不确定性数据管理技术研究综述

周傲英[1)] 金澈清[1)] 王国仁[2)] 李建中[3)]

[1)](华东师范大学软件学院 上海市高可信计算重点实验室 上海 200062)

[2)](东北大学信息科学与工程学院 沈阳 110004)

[3)](哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 随着数据采集和处理技术的进步,人们对数据的不确定性的认识也逐步深入.在诸如经济、军事、物流、金融、电信等领域的具体应用中,数据的不确定性普遍存在.不确定性数据的表现形式多种多样,它们可以以关系型数据、半结构化数据、流数据或移动对象数据等形式出现.目前,根据应用特点与数据形式差异,研究者已经提出了多种针对不确定数据的数据模型.这些不确定性数据模型的核心思想都源自于可能世界模型.可能世界模型从一个或多个不确定的数据源演化出诸多确定的数据库实例,称为可能世界实例,而且所有实例的概率之和等于 1.尽管可以首先分别为各个实例计算查询结果,然后合并中间结果以生成最终查询结果,但由于可能世界实例的数量远大于不确定性数据库的规模,这种方法并不可行.因此,必须运用排序、剪枝等启发式技术设计新型算法,以提高效率.文中介绍了不确定性数据管理技术的概念、特点与挑战,综述了数据模型、数据预处理与集成、存储与索引、查询处理等方面的工作.

关键词 不确定性数据;可能世界模型;数据集成;世系;不确定数据流

中图法分类号 TP393 **DOI 号**: 10.3724/SP.J.1016.2009.00001

A Survey on the Management of Uncertain Data

ZHOU Ao-Ying[1)] JIN Che-Qing[1)] WANG Guo-Ren[2)] LI Jian-Zhong[3)]

[1)](*Shanghai Key Laboratory of Trustworthy Computing*, *Software Engineering Institute*, *East China Normal University*, *Shanghai* 200062)

[2)](*School of Information Science and Engineering*, *Northeastern University*, *Shenyang* 110004)

[3)](*School of Computer Science and Technology*, *Harbin Institute of Technology*, *Harbin* 150001)

Abstract The importance of the data uncertainty was studied deeply with the rapid development in data gathering and processing in various fields, inclusive of economy, military, logistic, finance and telecommunication, etc. Uncertain data has many different styles, such as relational data, semistructured data, streaming data, and moving objects. According to scenarios and data characteristics, tens of data models have been developed, stemming from the core possible world model that contains a huge number of the possible world instances with the sum of probabilities equal to 1. However, the number of the possible world instances is far greater than the volume of the uncertain database, making it infeasible to combine medial results generated from all of possible world instances for the final query results. Thus, some heuristic techniques, such as ordering, pruning, must be used to reduce the computation cost for the high efficiency. This paper introduces the concepts, characteristics and challenges in uncertain data management, proposes the advance of the research on uncertain data management, including data model, preprocessing, integrating, storage, indexing, and query processing.

Keywords uncertain data; possible world model; data integration; lineage; uncertain stream

收稿日期:2008-09-05.本课题得到国家自然科学基金(60803020)、上海市重点学科建设项目(B412)资助.**周傲英**,男,1965 年生,教授,博士生导师,主要研究兴趣为数据管理与信息系统,包括:Web 数据管理、中文 Web 基础设施、Web 搜索与挖掘、数据流与数据挖掘、复杂事件处理与实时商务智能、不确定数据管理及其应用、数据密集的计算、分布存储与计算、对等计算及其数据管理、Web 服务计算等.**金澈清**(通信作者),男,1977 年生,博士,副教授,主要研究方向为数据流管理、不确定性数据管理技术等.E-mail: cqjin@sei.ecnu.edu.cn.**王国仁**,男,1966 年生,教授,博士生导师,主要研究领域为 XML 数据管理、生物信息学、分布与并行数据库、多媒体索引技术、并行计算等.**李建中**,男,1950 年生,教授,博士生导师,主要研究领域为数据库、并行计算等.

1 引 言

近四十年来,传统的确定性数据(deterministic data)管理技术得到了极大的发展,造就了一个数百亿的数据库产业.数据库技术和系统已经成为信息化社会基础设施建设的重要支撑.在传统数据库的应用中,数据的存在性和精确性均确定无疑.近年来,随着技术的进步和人们对数据采集和处理技术理解的不断深入,不确定性数据(uncertain data)得到了广泛的重视.在许多现实的应用中,例如经济、军事、物流、金融、电信等领域,数据的不确定性普遍存在,不确定性数据扮演着关键角色.传统的数据管理技术却无法有效管理不确定性数据,这就引发了学术界和工业界对研发新型的不确定性数据管理技术的兴趣.

不确定性数据的产生原因比较复杂.可能是原始数据本身不准确或是采用了粗粒度的数据集合,也可能是为了满足特殊应用目的或是在处理缺失值、数据集成过程中而产生的.

(1)原始数据不准确.这是产生不确定性数据最直接的因素.首先,物理仪器所采集的数据的准确度受仪器的精度制约.其次,在网络传输(特别是无线网络传输)过程中,数据的准确性受到带宽、传输延时、能量等因素影响.还有,在传感器网络应用[1-2]与 RFID 应用[3]等场合,周围环境也会影响原始数据的准确度.

(2)使用粗粒度数据集合.很明显,从粗粒度数据集合转换到细粒度数据集合的过程会引入不确定性.例如,假设某人口分布数据库以乡为基础单位记录全国的人口数量,而某应用却要求查询以村为基础单位的人口数量,查询结果就存在不确定性.

(3)满足特殊应用目的.出于隐私保护等特殊目的,某些应用无法获取原始的精确数据,而仅能够得到变换之后的不精确数据.

(4)处理缺失值.缺失值产生的原因很多,装备故障、无法获取信息、与其他字段不一致、历史原因等都可能产生缺失值.一种典型的处理方法是插值,插值之后的数据可看作服从特定概率分布.另外,也可以删除所有含缺失值的记录,但这个操作也在一定程度上变动了原始数据的分布特征.

(5)数据集成.不同数据源的数据信息可能是不一致的,在数据集成过程中就会引入不确定性.例如,Web 中含很多信息,但是由于页面更新等因素,许多页面的内容并不保持一致[4].

对某些应用而言,还可能同时存在多种不确定性.例如,基于位置的服务(Location-Based Service,LBS)[5]是移动计算领域的核心问题,在军事、通信、交通、服务业等方面有着广泛的应用.LBS 应用获取各移动对象的位置,为用户提供定制服务,该过程存在若干不确定性.首先,受技术手段(例如 GPS 技术)限制,移动对象的位置信息存在一定误差.其次,移动对象可能暂时不在服务区,导致 LBS 应用采集的数据存在缺失值情况.最后,某些查询要求保护用户的隐私信息,必须采用"位置隐私"等方式处理查询[6].

实际上,针对不确定性数据的研究工作已经有几十年历史了.从 20 世纪 80 年代末开始,针对概率数据库(probabilistic database)的研究工作就从未间断过[7-11].这类研究工作将不确定性引入到关系数据模型中去,取得了较大进展.近年来,针对不确定性数据的研究工作则在更广的范围内取得了更大的进展,即在更丰富的数据类型上处理更多种类的查询任务.图 1 描述了不确定性数据管理技术的典型框架,它包含 4 大部分:模型定义、预处理与集成、存储与索引和查询分析处理.

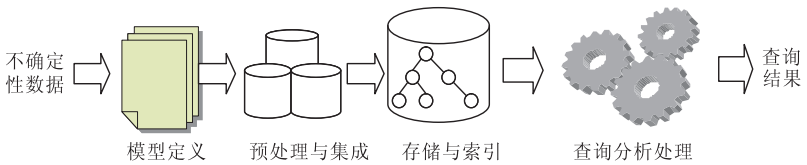


图 1 不确定性数据管理的框架

模型定义.定义与应用场景相匹配的数据模型是不确定性数据管理的首要任务.在不确定性数据管理领域,最常用的模型是可能世界模型(possible world model)[12-13].该模型从一个不确定性数据库演化出很多确定的数据库实例(称为可能世界实例),而且所有实例的概率之和为 1.不确定性数据的种类较多,例如关系型数据、半结构化数据、流数据、移动对象数据等,尽管存在许多与数据类型紧密相关的数据模型,但是这些模型最终都可以转化为可能世界模型.

预处理与集成. 某些应用需要为数据执行预处理操作，主动引入不确定性，从而达到信息隐藏和隐私保护的目的. 这种不确定性会降低查询结果质量，必须在查询质量与信息隐藏程度之间进行权衡. 当应用需要使用多个数据源时，数据不一致性问题就凸显出来. 这个问题在 WEB 上尤为突出. 数据集成所要应对的不确定性问题不仅包括原始数据不一致，还包括模式匹配不确定、待处理的查询语义不确定等多种因素.

存储与索引. 有效的存储和索引技术能够大幅提高数据管理效率. 尽管可以基于传统的关系型数据存储技术实现不确定性数据库的存储任务，但仍有必要开发新型存储技术，以提高特定查询任务(例如数据世系，data lineage)的执行效率. 概要数据结构(synopsis data structure)是存储流数据(data stream)[14-15]的典型技术. 不确定性数据与确定性数据的最大区别在于不确定性数据含有概率维度. 一部分查询任务仅使用基于非概率维度的索引. 例如，在处理不确定 Top-k 查询的过程中，往往只需对值维度以 ranking 函数创建索引. 另一类查询则需针对概率维度开发新的索引技术，例如，范围查询(Range query)、最近邻查询(Nearest Neighbor query)等. 当概率维度以概率密度函数(probabilistic density function，简称 pdf)描述而非概率值时，创建索引的难度更大.

