

不确定性数据管理的要求与挑战

李建中¹ 于戈² 周傲英³

¹哈尔滨工业大学

²东北大学

³华东师范大学

关键词：不确定性数据 可能世界模型

无所不在

近几十年来，数据管理技术发展迅猛，在国民经济建设中起到了突出作用。以Oracle、DB2、SQL Server等为代表的大型关系数据库管理系统（Relational Database Management System, RDBMS）更是诸多大型信息管理系统、客户关系管理软件不可或缺的核心部分。同时，以可扩展标记语言（Extensible Markup Language, XML）为代表的半结构化数据管理技术也在数据交换和缺乏严格结构的数据管理方面占据一席之地。上述技术均对数据质量、待处理数据的准确性要求非常高。当原始数据的质量不高时，需要先经过预处理过程提升数据质量。以部门人事管理系统为例，员工的个人资料、薪酬待遇和日常考核等信息必须准确。但在诸如经济、军事和电信等领域，数据的不确定性普遍存在，其存在性未知而且各属性值存在误差。尽管数据预处理能够提升原始数据集合的质量，但也可能会丧失原始数据集合的部分性质，导致无法返回高质量的查询结果^[1]。典型的应用背景如下。

传感器网络与射频识别电子标签

传感器网络（Sensor Network）^[2]与无线射频识别（Radio Frequency Identification, RFID）^[3]是两类新兴的数据收集和传输技术，在工业、军事等领域中有着重要应用。传感器

网络中分布着众多低成本的传感器节点，相互之间以无线网络方式通讯，可用来分析处理数据；无线射频识别利用它的阅读器以非接触方式读取附近的无线射频识别标签（RFID tag），改变了传统的基于条形码的识别方式。困扰这两类应用的难题就是数据质量问题。传感器节点体积小、功耗低、主要使用低成本电子元器件，因而采集到的数据精度不高；在实用应用中，无线射频识别阅读器的误读率高达30%~40%^[4]。另外，复杂多变的工作环境也会降低原始数据的质量。在无线网络环境中，数据传输的准确性受带宽、传输延时、能量等因素影响，并不稳定。当查询任务需同时考虑来自多个传感器或无线射频识别阅读器的数据时，数据可能不一致，从而增加了数据处理的难度。

互联网数据

互联网上的信息资源极为丰富，而且这些信息一直在不断地膨胀，乃至于有人将互联网称为史上最大规模的数据库。根据2009年1月中国互联网信息中心（China Internet Network Information Center, CNNIC）的调查报告，截至2008年底，中国网站总数为287.8万个，全国网页总数约为160.9亿，较2007年增长90%，网页字数为460,217,386,099KB。但是互联网数据的质量却不尽如人意。作为一个典型的分散管理系统，互联网中并不存在一个统一的信息

发布机构,各网站均可自由发布和维护信息。因此,当信息维护机构不同、信息更新不及时、工作人员误操作时,极易导致不同数据源(或者同一数据源内部)对同一对象描述的不一致。同时,互联网数据规模庞大,需要借助自然语言处理与识别技术从网页中自动抽取信息,因此所获得的结果也存在不准确性。

基于位置服务

基于位置的服务(Location-Based Service, LBS)是移动计算领域的核心问题。位置服务跟踪移动物体(或者用户),然后将物体(或用户)的位置在电子地图上定位,以此为基础提供空间信息服务。在这类应用中,移动物体的位置受到特定技术手段(例如GPS(Global Positioning System,全球定位系统)技术)制约,存在一定的误差。尽管这项误差会随着技术手段的提升而逐步缩小,但是“位置隐私”问题却显得日益突出。移动物体的位置信息非常重要,有些用户并不愿意公诸于众,以免带来麻烦。“位置隐私”的目的是降低位置的精度——在某时刻,移动物体并非在某一空间“点”上,而是在一个“区域”内,从而保护了隐私。与此同时,各互联网服务提供商仍然能够根据这项“区域”信息提供相应的服务,例如,查询移动对象附近的医院、宾馆等设施。

