华南理工大学

South China University of Technology

本科毕业设计（论文）说明书

不确定性数据的模型研究

学　　院　软件学院

专　　业　软件工程

学生姓名　江李科

指导教师　杜　卿

提交日期　2010 年 6 月 5 日

华 南 理 工 大 学

毕 业 设 计 （论文） 任 务 书

兹发给软件学院 06 级 3 班学生 江李科 毕业设计（论文）任务书，内容如下：

1.毕业设计（论文）题目：不确定性数据的模型研究

2.应完成的项目：

（1）2010 年 3 月 1 日前收集资料，拟定提纲并提交开题报告

（2）2010 年 5 月 15 日前完成论文初稿

（3）进行与论文题目相关的调研工作并收集相关的资料

（4）2010 年 6 月 1 日前参考外文文献资料并提交外文翻译译文

3.参考资料以及说明：

（1）李建中，于戈，周傲英. 不确定性数据管理的要求与挑战.中国计算机学会通讯.2009，5(4):6-15

（2）Nierman A, Jagadish H V. Pro TDB : Probabilistic data in XML/ / Proceedings of t he 28t h International Conference on Very Large Data Bases. Hong Kong , China , 2002 : 6462657

（3）Abiteboul S , Senellart P. Querying and updating probabilistic information in XML/ / Proceedings of t he 9t h International Conference on Extending Database Technology : Advances in Database Technology. Munich , 2006 : 105921068

（4）Senellart P , Abiteboul S. On t he complexity of managing probabilistic XML data/ / Proceedings of t he 26t h ACM SIGMOD2SIGACT2SIGART Symposium on Principles of Database Systems. Beijing , 2007 : 2832292

（5）Cohen S , Kimelfeld B , Sagiv Y. Incorporating const raint s in probabilistic XML/ / Proceedings of t he 27t h ACM SIGMODSIGACT2SIGART Symposium on Principles of Database Systems. Vancouver , 2008 : 1092118

（6）Abiteboul S , Kanellakis P , Grahne G. On t he representation and querying of set s of possible worlds. ACM SIGMOD Record , 1987 , 16 (3) : 34248

4.本毕业设计（论文）任务书于 2010 年 3 月 5 日发出，应于 2010 年 6 月 1 日前完成，然后提交毕业考试委员会进行答辩。

专业教研组（系）、研究所负责人_____审核_____年___月___日

指导教师_____签发 _____年___月___日

毕业设计（论文）评语：

毕业设计（论文）总评成绩：

毕业设计（论文）答辩负责人签字：

年　　月　　日

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。

我们以模型所针对处理的不确定数据的类型为依据，将模型分成了四大类。而可能世界模型则作为一种基本的模型在第二章中有介绍。当前，研究者已经为各种数据分别提出了不少模型，而且各种模型的差异也非常大。第二、三、四从模型的定义与优劣入手，分别详细介绍了可能世界模型、针对关系型数据的模型、针对结构化数据的模型。第五章简单介绍了针对数据流的模型以及针对多维数据的模型。

关键词：不确定性数据，模型，可能世界

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 model s have been developed, stemming from the core possible world model that contains a huge number of the possible world instances wit h the sum of probabilities equal to 1.

According the type of the uncertain data which are managed by the uncertain model, we group models into 4 types. And the possible world model is introduced in chapter 2 as a basic model. Currently, the researchers had developed many models, respectively, which are very different from each other. According to the definition and quality of the models, we have researched the possible world model, models for relational data, and models for semi-structured data , respectively , in chapter 2, chapter 3, chapter 4. And we also take a quick look at models for streaming data and models for multidimensional data in chapter 5.

Keyword: uncertain data, model, possible world

目 录

第一章 绪论

1.1 不确定性数据的背景

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[1]。 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。 但是，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。

1.2 不确定性数据的管理框架

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。 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 1-1 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[1]。

1.2.1 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)[2]。 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。

1.2.2 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。

1.2.3 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)[2]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 Neighborquery) 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。

1.2.4 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。

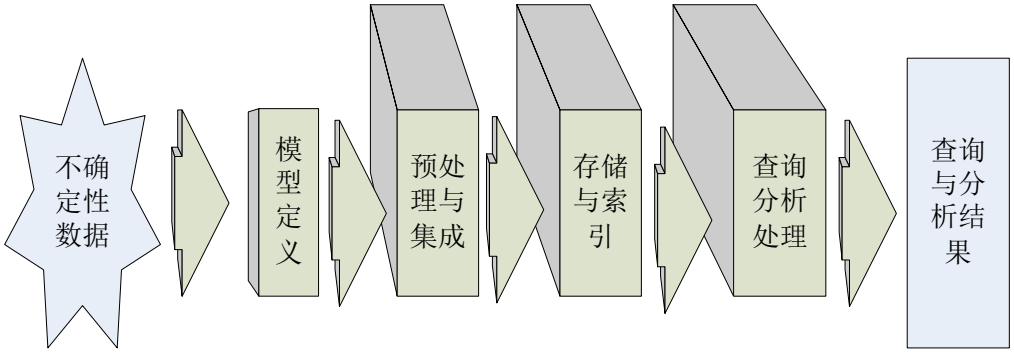


图 1-1 不确定性数据管理的框架

1.3 不确定性数据的模型

由于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），因此，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，相应地也出现了许多与数据类型紧密相关的数据模型。包括可能世界模型，针对关系数据库的模型（如，probabilistic ?-table、probabilistic c-table 等），针对半结构化数据的模型（如，P-文档模型，概率树，PXDB 模型等），针对数据流的模型（如，界标模型，滑动窗口模型等），针对多维数据的模型。

1.4 建模的要求与挑战

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建模技术面临以下挑战。

1.4.1 庞大的可能世界实例集合

毫无疑问，不确定性数据建模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（这也是不确定性数据管理所面临的最直接挑战）。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[3]。 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。

1.4.2 新出现的维度———概率维

不确定性数据建模时都会考虑到的问题是概率。概率在不确定性数据管理中扮演重要的角色。在为不确定性数据建模时，数据的概率维与不完整性也是研究者考虑的重要因素。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，迫切需要开发新模型、新技术进行处理。

1.4.3 不确定性数据管理的理论问题

作为不确定性数据管理的首要任务，不确定数据在建模技术方面仍然存在大量具体问题，特别是验证模型时的理论相关的问题。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。 模型中，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。

第二章 可能世界模型

2.1 可能世界模型的简介

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[4]。

可能世界模型使用了最直接的方式来描述不确定性数据。在该模型中，各元组的任一组合均构成一个可能世界实例，实例的概率值可以通过相关元组的概率计算得到。而且，所有实例的概率之和为1。

可能世界的优点在于它直观地反映了现实世界。它将不确定数据分成了不同的可能世界实例，而在每一个实例确定下来的同时它的数据也成为确定的了。换句话说，我们可以通过可能世界模型把一个不确定数据集转化为多个确定的数据集。