查询分析处理. 查询分析处理是不确定性数据管理的最终目标. 查询类型非常丰富，例如关系查询操作、数据世系、XML 处理、流数据查询、Ranking 查询、Skyline 查询、OLAP 分析、数据挖掘等. 尽管可以分别针对各个可能世界实例计算查询结果，再合并中间结果以生成最终查询结果，但由于可能世界实例的数量远大于不确定数据库的规模，该方法并不可行. 因此，必须采用排序、剪枝等启发式技术优化处理，以提高效率. 另外，由于输入数据具有不确定性，查询结果也往往是近似结果.

目前在研的主要项目

不确定数据管理正成为一个研究热点. 表 1 列举了一些知名大学以及公司的研究机构正在进行的相关科研项目的基本情况.

表 1　不确定性数据管理的相关研究项目

科研机构	项目名称	链接地址	描述
University of Puget Sound	proTDB	http://www.math.ups.edu/~anierman/umich/protdb/	主要研究概率半结构化数据的查询处理技术.
University of Toronto	Conquer	http://queens.db.toronto.edu/project/conquer/	主要研究针对不一致数据库(inconsistent database)的高效管理技术. 主要应用查询重写技术、实时和动态数据清洗技术.
Purdue University	Orion	http://orion.cs.purdue.edu/	曾用名：U-DBMS，是一个通用目的的不确定性数据库系统. 它支持离散或连续的概率密度函数；高效的访问不确定性数据的方法；优化连接、选择等操作；图形可视化界面.
Stanford University	Trio	http://infolab.stanford.edu/trio/	主要研究不确定数据管理技术，特别是针对不确定数据的世系分析技术. 它基于 ULDB 模型，使用 TriQL 语言.
Cornell University	MayBMS	http://www.cs.cornell.edu/database/maybms/	研究内容包括：查询语言、表示与存储技术、支持数据清洗、高效的查询处理、更新等.
University of Washington	MystiQ	http://www.cs.washington.edu/homes/suciu/project-mystiq.html	研究内容包括：数据模型、查询处理技术、关系代数计算等.
Intel/Berkeley	HeisenData	http://www.eecs.berkeley.edu/Research/Projects/Data/102060.html	该项目试图将现有的确定性数据管理框架与不确定性查询处理技术相结合，从而使得新增的功能模块能够嵌入到现有框架中，增强其性能.
University of Maryland	Prob DBs	http://www.cs.umd.edu/~vs/research.htm#pdb	研究概率聚集查询的计算方法，开发空间-时间概率数据库.
IBM Almaden	Avatar	http://www.almaden.ibm.com/cs/projects/avatar/	该项目有两大目标：(1)从非结构化数据中抽取结构化的信息；(2)基于这类信息构建下一代搜索和商业智能应用.

Ré 与 Suciu[16]观察到不确定性数据广泛出现在诸多应用之中，并总结了不确定性数据管理所面临的巨大挑战. Dalvi 和 Suciu[17]进一步从理论角度阐述不确定性数据管理的基础与挑战. Aggarwal 与 Yu[18]从算法与应用角度综述了不确定数据管理技术. Pei 等人[19]回顾了近期不确定性查询处理方面的进展，特别是他们自己的工作，包括范围查询、skyline 查询与 Ranking 查询等. 本文则以不确定性数据管理的框架为主线，综述了不确定性数据管理技术在数据模型、预处理与集成、存储与索引、查询分析处理等方面所取得的重要进展. 本文第 2～5 节分别介绍上述 4 个方面的内容；第 6 节总结全文.

2 数据模型与挑战

2.1 可能世界模型

不确定数据库建模的研究工作很多,可能世界模型则是应用最广泛的数据模型[12-13].在该模型中,各元组的任一合法组合均构成一个可能世界实例(instance),实例的概率值可以通过相关元组的概率计算得到.可能世界实例的数量远远高于不确定性数据库的规模,甚至是后者的指数倍,这也是不确定性数据管理技术所面临的最大难点.考虑如图 2 所示的一个例子.图 2(a)是一个不确定性数据库,包含 3 个元组,概率字段表示该元组的发生概率.元组之间可能独立也可能存在依赖关系.首先假设各个元组之间独立,则共有 $2^3=8$ 个可能世界实例,各实例的概率等于实例内元组的概率乘积与实例外元组的不发生概率的乘积,如图 2(b)所示.例如,可能世界实例{1,2}的发生概率为 $0.3\times0.7\times(1-0.6)=0.084$.某些场景下,元组之间并非独立,而是存在依赖关系,这种依赖关系可以用规则描述.假设规则为 $1\oplus3$,即元组 1 与元组 3 不能够同时发生,但可以同时不发生[20].总共有 6 个可能世界实例,如图 2(b)所示.可能世界实例{1}的发生概率为 $0.3\times(1-0.7)=0.09$,可能世界实例{2}的发生概率为 $(1-0.3-0.6)\times0.7=0.07$.

ID	信息	概率
1	A	0.3
2	B	0.7
3	C	0.6

(a) 一个不确定数据库样例

元组独立:
$PW=\{\{\},\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}$
$P(PW)=\{.084,.036,.196,.126,.084,.054,.294,.126\}$

依赖规则:$1\oplus3$
$PW=\{\{\},\{1\},\{3\},\{2\},\{1,2\},\{2,3\}\}$
$P(PW)=\{.03,.09,.18,.07,.21,.42\}$

(b) 可能世界

图 2 可能世界样例

在大多数应用中,不确定性可细分为存在级不确定性(Existential Uncertainty)和属性级不确定性(Attribute Level Uncertainty).存在级不确定性描述元组的存在与否,较为普遍.在图 2 中,各元组均具备存在级不确定性.属性级不确定性并不涉及整个元组的不确定性,而是以概率密度函数或统计参数(例如方差等)来描述特定属性的不确定性.例如,假设某传感器无法准确探测周围环境温度,典型的记录方式为:70% 的概率为 26℃,30% 的概率为 25℃.类似的记录均具有属性级不确定性.属性级不确定性往往比存在级不确定性更容易处理.有些时候,也可以将多个相关的元组视为单个含属性级不确定性的元组.例如,图 2(b)定义了依赖规则 $1\oplus3$,则元组 1 和 3 无法同时发生.可以将这两个元组视为单个元组,该元组有存在级不确定性,发生概率为 0.9;该元组的信息字段有属性级不确定性,由离散概率密度函数描述(信息=A 的概率为 1/3,信息=C 的概率为 2/3).

作为不确定性数据库建模的最核心思想,可能世界模型被广泛采纳于各种应用之中,并衍生出多种应用相关的模型,特别是针对关系型数据、半结构化数据、流数据和多维数据的模型.

2.2 针对关系型数据的模型

针对关系模型的扩展最为常见,包括 Probabilistic ?-table[9,11]、Probabilistic or-set table[11]、Probabilistic or-set-? table[11]、Probabilistic c-table[13]等.

Probabilistic ?-table 以一个独立的概率字段表示元组的概率,且各元组之间独立.一个特定的数据库实例(也即可能世界实例)的概率等于其所包含的元组的概率乘积和其所不包含的元组的不发生概率的乘积.图 3(a)所示的 Probabilistic ?-table 含 3 个字段 c1、c2 与概率字段,其中概率字段描述元组的发生概率.该表中有 2 个元组,可构成 4 个可能世界实例.Probabilistic ?-table 能够描述存在级不确定性,而 Probabilistic or-set table 则倾向于描述属性级不确定性.在 Probabilistic or-set table 中,元组的属性值被描述为多个候选值之间的"或"关系,可视为离散概率密度函数.以图 3(b)为例,元组 1 的 c2 字段既可取 2,也可取 3,其概率分别为 0.4 和 0.6;元组 2 的 c2 字段既可取 4 也可取 5,其概率分别是 0.2 与 0.8.Probabilistic or-set-? table[13,21]则是上述两种模型的综合体.例如,在图 3(c)中,元组 2 本身具有概率值,而且其 c2 字段既可取 4,也可取 5,概率分别是 0.2 和 0.8.Probabilistic c-table 的定义与 Probabilistic or-set table 比较类似,不同之处在于它是从 c table 衍生而出[22].部分学者也将 probabilistic or-set-? table 命名为 x-relation,它包含若干 x-tuple(无存在级不确定性)或者 maybe x-tuple(有存在级不确定性)[8,23].

c1	c2	概率
1	2	.5
2	3	.6

(a) Probabilistic ?-table

c1	c2
1	(⟨2, .4⟩, ⟨3, .6⟩)
2	(⟨4, .2⟩, ⟨5, .8⟩)