电信服务

电信行业的数据量庞大,包括用户通话数据、文件传输数据、日志数据以及电信增值服务的各类资源。这些原始数据一般都具有较高的质量。但是,由于数据规模过于庞大、数据产生速度极快,对数据的存储、查询和分析等提出了挑战。对于实时应用来说,可以首先对数据进行精简,然后再进行实时处理。例如,在分析网络日志的时候,可以在IP包头的信息以一定的采样比率获取之后,进行后续分析,

以降低路由器的负担。

数据挖掘应用

数据挖掘应用的目的是从大量纷繁芜杂的原始数据中获取知识。原始数据的质量能够在很大程度上决定数据挖掘任务的成功与否。当原始数据信息丰富、数据准确客观时,所获取的知识价值高;如果原始数据的质量并不理想,例如当存在缺失值、字段值有误差时,所获取的知识可能并无任何借鉴意义。缺失值产生的原因很多,例如物理设备故障、信息无法得到、数据不一致、历史原因等。数据预处理技术可以提升数据质量。数据预处理技术很多,例如可以对数据做插值处理,插值之后的数据可看作服从特定概率分布,此外还可以删除所有含缺失值的记录。但这些方法都会改变原始数据的自身特性。

金融服务

金融数据涵盖的范围很广,包括金融机构数据、企业自身数据、企业间交易数据、监管和审计数据等等。金融数据本身可能包含虚假信息,这些信息甚至可能是人为因素故意引入的。2008年的金融风暴以来,金融欺诈的案例屡见不鲜,对整个社会造成了严重影响。异常检测和预测分析是金融数据分析中的两个重要问题,必须考虑到虚假信息的因素。

挑战

与传统的面向确定性数据的管理技术相比,不确定性数据管理技术在以下几个方面面临着挑战。

差异显著的数据模型

不确定数据有两方面的内涵,即各元组本身存在性的不确定性和各元组属性值的不确定性。元组本身存在性的不确定性可用概率 p

描述：即该元组存在的概率是 p ，不存在的概率是 $1-p$ 。元组的属性值的不确定性有多种描述方式，最通用的方式是以概率密度函数描述属性值，也可以用一些统计值进行描述，例如方差等。传统数据模型无法准确描述不确定性数据，可能世界（Possible World）模型^[5]是描述不确定性数据的通用模型。该模型包含若干个可能世界实例，在各个实例中，一部分元组存在，剩余元组不存在。可能世界实例的发生概率等于实例内元组的概率乘积和实例外元组的不发生概率的乘积之积。所有可能世界实例的发生概率之和等于1。以图1为例。输入数据序列是3个相互独立的元组，存在概率分别是0.7、0.6和0.5，颜色表示各元组存在时的属性值。则共有 $2^3=8$ 个可能世界实例，各实例的发生概率依赖于所包含的元组集合。例如，仅包含紫、绿二球的的可能世界实例的发生概率等于 $(1-0.7) \times 0.6 \times 0.5 = 0.09$ 。

急剧攀升的问题复杂度

毫无疑问，管理不确定性数据所面对的最直接的挑战，就是相对于数据库规模呈指数倍的可能世界实例的数量。在图1中，当输入数据集合仅含3条记录时，就能够生成8个可能世界实例。那么假设元组独立的不确定数据库含 N 条记录，若各元组仅有存在级不确定性，可能世界的数目将达到 2^N 个；当各元组还有属性级不确定性时，可能世界的数目会远远超过 2^N 个。可见，简单列举所有可能世界实例的处理

开销惊人，更何况还需要进一步处理各项复杂的查询了。部分应用还需要考虑元组相斥的情况，即两个元组无法共存的情况，使得查询处理的复杂度进一步提升。在此情况下，“罗列可能世界实例，计算基于该实例的查询结果，整合各实例的查询结果生成最终的答案”的处理方式显然是不可行的，迫切需要结合各种剪切、排序等技术以快速计算查询结果。

