扫描,按迭代关系计算当前扫描记录的 Top-$K$ 概率.计算 $k$ 个记录之后,开始利用列表 $L_2$ 和当前扫描记录来估算所有未见记录的 global-Top$K$ 上限,当此上限比所有已扫描记录的 global-Top$K$ 概率小,则推断后面的记录位于 global-Top$K$的机会很小,停止扫描.此方法的缺点是要创建两个索引,但效率可提高 2 个数量级,创建索引的开销可以被查询效率的提高所缓冲;而且当有多个属性需要考虑时,该方法可扩展.

5.2.2 随机近似

在文献[11]中,求取 PT-$K$ 时,在可能世界空间取样可能世界实例.一个记录 $t$ 以其存在概率 $p(t)$出现在样本,生成规则 $\tau$以概率 $p(\tau)$出现其中一个记录,并且 $\tau$中每个记录 $t_i$ 出现的概率为 $\frac{p(t_i)}{p(\tau)}$.每取样一个可能世界实例,对出现在 Top-$K$ 中的记录 ID 累加 1,通过 Chernoff-Hoffding 不等式证明,样本规模和误差率存在直接关系.因此,

样本足够多的情况下,记录的累计频度接近该记录的真实 Top-$K$ 概率.

分值连续时,求解各不确定性 Top-$K$ 时需要在联合概率密度函数上进行积分运算[32].因此,U-$i$Ranks 的计算采用蒙特卡洛积分;而 U-Top$K$ 则用动态马尔可夫取样的方法,保证以高概率的方式取样到有效的可能世界,最后再通过多条马尔可夫链混合提高近似程度.

5.2.3 泊松近似

如果设变量 $X=\{X_1,X_2,\ldots,X_n\}$是独立随机变量,并且 $p(X_i=1)=p_i,p(X_i=0)=1-p_i$,则 $X$ 服从泊松分布.设记录 $t_1,t_2,\ldots,t_n$ 对应 $n$ 次泊松实验,那么记录 $t$ 在位置 $j$ 的位置概率 $p(t,j)$服从泊松分布.因为泊松分布是单峰的,并且在期望处达到最大值,因此一个记录 $t$ 在期望位置的概率会最大.记录 $t$ 的期望位置为 $u$+1,其中,$u$ 是排位在 $t$ 前的记录概率和,则基于 $k<<u$ 时 $t$ 在 Top-$K$ 的概率非常小这个事实,可以得到 $u$ 在满足一定条件下,$t$ 的 Top-$K$ 概率小于 $p$.扫描过程中不断计算 $u$,当 $u$ 满足条件时算法停止.

### 5.3 复杂性对比分析

确定性算法一般要比近似算法的复杂度高.在按排序函数分值顺序进行扩展但没有高效剪枝策略时,状态扩展方法的时间复杂度为 $O(n^k)$,实际应用中很难承受.动态规划需要求解问题具有最优子结构性质,通过存储和不断重用最优子结构解构造最终解,在牺牲一定空间代价的基础上大幅降低时间复杂度.例如,文献[31]在生成 Top-$K$ 向量分值分布时,使用动态编程的方法将时间复杂度降低到了 $O(kn)$(存在互斥规则时复杂度为 $O(kmn)$).在不确定性数据库大致满足泊松二项分布条件时,泊松二项递推方法可以在 $O\left(kn+k\sum span(R)\right)$ 求解 PT-$K$,其中,$span(R)$是生成规则的跨度.生成函数的方法可以有效利用生成树的结构,将求解记录位置概率的问题转化为求多项式系数的问题,将时间复杂度降低到 $O(n+\log n)$.从目前使用的各种确定性算法可以看到:(1) 高效率算法通常依赖问题的特殊性质.例如,使用动态规划问题必须具有最优子结构性质,使用泊松二项递推数据必须大致满足泊松二项分布等;(2) 高效率还需依赖有技巧的剪枝策略,而剪枝策略又依赖求解问题满足某些特殊性质.例如求解位置概率时,各位置所有记录概率和为 1、Top-$K$ 向量有效状态长度为 $k$ 等;(3) 不确定性数据库复杂时,算法不易扩展.例如存在生成规则时,需要重新设计算法.

