一种基于概率图模型的不确定性数据世系表示方法

岳 昆 刘惟一 朱运磊 张 伟
(云南大学信息学院计算机科学与工程系 昆明 650091)

摘 要 不确定性数据的世系分析是基于数据产生和演变的过程来跟踪数据不确定性的来源.为了有效地描述数据间复杂的相关性及不确定性,并从理论上保证世系分析中概率计算的正确性,文中研究了基于贝叶斯网这一重要的概率图模型的不确定性数据世系表示方法.以世系的布尔公式和不确定性数据本身为出发点,提出了将布尔公式等价转换为贝叶斯网的方法,并讨论了相应的条件独立性质和概率语义.案例研究和实验结果表明,文中的方法为世系分析提供了一种有效性的、可扩展的数据相关性表示和概率计算框架.

关键词 不确定性数据;世系表示;概率计算;概率图模型;贝叶斯网
中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2011.01897

A Probabilistic-Graphical-Model Based Approach for Representing Lineages in Uncertain Data

YUE Kun LIU Wei-Yi ZHU Yun-Lei ZHANG Wei
(*Department of Computer Science and Engineering, School of Information Science and Engineering, Yunnan University, Kunming* 650091)

Abstract Analyzing lineage (or called provenance) of uncertain data is to trace the origin of uncertainty based on the process of data production and evolution. To represent complex correlations and their uncertainties among uncertain data objects, and then guarantee the correctness of probability computations in lineage analysis theoretically, we study the method for representing lineages of uncertain data based on Bayesian network, an important probabilistic graphical model. Starting from the lineages' Boolean formula and the uncertain data, we propose the method to transform Boolean formulas into Bayesian network equivalently, and discuss the corresponding probabilistic semantics and properties. Case studies and experimental results show that the proposal in this paper provides an effective and extensible framework for representing data correlation and evaluating uncertainties in lineage analysis.

Keywords uncertain data; lineage representation; probabilistic evaluation; probabilistic graphical model; Bayesian network

1 引 言

数据规模不断扩大、共享范围越来越广、数据形式多样化,大量的不确定性数据随之产生,普遍存在于经济、军事、物流、金融、电信和科学计算等领域中、并且扮演着关键的角色.当前数据特点的新变化,使得不确定性数据管理成为了国内外研究的热

收稿日期:2011-07-18;最终修改稿收到日期:2011-08-19.本课题得到国家自然科学基金项目(61063009, 60933001)、国家教育部博士点基金新教师类课题(20105301120001)、国家教育部科学技术研究重点项目(211172)、国家"九七三"重点基础研究发展规划项目基金(2010CB328106)资助.**岳 昆**,男,1979 年生,博士,副教授,主要研究方向为不确定性数据管理、智能数据分析、不确定性人工智能与应用. E-mail: kyue@ynu.edu.cn. **刘惟一**,男,1950 年生,教授,博士生导师,主要研究领域为数据与知识工程. **朱运磊**,男,1987 年生,硕士研究生,主要研究方向为不确定性数据管理、不确定性人工智能. **张 伟**,男,1985 年生,硕士研究生,主要研究方向为不确定性数据管理、不确定性人工智能.

点，目前仍面临着不确定性建模、管理和优化等方面的挑战[1-2].

不确定性数据世系分析，是指基于数据产生和演变的过程来跟踪数据不确定性的来源，在各类不确定性数据的查询优化、集成及质量保证等方面，具有广泛的应用. 文献[3]对世系管理研究进行了综述，指出了不确定性数据世系分析研究的重要意义和面临的挑战. ULDB(Uncertainty-Lineage DataBase)是 Stanford 开发的面向世系分析的不确定性数据库[4]，是许多不确定性数据世系分析研究的基础. 目前，世系通常基于连接查询被表示为所涉及元组的布尔公式(图 1 给出了一个元组独立假设下不确定性数据世系的例子[4]). 基于表示世系的布尔公式计算查询结果的概率(如计算 $\lambda(31)$ 的概率)，称为世系概率计算，是世系分析研究的重要内容，有效支持查询处理的重要任务，也具有较大的挑战，成为了目前不确定性数据世系研究的焦点[4-6].

Attends

ID	person	day	prob
11	Garcia-Molina	Monday	0.8
12	Garcia-Molina	Wednesday	0.7
13	Ullman	Wednesday	0.6

Events

ID	day	event	prob
21	Monday	Reception	0.8
22	Tuesday	Museum	1.0
23	Wednesday	Banquet	0.9

$EventRoster=\pi_{person,\ event}(Attends \bowtie Events)$的查询结果

ID	person	event	prob	世系
31	Garcia-Molina	Reception	0.64	$\lambda(31)=11\wedge 21$
32	Garcia-Molina	Banquet	0.63	$\lambda(32)=12\wedge 23$
33	Ullman	Banquet	0.54	$\lambda(33)=13\wedge 23$

$EventAttendees=\pi_{person}(EventRoster)$的查询结果

ID	person	prob	世系
41	Garcia-Molina	0.8668	$\lambda(41)=(11\wedge 21)\vee(12\wedge 23)$
42	Ullman	0.54	$\lambda(42)=13\wedge 23$

图 1 不确定性数据及世系

Re 等人[5]针对大规模世系，提出了近似世系及其查询处理技术. Sarma 等人[6]针对 ULDB 查询处理，提出了基于世系表达式来计算查询结果概率的方法及优化策略. 这些方法将世系表示为布尔公式，并基于世系表达式计算查询结果的概率，但逻辑形式的布尔公式并不能有效地描述数据间复杂的相关性及数据相关的不确定性，也缺乏保证结果概率正确性的理论基础. 因此，如何针对表示为布尔公式的世系表达式，描述所涉及数据的相关性，并应用于世系表达式的高效概率计算，是这类研究的核心，即需要建立一种既能等价表达布尔公式信息、又能有效支持概率计算的模型.

幸运的是，以贝叶斯网(Bayesian Network，BN)为典型代表的概率图模型(Probabilistic Graphical Model，PGM)，是不确定性知识表示和推理的有效框架[7-8]，具有坚实的概率论理论基础和广泛的应用. Deshpande 等人[9]指出了 PGM 在数据库研究中的作用，基于 PGM 可有效地描述数据间的相关性和相互依赖. PGM 具有良好的概率语义和成熟的推理方法，为不确定性数据管理中概率计算这一关键问题提供了重要的工具，也是本文研究的基础.

近年来，基于 PGM 的不确定性数据管理已有许多研究工作，是目前用来描述不确定性数据库中数据相关性的重要方法. Wang 等人[10]基于 BN 扩展了概率关系模型，提出了直接支持概率数据库内部关系查询和概率模型上推理计算及查询优化的方法. Sen 等人[11]用 BN 表示不确定性数据库中元组间的相关性，从而避免了计算查询结果概率时的大量概率分布计算，也为高效的世系概率计算奠定了基础. PGM 也被用于世系的处理中，Kanagal 等人[12]基于连接树(Junction tree)及其上的概率推理算法，提出了可用于大规模世系表达式的概率计算方法，连接树就是 PGM 的一种等价表示. 这些方法基于 PGM 表示数据查询计划(或世系)对应的布尔公式，利用 PGM 的图形特征及其中的条件独立性[8-9]，大大简化概率的计算. 然而，Pearl[7]已给出重要结论：一个有向无环图只有满足了条件独立性、具有概率语义，才能视为 PGM 并进行概率推理. 但以上研究[10-12]并未给出满足条件独立性和概率语义的 PGM 构建方法，也未讨论图模型的条件独立性和概率语义，从而无法从理论和模型基础方面保证 PGM 与布尔公式具有等价的条件概率独立性，进而不能保证结果概率的正确性.