非同一般的概率维

直观来看，不确定性数据与确定性数据的差异并不大，仅多了一个概率维度。这是否意味着可以将概率维度当作一个普通维度，再利用传统技术进行处理？实际情况要更为复杂。概率维度对不确定性数据管理的影响非常深远，体现在查询定义、存储与索引、处理过程、结果呈现等各个环节之中。首先，部分查询定义可能拥有概率参数，例如Pt-k查询（一种top-k查询）需要一个概率参数 p ，仅返回成为top-k成员的概率超过 p 的元组集合^[6]。其次，传统的索引技术（例如B+树、R树等）无法有效索引不确定性数据，需要开发新的索引技术。再次，处理过程需要充分考虑概率因素，许多算法在执行过程中会优先考虑高概率的元组。最后，查询结果也会包含概率信息。因此，概率维度不是普通的维度，它的出现改变了传统的数据处理模式。

多样的数据形态

如前所述，不确定性数据在诸多应用中广泛出现。在各应用中，数据的描述方式各异。最早的数据形式是关系型数据，它在关系表中新增一个概率属性，描述该
















数据序列				可能世界实例		
时间	1	2	3			
彩球						
概率	0.7	0.6	0.5			
				1	  	0.21
				2	 	0.09
				3	 	0.14
				4	 	0.21
				5		0.06
				6		0.09
				7		0.14
				8		0.06

图1 可能世界模型实例

元组的存在概率,同时还可以借助于成熟的关系数据库处理引擎处理查询。其他重要的数据形式还包括半结构化数据(XML)、流数据、多维数据和空间数据等。引入概率信息之后,查询处理上述数据形式将会出现各种新问题。以半结构化数据为例,光是在如何描述不确定性半结构化数据这方面,就有多种模型,例如p-文档模型^[7]、概率树模型^[8]、PXDB模型^[9]等,更进一步的工作还包括查询、更新等。

丰富的查询类型

面向不确定性数据的查询任务丰富。大部分面向确定性数据的查询任务在不确定性数据环境中仍然具有现实意义,需要进行处理。一个比较有趣的现象是,在不确定性数据环境下,由于引入了概率维度,查询的种类反而会增加。元组的概率维度值从侧面反映了该元组的重要程度,因而影响着查询的定义。以top-k查询为例,在确定性数据处理领域,其意义清晰,返回秩函数的值最大的k个元组。但在不确定性数据管理领域,秩函数值仅是其中一项因素,概率值是表征元组重要性的另一因素。在此基础上,最近出现了多种面向不确定性数据的top-k查询,包括U-Topk、U-kRanks、PT-k和Pk-topk等^[10]。其他查询也存在类似现象。

如火如荼

面向不确定性数据的管理技术的研究工作并非最近开展起来的,只是在最近几年才在更广的范围内得到更多的关注。早在20世纪80年代末期,就有学者关注这方面的内容,当时关注的焦点是如何对关系数据模型进行扩展。

为了描述不确定性数据,可在关系表上额外增加一个概率字段,更复杂的结构还能够描述多个元组之间的相关性(主要是互斥)。数据管理系统可以接受类SQL的查询语言,并进行处理。这方面的研究工作对当前的不确定性数据管理技术的发展影响很大。现在很多不确定性数据管理系统,其底层的系统实现部分还是采用关系型数据库,同时拥有一个中间层接口,将一个类SQL语句转化为标准的SQL语句,并利用关系数据库管理引擎处理查询请求。典型的例子是斯坦福大学的Trio-One系统。

进入21世纪以来,不确定性数据管理方面的研究工作逐步向更广的方向发展,例如数据集成、数据挖掘、数据流处理、联机分析处理(On-Line Analysis Processing, OLAP)、半结构化数据等。