但是，可能世界模型也具其致命的缺陷：可能世界实例的规模远远大于不确定性数据库的规模，甚至是后者的指数倍。在下一小节中，我们将通过例子来说明这一点。

2.2 可能世界模型的举例与说明

考虑如图2-1的一个例子。 图2-1(a) 是一个不确定性数据库，包含3 个元组，每个元组都有三个字段（或者叫作属性），分别是ID字段、信息字段与概率字段。其中，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,如图2-1(b) 所示。此时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，例如，可能世界实例{1，3}的发生概率为 $0.3\times(1-0.7)\times0.6=0.054$。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。下面我们来构造一个存在依赖关系的场景。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[7]。总共有6个可能世界实例，如图2-1(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⊕3

PW = { {} , {1} , {3} , {2} , {1 ,2} , {2 ,3} }
$P(PW)$ = { .03 , .09 , .18 , .07 , .21 , .42}

(b) 可能世界

图 2-1　可能世界样例

图2-1 的例子只描述了数据存在级别的不确定性。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2-1 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（probabilistic density function，简称pdf）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。因此有些时候，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⊕3 ，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。

第三章 针对关系型数据的模型

3.1 Probabilistic ?-table 模型

针对关系模型的扩展最为常见，包括Probabilistic ?-table[5]、Probabilistic or-set table[6]、Probabilistic or-set-? table[6]、Probabilistic c-table[4]等。

Probabilistic ?-table 就一个针对关系型数据的模型。它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。

根据Probabilistic ?-table的定义我们可以知道，它所描述的存在级别的不确定性数据，对于属性级别的不确定性它就无能为力了。如图3-1(3)中所示，第二个元组存在的概率是0.6，因此我们可以推知，第二个元组不存在的概率为1 – 0.6 = 0.4。但是，对于第二个元组来说，属性c1 = 2，c2 = 3则是已经确定的了。对比第二章中可能世界模型的例子，可知道，Probabilistic ?-table也是一种可能世界模型，只不过它是针对关系型数据的。图3-1(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 $2^2 = 4$ 个可能世界实例。因此，Probabilistic ?-table模型也有着同可能世界相同的缺陷，即可能世界实例的数量是不确定性数据库的指数倍。而且，当它所描述的数据在属性级别上存在不确定性时，可能世界实例的规模还要大得多（如，在第一个元组中，c2=2,当c1=1的概率为0.1，c1=2的概率为0.5时，则第一个元组必须分裂与2个元组才能描述这种情况）。同时值得注意的是，在元组级别上存的概率相等，但是各属性的值的概率却不一定相等。

3.2 Probabilistic or-set table 模型

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。

当属性可以拥有多个概率值时，关系也就不可能符合第一范式。这必然导致出现复杂的代数计算与查询操作，这是此模型的一大不足之处。

以图3-1(b) 为例，第一个元组的c2字段既可取2 ，也可取3 ，其概率分别为0.4 和0.6 ；第二个元组的c2 字段既可取4 也可取5 ， 其概率分别是0.2 与0.8。

c1	c2	概率
1	2	0.5
2	3	0.6

(a) Probabilistic ?-table

c1	c2
1	(< 2 , 0.4 > , < 3 , 0.6 >)
2	(< 4 , 0.2 > , < 5 , 0.8 >)

(b) Probabilistic or-set table

c1	c2	概率
1	(< 2 , 0.4 > , < 3 , 0.6 >)	
2	(< 2 , 0.4 > , < 3 , 0.6 >)	0.8

(c) Probabilistic or-set-? table

图 3-1 基于关系数据的扩展模型

3.3 Probabilistic or-set-? Table 模型

Probabilistic or-set-? table[4]则是上述两种模型的综合体。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]。例如，在图3-1(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。

虽然此模型在表达能力上是前面两种模型（probabilistic ?-table 模型、probabilistic or-set table 模型）总和，但是它同时也还保留着它们的不足之处。如前面所述，当属性可以拥有多个概率值时，关系也就不可能符合第一范式。这必然导致出现复杂的代数计算与查询操作。

3.4 Probabilistic c-table 模型

Probabilistic c-table模型的定义与Probabilistic or-set Table模型比较类似，且同样是针对生母关系弄数据库的。不同之处在于前者是从c table衍生出来的[18]。

3.4.1 三个简单的表达系统

在文献[18]中介绍了三种表达系统。

第一个表达系统是建立在codd table的基础上的。Codd table可带有Codd空值@，表示相对应的属性的值为未知。这样，Codd表就可以描述不完整的数据（如下面的例子所述）。Codd表可以支持投影与选择[18]。

codd table的例子如图3-2。此表描述了生产某机器原部件的相关信息。从表中可以看出来：Smith是一个钉子生产者，住在伦敦，为机器生产提供（未知数量的）钉子；Brown则提供螺栓，但是Brown的住址与提供螺栓的数量都未知；Jones是螺母生产者，提供了40，000个螺母，但是他的住址未知。

提供者	住址	产品	数量
Smith	伦敦	钉子	@
Brown	@	螺栓	@
Jones	@	螺母	40,000

图 3-2 codd table样例

文献[18]阐述的第二个表达系统是建立在v table的基础上的。v table允许出现多个不同的空值或者变量。所以，与Codd table相比，它的表达能力更强了。此系统还有一个非常可取的特性，就是所有在v table上的关系运算子的操作与在一般（确定性数据）关系中的操作是一样的。也就是说，在处理概率属性域时可以将它们当作一般值。而且，在更新关系视图的操作中，v 表也显得相当自然。但是，系统也还存在不足之处，就是它不支持投影与任意选择。

此系统例子如图3-3。图中的v table中包含了三个不同的变量：x、y、z。此表要描述的信息是教师上课的安排情况。从表中可以看出，周二有编程的课程，但是任课老师（设为y）未知；FORTRAN的课程是由Smith教的，但是上课时间（设为z）未知；周一、周四都有数据库的课程，但是未能确定上课的老师（设为x）。对于数据库的课程，虽然不能确定是谁任课，但是通过查看表我们可确定的是：周一、周四两天教数据库的是同一个老师——这种表达能力是codd table所不具有的。

课程	教师	时间
数据库	*x*	星期一
编程	*y*	星期三
数据库	*x*	星期四
FORTRAN	Smith	*z*

图 3-3　v table样例

第三个系统是建立在c table上的。c table是在v 表的基础上加上了约束列形成的。这样，就使得c table不单可以通过变量来表达元组内（和元组之间）的约束，还可以在约束属性列（如图中的con列）直接设定约束。而且c table所支持操作也增多了。它所支持的操作有：投影、选择、联合、加入、更名。

例子如图3-4，与第一个系统的例子一样，图中的c table描述了生产某机器原部件的相关信息。与Codd table不同的是，c table增加了一个con属性列。此列是用来描述数据间的约束的。如表，可以知道，钉子要么是住在伦敦的Smith提供的，要么是住在纽约的Brown提供的。根据con属性列要求，当第二个元组提供者成立时，第一个元组不能存在。因为两个元组的con属性的值（要求）是互斥的。也就是说，Smith和Brown不能同时作为钉子的提供者存在。

