基于元组存在性的概率数据模型研究

陈 鹏
(北京语言大学 北京 100083)

摘 要 随着信息与通讯技术的快速发展,数据管理正面临着越来越多的挑战,其中之一就是数据的不确定性。提出一种基于元组存在性的概率数据模型(PDMET),将该模型与目前存在的模型相比较,并证明了该模型的完整性,同时提出相应的概率关系代数和概率数据库查询算法。

关键词 数据不确定,概率数据库,概率数据模型,概率关系代数

中图法分类号 TP393 **文献标识码** A

Probabilistic Data Model Based on Existence of Tuple

CHEN Peng
(Beijing Language and Culture University, Beijing 100083, China)

Abstract With the development of ICT, data management has meet more and more challenge, one of challenge is uncertainty of data. This paper proposed a probabilistic data model based on existence of tuple, compared with other models, this model was proved to be complete, and we proposed corresponding probabilistic relational algebra and query algorithm.

Keywords Uncertain data, Probabilistic database, Probabilistic data model, Probabilistic relational algebra

1 引言

在传统的数据管理研究过程中,主要研究的是一些确定性的问题。如某个数据项是否是某个数据查询的结果,答案要么是要么不是。然而在实际中,存在许多不确定的因素和模糊的界定,这种不确定性可能是由于有些数据来源不可靠(例如,通过数据抽取获取的数据);也有可能是查询条件的不确定性,包括使用关键词查询、查询解析的二义性等;还有可能是查询结果与查询条件的匹配存在不确定性或者说存在一些相似度匹配(查询结果的排序)。

与此同时,随着 Web2.0、传感网络和物联网络的兴起与应用,数据采集、存储、传输和处理等过程都存在一些不确定性。需要管理数据中的许多不精确性,这些不精确性包括很多种类型,具体有不匹配的数据值、不精确的查询、不一致的数据、适配有误的模式等等。例如:

1. 随着 Web2.0 应用的兴起,各类博客(Blog)、SNS(Social Networking Service,社交网络服务)、维客(Wiki)等应用主要通过广大网民主动参与、自主撰写,从而产生海量的信息资源。Google base、Flickr、ESP 游戏、Mechanical Turk 都是一些实例,这些由网民自主参与撰写的内容往往不能够与我们事先规定好的固定数据模式相一致。

2. 大规模的数据抽取系统从文本中自动抽取数据,这些数据往往也是不精确的。

3. 随着传感网络和 RFID 部署所形成的物联网络的兴起与应用,出现大量的来自真实物理世界的数据。由于传感数据和 RFID 部署的一些特点,这些来自真实世界的数据本质上就有不确定性的特征。

为解决和处理这些数据的不确定性,需要引入概率数据库的研究。实际上,概率数据库的研究可以追溯到 20 世纪 70 年代中期和 80 年代初,为了处理 Codd 所提出的原始模型中没有解决的其它类型的信息,出现了扩展关系数据库的需求。其中,不完整信息和不确定信息的引入就是研究者所提出的扩展之一。

当前,模型数据不确定性主要是使用概率数据模型,也就是说,通过概率相关的数学概念表示数据中的不确定、不完整性等问题。

对数据库中概率、不确定、不完整以及模糊数据的研究由来已久。随着网络的发展,不确定性数据的应用越来越广泛,因此,该领域的研究在近些年有了突破[7]。概率数据管理领域中的工作主要在以下方面有所不同:

1. 概率是与属性还是与元组相关联。

2. 所生成的关系是否符合第一范式。

3. 是否能够模型现实世界的关联。

Barbara 等人提出一种将概率与属性相关联的方法,该方法能够模型单个元组内的任意属性关系。然而所产生的关系不属于第一范式,而且每个查询操作的语义都非常混乱。最近,Choenni 等人[8]从概念上讨论了如何使用 Dempster-Schafer 理论使查询语义一致。

Cavallo、Pittarelli 和 Dey、Sarkar 提出并研究了能够显性捕获互斥关系的元组不确定性模型。最近,Andritsos 等人使用相似的模型来描述脏数据,同时为该模型开发查询处理算法。在一个系列论文中,Fuhr、Rolleke 提出使用元组概率来