下面介绍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。

1. 多伦多大学的Conquer项目¹。研究针对不一致数据库的管理技术。

2. 普度大学的Orion项目²。研究一个通用目的的不确定性数据库系统。

3. 斯坦福大学的Trio项目³。研究不确定数据的世系分析。

4. 康奈尔大学的MayBMS项目⁴。研究查询语言、表示与存储技术、支持数据清洗、高效的查询处理、更新等技术。

5. 华盛顿大学的MystiQ项目⁵。研究内容包括数据模型、关系代数计算、数据库理论等。

6. 英特尔/加州大学伯克利分校的HeisenData项目⁶。研究将不确定性查询技术以模块形式加入到确定性数据管理框架中,以增强系统性能。

¹ <http://queens.db.toronto.edu/project/conquer/>

² <http://orion.cs.purdue.edu/>

³ <http://infolab.stanford.edu/trio/>

⁴ <http://www.cs.cornell.edu/database/maybms/>

⁵ <http://www.cs.washington.edu/homes/suciu/project-mystiq.html>

⁶ <http://www.eecs.berkeley.edu/Research/Projects/Data/102060.html>

7. 马里兰大学的ProbDBs项目⁷。研究空间-时间概率数据库。

8. IBM Almaden的Avatar项目⁸。研究非结构化数据和商业智能等领域的不确定数据。

有兴趣的读者可以参阅一些综述文献了解更多相关知识。文献[11]介绍了不确定性数据的应用背景,并总结了不确定数据管理所面临的挑战。文献[12]从算法与应用角度综述了不确定数据管理技术。文献[13]进一步从理论角度阐述不确定数据管理的基础与挑战。裴健等人^[14]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top-k查询等。周傲英等人^[1]以不确定数据管理框架为主线,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、查询处理等各个方面所取得的重要进展。

近年来,有关不确定性数据管理的文章频频现诸于主流的国际学术会议与国际期刊,例如SIGMOD (Special Interest Group for the Management of Data)、VLDB (Very Large Data Bases)、ICDE (International Conference on Data Engineering)、TODS (Transactions on Database Systems)、VLDBJ (Very Large Data Base Journal)、TKDE (Transactions on Knowledge and Data Engineering)等。此外,一些主流的国际会议也积极召开一些小型的研讨会,讨论热点话题。近期关于不确定性数据管理方面的研讨会有:IEEE ICDM会议的DUNE 2007⁹,IEEE ICDE 会议的MOUND 2009¹⁰,VLDB会议的MUD 2007¹¹和MUD

2008¹²等。国际一流刊物VLDBJ¹³和TKDE¹⁴也将在2009年和2010年分别刊出一个关于该主题的特刊 (Special Issue)。因此,不确定性数据管理正成为一个研究热点。

春意盎然

不确定性数据管理技术的研究工作在很多方面都得到了很好的发展,包括数据集成、索引技术、半结构化数据、世系分析、关系代数处理、数据流分析、数据挖掘和联机分析处理等。

数据集成 数据集成技术是管理大规模、异构数据源不可或缺的技术。由于异构数据源的模式可能存在差异,因此需要制定规则,将异构数据源转换成某一共享的中介模式,并基于中介模式处理各项查询。董欣 (Xin Dong, 音译) 等人^[15]研究了面向不确定性数据的数据集成系统,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:不确定性数据源、不确定性模式映射和不确定性查询。首先,不确定性数据源是该数据集成系统最直接的动力,在很多情况下原始数据都可能不准确。其次,异构数据源与中介模式之间的映射关系可能存在不确定性,其可能的原因包括用户操作不熟练、机器自动匹配等。在实际应用中,利用半自动化工具自动生成中介模式也很常见,并非需要领域专家特别指定。最后,在很多应用中还需要处理不确定性查询。例如,在万维网应用中,以关键词形式提交查询非常普遍,这就是一个典型的不确定性查询。

⁷ <http://www.cs.umd.edu/~vs/research.htm#pdb>

⁸ <http://www.almaden.ibm.com/cs/projects/avatar/>

⁹ <http://www4.comp.polyu.edu.hk/~dmu07/>

¹⁰ <http://www.cse.ust.hk/~mound/index.htm>

¹¹ <http://mud.cs.utwente.nl/index2007.html>

¹² <http://mud.cs.utwente.nl/>

¹³ <http://www.cs.washington.edu/homes/suciu/cfp-final-vldb-j-probdb-2009.html>

¹⁴ <http://i.cs.hku.hk/~ckcheng/tkde-si/cfp.html>

索引技术 索引技术是数据管理技术的重要内容。关系型数据库往往采用B+树及其变种为一维数据建立索引；在多维数据管理领域或时间-空间数据管理领域，广泛使用R树以及其变种进行索引。这些索引技术均能够大幅提高查询处理速度。同理，在处理不确定性数据中也需要关注索引问题。在某些查询任务中，例如top-k查询，元组的概率值也非常重要，因此需要针对概率维度创建一维索引，此时传统索引技术有效。但传统的索引技术无法解决所有问题。当各元组的取值必须通过概率分布函数描述，且概率分布函数无法预先指定时，传统的索引技术就无法胜任了。目前较好的方法有概率阈值索引技术（PTI）^[16]和U-Tree技术^[17]。前者针对一维数据，后者针对多维空间数据。