近似算法通过损失少量精确度大幅提高查询效率.阈值逼近通过扫描按特定顺序排列的记录,不断更新某些特定值以逼近停止条件.虽然在算法结构上没有明显改进,但因为在保证一定精确度前提下限制了扫描记录的数量,仍然可以大幅提高查询效率.如 Expected Rank Top-$K$ 求解中,基本算法是一个复杂度为 $O(n\log n)$的排序算法,但因为可以及早逼近停止条件,在精确度 85%的前提下,最坏情况下需要扫描的记录也不超过数据库总记录的 20%.随机近似方法的时间复杂度只跟抽样量的大小有关.好的抽样方法可以在很小的抽样量下达到高精确度和小误差率,例如求解 PT-$K$ 的随机近似算法中,抽样量在 500 以上时,可以达到 97%的精确度和小于 5%的误差率.泊松近似利用泊松分布函数将记录期望位置 $u$ 的停止条件转化为一个关于 $k,p$ 的不等式,由于 $k,p$ 均为常量,其时间复杂度也为常数 $O(1)$.近似算法通常具有以下特点:(1) 精确度和误差率损失很小而查询效率大幅提高;(2) 精确度和误差率可控,与运行时间存在一个平衡关系.

由于不确定性 Top-$K$ 查询处理建立在不确定性数据模型和可能世界语义的基础上,而可能世界空间随着数据量呈指数增长,导致许多求解问题具有 NP 难度.因此,使用确定性算法不可避免地会遇到瓶颈.近似算法能以较高的精确度大幅提高求解速度,在求解不确定性 Top-$K$ 问题中有很好的发挥空间.

### 5.4 对生成规则的处理

以上讨论的不确定性 Top-$K$ 查询主要针对互斥规则进行处理,采用的方式多为压缩方式.即,将互斥规则 $\tau$ 内所有记录看成单一对象、某可能世界 $w_i$ 中,该规则的概率计算如下:

$$p_{w_i}(\tau)=\begin{cases}p(t), & \text{if } \tau\cap W=\{t\}\\ 1-\sum_{t_i\in\tau}p(t_i), & \text{if } \tau\cap W=\varnothing\end{cases}.$$

Global-Top$K$ 则采用推导每个记录 $t$ 的引导事件表来处理互斥规则.该方法中,每个记录 $t$ 的引导事件表只有一个属性 event,以一定概率存在.$t$ 以自身概率存在于其引导事件表中,与 $t$ 不同规则组的记录,排序函数分值

高于 $t$ 的按组求概率和,形成新记录放入引导事件表.这样处理的目的是使一个记录 $t$ 在不确定性数据库上的 global-Top$K$ 概率转化为在其引导事件表中的 global-Top$K$ 概率,而引导事件表总是不含互斥规则的.

生成规则的存在无疑会增加不确定性 Top-$K$ 查询处理的难度,从现有查询处理中看,互斥规则的规模、数量以及规则内记录的分布情况一定程度上还影响着查询算法的效率,在查询处理中是一个必须考虑的因素.

## 6 不确定性 Top-$K$ 查询在应用层面的研究

由于很多应用领域数据存在不确定性,如高维数据库、分布式系统、数据流等,因此,特定应用层面的不确定性 Top-$K$ 查询变得具有现实意义.高维数据库中不确定性 Top-$K$ 查询最主要的困难在于因维度提高带来的 Top-$K$ 语义变化和查询效率问题;而分布式系统中不确定性 Top-$K$ 查询除了研究如何平衡分值和概率、如何降低需要展开的可能空间数目外,第 3 个挑战是如何降低交互程度[39−41].文献[1,42]研究了在主流不确定性数据模型上如何设计低交互以及高效率的 Expected Rank 求解算法.不确定性数据流上的 Top-$K$ 查询[34,43]则对静止的不确定性数据模型和处理方式提出挑战,时间维的处理成为研究的焦点.