针对同时涉及逻辑知识与概率知识的情形，目前的研究通过逻辑知识来扩展 BN 的概率推理能力[14]，或者用 BN 来增强逻辑知识的表达能力[15]，而都未考虑从逻辑知识到概率模型的等价转换，进而都不能用来将表示世系的布尔公式等价地转换为可有效支持概率计算的 PGM. 因此，本文主要研究从表示世系的布尔公式到 BN 的等价转换方法，旨在为不确定性数据世系分析提供一种有效支持世系表示和概率计算的理论基础和知识框架. 总的来说，

本文的研究主要包括：

(1)从布尔公式到 BN 有向无环图结构的等价转换.

BN 是以随机变量为结点的有向无环图(Directed Acyclic Graph,DAG),每个结点有一张定量反映变量间依赖关系的概率参数表(Conditional Probability Table,CPT)[7-8],构建反映变量间条件概率独立性的 DAG,是构建 BN 的关键.为了以表示不确定性数据连接查询世系的布尔公式构建相应的 BN,本文首先将各布尔公式等价地表示为 Horn 子句[13],利用 Horn 子句与逻辑蕴含式之间的等价特征,重点研究并提出了从 Horn 子句到 DAG 的转换算法以及多个 Horn 子句所对应 DAG 的合并策略,将其称为世系图(Lineage Graph,LG).我们进一步讨论 LG 的条件独立性和概率语义,证明 LG 为概率图模型,从而保证可基于 LG 进行有效的概率推理.LG 等价地反映了世系表达式的逻辑语义和概率独立性,为基于 LG 进行世系概率计算奠定了模型层面上的基础.

(2)BN 概率参数的计算.

基于 LG 的图形特征和条件独立语义,针对 LG 本身的性质和 BN 概率推理的特点,本文通过为 LG 中各结点定义满足布尔公式的函数,给出了从 LG 和不确定性数据计算各结点 CPT 的方法,从而构建了等价反映世系表达式的 BN,称为世系贝叶斯网(Lineage Bayesian Network,LBN),弥补了现有基于 PGM 的不确定性数据管理方法在理论基础和模型构建方面存在的不足之处.

(3)案例研究和实验分析.

本文进一步给出了基于 LBN 进行世系概率计算的案例,同时对 LBN 的构建方法及相应的世系概率计算进行了实验测试.案例研究和实验结果表明,本文的方法为世系分析提供了一种有效性的数据相关性表示和概率计算框架.

本文第 2 节是研究的重点,给出从表示世系的布尔公式和不确定性数据构建 LBN 的方法;第 3 节给出基于 LBN 进行世系概率计算的案例研究;第 4 节给出实验结果;第 5 节总结全文,并展望将来的研究工作.

2　从世系表达式到贝叶斯网

假设不确定性数据库中元组相互独立,借鉴文献[4-6]中的相关描述,我们基于 x-relation 和元组级不确定性,以世系表达式为重点,首先定义带有世系的不确定性数据库,作为后续讨论的基础.

定义 1.　带世系的不确定性数据库 D 是一个三元组$(\overline{S},I(\overline{S}),\lambda)$,其中

(1)$\overline{S}=S_1,S_2,\cdots,S_n$ 为输入的基本 x-relation 集合,其中每个 S_i 是包含若干元组的多数据集.

(2)D 中每个元组有一个唯一的标识,$I(\overline{S})$ 为所有标识的集合.

(3)λ 为表示为布尔公式的世系函数.设 x-relation R 为 $\overline{S}$ 上的查询处理结果,$I(R)$ 为 R 中元组标识的集合,则 $t\in I(R)$,$\lambda(t)$:$I(\overline{S})\rightarrow I(R)$ 定义为 $\overline{S}$ 上的布尔公式,$\lambda(t)\rightarrow t$ 称为世系表达式.

注意到,$\lambda(t)\rightarrow t$ 反映了得到查询处理结果 t 的过程,$\lambda(t)$ 中仅包含 $I(\overline{S})$ 中元组标识的布尔公式,而不包含查询处理中间结果元组的标识.以元组标识作为原子公式,取值为 1(True)或 0(False),分别表示给定查询中是否包含该元组.

例 1.　针对图 1 中的连接查询 $EventRoster=\pi_{\text{person, event}}(Attends\bowtie Events)$ 和 $EventAttendees=\pi_{\text{person}}(EventRoster)$,结果元组 41 的世系表达式为 $(11\wedge 21)\vee(12\wedge 23)\rightarrow 41$.

为了从世系表达式 $\lambda(t)$ 和输入的不确定性数据 $\overline{S}$ 构建等价的 BN,以下 2.1 节和 2.2 节分别讨论 BN 的概率图模型结构构建和各结点概率参数的计算.

2.1　构建贝叶斯网的有向无环图结构

我们首先从世系表达式构建反映数据间依赖关系的 DAG,然后分别讨论其条件独立性和概率语义,证明该 DAG 是反映数据间条件独立性的概率模型,作为 BN 的 DAG 结构.

2.1.1　基于世系表达式构建依赖模型

Horn 子句是文字的析取式,其中带有最多一个正文字,基于 Horn 子句进行逻辑推理可得出唯一的结论,且 Horn 子句可以方便地转换为等价的逻辑蕴含式[14].因此,我们将 $\lambda(t)$ 中的元组标识(正文字)作为原子公式(其否定也是原子公式),为了将布尔公式转换为等价的 DAG,首先将 $\lambda(t)\rightarrow t$ 等价地转换为 Horn 子句的合取式,Horn 子句的集合记为 Σ.

例 2.　例 1 中的世系表达式 $(11\wedge 21)\vee(12\wedge 23)\rightarrow 41$,可做如下转换:$(11\wedge 21)\vee(12\wedge 23)\rightarrow 41=\overline{(11\wedge 21)\vee(12\wedge 23)\vee 41}=(\overline{(11\wedge 21)}\wedge\overline{(12\wedge 23)})\vee 41=(\overline{11}\vee\overline{21}\vee 41)\wedge(\overline{12}\vee\overline{23}\vee 41)$,转换结果为 Horn 子句 $(\overline{11}\vee\overline{21}\vee 41)$ 和 $(\overline{12}\vee\overline{23}\vee 41)$ 的合取式.

为了将世系表达式转换为满足条件独立语义的DAG模型，即用图模型表示查询处理中数据的相互依赖关系(即数据的相关性)，我们将其称为世系依赖模型，记为 M_Σ.