半结构化数据处理 半结构化数据模型能够有效描述缺乏严格模式结构的数据。众所周知，半结构化数据通常采用树状结构进行描述，各元素及元素的属性均能记录信息。含不确定性信息的半结构化数据也可以采用文档树进行描述，概率信息可以放在文档树的边上，各边还能够表述依赖关系。为了表述更为复杂的不确定性关系以优化查询、更新等操作，目前已经出现多种各具特色的描述模型，比较典型的有p-文档模型^[7]、概率树模型^[8]、PXDB模型^[9]等。

世系分析 数据的世系（Lineage或Provenance）是研究数据的产生以及演变过程的，不仅包括单一数据库的内在世系，也包括跨数据库的世系。尽管现有研究大多面向确定性数据，但是面向不确定性数据的世系分析仍非常重要。例如，在传感器网络中，各节点采集的原始数据精确度较低；在原始数据向中心服务器传输的过程中，数据不断被聚集、合并，以降低网络传输开销。但是，跟踪数据的演变仍有必要，若某节点出现故障，则该节点近期采集数据的可信度太低，需要清除影响。斯坦福大学的Trio项目组最早开始研究如何分

析不确定性数据的世系，提出了TriQL语言并开发Trio管理系统^[18]。

关系代数处理 研究工作开展得较早，从20世纪80年代后期开始到现在一直在延续。首先，在各关系表中添加属性字段以描述元组的不确定性；然后，数据库接受类SQL语言的查询语句，并且进行处理。这种类SQL语句往往可以转化为标准的SQL语句进行查询。查询结果中还包括概率字段以描述查询结果的准确性。一些近期的工作在于分析各查询操作的复杂度^[19]。

数据流分析 在部分新型应用中，数据规模宏大且产生速度极快，因此被称为数据流，例如互联网主干网路由器产生的日志数据。为处理数据流，必须设计单遍扫描算法，仅消耗少量内存，且能实时监测查询结果^[20]。近期的研究工作拓展到了不确定性数据流上，即流上的各元组都是不确定性元组。可以充分借鉴面向确定性数据流的工作成果，以设计针对不确定性数据流的新方法，包括计算 F_2 的pAMS方法^[21]、计算相异元素个数的pFM方法^[21]和生成聚类的ECF方法^[22]等。滑动窗口模型是数据流上的重要模型，仅考虑最近的N个元组，金澈清等人提出了在滑动窗口上计算top-k查询的新方法^[10]。最新的研究工作还包括对事件数据流的处理，特别是在RFID环境中^[23]。

联机分析处理 是数据仓库技术的重要组成部分，它使用多维数据模型，能分析数据仓库中各维度的数据信息。事实表存储数据信息，各个事实（fact）可被视为多维空间的一个点。在不确定性数据管理领域，各个事实可以并非是一个单独的点，而是跨越多个维度，成为一个“维度”。例如，某汽车销售商想借助联机分析处理技术分析各车型在各地的销售情况，则各车型被销售的城市就成为一个重要的维度。但是，如果仅仅知道该车在哪个省份销售，而不知道具体在哪个城市时，这条销售记录会跨越多个维度，成为不确定性数