陈 鹏(1979—),男,博士,讲师,主要研究方向为形式化验证、人工智能。

模型信息检索中的不确定性数据。上述工作的一个普遍假设是元组相互独立。

Cheng[6]和 Xia[9]将属性与(连续型)概率分布相关联,并提出相应的查询处理和索引技术。Trio 系统旨在将数据血统、数据不确定性和数据准确性统一在一个模型中进行描述,他们也研究了各种模型的完备性和封闭性等问题。

2 概率推理逻辑及语义

为了形式化概率数据库,需要一个能够推理概率的逻辑。Nilsson 在该领域进行过开创性的工作。Nilsson 提出一个一阶逻辑的语义泛化,其中语句的真值在 0 和 1 之间取值。他的逻辑中的真值对应于普通一阶逻辑语句的概率。Nilsson 通过可能世界分析来定义语句概率的概念。

Fagin、Halpern 和 Meggido 等人开发了一种推理概率的逻辑。他们的逻辑语言允许类似"E1 的概率少于 1/3"或者"E1 的概率至少为 E2 概率的 2 倍",其中 E1 和 E2 代表任意事件。他们考虑所有的事件是可度量的(也就是表示可度量集)。考虑推理一个论域和公式 ϕ 的一阶逻辑。该逻辑允许形如"$w(\phi)1/2$"的公式,它能够解释为"公式 ϕ 被满足的概率大于或者等于 1/2"。他们也证明非可度量的例子相当于使用 Dempster-Shafer 的信任函数代替概率度量。

一阶概率逻辑中的概率存在两种解释:第一,概率代表不确定性;第二,概率代表统计信息。对应这两种解释,考虑使用两种方法为这些逻辑赋予语义。

第一种方法,就是在数据库的取值范围上进行相应的概率分布。这种方法比较适合进行数据库的统计和概要描述。这种语义理解便于表示包括类似"随机选择的鸟能飞的概率大于 0.9"的统计信息。

第二种方法,就是在所有可能的世界上进行相应的概率分布。这种方法比较适合描述对数据的某个论断的可信程度。这种语义理解便于描述类似"Tweety 能飞的概率大于 0.9"的信任度公式。

分别给出这两种语义理解的形式化定义。

假定我们具有一个在某个领域推理的一阶逻辑,且该语言由预测符号和函数符号组成的集合 ϕ 构成。

定义 1 概率结构 PS1 是一个三元组(D,π,μ)。其中 D 代表数据库中各属性的取值范围,π 将 ϕ 中的预测和函数符号都赋予 D 中正确的值。μ 是 D 上的一个离散的概率函数。也就是说,$\mu:D\to[0,1]$,$\sum_{d\in D}\mu(d)=1$。同时,我们可以在乘积取值范围 D_n(由 D 中的 n 个元组的元素)定义一个离散的概率函数 μn,$\mu n(d_1,\cdots,d_n)=\mu(d_1)\times\cdots\times\mu(d_n)$。与此同时定义一个 valuation 函数,将每个对象变量映射到 D 中的一个元素,每个域变量映射到一个实数。

定义 2 概率结构式 PS2 是一个四元组(D,S,π,μ),其中 D 是数据库各属性的取值范围,S 是状态或者说可能世界的集合,$\forall s\in S,\pi(s)$将 ϕ 中的预测和函数符号都赋予 D 中正确的值。μ 是 S 上的一个离散的概率函数,$\mu:S\to[0,1]$,$\sum_{s\in S}\mu(s)=1$。

事实上,在实际应用过程中,根据实际的需要可以考虑将两种语义解释统一起来,从而充分利用各种概率信息(统计信息、物理规律以及缺省规则中归纳出的信任度等)进行逻辑推理。

3 基于元组存在性的概率数据模型

近些年,在国际上,数据库研究领域中的许多学者已经提出一些关于概率数据模型的表示方法,其中主要包括:C-tables[1]、基于独立元组模型[3]、数据血统与不确定统一模型[4]、概率关系图模型[5]。