定义 2. 设 F 为 Σ 中原子公式的集合，X 和 Y 分别是 Σ 中两个不相交的原子公式集，X 和 Y 对 Z 独立，记为 $\langle X|Z|Y\rangle_{M_\Sigma}$，$Z=F\backslash(X\cup Y)$，如果

(1)存在一个 Horn 子句的集合 $CS\subseteq\Sigma$，使得 F 中包含于 CS 中的原子公式也包含 X 和 Y；

(2)不存在 Σ 中的任何一个 Horn 子句 C_i，使得 C_i 包含 X 和 Y.

例 3. 若 $\Sigma=(C_1=\overline{A}\vee\overline{B}\vee C)\wedge(C_2=\overline{A}\vee D)\wedge(C_3=\overline{B}\vee D)$，$S'=\{\overline{A},\overline{B},C,D\}$，则存在 $CS=\{C_1,C_2,C_3\}$，使得 $X=C,Y=D$，有 $\{\overline{A},\overline{B},C,D\}\supseteq X\cup Y$，且不存在任何一个 $C_i(i=1,2,3)$ 包含 $X\cup Y$，因此 $Z=S'\backslash(X\cup Y)=\overline{A}\cup\overline{B}$，有 $\langle X|Z|Y\rangle_{M_\Sigma}$，即 $\langle C|\overline{A}\cup\overline{B}|D\rangle_{M_\Sigma}$.

由世系表达式 $A_1A_2\cdots A_n\rightarrow B$，可以得到等价的 Horn 子句 $C_i=(\overline{A_1}\vee\overline{A_2}\vee\cdots\vee\overline{A_n}\vee B)$，其中 B 为查询处理结果元组的标识，$A_1,A_2,\cdots,A_n$ 为输入元组的标识，是 C_i 中依赖关系的基础. 直观地，C_i 蕴含 $A_1\wedge A_2\wedge\cdots\wedge A_n\rightarrow A_i(i=1,2,\cdots,n)$ 和 $A_1\wedge A_2\wedge\cdots\wedge A_n\rightarrow B$. 因此，我们将 $C_i=(\overline{A_1}\vee\overline{A_2}\vee\cdots\vee\overline{A_n}\vee B)$ 等价地转换为 $C_j=(\overline{A_1}\vee\overline{A_2}\vee\cdots\vee\overline{A_n}\vee\overline{A_1\wedge A_2\wedge\cdots\wedge A_n}\vee B)$，将 $\overline{A_1\wedge A_2\wedge\cdots\wedge A_n}$ 称为 C_j 的核，记为 $Core_j$. 为了表示的方便，我们用 $A_1A_2\cdots A_n$ 表示 $A_1\wedge A_2\wedge\cdots\wedge A_n$.

C_j 的世系图(LG) $G=(V,E)$，其中 V 是结点的集合，E 是有向边的集合. G 中的每个结点对应 C_j 中的一个原子公式或原子公式的合取式，若 C_j 蕴含 $X\rightarrow Y$，那么 G 中就有一条从 X 指向 Y 的有向边. 根据上述思想，C_j 可转换为图 2 中与之逻辑上等价的 G_1 或 G_2，但是 G_1 不满足 PGM 的条件独立性，由后续讨论可知 G_1 不是一个概率模型. 因此，我们首先针对各 C_i 构建形如 G_2 的 LG，然后再将构建的 LG 进行合并，下面给出构建 LG 的方法.

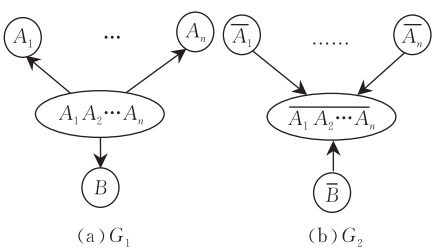


图 2 C_j 对应的两个世系图

考虑不确定性数据查询本身的特征，连接查询的安全性决定了是否可以得到正确的查询结果概率，不安全(Unsafe)的数据查询(中间结果不独立)处理是 #P 困难的问题[17]. 以两个 x-relation 的安全(Safe)连接查询为代表，如例 2 中的 $(11\wedge 21)\vee(12\wedge 23)$，基于上述构建 LG 的基本思想，核结点中包含参与连接的每个 x-relation 中的元组，对应于查询处理的中间结果，例如 $\overline{11\wedge 12}$ 和 $\overline{12\wedge 23}$，各 LG 中的核结点相互独立. 然而，对于形如 $(11\vee 13)\wedge 23\rightarrow 51$ 的不安全查询，由 $\overline{51}$ 共同指向的两个核结点 $\overline{12\wedge 23}$ 和 $\overline{13\wedge 23}$ 并不独立，这与概率模型的条件独立性[7]相违背.

文献[17]也针对引起不安全的数据引入了符号变量、给出了局部世系的概念，用 And-Or 图描述独立关系的中间结果. 因此，借鉴这一思路，为了基于 PGM 给出一种对于安全和不安全连接查询都适用的统一世系表示机制，且保证图模型的条件独立性，我们将参与某个连接查询、且属于同一个的 x-relation 的多个元组的析取式视为一个整体来考虑.

定理 1. 对于任意 S_i 与 $S_j(i\neq j,S_i,S_j\in\bar{S})$ 连接查询的世系函数 $\bigvee_{i,j}((t_{i1}\vee\cdots\vee t_{ik})\wedge(t_{j1}\vee\cdots\vee t_{jl}))$，令 $x=t_{i1}\vee\cdots\vee t_{ik}$，$y=t_{j1}\vee\cdots\vee t_{jk}$，则 LG 中对应于 $\bigvee_{i,j}(x\wedge y)$ 的任意两个核结点相互独立.

证明. 由于 x 和 y 分别是对 S_i 和 $S_j(i\neq j)$ 执行选择操作的结果，$x\cap y=\varnothing$，则对于任意两个核结点 $\overline{x_1\wedge y_1}$ 和 $\overline{x_2\wedge y_2}$，由于 $x_1\cap y_1=\varnothing$ 且 $x_2\cap y_2=\varnothing$，又有 $x_1\neq x_2$ 且 $y_1\neq y_2$，则核结点 $\overline{x_1\wedge y_1}$ 与 $\overline{x_2\wedge y_2}$ 相互独立. 证毕.

因此，我们首先按照定理 1 对 $\lambda(t)\rightarrow t$ 的 Horn 子句集进行预处理，例如 $(11\vee 13)\wedge 23\rightarrow 51$ 对应的 Horn 子句为 $\bar{x}\vee\overline{23}\vee 51$(其中 $x=11\vee 13$)，仍将 x 视为原子公式. 然后，通过算法 1 构建各 Horn 子句对应的 LG.

算法 1. 对每个 C_i 构建 LG.