### 6.1 高维不确定性Top-$K$查询

在高维数据库中,每个对象存在一个不确定区域,对象以一定概率分布存在于该区域[44,46].不确定区域的存在使度量标准——距离变得不再精确,具有广泛应用的 Range 查询(范围查询)、Top-$K$ NN 查询(最近邻查询)、Top-$K$ RNN 查询(逆最近邻查询)、Skyline 查询等呈现出新的特点[41,44,46,47],学者们根据这些新的特点开展了不同程度的研究.

Rang 查询、Top-$K$ (R)NN 查询在 LBS(基于位置查询)、GPS 位置追踪等应用中被广泛研究.当对象以一定概率分布于某不确定区域,两对象间距离不是一个固定值而变成了一个距离范围,除了期望距离等这类简单的标准,边际 NN 概率(某对象 $o$ 离另一对象 $q$ 的距离为最近的那些值的概率和)成为被广泛接受的距离衡量方式.不仅如此,根据查询对象与数据对象是否确定,查询处理方式也各不相同[44,45],具体有:(1) 查询对象确定、数据对象不确定;(2) 查询对象不确定、数据对象确定;(3) 查询对象和数据对象都不确定.

Skyline 查询有很长的研究历史.在一个多准则决策系统中,每个决策维存在不同分值函数,如果一对象 $u_1$ 在某一维分值优于另一对象 $u_2$,而同时在其他维都不比 $u_2$ 差,则 $u_1$ 支配 $u_2$,Skyline 查询那些不被任何对象支配的对象.不确定性的环境下,对象某属性依概率有多种取值情况,根据取值大小的差别,一个对象有可能在某一个可能世界中是 Skyline 点,在另一个可能世界中并不是或者根本没有出现.因此,对象的 Skyline 概率(一个对象在各可能世界中成为 Skyline 的概率和)成为 Skyline 查询研究的重点[46].Pei 定义了 $p$-Skyline 查询,设定一个概率阈值,查询那些 Skyline 概率大于该阈值的记录;Zhang 定义了 Top-$K$ sound 查询,查询具有最大 Skyline 概率的 $k$ 个记录.

这些查询中,有效的几何策略往往成为查询处理的关键[44,49]:(1) 为了降低 I/O 开销,对象读取基本上都采用基于距离序列增量获取的方式[44−46].Beskales 采用数据对象与查询对象的最小距离序列读取,Zhang 采用对象重心与原点的距离序列读取;(2) 为了降低 CPU 开销,根据对象的几何位置关系可以实现大规模剪枝.Beskales 将查询对象周围的对象按边界点位置关系分区,将分区边界积分作为分区内积分的上下限,从而降低大量积分带来的 CPU 开销;Zhang 将对象区域用 R-tree 索引,一旦某节点区域被其他对象支配,该节点内所有对象都被剪枝.

### 6.2 分布式环境中的不确定性Top-$K$查询

在大规模的 P2P 系统、传感器网络等分布式环境中,Top-$K$ 查询一直有着广泛的应用[39].而随着不确定性数据处理需求的出现,分布式环境中的不确定性 Top-$K$ 查询研究具有了重要的现实意义[40−42].分布式环境中,不确定性 Top-$K$ 查询处理面临的核心问题是如何在降低计算开销的同时最小化交互开销.Li 在文献[40]中针对如何在有集中式服务器的分布式环境中进行 Expected Rank 的问题展开了研究,Sun[42]则主要针对无集中式服务器的 P2P 分布式环境的不确定性 Top-$K$ 查询处理方法进行研究.