(b) Probabilistic or-set table

c1	c2	概率
1	(⟨2, .4⟩, ⟨3, .6⟩)	
2	(⟨4, .2⟩, ⟨5, .8⟩)	.8

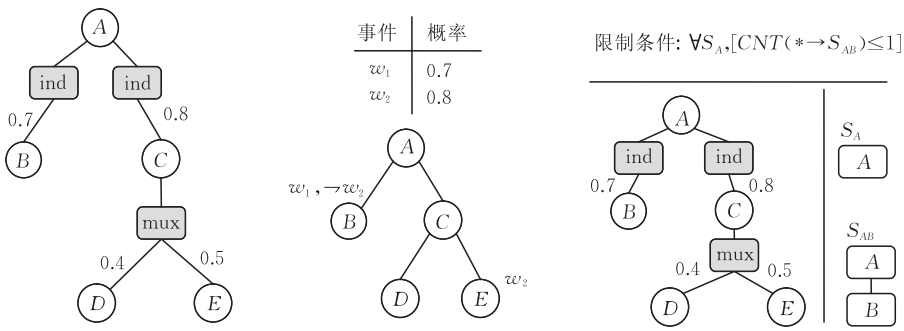
(c) Probabilistic or-set-? table

图 3　基于关系数据的扩展模型

2.3　针对半结构化数据的模型

半结构化数据模型(semistructured data model)能有效描述缺乏严格模式结构的数据[24]. 半结构化数据通常可以用文档树来描述. Dekhtyar 等人[25]提出了一种管理概率半结构化数据(probabilistic semistructured data)的方法,该方法以关系数据库技术为基础,支持丰富的代数查询. 更多的工作则是直接以文档树形式描述不确定性半结构化数据,例如 p-文档模型(p-document model)[26]、概率树模型(Probabilistic tree model)[27-28]、PXDB 模型[29]等.

p-文档模型[26]将概率值附加于文档树的边上,各节点的概率依赖于其祖先的概率,节点之间可以是互斥关系(mux)或相互独立(ind). 在图 4(a)所示的例子中,共有 5 个节点,4 条边. 边 A-B 与 A-C 独立,概率值分别为 0.7 和 0.8;边 C-D 与 C-E 互斥,概率值分别为 0.4 与 0.5. 此时,包含且仅包含节点 A、C、D 的子图的概率为 $(1-0.7)\times 0.8\times 0.4=0.096$;任意子图均不能同时包含 D、E 两个节点.



(a) p-document model　(b) Probabilistic tree model　(c) PXDB model

图 4　常用的不确定性 XML 模型

概率树模型是一个事件驱动的模型[27-28]. 它并不在各节点/边上附加概率值来描述不确定性,而是在各节点附加一系列事件变量,由外部事件的发生与否决定节点的存在性. 图 4(b)描述了一个概率树的例子,共有 2 个外部事件 w_1 和 w_2,其发生概率分别为 0.7 和 0.8. 节点 B 出现的前提条件是事件 w_1 发生且事件 w_2 不发生;节点 E 存在的前提条件是事件 w_2 发生. 由于节点 B 与 E 的存在条件互斥,不存在同时包含节点 B、E 的子图. 包含且仅包含 A、C、D 3 个节点的子图的概率为 $(1-0.8)\times(1-0.7)=0.06$,前提是 w_1、w_2 均不发生. 可以看出,概率树模型的表达能力强于 p-文档模型.

PXDB 模型[29]扩展了 p-文档模型,增加外部约束条件. 在图 4(c)所示的例子中,左下角是一个完整的 p-文档;右下角定义了两个子图 S_A 与 S_{AB},S_A 含节点 A,S_{AB} 含节点 A 与 B. 限制条件为:对于任意一个 S_A 子图,它所包含的 S_{AB} 子图的数量不能够超过 1 个. 因此,尽管 A-B 边在 p-文档中出现了 2 次,但它们无法同时出现在任意一个子图之中.

其他模型还包括 PXML 模型[30-31]、Keulen 等人的概率树模型[32]、PrXML 模型[33]等.

2.4　针对数据流的模型

在数据流模型中,数据到达的速度极快、数据规模极大,仅能够开发一次扫描算法,使用有限内存在线计算查询结果. 在不确定性数据流(Uncertain Data Stream, 或 Probabilistic Data Stream)中,各元组具有不确定性. 文献[34]假设各元组可以在一个离散域 B 中取多值,流上各元组的值是基于这些离散域的一个概率密度函数,例如某元组 t 被描述为(⟨i_1, p_1⟩, …, ⟨i_m, p_m⟩),则 $\forall 1\leq s\leq m$,有 $i_s\in B$, $Pr[i_s]=p_s$,且 $\sum_1^m p_s\leq 1$. 例如,考虑一个温度传感器产生的数据流,环境温度范围(离散域 B)是 $[-30,50]$,则可能的数据流为{(⟨20, 0.4⟩, ⟨22, 0.6⟩), (⟨22, 0.8⟩), (⟨21, 0.2⟩, ⟨23, 0.7⟩), …}. 部分学者将研究重点放在一个基本特例,即 $m=1$[35].

根据窗口定义不同,数据流模型可细分为界标模型、滑动窗口模型. 界标模型的范围从某固定时间

点至当前时间为止,滑动窗口模型仅考虑最新的 W 个元组[36].在各模型中,新元组的到达与旧元组的消逝均引发可能世界实例的大变迁.以上面的环境温度数据流为例,假设窗口大小 $W=2$,在时间点 2 时,需基于元组(〈20,0.4〉,〈22,0.6〉)和(〈22,0.8〉)构造可能世界实例,并回答查询;在时间点 3 时,则基于元组(〈22,0.8〉)和(〈21,0.2〉,〈23,0.7〉)构造可能世界实例,并回答查询;依此类推.

另外,在多数据流应用中,不同数据流上到达的元组之间可能存在相关性,必须整体考虑[37].

2.5 针对多维数据的模型

OLAP 提供了一种多维数据分析手段,能够快速得到复杂的查询统计结果.OLAP 中数据立方(Data Cube)的基本元素是 cuboid.在确定性多维数据模型中,各个事实(fact)必定属于某一个立方体中.但对于处理不精确数据的应用而言,各事实可能无法被准确地定位到立方体中.例如,考虑一个有关汽车销售的多维数据模型,它包括两个维度:city 与 automobile,分别表示购车城市与车体型号.city 维度是一个三级层次结构,国家→省→市.若仅仅知道某辆"奔驰车"是从"浙江北部城市"购买的话,由于"浙江北部城市"包含多个城市,该条记录是不确定性数据,无法存放到事实表中去.文献[38-39]提出了基于可能世界的多维数据模型,以处理这类不确定数据.在这种模型中,上述记录能够被存储于不确定性数据库中,可以基于可能世界语义执行 OLAP 操作(例如切块、上卷等).他们的后续工作也考虑到了元组之间存在相关性的情况[40].

2.6 要求与挑战

不确定数据管理技术采用与确定性数据管理技术截然不同的数据模型,这使得不确定性数据管理技术面临以下挑战:

(1)庞大的可能世界实例集合

毫无疑问,不确定性数据管理所面临的最直接的挑战就是其相对于数据库规模呈指数倍的可能世界实例的数量.假设某不确定性数据库含 N 条元组,各元组独立.当该数据库仅有存在级不确定性,可能世界的数目将达到 2^N 个;而若各个元组还拥有属性级不确定性时,可能世界的数目将远大于 2^N.如果查询要求访问所有的可能世界时,则这个查询开销将会是一个 #P 问题[41].因此,需要在查询的准确度与查询开销之间进行权衡,目标是以较小的计算开销获得高质量的近似结果.

(2)新出现的维度——概率维

概率在不确定性数据管理中扮演多重角色.输入数据可能有概率,表示元组自身或者某字段具有不确定性;输出结果可能有概率,表明该项结果的发生概率;查询定义可以有概率,用于约束查询结果;处理过程也与概率紧密相关.因此,概率维的出现极大地改变了传统的数据处理模式,迫切需要开发新技术进行处理.

(3)不确定性数据管理的理论问题

在不确定数据库管理技术方面仍然存在大量具体问题,特别是理论相关的问题[42].在高效计算复杂条件下的聚集查询(例如含有 HAVING 谓词的聚集查询)处理起来困难较大.灵活的约束条件能够提高数据质量,是不确定性数据管理的重要工具,但是当前仍不具备普遍接受的约束条件定义方式.

3 数据预处理和数据集成

数据预处理与集成是很多数据管理应用不可或缺的组成部分.在传统数据管理领域,数据预处理是针对不准确、不精确的数据进行数据清理、数据转换等处理,从而提升数据质量,最终能够被确定性管理技术所处理.例如,由于多种原因,RFID 读卡器的能够正确读取 RFID 标签的概率约为 60~70%左右,即超过 30%的数据被误读了[43].数据误读的原因很多,包括漏读、多读、脏数据等.因此,RFID 应用的一大关键模块就是数据清洗模块,它将这些不准确的数据转化为准确的数据,再进行后续处理.这种方法被广泛应用于面向不精确数据的数据管理领域.该方法的不足之处主要有两点,首先,从不精确数据到精确数据的转换过程会损失原始数据的部分特征,无法准确反映原始数据的全貌;其次,一种数据清洗技术往往针对特定的原始数据(例如,漏读产生的数据集合),而非对所有数据集合均有效,这使得直接将数据清洗技术从一个应用搬到其他应用的难度加大[3].