输入：世系表达式 $\lambda(t)\rightarrow t$

输出：各 C_i 对应的 LG G_i

步骤：

1. 初始化：

1.1. 将 $\lambda(t)\rightarrow t$ 等价地转换为满足定理 1 的 Horn 子句的合取式，得到各 Horn 子句的集合 Σ；

1.2. 将每个 $C_i=(\overline{A_1}\vee\overline{A_2}\vee\cdots\vee\overline{A_n}\vee B)$ 等价转换为 $C_j=(\overline{A_1}\vee\overline{A_2}\vee\cdots\vee\overline{A_n}\vee\overline{A_1\wedge A_2\wedge\cdots\wedge A_n}\vee B)$.

2. 对 Σ 中的每一个 Horn 子句，构建相应的 LG：

2.1. 针对 $C_j=(\overline{A_1}\vee\overline{A_2}\vee\cdots\vee\overline{A_n}\vee\overline{A_1\wedge A_2\wedge\cdots\wedge A_n}\vee B)$，构建形如图 3(a)中的 G_1；

2.2. 针对 $C_k=(\overline{A_i}\vee B)$，构建形如图 3(b)中的 G_2.

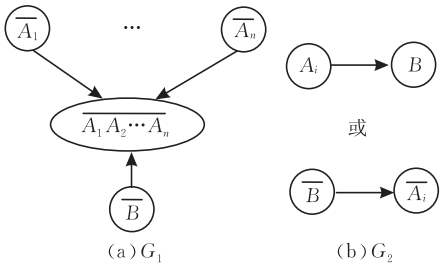


图 3　对应于 C_j 和 C_k 的子图

例 4.　针对世系表达式 $(11\wedge 21)\vee(12\wedge 23)\rightarrow 41$ 对应的 Horn 子句 $(\overline{11}\vee\overline{21}\vee 41)$ 和 $(\overline{12}\vee\overline{23}\vee 41)$，分别构建如图 4(a)和图 4(b)的 LG.

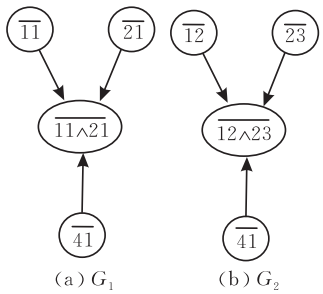


图 4　对应于 λ(41)→41 的两个 LG

注意到，指向核的结点包括两类原子公式对应的结点：连接所涉及 x-relation 中的元组标识的否定以及查询结果元组标识的否定. 为了计算结果元组 t 的概率，需考虑 Σ 中的所有 Horn 子句(具有合取关系)，即需将各 C_i 对应的 LG 合并为一个全局 LG. 合并 G_k 和 G_l 的基本思想如下：

(1) 对于查询结果元组 t，存在 $\bar{t}\rightarrow Core_i$ 和 $\bar{t}\rightarrow Core_j$，为了方便计算 t 的概率，我们将 $\bar{t}\rightarrow Core_i$ 和 $\bar{t}\rightarrow Core_j$ 分别等价转换为 $\overline{Core_i}\rightarrow t$ 和 $\overline{Core_j}\rightarrow t$，并引入结点 $\perp$、$\overline{Core_i}$ 和 $\overline{Core_j}$，$\perp=\overline{Core_i}\wedge Core_i$，$\perp$ 分别指向 $\overline{Core_i}$ 和 $Core_i$.

(2) 将 G_k 和 G_l 中的两个 $\perp$ 合并为一个 $\perp$，并指向原 $\perp$ 所指向的所有结点.

(3) 将 G_k 和 G_l 中的两个 t 合并为一个 t，并将原来指向 t 的所有结点指向合并后的 t.

需要说明的是，对任意 G_k 和 G_l，由定理 1 可知，不可能存在相同的 $\overline{A_i}$ 分别指向 G_k 和 G_l 的核结点. 引入的结点 $\perp=\overline{Core_i}\wedge Core_i$ 恒为 $False$ (即 0)，则 $\perp\rightarrow Core_i$ 和 $\perp\rightarrow\overline{Core_i}$ 都恒为 $True$ (即 1)，因此引入 $\perp$ 保证了 Horn 子句逻辑上的等价性.

例 5.　根据上述思想对图 4(a)和图 4(b)的 LG 进行合并，可以得到如图 5 所示的全局 LG.

若 $|\Sigma|$ 为 Horn 子句的个数，每个 Horn 子句至多包含 n 个原子公式，则算法 1 的计算时间为 $O(|\Sigma|\cdot n)$，对各 LG 进行合并的计算时间为 $O(|\Sigma|)$. 因此，我们可以在多项式时间内高效地构建 LG.

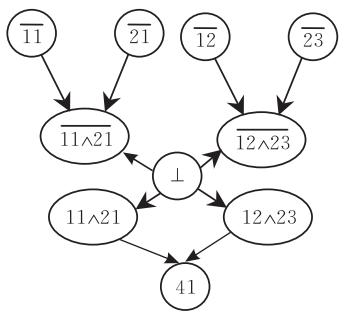


图 5　对应于 λ(41)→41 的全局 LG

如前所述，只有 LG 反映了 Σ 中数据的条件独立性，才可能作为拟构建 BN 的图结构. d-分离[7-8]是用 DAG 模型中描述变量间条件独立性的基本概念，下面首先给出 d-分离的定义，进而讨论 LG 的 d-分离性质，证明 LG 即为定义 2 中的依赖模型 M_Σ.

定义 3[7-8].　设 X，Y 和 Z 是 DAG $G=(V,E)$ 中不相交的结点集，称 X 和 Y 对于 Z 是 d-分离的，记为 $\langle X|Z|Y\rangle_G$，当且仅当从 X 中结点到 Y 中结点的路都不能被 Z 激活. DAG 中一条路被结点集 Z 激活，如果它满足：(1) 路上的每个汇聚结点(即两条有向边指向的公共结点)都在 Z 中，或者存在一条从汇聚结点到 Z 中的结点有向路；(2) 路上的非汇聚结点都不在 Z 中.

引理 1.　设 α 和 β 是 LG 中的两个非核结点，如果 α 和 β 包含于某个含有 α 的 Horn 子句 C_i 对应的子图 G_α 中，则 $\langle\alpha|V-\alpha-\beta|\beta\rangle_G$ 不成立.

证明.　由算法 1 和 LG 的合并策略可知，存在从 α 到 β 的路包含了核结点 $Core_\alpha$，该核结点在带有汇聚边的 G_α 中. 由定义 3，由于 $Core_\alpha\in V-\alpha-\beta$，因此 $\langle\alpha|V-\alpha-\beta|\beta\rangle_G$ 不成立.　证毕.

引理 2.　设 α 和 β 是 LG 中的两个非核结点，如果 α 和 β 不被包含于某个含有 α 的 Horn 子句 C_i 对应的子图 G_α 中，则 $\langle\alpha|V-\alpha-\beta|\beta\rangle_G$ 成立.