这些模型要么非常复杂,缺乏实用性[1];要么过于理想化,不能适应实际情况[2]。

我们的模型主要基于关系型数据库。

从可能世界语义的角度来看,一个概率数据库是一个概率空间 $PDB=(W,P)$,其结果是可能世界集合 $W=(I_1,\cdots,I_n)$。换句话说,这是一个函数 $P:W\to(0,1]$,$\sum_{I\in W}P(I)=1$。

基本符号定义:R 表示关系名,$Attr(R)$为属性,$r\subseteq U^k$ 是关系的一个实例,其中 k 是 R 的元数。$\overline{R}=R_1,\cdots,R_n$ 是一个数据库模式。

定义 3(不确定关系 R^p) 一个不确定性关系定义一个可能实例的集合,$I(R^p)=\{R_1,R_2,\cdots,R_n\}$,其中每个 R_i 是一个普通关系。

元组存在性随机变量。为了描述数据库中元组 t 存在的不确定性,将一个布尔随机变量 $t.e$ 与元组 t 关联。如果元组存在某个实例中,那么 $t.e$ 取值为 1(true),否则取值为 0(false)。

定义 4(元组依赖性因子 f) $f(X)$是一个在布尔型随机变量集合 $X=\{X_1,X_2,\cdots,X_n\}$上的一个函数,$f:X\to[0,1]$。其中,X_i 是元组存在性随机变量,n 表示 f 自变量的个数,称之为元组依赖性因子 f 的元数。

假定 f 的元数为 k,计算整个 f 函数的时间是 k 指数的,即 2^k。

元组依赖性因子主要用于编码概率数据库中各元组之间的依赖关系,可以将它作为度量依赖性的一个量化单位。

例 1 元组依赖性因子实例

S	A	B
s_1	a_1	1:0.6
s_2	a_1	2:0.4
s_3	a_2	1:0.6
s_4	a_2	2:0.4

其元组依赖性因子分别为:

$s_1.e$	$s_2.e$	fs_{12}
1	0	0.6
0	1	0.4

$s_3.e$	$s_4.e$	fs_{34}
1	0	0.6
0	1	0.4

定义 5(基于元组存在性的概率关系模型) 一个概率数据库 $\mathcal{D}$ 是二元组$(\mathcal{R},\mathcal{S})$,其中 $\mathcal{R}$ 是关系集合,$\mathcal{S}$代表 $\mathcal{R}$ 中所有元组存在性随机变量之上所定义的一个元组依赖性因子集合。

定义 6(概率数据模式的完整性) 如果对应一个给定模式的任何关系实例集合能够由数据模式 $\mathcal{M}$ 所表示的不确定关系所模型,那么称该数据 $\mathcal{M}$ 模式是完整的。

定理 1 基于元组存在性的概率关系模型是完整模型。

证明:给定实例集合$\{I_1,I_2,\cdots,I_n\}$,我们显示如何构造一个关系 $R=(T,f)$,使得 $I(R)=\{I_1,I_2,\cdots,I_n\}$。将 I_i 中出现的每个元组都对应一个元组标识符 $t\in T$。对于每个出现在 I_i 中不同的元组值,具有该元组值的 T 中的元组个数是在

任何实例中出现该值的最大数。对于任何 i，假定 $T_i \subseteq T$ 是出现在 I_i 中的元组标识符集合。$T_i=\{ti_1,ti_2,\cdots,ti_n\}$。假定 $T-T_i=\{si_1,si_2,\cdots,si_{mi}\}$。设定 factor 为：

$$\bigvee_{i=1}^{n} f(ti_1.e=1,ti_2.e=1,\cdots,tin.e=1,si_1.e=0,si_2.e=0,\cdots,si_{mi}.e=0)>0$$

$$\sum_{i=1}^{n} f(ti_1.e=1,ti_2.e=1\cdots,ti_n.e=1,si_1.e=0\ si_2.e=0,\cdots,si_{mi}.e=0)=1$$

f 的每个赋值都对应一个实例，上述的合取短语中的一个短语为真，f 才能满足。如果第 i 个短语为真，就获得实例 I_i。

例 2　基于元组存在性的概率关系实例

数据库可能实例($(I_1,0.2),(I_2,0.3),(I_3,0.5)$)包含如下实例：

I_1:[Amy,12/23/04,Standford,jay]