Li 提出在有集中式服务器的环境下,每个数据终端可按分值函数降序排列记录.由于 Expected Rank 是按照每条记录在各可能世界的期望位置排序,只要广播一条记录,计算它在所有数据终端记录序列的排位并求总和,便可得它的全局排位.集中式服务器则维护一个优先队列,从所有数据终端按期望位置值获取记录.为了最大程度地实现剪枝,类似于文献[10,30]中 Expected Rank 的求取方法,可以通过计算已扫描记录的期望排位上限和未扫描记录排位下限近似求解.低交互开销的实现,在于阈值的存在极大地限制了需要广播记录的数量及需要计算全局排位记录的数量.该方法同时提出,在数据终端计算能力有限时,通过近似模拟记录的局部排位数,可在保持低交互开销的同时将大部分计算开销转移至服务器端.

在 P2P 网络中,各数据终端存储和计算能力相近,没有一个集中式服务器可提供高性能计算和大数据量存储,因此对终端间交互更为依赖.Sun 提出,可通过建立不确定性四叉树索引获取不确定性数据的全局性信息[42].通过观察全局索引中各不确定对象区域在排序分值降序方向的位置关系,区分出一定在 Top-*K* 内、一定不在 Top-*K* 内以及以一定概率在 Top-*K* 内的 3 个记录集,分别为积极区域、消极区域以及不确定区域.空间剪枝后,积极区域记录数 $m$ 通常小于 $k$ 个,需要找出不确定区域中分值靠前的 $k–m$ 个.假设有 $p$ 个数据终端,在分布式剪枝阶段,每次每个数据终端只广播其第$(k–m)/p$ 个记录,接受其他 $p–1$ 个数据终端的记录后进行重排,每个数据终端按参与此次广播的 $p$ 个记录中最低者剪枝.几轮重复之后,将最终留下 $k–m$ 个记录.

根据不同分布式系统各自的特点,交互开销降低的方法各异,但增强全局性是一个技术导向,例如有集中式服务器中服务器端计算能力的充分利用以及 P2P 环境下使用全局索引进行初步剪枝.

### 6.3 数据流中的不确定性Top-*K*查询

在数据流上,由于记录到达的快速性、无序性以及数量的无限制性,要求查询处理算法必须具有一遍扫描、时间开销和空间开销都比较低等特点.确定性数据流上Top-*K*查询只需要维护大小为$k$的缓冲区,保存具有最高分值的 $k$ 个记录.新记录到达时,与缓存的 $k$ 个记录比较,将最低分值者替换掉即可.由于不确定性 Top-*K* 语义的复杂性,不确定性数据流上的 Top-*K* 查询变得复杂很多.关于不确定数据流上 Top-*K* 查询,目前主要有基于无限制数据流以及基于限定时间维数据流的研究.

无限制的不确定性数据流环境中,不考虑记录的消逝.随着记录不停到达,所有记录参与Top-*K*查询处理.Jin 在文献[43]中讨论了如何在无限制的不确定性数据流中进行 Expected Rank[10,30],并根据 Expected Rank 值获取前 $k$ 记录的 ER-Top*K* 查询.根据每个记录各不确定属性值概率分布特点,Jin 定义了记录之间的支配关系(例如记录 $t_1$=⟨(50,0.6),(70,0.4)⟩,$t_2$=⟨(20,0.4),(30,0.3)⟩,$t_1$ 支配 $t_2$),当一个记录被至少 $k$ 个记录支配时,该记录不可能在结果集中.没有被至少 $k$ 个记录支配的所有候选记录采用 domGraph 数据结构管理,domGraph 维护了每个候选记录的支配集、被支配集、当前状态以及在候选中的排位,通过计算候选的 Expect rank 分值得到解集.为了快速计算某记录的 Expect rank 分值,Cormode[10]提出可以通过累加比该记录属性值大的那些值的概率求得.基于此,Jin CQ 定义了一个二叉查找树 probTree,树的内部节点($v$,*prob*)表示当前比 $v$ 大的属性值的概率和为 *prob*,每次新记录到达都会使部分内部节点的 *prob* 更新.基于该二叉查找树,Expect Rank 分值直接可得.