证明.　由算法 1 和 LG 的合并策略可知，α 和 β 之间任意一条路 r_i 必包含子图 G_α 和 G_β 的核结点 $Core_\alpha$ 和 $Core_\beta$. (1) 若 r_i 包含 $\perp$，注意到 $\perp$ 是非汇聚结点且 $\perp\in V-\alpha-\beta$，由定义 3，r_i 不能被 $V-\alpha-\beta$ 激活. (2) 若 r_i 包含 $Core_\alpha$ 和 $Core_\beta$，由于 $Core_\alpha$ 和 $Core_\beta$ 不相邻，则 r_i 中必存在非核结点 $\theta(\theta\in V-\alpha-\beta)$ 为非汇聚结点，由定义 3，r_i 不能被 $V-\alpha-\beta$ 激活. 因此，$\langle\alpha|V-\alpha-\beta|\beta\rangle_G$ 成立.　证毕.

由引理 1 和引理 2,可以容易地得出结论:LG 即为定义 2 中的依赖模型 M_Σ,如定理 2 所述.

定理 2. 设 Σ 为世系表达式所对应 Horn 子句的集合,G 为相应的 LG,X 和 Y 是 Σ 中两个不相交的原子公式集,$Z=F\backslash X\cup Y$,F 为 Σ 中原子公式的集合.设 $\mathcal{X}$, $\mathcal{Y}$ 和 $\mathcal{Z}=V-\mathcal{X}-\mathcal{Y}$ 为 G 中与 X,Y 和 Z 对应的结点集.若 $\langle\mathcal{X}|\mathcal{Z}|\mathcal{Y}\rangle_G$ 成立,则 $\langle X|Z|Y\rangle_{M_\Sigma}$ 成立;若 $\langle\mathcal{X}|\mathcal{Z}|\mathcal{Y}\rangle_G$ 不成立,则 $\langle X|Z|Y\rangle_{M_\Sigma}$ 也不成立.即 $\langle\mathcal{X}|V-\mathcal{X}-\mathcal{Y}|\mathcal{Y}\rangle_G\Leftrightarrow\langle X|F-X-Y|Y\rangle_{M_\Sigma}$.

2.1.2 依赖模型的概率语义

由 PGM 的定义可知,仅当依赖模型为一个概率模型,才能基于该模型进行概率推理[7-8].也就是说,只有有向无环图 LG 仍为一个概率模型,才可基于 LG 进行概率推理,进而得到结果元组 t 的概率.Pearl[7] 已经证明了依赖模型作为概率模型的必要条件:对称律、分解律、弱归并律、收缩律和相交律.定理 3 给出了依赖模型 LG 所满足的概率模型条件.

定理 3. 设 Σ 为世系表达式所对应 Horn 子句的集合,X, Y 和 Z 是 Σ 中 3 个不相交的原子公式集.$\langle X|Z|Y\rangle_{M_\Sigma}$ 满足如下独立条件:

(1) 对称律 $\langle X|Z|Y\rangle_{M_\Sigma}\Leftrightarrow\langle Y|Z|X\rangle_{M_\Sigma}$;

(2) 分解律 $\langle X|Z|Y\cup W\rangle_{M_\Sigma}\Rightarrow\langle X|Z|Y\rangle_{M_\Sigma}\wedge\langle X|Z|W\rangle_{M_\Sigma}$;

(3) 弱归并律 $\langle X|Z|Y\cup W\rangle_{M_\Sigma}\Rightarrow\langle X|Z\cup W|Y\rangle_{M_\Sigma}$;

(4) 收缩律 $\langle X|Z|Y\rangle_{M_\Sigma}\vee\langle X|Z\cup Y|W\rangle_{M_\Sigma}\Rightarrow\langle X|Z|Y\cup W\rangle_{M_\Sigma}$;

(5) 相交律 $\langle X|Z\cup W|Y\rangle_{M_\Sigma}\vee\langle X|Z\cup Y|W\rangle_{M_\Sigma}\Rightarrow\langle X|Z|Y\cup W\rangle_{M_\Sigma}$.

证明.

(1) 由独立表达式可知,对称律成立.

(2) 假设 $\langle X|Z|Y\cup W\rangle_{M_\Sigma}$ 成立,由定义 2 可知,Σ 中存在 3 个 Horn 子句的集合 $F_1\supseteq X\cup Z$, $F_2\supseteq Y\cup Z$, $F_3\supseteq W\cup Z$,而不存在 $C_i\supseteq x,y$(或 $C_i\supseteq x,w$),其中 x, y 和 w 分别为 X, Y 和 W 中的原子公式.根据定义 2,如果 $\langle X|Z|Y\rangle_{M_\Sigma}$ 且 $\langle X|Z|W\rangle_{M_\Sigma}$ 成立,则 Σ 中存在 $F_1'\supseteq X\cup Z$, $F_2'\supseteq Y\cup Z$, $F_3'\supseteq W\cup Z$,而不存在 $C_j\supseteq x,y$, $C_k\supseteq x,w$.注意到,Σ 中不存在这样的 C_i, C_j 和 C_k,因此 $F_1=F_1'$, $F_2=F_2'$, $F_3=F_3'$,进而 $\langle X|Z|Y\cup W\rangle_{M_\Sigma}\Rightarrow\langle X|Z|Y\rangle_{M_\Sigma}\wedge\langle X|Z|W\rangle_{M_\Sigma}$ 成立.

(3) 由 $\langle X|Z|Y\cup W\rangle_{M_\Sigma}$ 可知,Σ 中存在 3 个 Horn 子句的集合 $F_1\supseteq X\cup Z$, $F_2\supseteq Z\cup Y$, $F_3\supseteq Z\cup W$,而不存在 $C_i\supseteq x,y$ (或 $C_i\supseteq x$, w).根据定义 2,若 $\langle X|Z\cup W|Y\rangle_{M_\Sigma}$ 成立,则 Σ 中存在两个 Horn 子句的集合 $F_1'\supseteq X\cup Z\cup W$, $F_2'\supseteq Z\cup W\cup Y$,而不存在 $C_j\supseteq x$, y.若 F_1 和 F_3 为 Σ 中的 Horn 子句集,那么 F_1' 也是一个 Horn 子句集;若 F_2 和 F_3 为 Σ 中的 Horn 子句集,那么 F_2' 也是一个 Horn 子句集.由于 C_i 不存在,C_j 也不存在.因此,$\langle X|Z|Y\cup W\rangle_{M_\Sigma}\Rightarrow\langle X|Z\cup W|Y\rangle_{M_\Sigma}$ 成立.

(4) 由 $\langle X|Z|Y\rangle_{M_\Sigma}\vee\langle X|Z\cup Y|W\rangle_{M_\Sigma}$ 可知,Σ 中存在 4 个 Horn 子句的集合 $F_1\supseteq X\cup Z$, $F_2\supseteq Y\cup Z$, $F_3\supseteq X\cup Z\cup Y$, $F_4\supseteq Z\cup Y\cup W$,而不存在 $C_i\supseteq x,y$, $C_j\supseteq x,w$.根据定义 2,如果 $\langle X|Z|Y\cup W\rangle_{M_\Sigma}$ 成立,则 Σ 中存在两个 Horn 子句的集合 $F_1'\supseteq X\cup Z$, $F_2'\supseteq Z\cup Y\cup W$,而不存在 $C_k\supseteq x,y$(或 $C_k\supseteq x,w$).注意到 $F_1=F_1'$, $F_4=F_2'$,由于 Σ 中不包含 C_i 和 C_j,则 Σ 中不包含 C_k.因此,$\langle X|Z|Y\rangle_{M_\Sigma}\vee\langle X|Z\cup Y|W\rangle_{M_\Sigma}\Rightarrow\langle X|Z|Y\cup W\rangle_{M_\Sigma}$ 成立.