提供者	住址	产品	Con(约束)
x	伦敦	钉子	*x* = Smith
Brown	纽约	钉子	*x* ≠ Smith

图 3-4　c table样例

3.4.2 Probabilistic c-table

在早期，数据库的概率模型研究比不完整性的研究要少。提出来的模型主要有两种。第一种模型中，关系的元组是相互独立的，每个元组都有给定的概率值与之相对应（例如，probabilistic ?-table模型与probabilistic or-set-? Table 模型）。第二种模型中，各个元组中的各个属性皆有其独立的分布（例如，probabilistic or-set table模型与probabilistic or-set-? Table模型）。这两种模型在查询答案时，都面临着相同的问题：元组的概率计算问题。为了解决这一问题，前人开发出了一些更为普遍的模型，在这些模型的行中增加了附加信息（如“事件表达式”、“路径”、“轨迹”）。同时，不少这种模型都不约而同地用到了c table中的约束。

Probabilistic c-table是在c-table的基础上再为各个变量添加上它们的概率分布。Probabilistic c-table是闭合的，而且是完全的[4]。

例子如图3-5，图中的表描述的是学生选修课程的信息。在表中，x表示未确定的选修课程。但是x的概率分布已在表的右边给出。例如，x取值为数学时的概率为0.3，类似地，物理为0.3，化学为0.4。表中t的取值不是0就是1（见图右边的等式），取值的概率分别是0.15与0.85。现在，我们开始分析表所表达的内容。Alice选修数学课程的概率是0.3，或者，类似地，物理为0.3，化学为0.4。而Bob与Alice选修了相同的课程。从表中可以看出来，Bob选修的课程只能是物理或者是化学。再看Theo，由于 $t=1$，所以Theo选修数学的概率是0.85。

学生	课程	Con(约束)
Alice	x	
Bob	x	x = 物理 $\vee$ x = 化学
Theo	数学	t = 1

x =	数学：0.3
	物理：0.3
	化学：0.4

t =	1：0.85
	0：0.15

图 3-5 Probabilistic c-table 样例

第四章 针对半结构化数据的模型

4.1 p-document 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。半结构化数据通常可以用文档树来描述。Dekhtyar 等人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) 、概率树模型（包括简单概率树模型、模糊树模型等）、PXDB模型，以及PXML模型、Keulen 等人的概率树模型、PrXML 模型等。

下面主要介绍比较常见的几种模型。首先介绍的是P-文档模型。

4.1.1 模型简介

P-文档模型（p-document model）[9]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，兄弟节点之间可以是互斥关系(mux) 或相互独立(ind) （如，图4-1所示）。

4.1.2 xml

传统的数据描述方法（如关系数据表）已经发展得相当成熟，但是随着结构化数据的应用越来越广泛，传统的数据描述方法也逐渐暴露了它的不足之处。因为，研究者便开始寻求新的针对结构化数据的方法。其中，文档树被广泛应用于各种各样的半结构化数据建模中。

文档树就是我们通常所用的 XML 文档。文献[9]用实例表明，xml 比嵌套的关系模型更为灵活，它允许数据在结构上存在更多的不同以及可以处理更为不完整的数据。

在描述半结构化数据时，概率关系数据模型的不足之处在于：

(1) 在实际应用中，往往是属性与概率相关联，而不是元组与概率相关联。在概率关系数据模型中，当属性具有多个概率值时，在元组级别则会出现“组合爆炸”（在可能世界模型的介绍中已有说明）；

(2) 当属性可以拥有多个概率值时，关系也就不可能符合第一范式。这必然导致出现复杂的代数计算与查询操作。

与嵌套的关系模型相比，XML 更为灵活，表达半结构数据也更为自然。它的优点在于：

(1) XML 数据是结构性的，而且它的结构可以在一定的范围内方便地变化。这样，XML 不单可以描述概率数据也可以描述不完整的数据。XML 的这一特性使得不确定性数据的表达更加自然；

(2) XML 描述数据时天生具有多重粒度性。大多数关系概率模型只能把概率关联到独立的元组上，因此元组的概率也就自然要成为关系集中的一个成员。这样容易造成数据的冗余。相对而言，在 xml 中，概率可以关联到元组上，也可以关联到属性级别上，对数据的描述更为精简，但表达能力却有增无减。

4.1.3 模型定义的相关问题与解决方法

XML 数据是结构化的，而且些结构是可变的。此特性使得 xml 可以更为自然地描述不确定性数据。但是要定义一个好的模型，还是要不少问题需要解决的。文献[9]中介绍了一些相关问题的解决：

(1) XML 天生具有多重粒度性，即它不单可以将元组与概率值相关联，也可将元组中的属性与概率值相关联。XML 本身规定了属性要值唯一，但是不确定性数据中属性的取值往往有多种可能。为解决这一问题，文献[9]将属性的描述放在了子元素中进行，以适应概率系统。如图 4-1 中，可以把 D、E 看作是同一个属性的不同取值。

(2) 概率计算的问题。计算源 XML 文档中各结点的概率值可以通过 Bayes’公式进行推导[9]，文献[9]也描述了各种组合条件下的概率计算，在此不一一详细说明，只举一简单的例子。在图 4-1 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 $(1-0.7)\times 0.8\times 0.4=0.096$。由于边 C-D 与 C-E 互斥，任意子图均不能同时包含 D、E 两个节点。

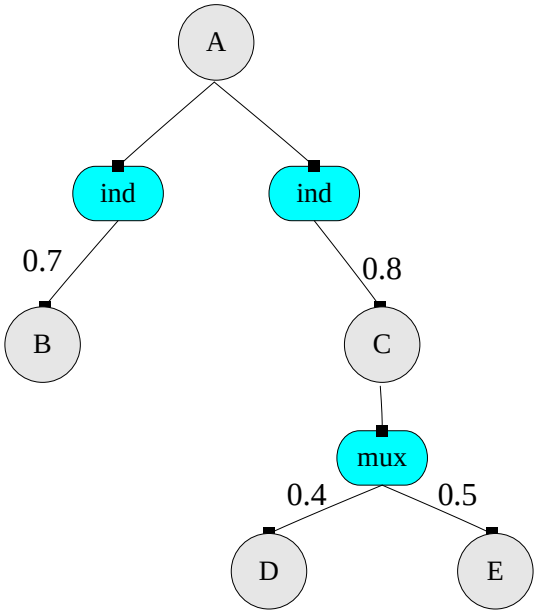


图 4-1　p-document model

4.2 概率树模型模型（probabilistic tree model）

4.2.1 模型快照

P-文档树模型虽然可以很好的描述半结构化数据。但是，通过p-文档树模型对数据进行复杂的条件查询时，就会变得非常复杂。它通过析取、合取等逻辑表达式，对复合约束条件的进行描述[10]，使得约束的描述既繁杂也不够自然。