而数据流对时间维的限定方式主要有两种:一种是随时间衰减的处理方式,一种是滑动窗口的处理方式.文献[34]中分析了不确定数据流上连续滑动窗口 Top-*K* 查询的特点,定义了滑动窗口 $W$ 内一定包含 Top-*K* 记录的最小子集——窗口的紧致集 $C(W)$.紧致集具有自包含性质,即,新记录 $t_{new}$ 分值低于紧致集内最低分值时,$C(W\cup t_{new})=C(W)$;新记录 $t_{new}$ 分值高于紧致集内最低分值时,$C(W\cup t_{new})\subseteq C(W)\cup t_{new}$.紧致集的自包含性质暗示了随窗口滑动时窗口紧致集的变化规律,将窗口的所有子窗口紧致集用队列保存,一旦窗口紧致集中有记录消逝,则以队列中更新的子窗口紧致集替换,便可实现紧致集的连续.这样,回答不确定性数据流上连续滑动窗口 Top-$k$ 查询,将只需要根据新到达记录维护窗口的紧致集,不需要保存窗口内所有记录,在紧致集规模远远小于窗口规模的时候,空间开销和时间开销大为降低.

无限制不确定数据流上 Top-*K* 查询由于考虑所有记录,因此更具有时间全局性;而考虑时间限制的不确定数据流上的 Top-*K* 查询更具有现时性.

## 7 总 结

不确定性数据在应用领域的广泛出现,推动了不确定性数据库的研究.从数据模型的研究到各种不确定性数据库的开发,处理的不确定性数据从结构化数据到半结构化数据再到本体,而查询的类型也由传统的查询类型不断延伸扩展.Top-*K* 查询由于概率的引入成为一个崭新的查询类型.不确定性 Top-*K* 查询从 2007 年提出到现在虽然只是短短几年,但已经有很多好的研究成果,总结起来有以下特点:(1) 涉及数据模型简单,几乎都是建立于主流的不确定性数据模型上;(2) 几乎都是从具体应用场景出发定义新的查询语义并给出处理方法,关于概率在 Top-*K* 查询中的作用和处理,没有统一规范;(3) 处理的方式更多地采用确定性算法附以有效的剪枝策略,用非确定性的算法,例如近似算法,去查询处理的研究还不深入;(4) 新定义的不确定性 Top-*K* 语义还没有广泛使用到应用领域.

作为一个崭新的领域,不确定性 Top-*K* 查询处理有很多工作需要展开:

(1) 有很多问题在简单的主流不确定性数据模型基础上并不能深刻描述;同时,不确定性数据库研究领域关于数据模型也已经有丰富的研究,需要拓展不确定性 Top-*K* 问题在更丰富的数据模型上如何进行查询处理;

(2) 概率的存在使分值和概率的平衡成为不确定性 Top-*K* 查询的焦点问题,研究更合理的平衡方式会使不确定性 Top-*K* 从语义到返回结果走向统一;

(3) 传统的确定性算法研究不确定性数据库有自身的局限,高效的非确定性算法研究有更广阔的前景;

(4) 虽然不确定性 Top-*K* 查询的应用会受到不确定性数据库研究进展的局限,但大量现实中存在的不确定性问题仍然可以在现有数据库技术的基础上利用不确定性 Top-*K* 查询的方法解决.

**致谢** 在此,向对本文给予支持和建议的同行,尤其是武汉大学计算机学院彭智勇教授领导的讨论组的老师和同学表示感谢.

**References**:

[1] Cao P, Wang Z. Efficient Top-*K* query calculation in distributed networks. In: Chaudhuri S, Kutten S, eds. Proc. of the PODC. New York: ACM Press, 2004. 206−215. [doi: 10.1145/1011767.1011798]

[2] Marian A, Bruno N, Gravano L. Evaluation Top-*K* queries over Web-accessible databases. ACM Trans. on Database Systems, 2004, 29(2):319−342. [doi: 10.1145/1005566.1005569]