据。文献[24]提出了构造EDB的方法来解决上述问题。他们的后续工作也考虑了不确定性数据之间含有相关性的情况^[25]。

数据挖掘 数据挖掘能从一堆纷繁芜杂的数据中抽取有用知识。当原始数据的精确度不高时,传统方法需要首先对数据进行清洗,利用插值、采样、回归、标准化等方法提高数据质量,然后再运行聚类算法。然而,数据清洗过程会丢失数据的部分特征,降低查询结果质量,因此如何在不确定性数据上直接做挖掘算法就显得很重要。聚类技术得到了最广泛的研究,包括k-means算法的改进版本UK-means算法^[26]、基于密度的FDBSCAN算法^[27],以及面向层次聚类问题的FOPTICS算法^[28]等。文献[29]提出了一种面向不确定性数据的分类算法,该算法基于支持向量机(Support vector machine, SVM)技术。文献[30]则给出了计算频繁项集合的新方法。

其他查询 针对一些重要查询的研究工作也不少,特别是top-k查询与skyline查询。在不确定性数据环境下,重要的top-k查询类型有U-Topk、U-kRanks、PT-k和Pk-topk等^[10]。Skyline查询是另一类查询,其目的是返回成为skyline的节点。裴健等人定义了概率skyline查询,并提出了自下而上和自上而下两种方法,充分运用定界、剪枝精化等启发式规则提高效率^[32]。连(Lian, 音译)与陈(Chen, 音译)等人则定义了在不确定性数据上的reverse skyline问题,并给出了解决方案^[33]。

方兴未艾

针对不确定性数据的管理的研究工作正处于

方兴未艾的状态。究其原因,主要有以下几点。

首先,最主要的原因是相关研究工作并非空穴来风,而是有着强大的应用基础。不确定性数据在很多应用(特别是新型引用)中广泛出现,迫切需要得到良好地解决。如前所述,典型的不确定性数据包括互联网数据、传感器网络数据、RFID数据、基于位置数据、隐私保护数据、互联网数据、金融数据等。事实上,这个列表还可以更长。另外,这些应用本身也在不断地发展演变过程中,亟需解决的问题层出不穷。以互联网数据为例,每年互联网数据都在急剧增长,还会出现新的数据形式,如论坛、博客等。这要求数据管理技术也能够不断更新。

其次,不确定性数据管理领域仍然有大量问题有待研究。例如,尽管针对关系型不确定性数据的研究工作已经有二十余年的历史,但有些核心的理论问题到现在仍有待解决。另外,针对其他数据形态的研究工作的开展时间都不算长,很多问题仅仅得到初步解决,而未能得到成熟的方案。这里面有待开展的工作非常多。

最后,针对不确定性数据的研究将丰富现有研究领域的研究内容。由于不确定性数据的普遍性,高效的不确定性数据管理技术对于许多学科均有积极意义,例如人工智能、机器学习、分布式计算、网络技术等。■