(5) 由 $\langle X|Z\cup W|Y\rangle_{M_\Sigma}\vee\langle X|Z\cup Y|W\rangle_{M_\Sigma}$ 可知,Σ 中存在 3 个 Horn 子句的集合 $F_1\supseteq X\cup Z\cup W$, $F_2\supseteq Z\cup W\cup Y$, $F_3\supseteq X\cup Z\cup Y$,而不存在 $C_i\supseteq x,y$, $C_j\supseteq x,w$.注意到 $F_1\supseteq F_1'$, $F_2=F_2'$,由于 Σ 中不包含 C_i 和 C_j,则 Σ 中不包含 C_k.因此,$\langle X|Z\cup W|Y\rangle_{M_\Sigma}\vee\langle X|Z\cup Y|W\rangle_{M_\Sigma}\Rightarrow\langle X|Z|Y\cup W\rangle_{M_\Sigma}$ 成立.

证毕.

由定理 3 可得出结论:依赖模型 LG 满足概率模型的条件,LG 既等价地反映了 $\lambda(t)\rightarrow t$ 中的逻辑语义,也等价地反映了其中蕴含的概率独立语义.上述讨论从理论上保证了所构建的 LG 为 PGM,我们以 LG 作为 BN 的 DAG 结构.

2.2 计算贝叶斯网的概率参数

为了计算 LG 中各结点的 CPT,与输入 x-relation 中相关元组对应的结点(即无父亲的结点),我们直接由输入的 x-relation 计算得到,例如:基于图 1 中的不确定性数据,可得到 $P(\overline{11}=1)=0.2$.而计算对应于查询处理中间结果的各结点的 CPT,是本节讨论的重点.根据 LG 的构建及合并方法,不难看出,LG 具有如下性质:

(1) 为了合并各 Horn 子句所对应的 LG,我们引入了保证逻辑上等价的 $\perp$ 结点,而这类结点对于描述 $\lambda(t)$ 中表达的连接查询、计算结果元组 t 的概率,并不产生直接影响.因此,在 CPT 计算和概率推理时,我们将忽略 LG 中的 $\perp$ 结点,图 6 给出了忽略图 5 中 $\perp$ 结点后的 LG 形式.

(2) 全局 LG 中,除 $\perp$ 结点之外,无父亲的结点对应于 x-relation 中的输入元组,无孩子的结点对

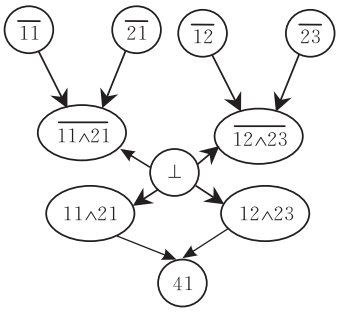


图 6　忽略了⊥的 LG 形式

应于查询结果元组 t，其它结点对应于查询处理的中间结果且相互独立(如定理 1 所述).

(3) 对于涉及多个 x-relation 的连接查询，多个父亲结点与其共同孩子结点之间为逻辑“或(OR)”的关系. 例如，图 5 中的 LG 包含如下“或”关系：$\overline{11}\vee\overline{21}\rightarrow\overline{11\wedge 21}$和$(11\wedge 21)\vee(12\wedge 23)\rightarrow 41$.

BN 的概率推理基于 DAG 结构和 CPT、通过信念传播(Belief propagation)完成[7-8]. 因此，为了构建中间结点的 CPT，体现父子结点取值之间的“或”关系，我们借鉴文献[11]中“因子(Factor)”参数的概念，给出通过父亲结点取值来导出孩子结点取值的函数，并反映基于 DAG 进行信念传播的思想，将其称为概率参数函数.

定义 4.　设 A 为 LG 中任意存在父亲的结点，A 的概率参数函数定义为

$$P(A=a\,|\,\boldsymbol{Pa}(\boldsymbol{A})=(a_1,a_2,\cdots,a_m))=f^{\mathrm{OR}}_{A,\boldsymbol{Pa}(\boldsymbol{A})}=\begin{cases}1, & \text{若 } a=a_1\vee a_2\vee\cdots\vee a_m\\ 0, & \text{其它}\end{cases},$$

其中：$\boldsymbol{Pa}(\boldsymbol{A})=A_1,A_2,\cdots,A_m$，为 A 的父亲结点集，$a,a_1,a_2,\cdots,a_m\in\{0,1\}$，分别为 $A,A_1,A_2,\cdots,A_m$ 的取值，$P(A=a\,|\,\boldsymbol{Pa}(\boldsymbol{A})=(a_1,a_2,\cdots,a_m))\in\{0,1\}$.

例 6.　对于表示为图 5 的 LG，根据定义 4，我们分别定义 $f^{\mathrm{OR}}_{\overline{11\wedge 21},\overline{11},\overline{21}}$，$f^{\mathrm{OR}}_{11\wedge 21,\overline{11},\overline{21}}$，$f^{\mathrm{OR}}_{\overline{12\wedge 23},\overline{12},\overline{23}}$，$f^{\mathrm{OR}}_{12\wedge 23,\overline{12},\overline{23}}$，$f^{\mathrm{OR}}_{41,11\wedge 21,12\wedge 23}$，部分结点的 CPT 如图 7 所示.

$P(\overline{11}=1)=0.2$
$P(\overline{11}=0)=0.8$

$P(\overline{21}=1)=0.2$
$P(\overline{21}=0)=0.8$

$P(\overline{12}=1)=0.3$
$P(\overline{12}=0)=0.7$

$P(\overline{23}=1)=0.1$
$P(\overline{23}=0)=0.9$

(a) 输入元组对应结点的CPT

$\overline{11\wedge 21}$ (布尔值)	11∧21 (布尔值)	$\overline{11}$ (布尔值)	$\overline{21}$ (布尔值)	$f^{\mathrm{OR}}_{\overline{11\wedge 21},\overline{11},\overline{21}}$ $f^{\mathrm{OR}}_{11\wedge 21,\overline{11},\overline{21}}$ (概率)
0	1	0	0	1
0	1	0	1	0
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	1
1	0	1	0	1
1	0	1	1	1

41 (布尔值)	11∧21 (布尔值)	12∧23 (布尔值)	$f^{\mathrm{OR}}_{41,11\wedge 21,12\wedge 23}$ (概率)
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

(b) 中间结点的CPT(概率参数函数)

图 7　LG 的 CPT

基于世系表达式和输入的不确定性数据，2.1 节和 2.2 节分别给出了构建 DAG 概率依赖模型、计算结点概率参数的方法，从而构建了描述世系表达式中所涉及数据间相互依赖关系的 PGM，我们将其称为世系贝叶斯网(Lineage Bayesian Network，LBN)，为世系分析提供了有效的概率推理模型基础.