数据预处理也包括将准确数据(或者高精度数据)转化为不精确的数据,从而达到隐私保护等特殊目的,典型的例子是基于位置的服务 LBS.作为移动计算领域的核心问题,LBS 在军事、通信、交通、服务业中均获得广泛应用.服务器利用 GPS 等技术获取移动对象的实时位置信息,并提供相应的服务.GPS 技术能够获取精度较高的位置信息,恶意用户完全能够根据某移动对象的运动轨迹推测出一些有

用的信息.例如,若某个对象每天早上沿相同路径移动,则一般来说起点就是家庭地址,而终点则是工作单位地址.

k-匿名模型(k-anonymity model)能够解决这种隐私保护问题[6].该模型最早应用于关系模型,关系中的属性被划分为准标识符(quasi-identifier)和敏感属性(sensitive attribute),使得任一准标识符至少包括 k 个不同元组.位置 k-匿名(location k-anonymity)则是 k-匿名模型在移动对象数据库上的扩展,当某消息被发送时,变换消息的空间信息,使其无法与其他 $k-1$ 条不同消息区分开来[44].Abul 等人[45]定义了(k,δ)-匿名问题((k-δ)-anonymity problem),在任一时刻,总能够找到 k 个对象,聚集在半径为 δ 的圆内,作者同时提出了 NWA 方法进行求解.

数据预处理还包括从原始的不确定数据库中构造一个新数据集合,并在此集合上计算查询结果.这个做法会降低查询结果的准确度,但是能够提高查询处理的效率[46].

数据集成是管理多自治与异构数据源的应用所需面对的普遍挑战[47].当前的数据集成系统仅是传统数据库的扩展,查询以结构化格式定义,数据以传统模型建模,例如关系模型和 XML 模型等.此外,系统也知道原始数据映射到中间模式的确切规则.然而,这些系统无法管理不确定性数据.Philippi 和 Kohler[48]认为,针对生命科学数据库的数据集成系统的最大挑战在于数据的不精确性:数据没有统一的概念模式,数据不完整,缺乏部分信息,且有不确定性.事实上,这种情况也在其他许多领域广泛存在.

Xin 等人最早研究了针对不确定性数据库的数据集成系统,他们认为一个数据集成系统需要在三个层次上处理不确定性[49]:不确定性数据源、不确定性模式映射、不确定性查询.系统的基本框架结构如图 5 所示.

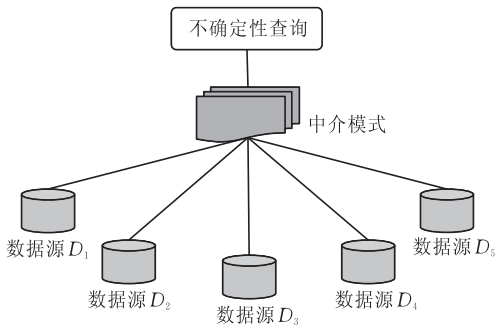


图 5　面向不确定数据的数据集成系统架构

(1)不确定性数据源.不确定性数据源是该数据集成系统最直接的动力.很多情况都可能产生不确定性数据.例如,当数据从非结构化数据源或者半结构化数据源中自动抽取出来时,会引入不确定性;当数据从某些不可靠的或者过时的站点获取时,也会引入不确定性.

(2)不确定性模式映射.数据集成系统利用模式映射技术从多个原始数据源的模式构造中介模式(mediated schema),再利用中介模式回答查询.事实上,中介模式也可能存在不准确性.原因很多,① 用户并非熟练地进行精确映射管理,比如在个人信息管理中;② 对某些字段的理解不充分,因此无法正确映射,例如生物信息;③ 超大数据规模阻止了产生和维护精确映射的可能,例如 Web 数据集成.实际应用中,中介模式往往通过半自动化工具生成,而非由领域专家特别指定.

(3)不确定性查询.查询也可能具备不确定性.特别是在 Web 应用中,查询往往以"关键词"形式被提交,而并非一个定义规范的结构化查询.系统需要将这些查询转化成某些结构化形式,使得它们能够在这些数据源上重新定义.在这个步骤,系统可能会产生多个候选的结构化查询,并且拥有一些不确定性.

4　存储和索引

目前,传统的关系型数据存储技术仍然是实现不确定性数据库的存储任务的主流技术.例如,Orion 项目[50]由 C 语言和 PL/pgSQL 实现,运行于 PostgreSQL 之上;MystiQ 项目[51]具有较好的层次结构,支持 PostgreSQL、Sql Server、DB2 等关系型数据库;MayBMS[52]运行于 PostgreSQL 之上;Trio 项目的初版原型系统 Trio-One 基于标准的关系数据库(postgreSQL)[53]等.图 6 描述了 Trio-One 系统的架构.该架构共有 3 层,用户界面层、Trio 接口与转换层、关系数据库管理系统.用户界面层包括命令行界面与图形用户界面,并将指令以 TriQL 语言(Trio 项目的查询语言)的形式传递给下一层.中间层是通过 Python 实现的 Trio 接口和转换器,它将来自用户界面层的 TriQL 指令翻译成标准的 SQL 语句,发送底层的关系数据库管理系统.关系数据库管理系统存储了一些必要的元数据、存储过程、编码数据表和世系表等,它处理来自中间层的 SQL 语句,将查询结果经由中间层送到用户界面层,并呈现给终端用户.

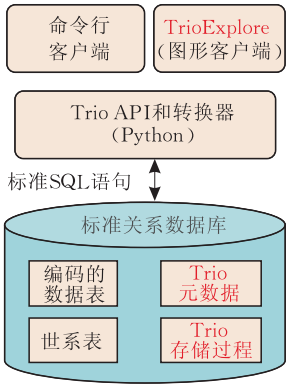


图 6 Trio-One 的系统架构[53]

关系型的存储技术比较直接,但是无法有效处理部分特殊查询,例如数据世系等. 一些研究小组逐渐认识到应该开发新的存储技术. 例如,Trio 项目组认为一些数据库物理设计问题非常重要,包括数据呈现(Data layout)、索引(indexing)、划分(partitioning)、物化视图(materialized view)等. 因此,他们期望新版的 Trio 系统不再是简单地基于现有 RDBMS 之上,而是能够将一些针对性的设计理念融入到数据库物理设计中去,以提高查询处理性能[23].

另外,对于不确定性数据流而言,其存储任务仍旧是构建基于内存的各种概要数据结构,而非基于磁盘的数据结构,以便实时计算查询结果. 半结构化数据的通用存储形式是文档树.

有效的索引能够大幅提高查询效率. 不确定性数据含有概率维度,有些查询仅使用非概率维度的索引,而有些查询却需要对概率维度进行索引. 例如,处理不确定 Top-k 查询过程就仅需以 ranking 函数对特定值维度进行索引[20,54],而处理范围查询[55-58]、最近邻查询(Nearest Neighbor query, NN query)[57,59]等则迫切需要对概率维度进行索引[56].

在移动对象数据库等应用中,物体在某时刻的位置可能并非确定,仅知道在一个较大的区域之内,一般用概率密度函数(pdf)描述. R 树[60]以及其变种是为高维数据创建索引的重要手段,一种方法是以 pdf 的覆盖范围作为该物体的实际空间尺寸,并建立索引. 但是这个方法的可行性不高,原因在于 pdf 所覆盖的范围可能很大,而物体最可能出现的地方却是其中的一小块地方. 构建索引必须考虑 pdf 的分布特征.

概率阈值索引(Probability Threshold Indexing, PTI)[55]能够实现对一维数据的索引. 假设数据库中各元组的属性值对应一个一维区间$[a,b]$,可以设置多个 x-bound. 各个 x-bound 由两根线组成,在线的左边或右边,其总概率均不超过 x. 令 M_i 表示第 i 个元组的 MBR,$M_i.lb(x)$和 $M_i.rb(x)$分别表示 x-bound 的左边和右边值,L_i和 R_i分别表示最左边和最右边的值,f_i表示该元组的概率密度函数,则我们有 $\int_{L_i}^{M_i.lb(x)} f_i(y)\mathrm{d}y \leq x$ 和 $\int_{M_i.rb(x)}^{R_i} f_i(y)\mathrm{d}y \leq x$. 该做法能够避免一些额外的计算.

U-tree 可被视为是 PTI 的多维扩展版本[56,61]. 令 o 表示任一对象,$o.ur$ 表示该对象的范围,$o.pdf(\cdot)$表示该对象的概率密度函数. 当位置 x 在 $o.ur$ 之内时,$o.pdf(x)>0$;否则,$o.pdf(x)=0$. 该方法基于概率约束区域(Probabilistically Constrained Region,PCR)技术,首先将区域划分成若干个规则的矩形,各个矩形分别和一个概率相关,然后利用 U-tree(类似于 R 树)对这些 PCR 进行索引. 这些 PCR 有助于在查询过程中进行剪枝、验证等操作. 图 7 描述了一个不确定对象 o 的示意图. 整个凸多边形就是 $o.ur$,它被 4 条线(l_{1-},l_{1+},l_{2-},l_{2+})切出中间一块空白的矩形,即 PCR,满足条件:在 l_{1-} 左边、l_{1+} 右边、l_{2+} 上面、l_{2-} 下边的区域的发生概率各为 0.2. 例如,针对 l_{1-} 可以有

$$\int_{-\infty}^{l_{1-}}\int_{-\infty}^{\infty} pdf(x,y)\mathrm{d}y\mathrm{d}x=0.2.$$

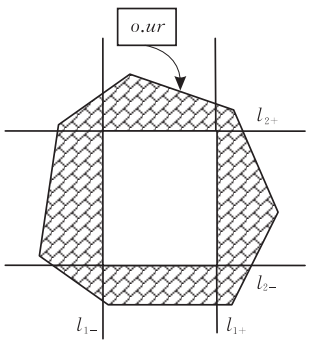


图 7 PCR 例子

Ljosa 等人[57]提出了 APLA-tree 技术,创建多维索引,他们同时也针对 kNN 查询进行了优化.