李建中

中国计算机学会理事、高级会员。哈尔滨工业大学计算机科学与技术学院教授。主要研究方向为无线传感器网络、并行数据库等。lijzh@hit.edu.cn



于戈

中国计算机学会监事、高级会员。东北大学教授。主要研究方向为数据库理论与技术、分布式信息系统。yuge@mail.neu.edu.cn



周傲英

中国计算机学会高级会员。华东师范大学教授。主要研究方向为数据密集计算系统的数据管理、数据挖掘与数据流分析。ayzhou@sei.ecnu.edu.cn

参考文献

- [1] 周傲英, 金澈清, 王国仁, 李建中. 不确定性数据管理技术研究综述. 计算机学报, 31(1), 2009
- [2] 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] 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] Shawn R. Jeffery, Minos Garofalakis, Michael J. Franklin. Adaptive Cleaning for RFID Data Streams// Proceedings of the 32nd international conference on Very large data bases . Seoul, 2006:163~174
- [5] Todd J. Green, Val Tannen. Models for incomplete and probabilistic information. IEEE Date Engineering Bulletin, 2006, 29(1):17~24
- [6] Ming Hua, Jian Pei, Wenjie Zhang, Xuemin Lin. 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
- [7] A. Nierman and H. V. Jagadish. ProTDB: Probabilistic Data in XML//Proceedings of the 28th international conference on Very Large Data Bases. Hong Kong, 2002:646~657
- [8] P. Senellart, S. Abiteboul. 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
- [9] Sara Cohen, Benny Kimelfeld, Yehoshua Sagiv. Incorporating Constraints in Probabilistic XML// Proceedings of the 27th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. Vancouver, 2008:109~118
- [10] Cheqing Jin, Ke Yi, Lei Chen, Jeffrey Xu Yu, Xuemin Lin. Sliding-window Top-k Queries on uncertain Streams. In Proceedings of VLDB endowment. 2008, 1(1):301~312
- [11] Christopher Ré, Dan Suciu. Management of Data with Uncertainties//Proceedings of the 16th ACM conference on Conference on information and knowledge management. Lisbon, 2007:3~8
- [12] C. C. Aggarwal and P. S. Yu. A Survey of Uncertain Data Algorithms and Applications. IBM research report. October 31, 2007
- [13] Nilesh Dalvi, Dan Suciu. 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
- [14] Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin. Query Answering Techniques on Uncertain and Probabilistic Data: tutorial summary//Proceedings of 2008 ACM SIGMOD International Conference on Management of Data. Vancouver, 2008: 1357~1364
- [15] Xin Dong, Alon Y. Halevy, Cong Yu. Data Integration with Uncertainty//Proceedings of 33rd International Conference on Very Large Data bases. Vienna, 2007:687~698
- [16] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. S. Vitter. 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
- [17] Y. Tao, X. Xiao, and R. Cheng. Range search on multidimensional uncertain data. ACM Transactions on Database Systems, 2007, 32(3):15
- [18] O. Benjelloun, A. Das Sarma, A. Halevy, M. Theobald, and J. Widom. Databases with Uncertainty and Lineage. VLDB Journal, 2008, 17(2):243~264
- [19] N. Dalvi and D. Suciu. 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
- [20] 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
- [21] G. Cormode, M. Garofalakis. Sketching Probabilistic Data Streams//Proceedings of the 2007 ACM SIGMOD international conference on Management of data. Beijing, 2007:281~292
- [22] C. C. Aggarwal, P. S. Yu. A Framework for Clustering Uncertain Data Streams//Proceedings of the 24th IEEE International Conference on Data Engineering. Cancun, 2008:150~159
- [23] Christopher Ré, Julie Letchner, Magdalena Balazinska and Dan Suciu. Event Queries on Correlated Probabilistic Streams//Proceedings of the 27th ACM SIGMOD international conference on Management of data. Vancouver, 2008:715~728
- [24] Doug Burdick, Prasad M. Deshpande, T.S. Jayram, Raghu Ramakrishnan, Shivakumar Vaithyanathan. OLAP over Uncertain and Imprecise Data. The VLDB Journal, 2007, 16(1):123~144

- [25] Doug Burdick, AnHai Doan, Raghu Ramakrishnan, Shivakumar Vaithyanathan. Olap over Imprecise Data with Domain Constraints. Proceedings of the 33rd international conference on Very large data bases. Vienna, 2007:39~50
- [26] Wang Kay Ngai, Ben Kao, Chun Kit Chui, Reynold Cheng, Michael Chau, Kevin Y. Yip. Efficient Clustering of Uncertain Data//Proceedings of the 6th International Conference on Data Mining. Hong Kong,2006:436:445
- [27] Hans-Peter Kriegel, Martin Pfeifle. Density-Based Clustering of Uncertain Data Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining. Chicago, 2005:672~677
- [28] Hans-Peter Kriegel, Martin Pfeifle. Hierarchical Density-Based Clustering of Uncertain Data//Proceedings of 5th International Conference on Data Mining. Houston, 2005:689~692
- [29] Jinbo Bi, Tong Zhang. Support Vector Classification with Input Data Uncertainty. Advances in Neural Information Processing Systems 17. Vancouver, 2005:161~168
- [30] Chun-Kit Chui, Ben Kao¹, Edward Hung². Mining Frequent Itemsets from Uncertain Data. In Proc. Of PAKDD, 2007
- [31] J. Pei, B. Jiang, X. Lin, Y. Yuan. Probabilistic Skylines on Uncertain Data// Proceedings of the 33rd international conference on Very large data bases. Vienna, 2007:15~26
- [32] X. Lian, L. Chen. 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