[3] Ilyas IF, Beskales G, Soliman MA. A survey of Top-*K* query processing techniques in relational database systems. ACM Computing Surveys, 2008,40(4):11:1−11:57. [doi: 10.1145/1391729.1391730]

[4] Boulos J, Dalvi D, Mandhani B, Mathur B, Re C, Suciu D. MYSTIQ: A system for finding more answers by using probabilities. In: Özcan F, ed. Proc. of the SIGMOD. New York: ACM Press, 2005. 891−893. [doi: 10.1145/1066157.1066277]

[5] Widom J. Trio: A system for integrated management of data, accuracy, and lineage. In: Proc. of the CIDR. Asilomar: Online Proc., 2005. 262−276.

[6] Benjelloun O, Sarma AD, Halevy AY, Widom J. Uldbs: Databases with uncertainty and lineage. In: Dayal U, Whang KY, Lomet DB, Alonso G, Lohman GM, Kersten ML, Cha SK, Kim YK, eds. Proc. of the VLDB. San Francisco: Morgan Kaufmann Publishers, 2006. 953−964.

[7] Singh S, Mayfield C, Mittal S, Prabhakar S, Hambrusch S, Shah R. Orion 2.0: Native support for uncertain data. In: Tsong J, Wang L, eds. Proc. of the SIGMOD. New York: ACM Press, 2008. 1239−1242. [doi: 10.1145/1376616.1376744]

[8] Ré C, Dalvi NN, Suciu D. Efficient Top-*K* query evaluation on probabilistic data. In: Chirkova R, Dogac A, Özsu MT, Sellis TM, eds. Proc. of the Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2007. 886−895. [doi: 10.1109/ICDE.2007.367934]

[9] Soliman MA, Ilyas IF, Chang KCC. Top-*K* query processing in uncertain databases. In: Chirkova R, Dogac A, Özsu MT, Sellis TM, eds. Proc. of the Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2007. 896−905. [doi: 10.1109/ICDE.2007.367935]

[10] Cormode G, Li F, Yi K. Semantics of ranking queries for probabilistic data and expected ranks. In: Ioannidis YE, Lee DL, Ng RT, eds. Proc. of the Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2009. 305−316. [doi: 10.1109/ICDE.2009.75]

[11] Hua M, Pei J, Zhang WJ, Lin XM. Ranking queries on uncertain data: A probabilistic threshold approach. In: Tsong J, Wang L, eds. Proc. of the SIGMOD. New York: ACM Press, 2008. 673−686. [doi: 10.1145/1376616.1376685]

[12] Hua M, Pei J, Zhang WJ, Lin XM. Efficiently answering probabilistic threshold Top-*k* queries on uncertain data. In: Alonso G, Blakeley JA, Chen ALP, eds. Proc. of the Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2008. 1403−1405. [doi: 10.1109/ICDE.2008.4497570]

[13] Hua M, Pei J. Ranking queries on uncertain data. The VLDB Journal, 2011,20(1):129−153. [doi: 10.1007/s00778-010-0196-4]

[14] Li J, Saha B, Deshpande A. A unified approach to ranking in probabilistic databases. The VLDB Journal, 2011,20(2):249−275. [doi: 10.1007/s00778-011-0220-3]

[15] Zhang X, Chomicki J. On the semantics and evaluation of Top-*K* queries in probabilistic databases. In: Proc. of the Int'l Conf. on Data Engineering Workshop. Washington: IEEE Computer Society Press, 2008. 556−563. [doi: 10.1109/ICDEW.2008.4498380]

[16] Xiang L, Chen L. Probabilistic ranked queries in uncertain databases. In: Teubner J, ed. Proc. of the EDBT. New York: ACM Press, 2008. 511−522. [doi: 10.1145/1353343.1353406]

[17] Imielinski T, Jr. WL. Incomplete information in relational databases. Journal of the ACM, 1984,31(4):761−791. [doi: 10.1145/1634.1886]

[18] Abiteboul S, Kanellakis PC, Grahne P. On the representation and querying of sets of possible worlds. In: Dayal U, Traiger IL, eds. Proc. of the SIGMOD. New York: ACM Press, 1987. 34−48. [doi: 10.1145/38714。38724]

[19] Cavallo R, Pittarelli M. The theory of probabilistic databases. In: Stocker PM, Kent W, Hammersley P, eds. Proc. of the VLDB. San Francisco: Morgan Kaufmann Publishers, 1987. 71−81.