5 查询与分析

5.1 关系代数处理

关系型不确定性数据管理技术的研究工作起源较早,从 20 世纪 80 年代后期到现在一直在延续. 根据存在级不确定性与属性级不确定性,细分的数据模型就有 Probabilistic ?-table[9,11]、Probabilistic or-set table[11]、Probabilistic c-table 等. 一般可以向数

据库提交类似于 SQL 的查询语句，从而返回查询结果.这其中如何计算查询结果的概率值是一个大的研究问题.

Andritsos 等人提出了查询重写技术来处理查询[62].给定一个 SPJ 查询："select A_1，…，A_n from R_1，…，R_m where W"，可以将其转化为"select A_1，…，A_n，$sum(R_1.prob.*\cdots*R_m.prob)$ from R_1，…，R_m where W group by A_1，…，A_n".新的查询语句中包含了结果查询的产生概率值.

Sen 和 Deshpande 等人则利用构建图模型的方法来处理 SQL 语句[63].数据库上的查询评价问题看作是基于概率图模型上的推断问题.文中应用概率图模型来刻画元组间的相关关系，并应用概率图模型来分解联合概率分布，通过分解后因子表达式中的条件概率分布计算并返回查询概率.

要完整地计算出查询结果有时是非常困难的.研究结果表明，一个含连接操作的查询要么相对于概率数据库在多项式复杂度的时间内被计算出来，要么相对于概率数据库是 ♯P-完全问题[41,64].

文献[65]研究了近似查询及其复杂性.

5.2　世系分析

数据的世系(lineage 或者 provenance)是指数据产生并随着时间推移而演变的整个过程[66].现有工作大多针对确定性数据库，但很多应用(特别是超大规模应用)往往需面对不精确数据集合，而非精确的数据集合，包括科学数据管理、传感器数据管理、近似查询处理、隐私保护等[67].例如，一些大型科研项目必须由多个科研小组分工合作完成，一部分小组的任务是制造并记录大量原始实验数据，另一部分学者基于原始数据、外部数据与中间数据进行分析，生成新数据集合，乃至最终获取查询结果.在整个数据演化过程中，各环节所产生的不确定性不断传递、放大，从而能够极大地影响最终查询结果的质量.

Trio 项目最早研究如何整合不确定性与数据的世系[67].该项目定义了 ULDB(Uncertainty-Lineage Database)，面向世系分析的不确定性数据库，并证明它是完全的(complete)；提出了执行关系操作的算法[23,68].

图 8 是一个关于犯罪现场数据的世系的例子.图 8(a)是目击调查表，张三与李四是两个目击证人，张三看到的可能是宝马或者奔驰，李四看到的是奔驰.图 8(b)是驾驶记录表，记录在某个时段内通过某区域的所有宝马与奔驰车辆以及车主.图 8(c)是通过上述两个表生成的 3 项指控记录.例如，指控记录 42 的世系为(42,1)={(21,2),(32,1)}，表示张三指控赵六，证据是目击调查表的记录 21 的第 2 项与驾驶记录表的记录 32.但由于目击调查表存在不确定性，这项指控本身也存在不确定性.

ID	目击调查 (目击者，车型)	
21	(张三，宝马)‖(张三，奔驰)	?
22	(李四，奔驰)	?

(a) 目击调查表

ID	驾驶记录 (驾驶员，车型)
31	(王五，宝马)
32	(赵六，奔驰)

(b) 驾驶记录表

ID	指控 (目击者，驾驶员)		
41	(张三，王五)	?	(41,1)={(21,1),(31,1)}
42	(张三，赵六)	?	(42,1)={(21,2),(32,1)}
43	(李四，赵六)	?	(43,1)={(22,1),(32,1)}

(c) 指控表

图 8　一个数据世系的例子

当数据的世系比较复杂时，如何计算查询结果的概率就成为一件困难的事情.当不考虑概率因素时，可以为某一个数据世系设计多种查询计划，并且得到相同结果.但是当存在概率因素时，不同查询计划返回的查询结果的概率值却可能不同.其原因是在设计查询计划的过程中未考虑到数据的相关性特征，导致重复计算[23].文献[69]提出了一种数据计算与概率计算解偶的技术，使两者可以分开计算，这样一方面概率值可以采用传统的关系型数据库方法进行计算，另一方面，如果用户并不关注查询结果的概率，则可以节约计算概率值的开销.

在一些大型应用(特别是生物数据库)中，元组的世系可能非常庞大，甚至高达 10MB，而其中大部分数据仅对结果产生较小的影响.文献[70]就考虑以高效的方法近似描述一个元组的数据世系.

5.3　Top-k 查询

面向确定性数据库的 top-k 查询的定义非常清晰：返回 ranking 函数值最大的 k 个元组.但是在不确定性数据库上却存在多种定义方法，例如 U-Top$k$[54]、U-kRanks[54]、PT-k[20]和 Pk-top$k$[36]查询等.U-Top$k$ 查询返回一个长度为 k 的元组矢量，它在所有可能世界中的发生概率最大；U-kRanks 查询返回在各个级别中出现的总概率最大的元组；PT-k 首先定义一个阈值 p，返回所有在可能世界实

例中成为 top-k 的总概率超过阈值的元组；Pk-top$k$ 则返回在所有可能世界实例中成为 top-k 的总概率最大的 k 个元组. 假设一个不确定数据库含有 4 个元组，即$\{t_1=(5,0.8), t_2=(6,0.5), t_3=(8,0.4), t_4=(2,0.4)\}$. 当 $k=2$ 时，U-top$k$ 返回(t_2,t_1)，U-kRanks返回(t_3,t_1)，PT-k 返回(t_1,t_2,t_3)；当 $p=0.3$ 时，Pk-top$k$ 返回(t_1,t_2).

Soliman 等人提出了基于搜索空间的方法来处理 U-Top$k$ 查询与 U-kRanks 查询[54]. 各元组首先按照 ranking 函数从大到小进行排序，然后不断地构造搜索空间，缩小空间的范围，最终获得查询结果. Cheng 等人针对 U-Top$k$ 查询提出了一种动态维护的结构，支持元组的插入与删除[71]. Hua 等人针对 PT-k 查询提出了构造 dominant 集合的方法[20]. 在其后续工作中，也提出了近似解决方案[72]. 上述算法均需要预先对数据进行排序. Jin 等人[36]提出了面向数据流应用的 Top-k 查询处理算法，不仅能够解决上述 4 种查询，而且是一次遍历算法，能够处理数据流应用.

上述 4 种查询均可视为是从单个数据库表中获取的数据. Ré 等人[73]则研究了在多表之间做连接操作的情况. 各个表的数据并不精确，存在不一致性. 他们的基本想法是并行地运行多个 Monte-Carlo 模拟器，每一个对应于一个候选的答案，再计算各个候选答案的近似概率.

5.4 Skyline 查询

Skyline 查询[74]能用于解决多准则决策(Muli-Criteria Decision-Making，MCDM)问题. 给定一个确定性的 n-维数据集合 D，任一点 d 可被表示为$(d.D_1,\cdots,d.D_n)$. Skyline 查询返回数据集合 S，$S\subseteq D$，则 $\forall u\in S$，不存在其它点 v，满足(1)对于任一维度 $i(1\leq i\leq n)$，$u.D_i\leq v.D_i$；(2)存在一个维度 $j(1\leq j\leq n)$ 使得 $u.D_j<v.D_j$.

近来，面向不确定性数据的 skyline 查询处理问题也得到了关注. 各个元组的值并不确定，以概率密度函数描述. Pei 等人根据可能世界模型定义了概率 skyline 查询[75]. 不确定性数据库会衍生出很多可能世界实例，各元组在各可能世界实例中可能是 skyline 点，也可能不是 skyline 点. 由此，p-skyline 查询$(0\leq p\leq 1)$被定义为返回所有成为 skyline 点的概率超过 p 的数据点. 文献[75]同时提出了两种解决方法：自下而上方法(bottom-up method)和自上而下方法(top-down method)，分别采用不同的定界、剪枝、精化等启发式规则进行迭代处理.

Lian 与 Chen 等人则考虑了如何在不确定数据集合上处理 reverse skyline 查询[76]的问题[77]. 确定性 reverse skyline 查询返回在数据库中所有的动态 skyline 包含给定查询点的数据点. 相应的，概率 reverse skyline 查询(Probabilistic Reverse Skyline，PRS)被定义为：给定一个概率阈值 $p(0\leq p\leq 1)$和一个查询对象 q，返回所有对象 v，使得对象 q 为 v 的动态 skyline 点的概率不低于阈值 p. 文献[77]将每一个数据对象看作是一个不确定区域，应用确定情形下的 BBS 算法[78]进行空间剪枝，应用用户定义的概率阈值 p 进行概率剪枝，得到候选的 PRS 点进行精化，最后输出查询结果.

5.5 数据流

在不确定性数据流中，各元组以概率形式表达不确定性，可以是单一的概率值，也可以是复杂的概率密度函数. 数据流模型中，数据到达速率极快，数据量极大，要求设计单遍扫描算法，以低空间复杂度实时处理查询.

传统的面向确定性数据流的方法经过改装之后能够应用于不确定性数据流应用之中. 例如，AMS sketch[79]能用于处理聚集查询，特别是 F_2 问题；FM sketch[80]能用于求解数据流上的相异元素个数；Cluster Feature 被广泛用于设计各种在线聚类算法[81-82]. Cormode 等人发展了 AMS sketch 和 FM sketch 方法，引入了概率参数，构造了 pAMS 结构和 pFM 结构，能够处理不确定性数据流上的相应查询[35]. Aggarwal 和 Yu 改进了 CF 方法，提出了 ECF(Error-based CF)方法，处理数据流上的聚类问题[83].