I_2:[Amy,12/23/04,Standford,jay]

[Amy,12/23/04,Standford,crow]

I_3:[Amy,12/23/04,Standford,jay]

[Amy,12/23/04,Standford,raven]

总共包含了如下 3 个元组：

t_1:[Amy,12/23/04,Standford,jay]

t_2:[Amy,12/23/04,Standford,crow]

t_3:[Amy,12/23/04,Standford,raven]

那么，基于元组存在性的概率关系 R^p 为：

	t_1	t_2	t_3	f
I_1	1	0	0	0.2
I_2	1	1	0	0.3
I_3	1	0	1	0.5

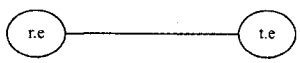
实际上，我们在证明过程中，使用到了完整元组存在性依赖因子。所谓完整元组存在性依赖因子，就是因子的定义域是所有元组存在性随机变量。

4　概率关系代数

主要介绍基于元组存在性的选择、笛卡尔乘积和投影 3 个主要的关系代数操作。在本节中所指的关系 R 都是代表概率关系。

1. 选择(σ)

假定 $\sigma_c(R)$代表我们感兴趣的查询，其中 c 表示选择操作符预测。元组 $t\in R$。如果没有一个能够满足条件 c，那么在结果元组中不出现 t。即使有一个元组 t 能够满足条件 c，我们产生一个中间元组 r，建立一个新的布尔变量 $r.e$ 与元组存在性相关联。可以确定 $r.e$ 与 $t.e$ 存在关联关系，即 $r.e$ 依赖 $t.e$。



为了刻画 $r.e$ 与 $t.e$ 的关联关系，在已知的 $f_t(t.e)$的基础上定义 $f_{tr}(r.e,t.e)$。不失一般性，我们假定

t.e	f_t
0	P
1	1−p

那么，

t.e	r.e	f_{tr}
0	0	P
1	1	1−p

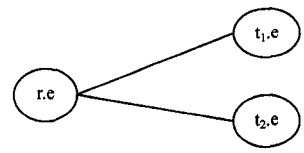
因此

r.e	f_r
0	p
1	1−p

所以，$\Pr(r.e=1)=\Pr(t.e=1)$。

2. 笛卡尔乘积(×)

假定 R_1 和 R_2 是笛卡尔乘积的两个关系。r 表示两个元组 t_1 和 t_2 的联合结果，其中 $t_1\in R_1,t_2\in R_2$。$r.e$ 表示元组 r 的存在性随机变量。$r.e$ 与 $t_1.e$ 和 $t_2.e$ 存在关联关系。



此时，存在两种情况：

1)$t_1.e$ 和 $t_2.e$ 相互独立

$t_1.e$	f_{t_1}
0	1−p
1	P

$t_2.e$	f_{t_2}
0	1−q
1	q

这样，

$t_1.e$	$t_2.e$	$f_{t_1t_2}$
0	0	(1−p)(1−q)
0	1	(1−p) q
1	0	p(1−q)
1	1	pq

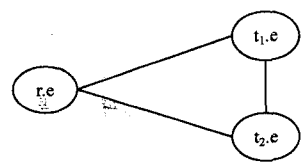
那么，

$t_1.e$	$t_2.e$	r.e	$f_{t_1t_2}r$
0	0	0	(1−p)(1−q)
0	1	0	(1−p) q
1	0	0	p(1−q)
1	1	1	pq