3　案例研究

针对文献[6]中的经典案例，即本文图 1 中的不确定性数据和例 1 中的世系表达式 $(11\wedge 21)\vee(12\wedge 23)\rightarrow 41$，基于本文中基于 PGM 的世系表示方法，LBN 的 DAG 和 CPT 分别如图 6 和图 7 所示，本节给出将所构建的 LBN 应用于世系概率计算的案例. 需要计算结果元组 41 的概率，即变量 41 取值为 1(True)时的边缘概率分布 $P(41=1)$，基于 LBN 可以得到如下计算公式(由于篇幅的限制，我们忽略了值为 0 的所有项)：

$$P(41=1)=\sum_{\overline{11}}\sum_{\overline{21}}\sum_{\overline{12}}\sum_{\overline{23}}\sum_{\overline{11\wedge 21}}\sum_{\overline{12\wedge 23}}P(\overline{11})\cdot P(\overline{21})\cdot P(\overline{12})\cdot P(\overline{23})\cdot P(\overline{11\wedge 21}\,|\,\overline{11},\overline{21})\cdot P(\overline{12\wedge 23}\,|\,\overline{12},\overline{23})\cdot P(41=1\,|\,11\wedge 21,12\wedge 23)=$$

$0.8\times0.8\times0.7\times0.9\times1\times1\times1+\cdots+$
$0.8\times0.8\times0.7\times0.1\times1\times1\times1+\cdots+$
$0.8\times0.8\times0.3\times0.9\times1\times1\times1+\cdots+$
$0.8\times0.8\times0.3\times0.1\times1\times1\times1+\cdots+$
$0.8\times0.2\times0.7\times0.9\times1\times1\times1+\cdots+$
$0.2\times0.8\times0.7\times0.9\times1\times1\times1+\cdots+$
$0.2\times0.2\times0.7\times0.9\times1\times1\times1=0.8668.$

可见,基于 LBN 的世系概率计算结果与文献[6]中的结果相同,这从一定程度上验证了本文基于 PGM 的不确定性数据世系表示方法是正确的、可行的.同时,由以上 $P(41=1)$的计算过程不难看出,相对于直接基于可能世界的方法,LBN 中变量间的条件独立性使得边缘概率的计算得到很大程度的简化,大大减少了所涉及概率分布的数量,这正是基于 PGM 进行世系分析的优势所在.

4 实验结果

第 2 节给出了 LBN 构建算法的时间复杂度分析和 LBN 作为 PGM 的正确性证明,除此之外,为了进一步测试本文方法的有效性,我们实现了 LBN 构建以及基于 LBN 的世系概率计算方法,测试了 LBN 构建方法的效率、基于 LBN 表示世系的空间开销以及基于 LBN 进行世系概率计算的性能.实验环境如下:Intel Pentium(R) Dual-Core 3.00 GHz 处理器,2 GB 内存,Window XP Professional 操作系统,使用 MySQL5.1 存储不确定性数据和 LBN 的 CPT,使用 Java 语言编写程序.

首先,我们基于 TPC-H 基准数据集[18]中的 Orders 表和 Lineitems 表(分别简记为 O 和 L),针对 O 和 L 的一对一、一对多和多对多连接查询,测试了连接所涉及元组总数增加时构建 LBN 的效率,如图 8 所示.可以看出,构建 LBN 所需时间对元组总数并不敏感.元组数量相同时,多对多情形下构建 LBN 耗时最多,一对一情形下耗时最少,因为需要针对一对多和多对多的情形进行如定理 1 所描述的预处理,这成为了影响 LBN 构建效率的瓶颈.

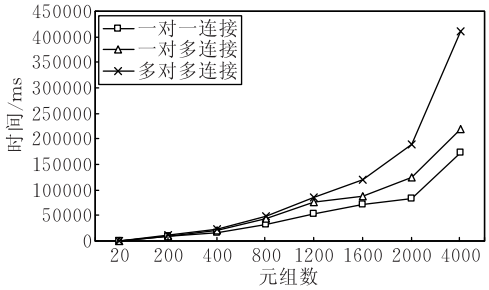


图 8 构建 LBN 的效率

然后,我们测试了用 LBN 表示不确定性数据连接查询世系的空间开销,并与直接基于布尔公式的世系表示的空间开销进行了比较.基于本文第 2 节中的从 Horn 子句到 DAG 的转换、DAG 的合并等步骤,我们可得到表示给定世系表达式的 LBN,作为最终表示世系信息的概率图模型.我们将表示世系的布尔公式及相应 LBN 的 DAG 结构存储在文件中,测试了连接操作所涉及元组总数增加时文件大小增加的趋势,如图 9 所示(采用了对数刻度).可以看出,对于表示同一世系信息的布尔公式和 LBN,后者所耗费的空间远远高于前者,这与布尔公式及 LBN 本身的表示机制而导致的空间开销差异相一致,因为后者不仅需要描述参与连接查询的输入元组、中间结果和结果元组,并且还要以 DAG 描述它们之间的相互关系.然而,基于构建得到的 LBN 可进行任意正确、高效的概率推理,而不需在每次世系分析时都为给定的布尔公式构建支持概率计算的模型,因此,LBN 与给定的世系表达式一一对应,LBN 的空间开销保证了方法统一且处理高效的世系分析.

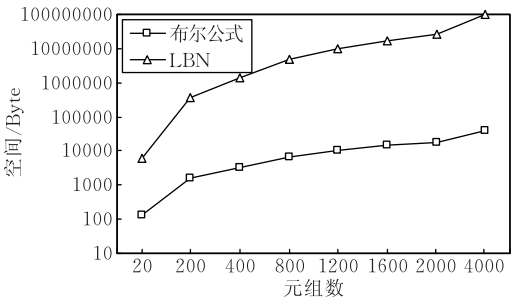
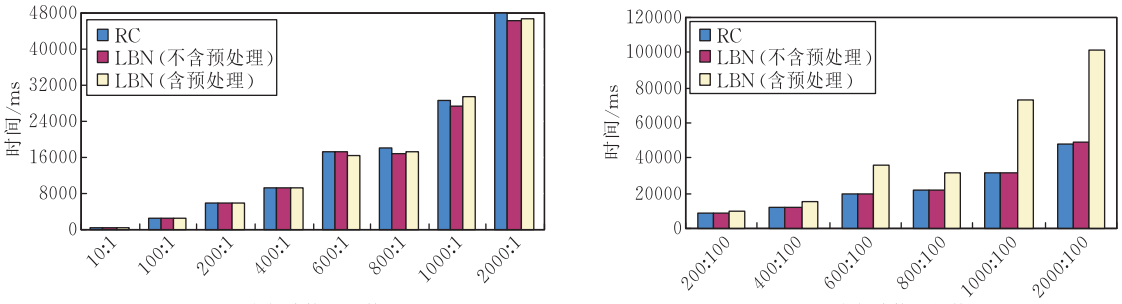


图 9 LBN 的空间开销

接着,我们分别考虑 O 表和 L 表的一对多和多对多连接查询且查询结果中只包含 1 个元组的情形,测试了世系概率计算(记为 LBN)的效率,并将其与文献[6]中信念值计算方法(记为 RC)的效率进行了比较,如图 10 所示.可以看出,LBN 和 RC 的效率随着参与连接的元组数量增加都不敏感,且计算时间增加的趋势基本一致.然而,预处理也是影响基于 LBN 进行世系概率计算效率的瓶颈.