文献[34]最早在数据流上计算简单的聚集函数，特别是 AVG 函数. 他们的方法比较复杂，主要采用生成函数技术(generating functions)，难以进行扩展以解决其它问题. 他们的后续工作[84]能够解决更多的聚集查询问题，包括 F_0 和 F_2，并且提高了 AVG 函数的查询效率. Zhang 等人定义了在不确定性数据流上查询频繁元素的问题，并设计解决方法[85].

Jin 等人最早提出了面向滑动窗口模型的查询处理方法[36]. 如前所述，存在各种 top-k 查询. 他们提出了一种针对 top-k 查询的框架. 首先，可以针对各种 top-k 查询设计 compact set，各个 compact set 含有一部分数据. 这个 compact set 具有两个特性：(1)能够计算 top-k 查询结果；(2)能够增量维护. 但是 compact set 仍然不足以回答滑动窗口模型，因此，

可以将多个 compact sets 进行压缩，降低空间复杂度与时间复杂度. 他们提出的 SCSQ-buffer 策略在时间复杂度与空间复杂度上均是优秀的.

事件数据流是一种重要的数据流. 传统方法大多仅能够处理准确的数据流，例如 Cayuga[86]、SASE[87]、SnoopIB[88]等. 文献[89]能够处理概率数据流上的查询分析，但是他们的工作主要集中于少数几种查询，例如选择、映射和聚集查询等. Lahar 系统则能够处理不精确的数据流，特别适用于 RFID 等环境中，能够处理误读数据、冲突数据、粒度不匹配等不精确信息[37].

5.6 OLAP

OLAP 技术使用多维模型. 事实表(fact table)中的各个事实(fact)可被视为多维空间的一个点. 然而，当存在不确定因素时，各事实记录并不表现为一个点，而是跨越多个维度值，成为多维空间里的一个“区域”. 例如，可以以“籍贯”为维度对在校学生进行 OLAP 分析. 籍贯是层次型的，包括省级与区县级. 若学生均已准确登记其区县级的籍贯信息，则各事实都是空间的一个点；若某学生仅登记其籍贯为“浙江省”，则该事实表现为空间里的“区域”.

Burdick 等人[38-40,90]基于可能世界模型来处理上述不精确的 OLAP 查询. 根据可能世界模型的语义，一个含不精确事实的数据库可被描述为多个可能世界，各可能世界仅含有精确事实，最终的查询结果能够从这些可能世界实例中获取. 例如，首先在一个不精确的数据库 D 上生成所有可能世界实例 $w_1,\cdots,w_n$，各可能世界实例 w_i 的发生概率为 $p(w_i)$；在各可能世界实例上分别执行查询 Q，得到 $Q(w_i)$；最后，对这些查询结果进行聚集，得到 $\sum p(w_i)\times Q(w_i)$.

当数据集合较大时，通过枚举所有可能世界实例的方法并不可行. Burdick 等人首先研究了在数据独立情况下的优化方法. 该方法首先构造 EDB(Extended DataBase)，然后再单遍扫描 EDB，快速计算查询结果[39]. 文献[90]则改进了计算 EDB 的方法，进一步提高查询效率.

数据独立的情况较易处理，而若数据之间存在约束条件时则更为复杂. 典型的约束条件如：“张三和李四的籍贯都是浙江省，但是他俩来自不同地级市”，则不可能存在如下可能世界实例：{(张三，文成县)，(李四，平阳县)，…}①. 首先在一个不精确的数据库 D 上生成所有可能世界实例 $w_1,\cdots,w_n$，然后运用规则，仅剩下 m 个有效的实例，各可能世界实例 w_i 的发生概率为 $p(w_i)$；对各有效可能世界实例上生成的查询结果进行聚集，得到 $\sum p(w_i)\times Q(w_i)$. Burdick 等人[40]提出了优化方法，基于原始数据库 D 构造 MDB(Marginal DataBase)，然后能够利用 MDB 以单遍扫描的方式迅速获得查询结果. 在 MDB 中，各个事实均被首先计算了边缘概率值，以加速查询处理.

5.7 数据挖掘

面向不确定性数据的挖掘算法越来越引起人们的关注，主要研究内容包括聚类技术和分类技术. 数据的不确定性能够显著影响数据挖掘应用的结果.

聚类技术得到了广泛的研究. Kriegel 等人意识到不确定因素会影响两个元组之间的距离，因此重新定义了距离公式. 令 $p(\overline{X},\overline{Y})$ 表示元组 $\overline{X}$ 和 $\overline{Y}$ 之间的距离的概率密度函数，则 $\overline{X}$ 与 $\overline{Y}$ 间的距离在 (a,b) 之间的概率为 $p(a\leq d(\overline{X},\overline{Y})\leq b)=\int_a^b p(\overline{X},\overline{Y})(z)\,\mathrm{d}z$. 基于上述距离公式，他们改进了 DBSCAN[91]算法，提出了 FDBSCAN 算法[92]；改进了 OPTICS[93]算法，提出了 FOPTICS 算法，解决层次聚类问题[94]. Ngai 等人提出了 UK-means 算法[95]，基本思想与 k-means 类似：各数据点将被距离最近的簇吸收. 为了提高计算效率，UK-means 算法将数据点可能出现的区域用最小边界矩形(MBR)描述，设计剪枝策略减少运算量. 文献[96]提出了针对移动对象的聚类方法，主要思路也是计算从各个对象到簇中心点的期望距离. 文献[97]提出的方法采用一个函数来计算不确定的点到任意一个中心的距离的期望值，然后再运用传统的聚类方法进行计算.

上述工作均面向传统的静态数据库，无法适应高速的数据流模型. Aggarwal 与 Yu 将 CF 结构扩展为 ECF 结构，增加了描述不确定性的组成部分，以解决数据流上的聚类问题[83]. 算法主要改进了新到点与簇中心点的距离与每个簇的接收半径的计算过程.

Bi 等人[98]提出了一种面向不确定性数据库的分类算法，它基于支持向量机(Support Vector Machine，SVM)技术.

6 总　结

近年来，不确定性数据广泛出现在诸多应用领

① 文成县与平阳县均属于浙江省温州市.

域之中，例如传感器网络、RFID 应用、数据集成、LBS、Web 应用等，为数据管理技术提出了如下要求.(1)丰富的数据类型，包括结构化的关系数据、半结构化数据、流数据、移动对象数据以及其他领域相关数据等.(2)广泛性，数据处理的各个环节均可能存在不确定性，包括原始数据采集、预处理与集成、存储与索引、查询分析处理等.(3)繁多的查询类别，例如关系操作、数据世系、数据挖掘、模式匹配等.(4)概率维的冲击，概率维与普通数据维不同，会极大地影响查询定义、查询处理、结果定义等关键环节.

可能世界模型是不确定性数据管理领域中最通用的数据模型.尽管存在针对不同应用的具体模型，但这些模型通常均可转化为可能世界模型，如第 2 节所示.可能世界模型将问题域划分成若干可能世界实例，所有可能世界实例的发生概率之和为 1，但各可能世界实例内部不存在不确定性.因此，可以在各可能世界实例上分别处理查询，然后再合并生成最终的查询结果.而事实上，由于可能世界实例的数量过于庞大，无法利用上述方法进行计算.目前，典型的不确定性解决方案大量采用排序、剪枝等启发式技术，从而大大减少计算量，快速获取查询结果.

在不确定性数据管理领域仍然存在很多工作有待完成：

(1)许多在确定性数据管理领域所遇到的问题在不确定数据管理领域也非常重要，可以将这些查询问题搬到不确定数据管理领域上，以寻求解决方案.

(2)由于概率维的存在，不确定数据管理领域存在一些特有的查询问题，需要找出这些查询并设计处理算法.例如，现有 4 种不确定 top-k 查询，包括 U-Top$k$、U-kRanks、PT-k、Pk-Top$k$.

(3)寻求在新模型定义下的查询处理方法.尽管大家都使用可能世界模型，但是不同应用场景下的模型仍然有所区别.事实上，不确定性 XML 处理技术的提升总是伴随着不确定性 XML 模型定义的更新.

(4)需要针对不确定性查询设计更优秀的算法.