[20] Lakshmanan LVS, Leone N, Ross RB, Subrahmanian VS. Probview: A flexible probabilistic database system. ACM Trans. on Database System, 1997,22(3):419−469. [doi: 10.1145/261124。261131]

[21] Dalvi N, Suciu D. Management of probabilistic data: Foundations and challenges. In: Libkin L, ed. Proc of the PODS. New York: ACM Press, 2007. 1−12. [doi: 10.1145/1265530.1265531]

[22] Green TJ, Tannen V, Models for incomplete and probabilistic information. IEEE Data Engineering Bulletin, 2006,29(1):17−24. [doi: 10.1007/11896548\24]

[23] Suciu D, Dalvi N. Foundations of probabilistic answers to queries. In: Özcan F, ed. Proc. of the SIGMOD. New York: ACM Press, 2005. 963. [doi: 10.1145/1066157.1066303]

[24] Dalvi N, Re C, Suciu D. Probabilistic databases: Diamonds in the dirt. Communications of the ACM, 2009,52(7):86−94. [doi: 10.1145/1538788.1538810]

[25] Sarma AD, Benjelloun O, Halevy AY, Widom J. Working models for uncertain data. In: Liu L, Reuter A, Whang KY, Zhang JJ, eds. Proc. of the Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2006. 7. [doi: 10.1109/ICDE. 2006. 174]

[26] Zhou AY, Jin CQ, Wang GR, Li J. A survey on the management of uncertain data. Chinese Journal of Computers, 2009,32(1):1−16 (in Chinese with English abstract).

[27] Fuhr N, Rölleke T. A probabilistic relational algebra for the integration of information retrieval and database systems. ACM Trans. on Information Systems, 1997,15(1):32−66. [doi: 10.1145/239041.239045]

[28] Sen P, Deshpande A, Getoor L. PRDB: Managing and exploiting rich correlations in probabilistic databases. The Journal of VLDB, 2009,18(5):1065−1090. [doi: 10.1007/s00778-009-0153-2]

[29] Zhang F, Ma ZM, Yanhui L, Wang X. Formal semantics-preserving translation from fuzzy ER model to fuzzy OWL DL ontology. Lecture Notes in Computer Science, 2008,5367:46−60. [doi: 10.1007/978-3-540-89704-0_4]

[30] Jestes J, Cormode G, Li FF, Yi K. Semantics of ranking queries for probabilistic data. IEEE Trans. on Knowledge and Data Engineering, 2010,23(12):1903−1917. [doi: 10.1109/TKDE.2010.192]

[31] Ge T, Zdonik S, Madden S. Top-*K* queries on uncertain data: on score distribution and typical answers. In: Çetintemel U, Zdonik SB, Kossmann D, Tatbul N, eds. Proc. of the SIGMOD. New York: ACM Press, 2009. 375−388. [doi: 10.1145/1559845.1559886]

[32] Soliman MA, Ilyas IF. Ranking with uncertain scores. In: Ioannidis YE, Lee DL, Ng RT, eds. Proc. of the ICDE. Washington: IEEE Computer Society Press, 2009. 317−328. [doi: 10.1109/ICDE.2009.102]

[33] Zhang X, Chomicki J. On the semantics and evaluation of Top-*K* queries in probabilistic databases. Journal of Distributed and Parallel Databases, 2009,26(1):67−126. [doi: 10.1007/s10619-009-7050-y]

[34] Jin CQ, Yi K, Chen L, Yu X, Lin XM. Sliding window Top-*K* queries on uncertain streams. The VLDB Journal, 2010,19(3): 411−435. [doi: 10.1007/s00778-009-0171-0]

[35] Li J, Deshpande A. Ranking continuous probabilistic datasets. Journal Proc. of VLDB Endowment, 2010,3(1-2):638−649.