从更好地表达约束条件的角度来考虑，概率树模型的确有p-文档树所不可及优点。概率树模型是一个事件驱动的模型[10]。 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4-5(b) 描述了一个概率树的例子，共有2 个外部事件*w1* 和*w2* ，其发生概率分别为0.8和0.7。节点B 出现的前提条件是事件*w1*发生且事件*w2*不发生；节点D存在的前提条件是事件*w2* 发生。 由于节点B与D的存在条件互斥，不存在同时包含节点B 、D 的子图。包含且仅包含A、C、D 3个节点的子图的概率为 $(1-0.8)\times 0.7=0.14$，前提是*w1* 、*w2*均不发生。可以看出， 概率树模型的表达能力强于p2文档模型。

4.2.2 模型的定义

在上一小节中，我们提到的模型是属于概率树模型的一种，叫做模糊树模型（fuzzy tree model）[10]。在介绍模糊树模型之前，值得一提的是一种更为直观的概率树模型——简单概率树模型（simple probabilistic tree model，简称SP tree）。图4-2是一个可能世界集，它对应的一个简单概率树模型描述如图4-3。

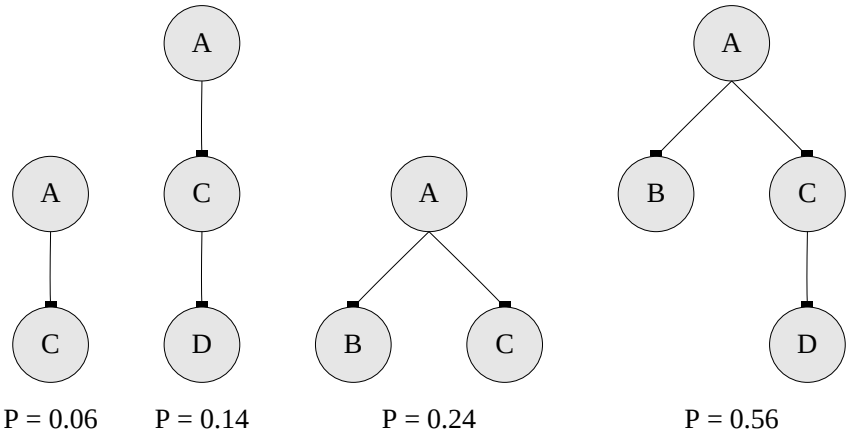


图 4-2 可能世界集样例

与P-文档树相似，简单概率树也是将概率附加于文档树的边。只不过，简单概率树是通过子图来表达条件约束以及其对应的概率的，因此具有更强的、更自然的约束表达能力。在简单概率树模型中，查询的结果是一子树（如图4-2中所示）。选择集合中的任一子图t_X，则我们可以通过此模型得到t_X概率

$$P_X = \prod_{s\in X}\pi(s) * \prod_{s\in S-X}(1-\pi(s)) \qquad (4\text{-}1)$$

其中，S表示所有可能的结点集合（有限），$\pi: S-\{r\}->]0;1]$表示取结点的概率值。从图4-3中得出子图的概率$P_{A->C} = (1-0.8)*(1-0.7) = 0.06$。值得注意的是，当文档树的某边上没有标明存在的概率时，其默认值为1（例如，边A->C），也就是说，它是必然存在的。如图4-2的可能世界样例中，A->C存在于各个可能世界实例中。

遗憾的是，简单概率树模型对不确定数据本身的描述能力却是有限的。图4-4是一可能世界集合，但是，它却无法用简单概率树模型描述出来。更致命的是，简单概率树模型在数据更新操作时不是闭合的[10]。

为了解决简单概率树存在的缺陷，有人提出了模糊树模型[10]。模糊树模型产生于前面所介绍的简单概率树模型，其为简单概率树模型增加了条件（我们可以称之概率条件），从而使之更具可用性。

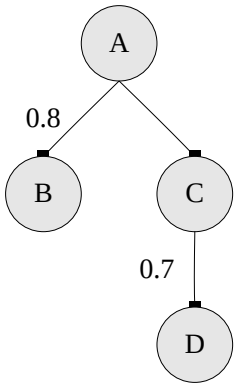


图 4-3 SP tree样例

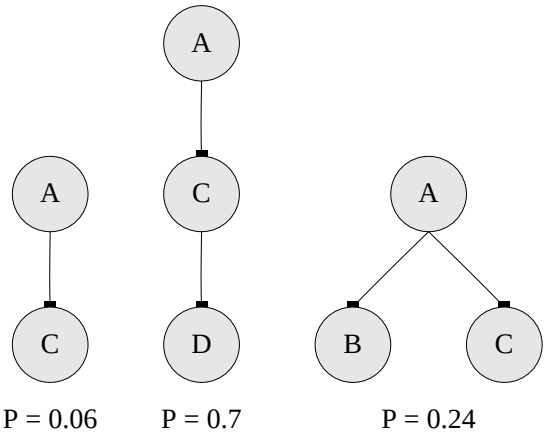
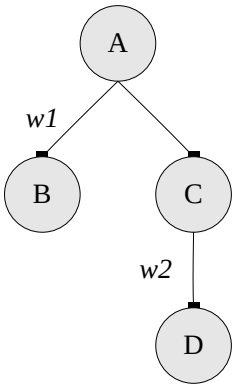


图 4-4 更复杂的可能世界集样例

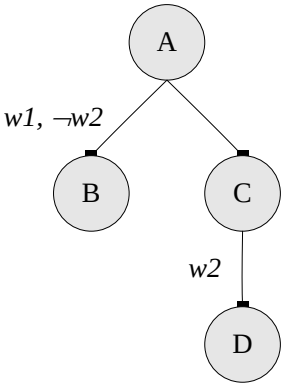
文献[10]运用了概率事件（probabilistic event）来对此“条件”进行定义。给出一事件名集合W，一概率分布π为W中的各元素指定概率值，值的范围为]0; 1]。一个事件条件(event condition)是事件原子（event atoms）的集合，其格式为w或者$\neg w$。 w是W中的一个事件。查询集合中的任一子图t_X，令$cond$表示所查询的子图的概率事件的集合，$\pi: S-\{r\}->]0;1]$表示取结点的概率值。当存在一事件w，且w与$\neg w$同时存在于$cond$中时，则查询返回值为0；否则，查询返回值为$P_X=\prod_{w\in cond}\pi(w)*\prod_{\neg w\in cond}(1-\pi(w))$。

表面上看，模糊树模型只是将简单概率树模型边上的概率值用事件概要替代掉。其实，模糊树的定义并不是一个简单的替换拼凑过程。从图4-5我们可以看出来，模糊由一文档树与一事件条件的集合组成。这样，概率条件（附加在文档树边上的）就不单可以是单个事件，也可以是几个概率事件的组合，显得非常灵活。明显地，模型的表达能力的确变得更强了。不单如此，模型的数据更新也由不闭合（简单树模型）变成了闭合的[10]。如图4-5分别描述了图4-2与图4-4中的两个可能世界分集合。其中，图4-5右边的文档树在$A->B$边上增加了$\neg w2$事件。因为w与$\neg w$同时存在的概率为0，所以$A->B$与$A->C->D$就不可能存在于同一个子图中（与图4-4中可能世界集一致）。简单概率树模型所不能描述的集合，模糊树模型轻易就可以描述出来，这足见模糊树模型表达能力远高于简单概率树模型。