所以，$\Pr(r.e=1)=pq$。

2)$t_1.e$ 和 $t_2.e$ 存在依赖关系

如果 $t_1.e$ 和 $t_2.e$ 存在依赖关系，那么一定存在一个依赖性因子 $f_{t_1t_2}$ 刻画它们之间的关联关系。



不失一般性，假定 $f_{t_1t_2}$ 定义如下：

$t_1.e$	$t_2.e$	$f_{t_1t_2}$
0	0	p
0	1	q
1	0	m
1	1	1−p−q−m

其中，$1\geqslant p\geqslant 0,1\geqslant q\geqslant 0,1\geqslant m\geqslant 0,1\geqslant p+q+m\geqslant 0$。

那么，

$t_1.e$	$t_2.e$	$r.e$	$f_{t_1t_2}r$
0	0	0	p
0	1	0	q
1	0	0	m
1	1	1	1−p−q−m

所以，$\Pr(r.e=1)=\Pr(t_1.e=1 \wedge t_2.e=1)$。

3. 投影(Π)

假定$\Pi_a(R)$代表我们感兴趣的操作，其中$a \subseteq attr(R)$代表我们希望投影的属性集合。r代表投影$t \in R$的结果。其中假定$Tr=\{t_1,t_2,\cdots,t_k\}$为元组集合，$\Pi_a t_1=\Pi_a t_2=\cdots=\Pi_a t_k$，那么$\Pr(r.e=1)=\Pr(t_1.e=1 \vee t_2.e=1 \vee \cdots \vee t_k.e=1)$。

假定，$\Pr(t_1=1)=p_1$，$\Pr(t_2=1)=p_2$，$\cdots$，$\Pr(t_k=1)=p_k$。

讨论两个特殊的例子：

1) Tr中的所有元组相斥

$\Pr(r.e=1)=p_1+p_2+\cdots+p_k$

2) Tr中的所有元组相互独立

$\Pr(r.e=1)=1-(1-p_1)(1-p_2)\cdots(1-p_k)$

为了有效地计算投影(Π)，需要降低计算的复杂度。如果充分利用元组之间的依赖关系，就可以降低计算复杂度。

算法的思路：在Tr中通过相互独立的元组集合，假定存在n个独立的元组集合$\{T_{r1},T_{r2},\cdots,T_m\}$($T_{r1},T_{r2},\cdots,T_m$为$Tr$的一个划分，见图1)，其中$\forall t_i \in T_{ri}$，$\forall t_2 \in T_{rj}$，$i \neq j$，那么，$t_1$与$t_2$相互独立。将$t_1.e=1 \vee t_2.e=1 \vee \cdots \vee t_w.e=1$记为$T_{ri}=1$，其中$T_{ri}=\{t_1,t_2,\cdots,t_w\}$。

$$\begin{aligned}\Pr(r.e=1)&=\Pr(t_1.e=1 \vee t_2.e=1 \vee \cdots \vee t_k.e=1)\\&=\Pr((T_{r1}=1)(T_{r2}=1)\cdots(T_m=1))\\&=1-(1-\Pr(T_{r1}=1))(1-\Pr(T_{r2}=1))\cdots(1-\Pr(T_m=1))\end{aligned}$$

通过算法，将投影算法的复杂性从2^k降到2^q，其中$k=|T_r|$，$q=\mathrm{Max}(|T_{r1}|,|T_{r2}|,\cdots,T_m|)$，从而实现投影算法的优化。

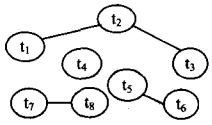


图1 元组依赖图

5 概率数据库的查询处理

概率数据库在查询处理方面与传统的数据库存在很大的区别。

当数据库R出现不完整和不确定信息的时候，为了说明数据库的语义，需要为数据库关联一个表示函数(representation function)$REP(R)$，这个函数是由关系所表示的可能世界集合。一般而言，表示函数允许将概率数据库与一个传统数据库集合相关联，其中每个传统数据库都是它的可能世界。世界的真实状态由其中一个可能世界所表示，然而，并不确定具体是哪个。