综上,本节给出的实验结果表明,本文基于 PGM 的不确定性数据世系表示方法,在模型构建、空间开销和应用于世系概率计算等方面,都有较好的性能,从一定程度上验证了本文所提出方法的有效性.然而,针对预处理的高效算法和优化策略,是我们将要重点研究并进行测试的内容.



(a) O表和L一对多连接　　(b) O表和L多对多连接

图 10　基于 LBN 的世系概率计算效率

5　总结与展望

以有效支持不确定性数据世系分析为目标，以不确定性的跟踪和计算为核心，本文基于元组独立假设提出了一种基于概率图模型的世系表示方法，给出了将布尔公式等价转换为贝叶斯网的算法，适用于数据处理结果包含任意多个元组、以及世系表达式中包含任意多个布尔公式的情形. 理论分析、案例研究和实验结果表明，本文的方法为世系分析提供了一种有效性的数据相关性表示和概率计算框架. 然而，缺乏高效的预处理策略，是本文方法的主要不足之处.

直接计算不确定性数据查询结果元组的概率往往是 #P 困难的[4,6]，由第 3 节中的案例可以看出，虽然基于本文的方法可以提高世系分析的效率，但基于 LBN 的精确概率推理计算仍具有指数级的时间复杂度，因此，对精确概率推理算法的优化或进行近似推理，是我们正在开展的工作. 此外，将世系表示为 LBN，除了世系概率计算之外，仍可以利用 LBN 之上的后验概率推理进行任意形式的世系分析，进而追踪数据来源的不确定性、数据审核与质量评估；针对元组不独立的情形，讨论世系的有效表示和概率计算方法，这些是我们将要开展的工作.

参　考　文　献

[1] Aggarwal C, Yu P. A survey of uncertain data algorithms and applications. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(5): 609-623

[2] Zhou Ao-Ying, Jin Che-Qing, Wang Guo-Ren, Li Jian-Zhong. A survey on the management of uncertain data. Chinese Journal of Computers, 2009, 32(1): 1-16(in Chinese)
(周傲英，金澈清，王国仁，李建中. 不确定性数据管理技术研究综述. 计算机学报，2009，32(1)：1-16)

[3] Gao Ming, Jin Che-Qing, Wang Xiao-Ling, Tian Xiu-Xia, Zhou Ao-Ying. A survey on management of data provenance. Chinese Journal of Computers, 2010, 33(3): 373-389 (in Chinese)
(高明，金澈清，王晓玲，田秀霞，周傲英. 数据世系管理技术研究综述. 计算机学报，2010，33(3)：373-389)

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

[5] Re C, Suciu D. Approximate linage for probabilistic databases. PVLDB, 2008, 1(1): 797-808

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

[7] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann Publishers, 1988

[8] Liu Wei-Yi, Li Wei-Hua, Yue Kun. Intelligent Data Analysis. Beijing: Science Press, 2007(in Chinese)
(刘惟一，李维华，岳昆. 智能数据分析. 北京：科学出版社，2007)

[9] Deshpande A, Sarawagi S. Probabilistic graphical models and their role in databases//Proceedings of the 33rd International Conference on Very Large Data Bases, University of Vienna. Austria, 2007: 1435-1436

[10] Wang D, Michelakis E, Garofalakis M, Hellerstein J. BayesStore: Managing large, uncertain data repositories with probabilistic graphical models. PVLDB, 2008, 1(1): 340-351

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

[12] Kanagal B, Deshpande A. Lineage processing over correlated probabilistic databases//Proceedings of the ACM SIGMOD International Conference on Management of Data. Indianapolis, Indiana, USA, 2010: 675-686

[13] Russel S, Norvig P. Artificial Intelligence—A Modern Approach. Boston: Pearson Education, Publishing as Prentice-Hall, 2002

[14] Poole D. Probabilistic Horn abduction and Bayesian networks. Artificial Intelligence, 1993, 64(1): 81-129

[15] Jaeger M. Relational Bayesian networks//Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence. Brown University, Providence, Rhode Island, USA, 1997: 266-273

[16] Laskey K, Costa P, Janssen T. Probabilistic ontologies for knowledge fusion//Proceedings of the 11th International Conference on Information Fusion. Cologne, Germany, 2008: 1-8

[17] Jha A, Olteanu D, Suciu D. Bridging the gap between intensional and extensional query evaluation in probabilistic database//Proceedings of the 13rd International Conference on Extending Database Technology. Lausanne, Switzerland, 2010: 323-334

[18] Transaction Processing Council (TPC). TPC Benchmark H: Standard specification. http://www.tpc.org/tpch, 2006

YUE Kun, born in 1979, Ph. D., associate professor. His research interests include uncertain data management, intelligent data analysis, uncertainty in artificial intelligence and its applications.

LIU Wei-Yi, born in 1950, professor, Ph. D. supervisor. His research interests include data and knowledge engineering.

ZHU Yun-Lei, born in 1987, M. S. candidate. His research interests include uncertain data management, uncertainty in artificial intelligence.

ZHANG Wei, born in 1985, M. S. candidate. His Interests include uncertain data management, uncertainty in artificial intelligence.

Background

This paper discuss the representation of lineages or provenances in uncertain data that belongs to the database category. Uncertain data management is the current subject of intense debate. Analyzing lineage is to trace the origin of uncertainty based on process of data production and evolution. Lineages are widely used in several applications of uncertain data, such as query optimization, integration and quality endurance etc. In current research, lineages were represented as Boolean formulas and the probabilities of conjunctive query results were evaluated. Complex correlations and their uncertainties cannot be represented by the logical Boolean formulas, and there is no theoretical basis to guarantee the correctness of result probabilities.

In this paper, we focused on the method for representing lineages of uncertain data based on Bayesian network, an important and popular probabilistic graphical model, which is adopted as the framework for representing and computing uncertainties. The proposal in this paper can provide an effective and extensible framework for representing data correlation and evaluating uncertainties in lineage analysis. Based on the lineage representation given in this paper, we can make probabilistic computations from input uncertain data to the result of data evolution, on which the current works focus. Additionally and more generally, we can also make probabilistic inferences from the result to the participating inputs or any marginal probabilities in the evolution process of uncertain data.

The work in this paper is supported by the National Natural Science Foundation of China (Nos. 61063009, 60933001), the Ph. D. Programs Foundation of Ministry of Education of China (No. 20105301120001), the Foundation for Key Program of Ministry of Education of China (No. 211172), and National Basic Research (973) Program (No. 2010CB328106). These projects aim to propose a series of methods for representing, inferring and evaluating lineages in uncertain data, centered on inferring uncertainties. This paper gives the method for lineage representation based on the probabilistic graphical model, which provides a solid basis for the above issues.