Event	Proba.
w1	0.8
w2	0.7

(a)



Event	Proba.
w1	0.8
w2	0.7

(b)

图 4-5 fuzzy tree样例

4.2.3 模型的不足之处

虽然与简单概率树模型相比，模糊树模型改进了很多，但是模糊树模型还是具缺陷的，如下：

(1) 在查询时，模糊树模型的开销是很大的。文献[17]已经证明，不平凡布尔树在此模型中的查询开销是#P-complete。特别要指出的是，文献[11]中考察的“单调(monotonic)”查询是不允许映射的。

(2) 对于简单的更新（插入、删除操作不对其它的分支存在依赖性），模糊树模型不会像简单树模型那样产生指数生长的开销。但是，复杂的更新的开销依然相当大[10]。

4.3 PXDB 模型

4.3.1 PXDB 模型引入

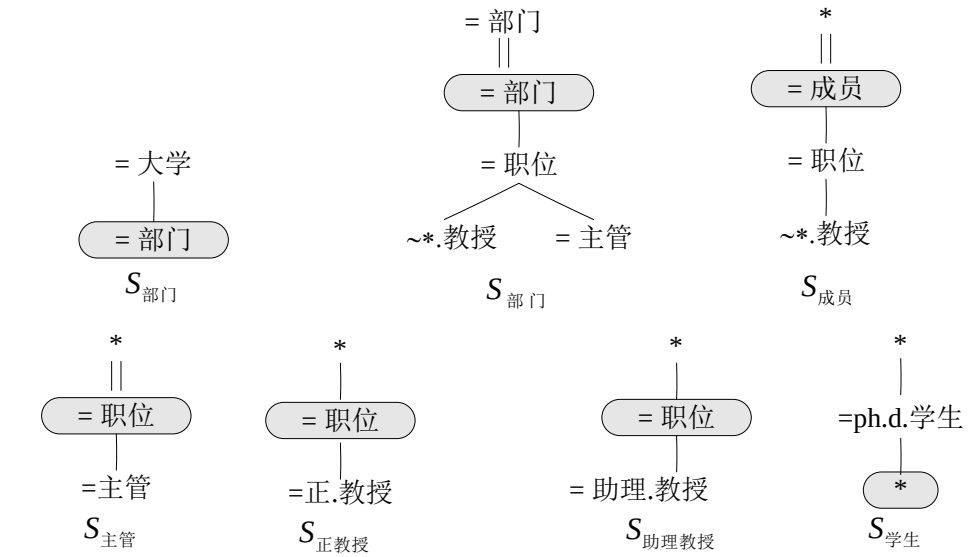
新模型的建立往往是一个取长补短的过程，PXDB的建立也不例外。下面我们先回顾一下之前介绍的两种模型：P-文档模型与概率树模型（以模糊树模型为典型代表）。P-文档为XML构造了一个概率模型（在文献[9]有描述）。在P-文档中，概率附加在文档树的边上。结点的概率只能依赖于其祖先结点的概率。例如，在图4-7中，一个P-文档描述了一个学院的部门。David是Lisa的一个Ph.D.学生（见树的中部）的概率依赖于David结点之上的结点，而与文档中的其它数据项无关。概率树，概率XML数据的另一种模型，在文献[10]有描述。在此模型中，概率事件与结点相关联。这样能够清楚地表达结点间复杂的概率依赖关系，如前面部分所述。

在当前的模型中，文献[17]清晰地描述了模型在查询评价的效率与概率依赖的表达能力之间的权衡。一方面，针对P-文档，文献[17]已经为映射方式的评估细枝查询提出了高效的算法。另一方面，概率树模型的情况却很不一样。文献[17]中已证明，在概率树模型中查询评估平凡布尔树的复杂度为#P-complete（目前为止，这种复杂程度是难以达到应用水平）。特别要关注的是，文献[11]考察的“单调（monotonic）”查询是不允许映射的。因此，典型地，要么是查询评价是难以处理的，要么是模型的表达能力有限。

再考虑另外的模型，Bayesian网络被普遍地用于描述概率。然而，怎么将约束与查询有效地转化到Bayesian网络还是不明了的。此处，对于本论文中所描述的易处理的结果，它也不大可能会存在类似的结果，因为在Bayesian网络中决定（或者近似）每一个简单事件的概率是不容易处理的。

因此我们将引入PXDBs，一种新颖的概率XML模型定义方法。在我们的模型中，一个概率数据库包括了一个P-文档和一个约束集。与模糊树相似，它的数据描述与约束（模糊树上的概率条件）在一定程度上是分离的。传统数据库的观点最先认为，约束是重要的，因为它一个是维持数据完整性的机制。更为有意义的是，约束在概率方面也是很重要的，因为它悄悄地捕获了数据项中自然的概率依赖。再且，约束可以直接采自有关真实世界需求的用户知识，而且，很有望可以被确切地阐述。本论文首先考虑的是概率数据与丰富的约束语言的结合。