参 考 文 献

[1] Deshpande A, Guestrin C, Madden S, Hellerstein J M, Hong W. Model-driven data acquisition in sensor networks//Proceedings of the 30th International Conference on Very Large Data Bases. Toronto, 2004: 588-599

[2] Li Jian-Zhong, Li Jin-Bao, Shi Sheng-Fei. Concepts, issues and advance of sensor networks and data management of sensor networks. Journal of Software, 2003, 14(10): 1717-1727(in Chinese)
(李建中，李金宝，石胜飞. 传感器网络及其数据管理的概念、问题与进展. 软件学报，2003，14(10)：1717-1727)

[3] Gu Yu, Yu Ge, Zhang Tian-Cheng. RFID complex event processing techniques. Journal of Frontiers of Computer Science and Technology, 2007, 1(3): 255-267(in Chinese)
(谷峪，于戈，张天成. RFID 复杂事件处理技术. 计算机科学与探索，2007，1(3)：255-267)

[4] Madhavan J, Cohen S, Xin D, Halevy A, Jeffery S, Ko D, Yu C. Web-scale data integration: You can afford to pay as you go//Proceedings of the 33rd Biennial Conference on Innovative Data Systems Research. Asilomar, 2007: 342-350

[5] Liu Ling. From data privacy to location privacy: Models and algorithms (tutorial)//Proceedings of the 33rd International Conference on Very Large Data bases. Vienna, 2007: 1429-1430

[6] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information (abstract)//Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Seattle, 1998: 188

[7] Cavallo R, Pittarelli M. The theory of probabilistic databases//Proceedings of the 13th International Conference on Very Large Data Bases. Brighton, 1987: 71-81

[8] Barbara D, Garcia-Molina H, Porter D. The management of probabilistic data. IEEE Transactions on Knowledge and Data Engineering, 1992, 4(5): 487-502

[9] Fuhr N, Rolleke T. A probabilistic relational algebra for the integration of information retrieval and database systems. ACM Transactions on Information Systems, 1997, 15(1): 32-66

[10] Zimanyi E. Query evaluation in probabilistic databases. Theoretical Computer Science, 1997, 171(1-2): 179-219

[11] Lakshmanan L V S, Leone N, Ross R, Subrahmanian V S. ProbView: A flexible database system. ACM Transactions on Database Systems, 1997, 22(3): 419-469

[12] Abiteboul S, Kanellakis P, Grahne G. On the representation and querying of sets of possible worlds. ACM SIGMOD Record, 1987, 16(3): 34-48

[13] Green T J, Tannen V. Models for incomplete and probabilistic information. IEEE Date Engineering Bulletin, 2006, 29(1): 17-24

[14] Babcock B, Babu S, Datar M, Motwani R, Widom J. Models and issues in data stream systems//Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Madison, 2002: 1-16

[15] Jin Che-Qing, Qian Wei-Ning, Zhou Ao-Ying. Analysis and management of streaming data: A survey. Journal of Software, 2004, 15(8): 1172-1181(in Chinese)

(金澈清，钱卫宁，周傲英. 流数据分析与管理综述. 软件学报，2004，15(8)：1172-1181)

[16] Ré C，Suciu D. Management of data with uncertainties//Proceedings of the 16th ACM Conference on Information and Knowledge Management. Lisbon，2007：3-8

[17] Dalvi N，Suciu D. Management of probabilistic data foundations and challenges//Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Beijing，2007：1-12

[18] Aggarwal C C，Yu P S. A survey of uncertain data algorithms and applications. IBM Research Report，October 31，2007

[19] Pei J，Hua M，Tao Y F，Lin X M. Query answering techniques on uncertain and probabilistic data：Tutorial summary//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. Vancouver，2008：1357-1364

[20] Hua M，Pei J，Zhang W J，Lin X M. Efficiently answering probabilistic threshold top-*k* queries on uncertain data//Proceedings of the 24th IEEE International Conference on Data Engineering. 2008：1403-1405

[21] Sarma A D，Benjelloun O，Halevy A，Widom J. Working models for uncertain data//Proceedings of the 22nd IEEE International Conference on Data Engineering. Atlanta，2006：7

[22] Imieliński O，Jr W L. Incomplete information in relational databases. Journal of ACM，1984，31(4)：761-791

[23] Benjelloun O，Sarma A D，Halevy A，Widom J. ULDBs：Databases with uncertainty and lineage//Proceedings of the 32nd International Conference on Very Large Data Bases. Seoul，2006：953-964

[24] Abiteboul S，Buneman P，Suciu D. Data on the Web：From Relations to Semistructured Data and XML. San Francisco，CA：Morgan Kaufmann，1999

[25] Dekhtyar A，Goldsmith J，Hawkes S R. Semistructured probabilistic databases//Proceedings of the 13th International Conference on Statistical and Scientific Database Management. Tokyo，2001：36-45

[26] Nierman A，Jagadish H V. ProTDB：Probabilistic data in XML//Proceedings of the 28th International Conference on Very Large Data Bases. Hong Kong，China，2002：646-657

[27] Abiteboul S，Senellart P. Querying and updating probabilistic information in XML//Proceedings of the 9th International Conference on Extending Database Technology：Advances in Database Technology. Munich，2006：1059-1068

[28] Senellart P，Abiteboul S. On the complexity of managing probabilistic XML data//Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Beijing，2007：283-292

[29] Cohen S，Kimelfeld B，Sagiv Y. Incorporating constraints in probabilistic XML//Proceedings of the 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Vancouver，2008：109-118

[30] Hung E，Getoor L，Subrahmanian V S. PXML：A probabilistic semistructured data model and algebra//Proceedings of the 19th IEEE International Conference on Data Engineering. Bangalore，2003：467-478

[31] Hung E，Getoor L，Subrahmanian V S. Probabilistic interval XML. ACM Transactions on Computational Logic，2007，8(4)：24

[32] van Keulen M，de Keijzer A，Alink W. A probabilistic XML approach to data integration//Proceedings of the 21st IEEE International Conference on Data Engineering. Tokyo，2005：459-470

[33] Kimelfeld B，Kosharovsky Y，Sagiv Y. Query efficiency in probabilistic XML models//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. Vancouver，2008：701-714

[34] Jayram T S，Kale S，Vee E. Efficient aggregation algorithms for probabilistic data//Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. New Orleans，2007：346-355

[35] Cormode G，Garofalakis M. Sketching probabilistic data streams//Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. Beijing，2007：281-292

[36] Jin Che-Qing，Yi Ke，Chen Lei，Yu Xu，Lin Xue-Min. Sliding-window top-*k* queries on uncertain Streams. Proceedings of the VLDB Endowment，2008，1(1)：301-312

[37] Ré C，Letchner J，Balazinska M，Suciu D. Event queries on correlated probabilistic streams//Proceedings of the 27th ACM SIGMOD International Conference on Management of Data. Vancouver，2008：715-728

[38] Burdick D，Deshpande P M，Jayram T S，Ramakrishnan R，Vaithyanathan S. OLAP over uncertain and imprecise data//Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim，2005：970- 981

[39] Burdick D，Deshpande P M，Jayram T S，Ramakrishnan R，Vaithyanathan S. OLAP over uncertain and imprecise data. The VLDB Journal，2007，16(1)：123-144

[40] Burdick D，Doan A，Ramakrishnan R，Vaithyanathan S. Olap over imprecise data with domain constraints//Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna，2007：39-50

[41] Dalvi N，Suciu D. The dichotomy of conjunctive queries on probabilistic structures//Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Beijing，2007：293-302

[42] Dalvi N，Suciu D. Efficient query evaluation on probabilistic databases//Proceedings of the 30th International Conference on Very Large Data Bases. Toronto，2004：864-875

[43] Jeffery S R，Garofalakis M，Franklin M J. Adaptive cleaning for RFID data streams//Proceedings of the 32nd International Conference on Very Large Data Bases. Seoul，2006：163-174

[44] Gruteser M, Grunwald D. Anonymous usage of location-based services through spatial and temporal cloaking//Proceedings of the 1st International Conference on Mobile Systems, Applications and Services. San Francisco, 2003: 31-42

[45] Abul O, Bonchi F, Nanni M. Never walk alone: Uncertainty for anonymity in moving objects databases//Proceedings of the 24th IEEE International Conference on Data Engineering. Cancun, 2008: 376-385

[46] Cheng R, Chen Jinchuan, Xie Xike. Cleaning uncertain data with quality guarantees. Proceedings of the VLDB Endowment, 2008, 1(1): 722-735

[47] Halevy A, Rajaraman A, Ordille J. Data integration: The teenage years//Proceedings of the 32nd International Conference on Very Large Data Bases. Seoul, 2006: 9-16

[48] Philippi S, Kohler J. Addressing the problems with life-science databases for traditional uses and systems biology. Nature Reviews Genetics, 2006, (7): 481-488

[49] Xin D, Halevy A, Yu C. Data integration with uncertainty//Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, 2007: 687-698

[50] Singh S, Cheng R, Prabhakar S. U-DBMS: A database system for managing constantly-evolving data//Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim, 2005: 1271-1274

[51] Boulos J, Dalvi N, Mandhani B, Mathur S, Ré C, Suciu D. Mystiq: A system for finding more answers by using probabilities//Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. Baltimore, 2005: 891-893

[52] Antova L, Koch C, Olteanu D. MayBMS: Managing incomplete information with probabilistic world-set decompositions//Proceedings of the 23rd International Conference on Data Engineering. Istanbul, 2007: 1479-1480

[53] Mutsuzaki M, Theobald M, de Keijzer A, Widom J, Agrawal P, Benjelloun O, Sarma A D, Murthy R, Sugihara T. TrioOne: Layering uncertainty and lineage on a conventional DBMS//Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research. Asilomar, 2007: 269-274

[54] Soliman M A, Ilyas I F, Chang K C. Top-*k* query processing in uncertain databases//Proceedings of the 23rd IEEE International Conference on Data Engineering. Istanbul, 2007: 896-905

[55] Cheng R, Xia Y, Prabhakar S, Shah R, Vitter J S. Efficient indexing methods for probabilistic threshold queries over uncertain data//Proceedings of the 30th International Conference on Very Large Data Bases. Toronto, 2004: 876-887

[56] Tao Y, Cheng R, Xiao X, Ngai W K, Kao B, Prabhakar S. Indexing multi-dimensional uncertain data with arbitrary probability density functions//Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim, 2005: 922-933

[57] Ljosa V, Singh A K. APLA: Indexing Arbitrary Probability Distributions//Proceedings of the 23rd IEEE International Conference on Data Engineering. Istanbul, 2007: 946-955