[36] Ge T, Zdonik S. Handling uncertain data in array database systems. In: Alonso G, Blakeley JA, Chen ALP, eds. Proc. of the Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2008. 1140−1149. [doi: 10.1109/ICDE.2008.449 7523]

[37] Fagin R, Kumar R, Sivakumar D. Comparing Top-*K* lists. SIAM Journal on Discrete Math, 2003,17(1):134−160. [doi: 0.1137/S08954 80102412856]

[38] Yi K, Li F, Efficient processing of Top-*k* queries in uncertain databases. IEEE Trans. on Knowledge and Data Engineering, 2008, 20(12):1669−1682. [doi: 10.1109/TKDE.2008.90]

[39] Wang GR, Yuan Y, Sun YJ. PeerLearning: A content-based e-learning material sharing system based on P2P network. World Wide Web, 2011,13(3):275−305. [doi: 10.1007/s11280-010-0086-0]

[40] Li F, Yi K, Jestes J. Ranking distributed probabilistic data. In: Çetintemel U, Zdonik SB, Kossmann D, Tatbul N, eds. Proc. of the SIGMOD, New York: ACM Press, 2009. 361−374. [doi: 10.1145/1559845.1559885]

[41] El-Desouky AI, Ali HA, Abdul-Azeem YM. Ranking distributed uncertain database systems: Discussion and analysis. In: Proc. of the Int'l Conf. on Computer Engineering and Systems. 2010. 295−300. [doi: 10.1109/ICCES.2010.5674872]

[42] Sun YJ, Yuan Y, Wang GR. Top-*K* query processing over uncertain data in distributed environments. World Wide Web, On Line First TM, 2011. 1−18. [doi: 10.1007/s11280-011-0141-5]

[43] Jin CQ, Gao M, Zhou AY. Handling ER-Top *K* query on uncertain streams. Lecture Notes in Computer Science, 2011,6587: 326−340. [doi: 10.1007/978-3-642-20149-3_25]

[44] Beskales G, Soliman MA, IIyas IF. Efficient search for the Top-*k* probable nearest neighbors in uncertain databases. Journal Proc. of VLDB Endowment, 2008,1(1):326−339. [doi: 10.1145/1453856.1453895]

[45] Wang CH, Li YY, Jia HY. Top-*K* ranking for uncertain data. In: Proc. of the Fuzzy Systems and Knowledge Discovery. 2010. 36−368. [doi: 10.1109/FSKD.2010.5569645]

[46] Zhang Y, Zhang WJ, Lin XM, Jiang B, Pei J. Ranking uncertain sky: The probabilistic Top-*K* skyline operator. In: Proc. of the Information Systems. 2011. 898−915. [doi: 10.1016/j.is.2011.03.008]

[47] Lian X, Chen L. Shooting Top-*k* stars in uncertain databases. Journal of VLDB, 2011,20(6):819−840. [doi: 10.1007/s00778-011-0225-y]

[48] Wang CH, Yuan LY, You JH, Zaiane OB. On pruning for Top-*k* ranking in uncertain databases. Journal Proc. of VLDB Endowment, 2011,4(10):598−609.

[49] Wang CH. Top-*K* ranking with uncertain data [Ph.D. Thesis]. Edmonton: University of Alberta, 2012.

**附中文参考文献:**

[26] 周傲英,金澈清,王国仁,李建中.不确定性数据管理技术研究综述.计算机学报,2009,32(1):1−16.

**李文凤**(1983－),女,河南新乡人,博士生,主要研究领域为不确定性数据管理.

**李德毅**(1944－),男,博士,教授,博士生导师,中国工程院院士,主要研究领域为人工智能,人工指挥自动化.

**彭智勇**(1963－)男,博士,教授,博士生导师,主要研究领域为数据库理论和技术.