可见，PXDB 模型[12]扩展了P-文档模型，增加外部约束条件。



$C1: \forall S_{部门}[* -> C_{NT}(S_{主管}) \le 1]$

$C2: \forall S_{部门}[C_{NT}(S_{成员}) > 2 -> T_{主管}]$

$C3: \forall S_{成员}[T_{主管} -> T_{正教授}]$

$C4: \forall S_{成员}[\mathrm{T}_{部门} -> C_{NT}(S_{学生})]$

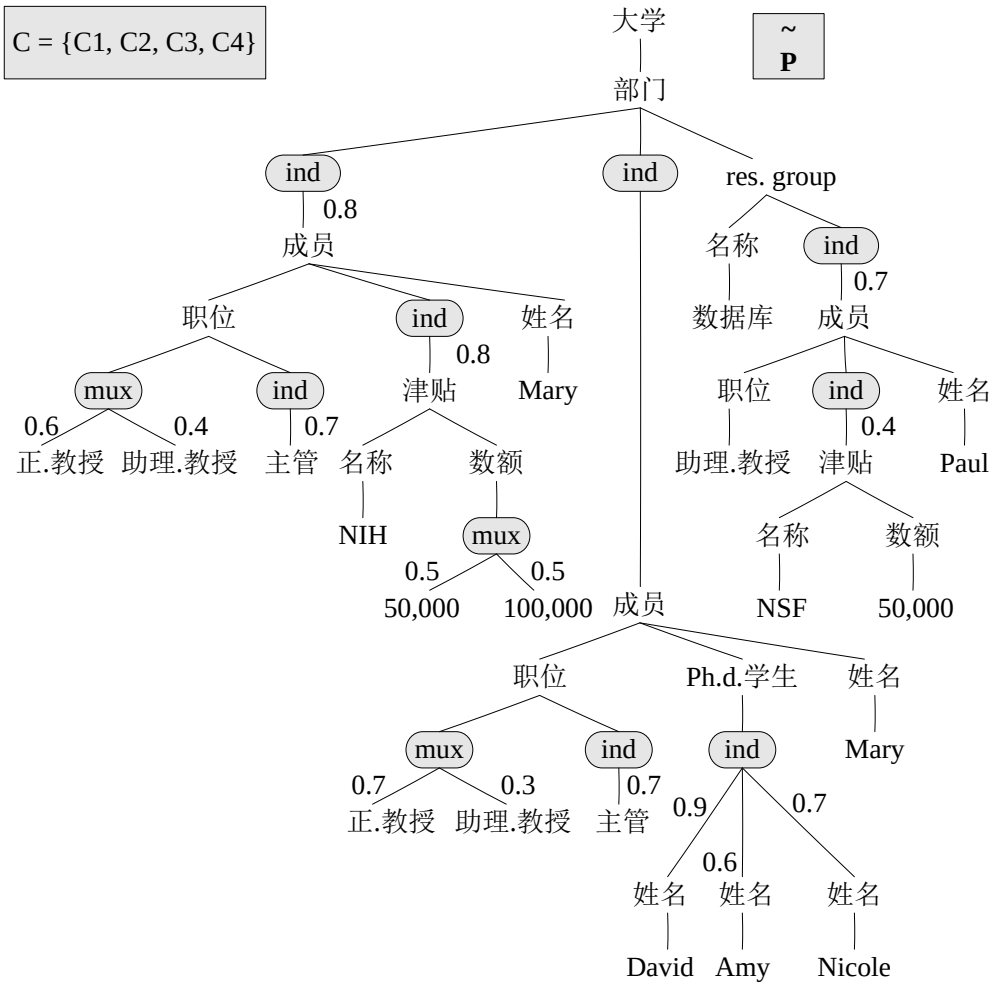


图 4-6 一个描述大学部门信息的PXDB实例

4.3.2 模型定义

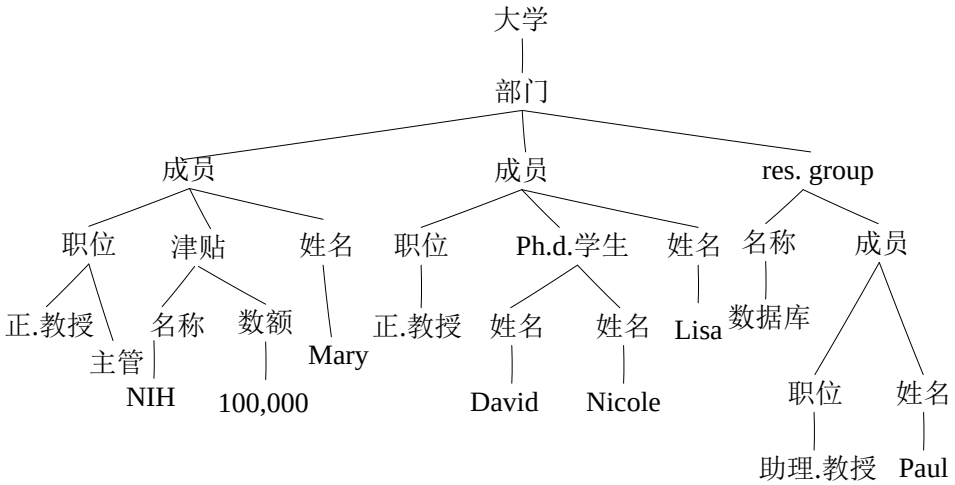


图 4-7　PXDB的一个随机实例

简单来说，一个PXDB就是一个P-文档与一个约束集合的组合。正式来说，PXDB就是一个由所有满足约束条件的文档组成的一个子空间。接下来，我们给出它正式的定义。

如果 $\Pr(P \models C) > 0$，则我们说一个P-文档 $\tilde{P}$ 与约束集 C 是一致的。一个PXDB是用对 $(\tilde{P}, C)$ 来定义的一个概率空间。所以 $\tilde{P}$ 是一个P-文档，C 是一约束集并且 $\tilde{P}$ 与 C 是一致的。PXDB $\tilde{D} = (\tilde{P}, C)$ 由 $\tilde{P}$ 中所有满足如下条件的文档 d 组成。要满足的条件为：$d \models C$，且 $\Pr(P = d) > 0$。$\tilde{D} = (\tilde{P}, C)$ 就是 $\tilde{P}$ 满足 C 的概率分布。也就是说，$\tilde{D}$ 中一个文档 d 的概率是 $\Pr(P = d \mid P \models C)$。

例如，图4-6描述了PXDB $\tilde{D} = (\tilde{P}, C)$，其中 $C = \{C_1, ..., C_4\}$。D 标记为一个随机变量，此变量代表了一个概率空间 $\tilde{D}$ 的一个文档。对于所有的文档 d （通过条件概率的定义），

$$\Pr(D = d) = \begin{cases} \dfrac{\Pr(P = d)}{\Pr(P \models C)}, if: d \models C; \\ 0, otherwise. \end{cases} \tag{4-2}$$

第二个例子，为了阐明一个PXDB中结点间存在的依赖关系，让我们考虑Ph.D.学生出现在一个随机文档中的事件。但是现在的概率空间是 $\tilde{D}$ 而不是 $\tilde{P}$ （都在图4-6中有描述）。由于 C_4，此事件依赖于其它两个事件：Lisa是正教授还是助理教授，以及她是否还有其它Ph.D.学生。依次地，C_3 暗示了前面的事件Lisa是否是一个主管。但是，由于 C_1，当Mary是主管时Lisa不能是一个主管。Mary是不是主管，依赖于她的职位（也就是说，是正教授还是助理教授），以及，由于 C_2，也依赖于Paul是否作为一个成员存在（否则，将少于三个成员，也因此不需要有主管）。所有，这似乎没有一种明显的方法可以计算Amy出现在 D

中。

上面的例子显示，一个PXDB的数据项间的概率依赖远远超出P-文档的树结构。特别是，它们所涉及的结点不一定是祖先—后代关系。

4.3.3 c-formulae

要解决满足约束与查询评估问题，就要首先能够计算文档树满足约束集要求的概率。我们通过c-formulae来完在这项工作。c-formulae囊括了约束与布尔查询，而且能够表达约束集（C）、约束集与匹配的合取($C \wedge T'$)。直观地，一个新的c-formulae是在已有的c-formulae上通过三步构建的。第一步，我们增加一个模式，并将c-formulae附加到其每个结点上。（当的一个结点到映射 d 上的一个结点时，子树必须满足附加于上的c-formulae。第二步，我们将所增加的模式转化为s-formulae，其中s-formulae为一个普通的选择器。第三步，我们使用一个或者更多的s-formulae来生成c-formulae，其中，c-formulae是普通的约束。然后，我们给出所增加模式的相互递归的定义，s-formulae和c-formulae。这样，公式要么引用c-formulae，要么引用s-formulae。