形式化来说，假定f是一个多集合关系R上的任意关系表达式，也就是说$R=\langle R_1,\cdots,R_k\rangle$。定义扩展于$\overline{f}(R)$满足：

$$REP(\overline{f}(R))=f(REP(R))=\{f(r) \mid r \in REP(R)\}$$

实际上，因为世界的实际状态r^*在$REP(R)$上，因此$f(r^*)$也是$f(REP(R))$上的一个可能世界。

概率关系上的操作如图2所示。

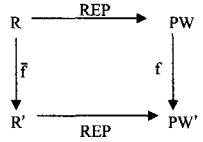


图2 概率关系上的操作

5.1 概率数据库查询的内涵与外延语义

基本符号定义：R表示关系名，$Attr(R)$为属性，$r \subseteq Uk$是关系的一个实例，其中k是R的维数。$\overline{R}=R_1,\cdots,R_n$是一个数据库模式，$D$代表数据库实例。$\Gamma \vdash D$表示$D$满足函数依赖性。

概率事件。假定AE是符号集合，$\Pr: AE \to [0,1]$是一个概率函数。AE中的每个元素是一个基本事件，假定所有的基本事件都是相互独立的。$\perp \in AE$代表不可能事件，$\Pr(\perp)=0$。一个复杂事件是使用$\vee$，$\wedge$将原子事件组合而成的。E代表所有复杂事件的集合。对于每个复杂事件e，假定$\Pr(e)$是它的概率。

概率数据库。一个概率关系是具有事件属性E的关系，事件属性的值是复杂事件。增加上标p来表示“概率”，即R^p，r^p，$\overline{R}^p$，Γ^p。考虑R^p，R是移出事件属性的确定性部分：$Attr(R)=Attr(R^p)-\{E\}$。用户只能看到R，但是系统需要访问事件属性$R^p.E$。函数依赖性Γp总是包含$Attr(R) \to R^p.E$。

也就是说，每个关系R^p确保我们没有将两个不同的事件e_1和e_2赋予相同的元组t(相反，我们希望将$e_1 \vee e_2$与元组t相关联)。

除了上述的描述，我们考虑一个函数表示，其中一个类型R^p的概率实例r^p，按照如下的函数描述$e_R: U^k \to E$，其中k是R的维数。当t与事件e一起出现在r^p中，那么$e_R(t)=e$，否则$e_R(t)=\perp$。同样地，可以通过e_R收集所有$e_R(t) \neq \perp$的元组t，从而恢复r^p。

1. 概率数据库查询的内涵语义

查询概率数据库的一种方法是使用复杂事件。首先将查询使用σ，Π，$\times$表达为一个查询计划，然后修改每个操作符，从而计算中间结果的事件属性E：分别使用σ^i，Π^i，$\times^i$表示修改后的操作符。

$$e_{\sigma_c^i(p)}(t)=\begin{cases}e_p(t), & \text{如果 } c(t) \text{是正确的}\\ \perp, & \text{如果 } c(t) \text{是错误的}\end{cases}$$

$$e_{\Pi_A^i(p)}(t)=\vee_{t':\Pi_A^i(t')=t} e_p(t')$$

$$e_{p\times^i p'}(t,t')=e_p(t) \wedge e_p'(t')$$

使用内涵语义来计算排序概率是非常不实际的。首先，在$q^i(D^p)$中事件表达式能够变得非常大。在最差的情况下，这个表达式的尺寸能够变得与数据库的规模相一致。这极大地增加了查询操作符的复杂性，使优化的工作更加复杂。第二，对于每个元组t，必须计算元组所对应的事件e的概率$\Pr(e)$，这是一个#P-完全问题。

2. 概率数据库查询的外延语义

我们现在将查询操作符修改为用于计算概率而不是复杂

事件：使用 σ^e，Π^e，$\times^e$ 表示修改的操作符。这是更加具有效率的，因为它涉及实数而不是事件表达式。我们通过迭代查询计划 p 结构为每个元组 t 定义一个数字 $\Pr(t)[0,1]$。

$$\Pr_{\sigma_c^e(p)}(t)=\begin{cases}\Pr_p(t), & \text{如果 } c(t) \text{是正确的}\\ 0, & \text{如果 } c(t) \text{是错误的}\end{cases}$$

$$\Pr_{\Pi_A^e(p)}(t)=1-\prod_{t',\Pi_A^i(t')=t}(1-\Pr(t'))$$

$$\Pr_{p\times ip'}(t,t')=\Pr p(t)\wedge \Pr p'(t')$$

这样，$P^e(D^p)$ 是一个外延概率关系，称之为查询计划 p 的外延语义。如果知道 $P^e(D^p)=qrank(D^p)$，那么在外延语义下简单执行计划就可。然而，不幸的是，$p^e(D^p)$ 依赖为 q 所选择的特定计划 p。