[58] Chen J, Cheng R. Efficient evaluation of imprecise location-dependent queries//Proceedings of the 23rd IEEE International Conference on Data Engineering. Istanbul, 2007: 586-595

[59] Cheng R, Chen J, Mokbel M, Chow C. Probabilistic verifiers: Evaluating constrained nearest-neighbor queries over uncertain data//Proceedings of the 24th IEEE International Conference on Data Engineering. Cancun, 2008: 973-982

[60] Guttman A. R-trees: A dynamic index structure for spatial searching//Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data. Boston, 1984: 47-57

[61] Tao Y, Xiao X, Cheng R. Range search on multidimensional uncertain data. ACM Transactions on Database Systems, 2007, 32(3): 15

[62] Andritsos P, Fuxman A, Miller R J. Clean answers over dirty databases: A probabilistic approach//Proceedings of the 22nd IEEE International Conference on Data Engineering. Atlanta, 2006: 30

[63] Sen P, Deshpande A. Representing and querying correlated tuples in probabilistic databases//Proceedings of the 23rd IEEE International Conference on Data Engineering. Istanbul, 2007: 596-605

[64] Gradel E, Gurevich Y, Hirsch C. The complexity of query reliability//Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Seattle, 1998: 227-234

[65] Koch C. Approximating predicates and expressive queries on probabilistic databases//Proceedings of the 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Vancouver, 2008: 99-108

[66] Woodruff A, Stonebraker M. Supporting fine-grained data lineage in a database visualization environment//Proceedings of the 13th IEEE International Conference on Data Engineering. Birmingham, 1997: 91-102

[67] Widom J. Trio: A system for integrated management of data, accuracy, and lineage//Proceedings of the 2nd Biennial Conference on Innovative Data Systems Research. Asilomar, 2005: 262-276

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

[69] Sarma A D, Theobald M, Widom J. Exploiting lineage for confidence computation in uncertain and probabilistic databases//Proceedings of the 24th IEEE International Conference on Data Engineering. Cancun, 2008: 1023-1032

[70] Ré C, Suciu D. Approximate lineage for probabilistic databases. Proceedings of the VLDB Endowment, 2008, 1(1): 797-808

[71] Chen J, Yi K. Dynamic structures for top-k queries on uncertain data//Proceedings of the 18th International Symposium on Algorithms and Computation. Sendai, 2007: 427-438

[72] Hua Ming, Pei Jian, Zhang Wenjie, Lin Xuemin. Ranking queries on uncertain data: A probabilistic threshold approach//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. Vancouver, 2008: 673-686

[73] Ré C, Dalvi N, Suciu D. Efficient top-k query evaluation on probabilistic data//Proceedings of the 23rd IEEE International Conference on Data Engineering. Istanbul, 2007: 886-895

[74] Borzsonyi S, Kossmann D, Stocker K. The skyline operator//Proceedings of the 17th IEEE International Conference on Data Engineering. Heidelberg, 2001: 421-430

[75] Pei J, Jiang B, Lin X, Yuan Y. Probabilistic skylines on uncertain data//Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, 2007: 15-26

[76] Dellis E, Seeger B. Efficient computation of reverse skyline queries//Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, 2007: 291-302

[77] Lian X, Chen L. Monochromatic and bichromatic reverse skyline search over uncertain databases//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. Vancouver, 2008: 213-226

[78] Deng K, Zhou X, Shen H T. Multi-source skyline query processing in road networks//Proceedings of the 23rd International Conference on Data Engineering. Istanbul, 2007: 796-805

[79] Alon N, Matias Y, Szegedy M. The space complexity of approximating the frequency moments//Proceedings of the 28th Annual ACM Symposium on Theory of Computing. Philadelphia, 1996: 20-29

[80] Flajolet P, Martin G N. Probabilistic counting algorithms for database applications. Journal of Computer and System Sciences, 1985, 31(2): 182-209

[81] Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases//Proceedings of the 15th ACM SIGMOD International Conference on Management of Data. Montreal, 1996: 103-114

[82] Aggarwal C C, Han J, Wang J, Yu P S. A framework for clustering evolving data streams//Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. Berlin, 2003: 81-92

[83] Aggarwal C C, Yu P S. A framework for clustering uncertain data streams//Proceedings of the 24th IEEE International Conference on Data Engineering. Cancun, 2008: 150-159

[84] Jayram T S, McGregor A, Muthukrishnan S, Vee E. Estimating statistical aggregates on probabilistic data streams//Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Beijing, 2008: 243-252

[85] Zhang Qin, Li Fei-Fei, Yi Ke. Finding frequent items in probabilistic data//Proceedings of the 27th ACM SIGMOD International Conference on Management of Data. Vancouver, 2008: 819-832

[86] Demers A J, Gehrke J, Hong M, Riedewald M, White W M. Towards expressive publish/subscribe systems//Proceedings of the 9th International Conference on Extending Database Technology: Advances in Database Technology. Munich, 2006: 627-644

[87] Wu E, Diao Y, Rizvi S. High-performance complex event processing over streams//Proceedings of the ACM SIGMOD International Conference on Management of Data. Chicago, 2006: 407-418

[88] Adaikkalavan R, Chakravarthy S. Snoopib: Interval-based event specification and detection for active databases. Data Knowledge Engineering, 2006, 59(1): 139-165

[89] Kanagal B, Deshpande A. Online filtering, smoothing and probabilistic modeling of streaming data//Proceedings of the 24th IEEE International Conference on Data Engineering. Cancun, 2008: 1160-1169

[90] Burdick D, Deshpande P M, Jayram T S, Ramakrishnan R, Vaithyanathan S. Efficient allocation algorithms for OLAP over imprecise data//Proceedings of the 32nd International Conference on Very Large Data Bases. Seoul, 2006: 391-402

[91] Ester M, Kriegel H P, Sander J, Xu X. A density based algorithm for discovering clusters in large spatial databases with noise//Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland, 1996: 226-231

[92] Kriegel H P, Pfeifle M. Density-based clustering of uncertain data//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago, 2005: 672-677

[93] Ankerst M, Breunig M M, Kriegel H P, Sander J. OPTICS: Ordering points to identify the clustering structure//Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. Philadelphia, 1999: 49-60

[94] Kriegel H P, Pfeifle M. Hierarchical density-based clustering of uncertain data//Proceedings of the 5th International Conference on Data Mining. Houston, 2005: 689-692

[95] Ngai W K, Kao B, Chui C K, Cheng R, Chau M, Yip K Y. Efficient clustering of uncertain data//Proceedings of the 6th International Conference on Data Mining. Hong Kong, China, 2006: 436-445

[96] Chau M, Cheng R, Kao B, Ngai J. Uncertain data mining: An example in clustering location data//Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore, 2006: 199-204

[97] Cormode G, McGregor A. Approximation algorithms for clustering uncertain data//Proceedings of the 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Vancouver, 2008: 191-200

[98] Bi Jinbo, Zhang Tong. Support vector classification with input data uncertainty//Advances in Neural Information Processing Systems. Vancouver, 2005, 17: 161-168

ZHOU Ao-Ying, born in 1965, professor, Ph. D. supervisor. His research interests focus on data management and information system, inclusive of Web data management, Chinese Web infrastructure, Web searching and mining, data streaming and mining, complex event processing and real-time business intelligence, uncertain data management and applications, data-intensive computing, distributed storage and computing, peer to peer computing and management, Web service.

JIN Che-Qing, born in 1977, Ph. D., associate professor. His research interests include data stream management, uncertain data management.

WANG Guo-Ren, born in 1966, professor, Ph. D. supervisor. His research interests include XML management, bioinformatics, distributed database, and parallel computing.

LI Jian-Zhong, born in 1950, professor, Ph. D. supervisor. His research interests include database, parallel computing.

Background

This paper surveys the recent research work on uncertain data management that belongs to the database category. Data uncertainty widely appears in various applications, inclusive of economy, military, logistic, finance and telecommunication etc. The reasons for uncertain data include, but are not limited to the following: Imprecise data caused by the physical devise, network or environment; Using a coarse-grained dataset; To meet the special application requirement; Incomplete dataset; Data integration. Thus, it is critical to develop new techniques to manage such uncertain database.

The research of management on uncertain database starts from the late 80's last century, and becomes a very hot field today. The work in the early stage focused on extending the relational model with an additional probability field to process SQL like queries, but now it has been developed to a quite boarder range. Besides the relational data, new data types such as semistructured data, streaming data, and moving objects are also studied intensively, which leads to numerous novel sophistical query processing issues. However, neither the traditional techniques for deterministic data, nor the techniques for probabilistic relational database are capable of handling such query tasks efficiently.

There are already a few survey papers on management of uncertain database with different emphasis recently. Ré and Suciu summarized some big challenges in this field in 2007. Dalvi and Suciu pointed out the foundation and challenges with the analysis in theory in 2007. Aggarwal and Yu focused on algorithms and applications. The literature by Pei et al. mainly aimed at their own work.

Contrarily, this paper surveys present work according to a general way of processing uncertain database, including modeling, preprocessing and cleaning, storage and indexing, and query processing. At first, several uncertain models for different data types are proposed, stemming from the core *possible world* semantics, following which the concepts for the data preprocessing and cleaning are also introduced. After outlining the storage and indexing techniques, the work for concrete query tasks are listed, inclusive of relational operator, data lineage, skyline query, ranking query, stream query, OLAP, and data mining.