4.3.4 模型评价

PXDBs提供一套自然的、以数据库为驱动的方法，能够高效地描述概率数据。尽管P-文档本身有其局限性，但是，约束的增加使得它可以描述概率数据间复杂的依赖关系。尽管数据项间的依赖关系是复杂的，但是对于c-fomulae我们却可以高效地解决“满足约束”、“查询评估”和“采样”三大问题。此处的c-fomulae是指一种丰富的语言，包括了以下的聚合函数：计数，取最大值，取最小值和比率。

在文献[17]中，作者研究了概率XML的各种组合与自然扩展[13]。文献[17]描述的重点是模型的表达性，以及查询评估（形如模式投影）的易处理性。此工作也研发、实践和优化了一个模型的查询处理器。

在文献[17]的模型研究中，给出了一个清晰的权衡——查询评估的效率与模型中概率依赖性的表达能力之间的权衡。尤其是，高效的评估只能在分布结点是相互独立的情况下出现。在文献[17]中也展示了取近似值技术克服了固有的权衡，如下所述。对于表达能力很强的模型，查询评估也可以高效地（乘法的）取其近似值（文献[17]中有考虑过）。

模型PXDBs采用了一种完全不同的方法来克服上述权衡中固有的局限性。它之所以可以做到这样，是因为它在描述概率的依赖性采用了一个固定的约束集，而不是在很多分布结点间指定它们的关系（文献[10]中的做法）。以后可能会描述这样的一种情况，它不能被前者处理，但是它可以被轻易地构造且是简单明了的（因此，假设是固定的实际的）。此

外，对于文献[10]中的模型，即使是查询的反面（是计数的一个有一定限制的形式），也不能有效地取其评估的近似值（令其严格独立）。

第五章 其它模型

5.1 针对数据流的模型

5.1.1 针对数据流的模型

过去的网络服务提供商主要用传统的网络流量监控方法，就是采用离线分析，即先保存，在对其进行数据挖掘等处理。但实际中的一些需求是：用户要知道当前网络中的流量现状。进而根据当前状况采取可以提高服务性能的有利措施或者对有害行为进行预警等。而要在监控对象是实时、大量流数据的前提下，实现以上的需求，传统的方法是不合适的。

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。近来，对不确定数据流的模型研究越来越引起重视，但是相关的成果却不算多。下面，我们来看一个比较一般的例子。

5.1.2 一个常用模型的定义

文献[14]的定义了一个比较常用的模型。模型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。其相关的定义如下：

不确定域的定义：令B表示一个离散基本域，令$\perp$作为特殊的记号来表示不属于B。关于B的一个不确定域U就是所有分布函数的集合，或者关于$B \cup \{\perp\}$的一个不确定域U是所有pdf的集合。

可能数据流的定义：给定一个概率数据流$P = \theta_1, \theta_2, ..., \theta_n$，我们考虑下面的定义的确定数据流的概率分布。单独地，对于每一个$i = 1, 2, ..., n$（按顺序），我们首先根据pdf θ_i来选择一个元素$b \in B \cup \{\perp\}$。如果$b = \perp$，我们不输出任何东西；否则，我们输出b；所得到的确定数据流是产生自上面的实验的基本元素序列。可以用这种方式生产的确定数据流叫做“产生自P的可能数据流”。我们根据上述产生可能数据流S的概率，将概率值与每个S相关联。值得注意的是，它在可能数据流的集合上定义了一个合法的概率分布。

例如某元组t 被描述为$(< i_1, p_1 >, ..., < i_m, p_m >)$，则$\forall 1 \le s \le m$，有$i_s \in B$， $\Pr[i_s] = p_s$，$\sum^{m} p_s \le 1$。 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$[-30, 50]$[1]，则可能的数据流为{< (2> 0 < , 0 > . 4 , , >。部分学者将研究重点放在

一个基本特例，即m = 1[15]。

5.1.3 相关窗口的分类

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[15]。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 ，014〉，〈22 ，016〉) 和(〈22 ，018〉) 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 ，018〉) 和(〈21 ，012〉，〈23 ，017〉) 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。

5.2 针对多维数据的模型

5.2.1 关于 OLAP

当今的数据处理大致可以分成两大类：联机事务处理OLTP（on-line transaction processing）、联机分析处理OLAP（On-Line Analytical Processing）。OLTP是传统的关系型数据库的主要应用，主要是基本的、日常的事务处理，例如银行交易。OLAP是数据仓库系统的主要应用，支持复杂的分析操作，侧重决策支持，并且提供直观易懂的查询结果。

联机分析处理 (OLAP) 的概念最早是由关系数据库之父E.F.Codd于1993年提出的，他同时提出了关于OLAP的12条准则。OLAP的提出引起了很大的反响，OLAP作为一类产品同联机事务处理 (OLTP) 明显区分开来。

随着数据库技术的发展和应用，数据库存储的数据量从20世纪80年代的兆（M）字节及千兆（G）字节过渡到现在的兆兆（T）字节和千兆兆（P）字节，同时，用户的查询需求也越来越复杂，涉及的已不仅是查询或操纵一张关系表中的一条或几条记录，而且要对多张表中千万条记录的数据进行数据分析和信息综合，关系数据库系统已不能全部满足这一要求。操作型应用和分析型应用，特别是在性能上难以两全，人们常常在关系数据库中放宽了对冗余的限制，引入了统计及综合数据，但这些统计综合数据的应用逻辑是分散而杂乱的、非系统化的，因此分析功能有限，不灵活，维护困难。在国外，不少软件厂商采取了发展其前端产品来弥补关系数据库管理系统支持的不足，他们通过专门的数据综合引擎，辅之以更加直观的数据访问界面，力图统一分散的公共应用逻辑，在短时间内响应非数据处理专业人员的复杂查询要求。

5.2.2 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。

5.2.3 相关模型

有人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[16]。

第六章 总结

6.1 内容总结

不确定性数据模型的定义是不确定性数据管理的首要任务。定义了合理的模型有利于数据的查询、更新等操作。同时，也利于数据的存储与索引。

当前，研究人员已经提出了相当数量的不确定性数据模型。并且，不确定性数据的分类也越来越细，例如文献[1]中提到的，将不确定性数据分为四大类——关系型不确定数据、半结构化数据、不确定数据流、不确定多维数据。随着分类的细化，各种数据模型的也越来有其针对性，即针对某一种类型的数据来建立一种适用的算法，而不是试图用一种模型来描述所有的类型的数据。目前，大多数的模型研究人员都把精力放在了针对半结构的数据建模上，并且建立了不少模型。例如P-文档、概率树、PXDB模型，以及PXML模型、PrXML模型等。