定义 7 考虑一个模式 R^p，Γ^p。当对于模式中的所有概率数据库 D^p，$pe(D^p)=qrank(D^p)$，p 是查询 q 的查询计划，那么我们称查询 q 的查询计划 p 是安全的。

5.2 查询优化算法

经过上述两种语义的分析，我们的概率数据库查询处理算法的核心思路是：在获得查询 q 时，首先尽可能地获取安全查询计划，如果获取成功，就采用外延语义计算概率；否则，当不能够采取外延语义的时候，就使用第 5 节所介绍的操作代数算法进行查询处理。

这样，查询优化算法的核心便是寻找安全查询计划。

首先假定我们的查询是联合查询。

符号定义如下：

$Rels(q)=\{R_1,\cdots,R_k\}$，表示所有出现在查询 q 中的关系名。假定每个关系名在查询中至多出现一次。

$PRels(q)$ 表示出现在查询 q 中的概率关系名，$PRels(q)\subseteq Rels(q)$。

$Attr(q)$ 表示 q 中所有关系的属性。使用 $R_i.A$ 来表示属性。

$Head(q)$ 是 q 的头属性，$Head(q)\subseteq Attr(q)$。

在 $Attr(q)$ 上定义函数依赖 $\Gamma^p(q)$：

1）在 Γ^p 中的每个 FD 都出现在 $\Gamma^p(q)$。

2）对于每个 join 预测，$R_i.A=R_j.B$，存在 $R_i.A\rightarrow R_j.B$ 和 $R_j.B\rightarrow R_i.A$。

3）对于每个选择预测，$R_i.A=c,\emptyset\rightarrow R_i.A$。

我们寻找一个安全计划 p，也即正确计算概率的查询计划。在 p 中的每个操作符必须是安全的。

假定 q_1，q_2 是两个查询，$op\{\sigma,\Pi,\times\}$ 是关系操作符。考虑新的查询 $op(q_1,q_2)$。如果 $\forall D^p$，$\Gamma^p\vdash D^p$，$op^e(\Pr(q_1^i(D^p)),\Pr(q_2^i(D^p)))=\Pr(op^i(q_1^i(D^p)),q_2^i(D^p))$，我们说，$op^e$ 是安全的。换言之，op 是安全的，当为 op^e 的输入是给定的正确概率，那么它能为输出元组计算正确的概率。

定理 2 假定 q,q' 是联合查询

1）σ_c^e 在 $\sigma_c(q)$ 上总是安全的。

2）$\times^e$ 在 $q\times q'$ 上总是安全的。

3）对于每个 $Rp\in PRels(q)$，如果以下的函数依赖能够从 $\Gamma p(q)$ 中推断出来，那么称 $\Pi^e_{A_1,\cdots,A_k}$ 在 $\Pi^e_{A_1,\cdots,A_k}(q)$ 中是安全的。

$A_1,\cdots,A_k,R^p.E\rightarrow Head(q)$

证明：

1）依据定义就可以得出。

2）由于我们假定查询中所有关系都是唯一的，因此在 q 和 q' 的输出中的复杂事件包含着不同的原子事件。这样，给定一个元组 $tjoin=(t,t')\in q\times q'$，$\Pr(t_{join}.E)=\Pr(t.E\wedge t'.E)=\Pr(t.E)\Pr(t'.E)$，因此联合操作是安全的。