不确定性数据模型的研究主要是专注于以下三个方面的研究上。

第一方面是模型对不确定数据的描述能力。因为概率维的出现，传统的数据处理模式的不足之处越来越明显，建立新的数据模型成为迫切的需要。同时，不确定性数据的不完整性也是传统的数据模型所难以描述的，这样的情况在半结构化数据的处理中最为典型。不确定性数据的应用已是不可阻挡趋势，因此，建立与之相适应的数据模型也是势在必行。而建模的第一步，首先要考虑新模型对不确定数据描述能力。当前，在描述不确定性数据方面，模型已经相对成熟。例如，基于XML不确定模型，就充分利用了XML的优点，可以很自然地描述半结构化数据。而对于传统的关系型数据则最早被研究，并建立了不少出色的模型，如probabilistic c-table模型，probabilistic or-set-? Table模型等。

第二方面则是模型的约束表达能力。早在传统的确定数据模型的研究中，研究者们就意识到约束的重要性。在传统的确定数据模型中，约束主要是指保持完整性的约束，也包括用户自定义的约束等。而在不确定性数据模型中，数据的不完整性是被允许的。它的约束则与传统模型中的用户自定义约束更为相似。早期的不确定性数据模型并没有很好的考虑到模型约束的表达能力。但是，随着模型研究的深入，不确定性数据模型的约束表达能力也越来越受到研究者的重视。例如，probabilistic c-table模型[4]，模糊树模型[10]以及PXDB[12]模型等都有相当好的约束表达能力。

第三方面则是数据的查询评估效率问题。这是不确定性数据研究的一个比较严峻的挑战。因为在一般情况下，一个不确定性数据模型的表达能力越是强大，它的查询评估就越是复杂，反之亦然。例如，P-文档模型，虽然已有不少高效的查询算法，但是它的表达能

力却是很有限的；相反，模糊树模型的表达能力足以达到应用的水平，但是在其上的查询评估的复杂度却是#P-complete的，难以达到应用的水平。当然，这个问题也并不是不可解决的。例如，PXDB模型就在表达能力与查询效率两者间作了一个很好的权衡[12]。相信，随着研究的不断深入，会有更好的模型出现。

另外，值得一提的是，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。

6.2 展望

不确定性数据模型的研究一般被认为是从模型级别来描述数据的不确定性。近来，不确定性数据模型得到一定程度的发展，但是不确定性数据的研究任重而道远。在未来的研究中，以下两个方面将是模型研究的重点。一是，模型的表达能力改善（包括数据的表达能力与约束的表达能力）。尽管当前的一些模型的表达能力已相当的强了，但是表的方式却还显得复杂。所以，寻求更方便简洁的表达方式仍是研究的重点。二是，模型的查询效率问题。查询效率是不确定性数据模型的一个大难题。提高查询评估的效率将是未来不确定性数据建模的一大重点，也是一大挑战。

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 计算机科学报 第32卷 第1期 2009年1月

[2] Babcock B , Babu S , Datar M , Motwani R , Widom J . Models and issues in data st ream systems/ / Proceedings of the21st ACM SIGMOD2SIGACT2SIGART Symposium on Principles of Database Systems. Madison , 2002 : 1216

[3] Dalvi N , Suciu D. The dichotomy of conjunctive queries on probabilistic structures/ / Proceedings of t he 26t h ACM SIGMOD2SIGACT2SIGART Symposium on Principles of Database Systems. Beijing , 2007 : 2932302

[4] Green T J , Tannen V. Models for incomplete and probabilistic information. IEEE Date Engineering Bulletin , 2006 , 29 (1) : 17224

[5] NORBERT FUHR and THOMAS RO¨ LLEKE University of Dortmund. A Probabilistic Relational Algebra for the Integration of Information Retrieval and Database Systems

[6] Lakshmanan L V S , Leone N , Ross R , Subrahmanian V S.ProbView : A flexible database system. ACM Transactions on Database Systems , 1997 , 22 (3) : 4192469

[7] Hua M , Pei J , Zhang W J , Lin X M. Efficiently answering probabilistic t hreshold top2*k* queries on uncertain data/ / Proceedings of t he 24t h IEEE International Conference on Data Engineering. 2008 : 140321405

[8] Barbara D , Garcia2Molina H , Porter D. The management of probabilistic data. IEEE Transactions on Knowledge and Data Engineering , 1992 , 4 (5) : 4872502

[9] Nierman A, Jagadish H V. Pro TDB : Probabilistic data in XML/ / Proceedings of t he 28t h International Conference on Very Large Data Bases. Hong Kong , China , 2002 : 6462657

[10] Abiteboul S , Senellart P. Querying and updating probabilistic information in XML/ / Proceedings of t he 9t h International Conference on Extending Database Technology : Advances in Database Technology. Munich , 2006 : 105921068

[11] Senellart P , Abiteboul S. On t he complexity of managing probabilistic XML data/ / Proceedings of t he 26t h ACM SIGMOD2SIGACT2SIGART Symposium on Principles of Database Systems. Beijing , 2007 : 2832292

[12] Cohen S , Kimelfeld B , Sagiv Y. Incorporating const raint s in probabilistic XML/ / Proceedings of t he 27t h ACM SIGMODSIGACT2SIGART Symposium on Principles of Database Systems. Vancouver , 2008 : 1092118

[13] B. Kimelfeld and Y. Sagiv. Matching twigs in probabilistic XML. In *VLDB*, 2007.

[14] J ayram T S , Kale S , Vee E. Efficient aggregation algorit hms for probabilistic data/ / Proceedings of t he 18t h Annual ACMSIAM Symposium on Discrete Algorit hms. New Orleans , 2007 : 3462355

[15] Cormode G, Garofalakis M. Sketching probabilistic data streams/ / Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. Beijing , 2007 :2812292

[16] Burdick D , Doan A , Ramakrishnan R , Vait hyanat han S.Olap over imprecise data wit h domain const raint s/ / Proceedings of t he 33rd International Conference on Very Large Data Bases. Vienna , 2007 : 39250

[17] B. Kimelfeld, Y. Kosharovski, and Y. Sagiv. Query efficiency in probabilistic XML models. In *SIGMOD*, 2008.

[18] TOMASZ IMIELII~ISKI AND WITOLD LIPSKI~ JR. Pohsh Academy of Sctences, Warsaw, Poland Incomplete Information in Relational Databases

致谢

本文是在我的导师的悉心指导下完成的。从论文的选题、研究思路的确定、论文的撰写直到论文修改的整个过程中，倾注了大量的心血和精力。导师认真求实的作风和积极向上的人生态度对我影响颇深，在此谨向她表示衷心的感谢和敬意。

毕业设计过程中，同队的同学也让我受益匪浅。我们一起收集资料，一起讨论问题，一起听相关的讲座，建立了真挚的友谊。感谢他们为我的学习、研究创造了良好的氛围，在生活中给予的帮助和鼓励，共同分享了艰辛和欢乐，为我的大学生活留下了美好的回忆。

再次，感谢我的家人。感谢父母为我付出的许多许多。是他们在物质与精神上的支持，让我得以专心于学业，是他们的孜孜不倦的身影让我不敢放松学习。

最后感谢，是这所学校赋予我青春最宝贵的财富！感谢一直关心、爱护我的老师、同学们。