3）对于投影的输出元组 t，考虑映射到 t 的输入元组集合 S_t。当且仅当 S_t 中的元组所对应的复杂事件相互独立，操作符才是安全的。这样对于 $A_1,\cdots,A_k$ 中具有相同值的所有元组，没有基本事件出现在具有 $Head(q)$ 不同值的两个元组之间，也即存在如下函数依赖性：$A_1,\cdots,A_k,R^p.E\rightarrow Head(q)$。

显然，一个只包含安全操作符的计划 p 也是安全的。反之亦然。

定义 8（分离关系） 对于一个查询 φ，当两个关系 R_i 和 R_j 中没有任何表的列与另一个表的列进行 join 操作，那么我们称 R_i 和 R_j 是分离的。

定义 9（分离） 两个关系集合 S_i 和 S_j 称为在查询中形成一个分离，如果：

1）它们划分了 $PRels(\varphi)$ 集合。

2）对于 $R_i\in S_i$ 和 $R_j\in S_j$，它们必须是分离关系。

安全查询计划的算法描述如下：

```
Algorithm SAFE-PLAN(q)
If Head(q)= Attr(q) then
  Return any plan p for q
  (p is projection-free, hence safe)
End if
For A∈(Attr(q)−Head(q)) do
  Let qA be the query obtained from q by adding A to the head variable
  If Π_head(q)(q_A) is a safe operator then
    Return Π_head(q)(q_A)(SAFE−PLAN(q_A))
  End if
End for
Sqlit q into q1 ⋈_c q2
If no such split exists then
Return error("no safe plans exist")
End if
Return SAFE-PLAN(q1) ⋈_c SAFE-PLAN(q2)
```

该算法是一个寻找安全计划的优化算法，它采用自上而下的方法，如下所述。首先，它尽力在查询计划中最后做所有的安全投影。当没有更多的安全投影，那么它尽力进行一个 join 操作，将 q 分离为 $q_1\bowtie_c q_2$。由于 $\bowtie_c$ 是查询计划中的最后操作，在 c 中的所有属性一定存在 $Head(q)$，因此，$Rels(q_1)$ 和 $Rels(q_2)$ 必须形成一个分离。

给定一个查询 q，一个分离能够按照如下的方式发现。构造一个图 G，它的节点是 $Rels(q)$，它的边是所有的 (R_i,R_j)，其中，q 中具有联合条件 $R_i.A=R_j.B$，$R_i.A$ 和 $R_j.B$ 均出现在 $Head(q)$。发现 G 的连通组件，选择 q_1 和 q_2 是这些连通图的一个划分：这定义了 $Rels(q_i)$ 和 $Attr(q_i)$，$i=1,2$；定义 $Head(q_i)=Head(q)\cap Attr(q_i)$，$i=1,2$。如果 G 是一个连通图，那么查询就没有安全计划。如果 G 具有多个连通组件，我们就有多个分裂 q 的选择。

结束语 数据的不确定性已经成为当下数据管理所面临的一个重大挑战之一，是包括 Web 数据管理、数据集成和深度 Web 等诸多领域的研究热点。

本论文的主要贡献包括：

1) 针对数据的不确定性特点，提出一种基于元组存在性的概率数据模型。

2) 结合基于元组存在性的概率数据模型，提出相关的概率关系代数和概率数据库查询处理算法。

参 考 文 献

[1] Imielinski T, Lipski W. Incomplete information in relational databases[C]//Journal of the ACM. October 1984, 31: 761-791

[2] Dalvi N, Suciu D. Management of Probabilistic Data: Foundations and Challenges[C]//PODS. 2007: 1-12

[3] Benjelloun O, Sarma A D, Halevy A, et al. ULDBs: Databases with uncertainty and lineage[C]//VLDB. 2006: 953-964

[4] Sen P, Deshpande A. Representing and querying correlated tuples in probabilistic databases[C]//ICDE. 2007

[5] Cavallo R, Pittarelli M. The theory of probabilistic databases[C]// Proceedings of 13th Int. Conf. on Very Large Data Bases (VLDB'87). Brighton, England, September 1987: 71-81

[6] Cheng R, Kalashnikov D, Prabhakar S. Evaluating probabilistic queries over imprecise data[C]// International Conference on Management of Data. 2003

[7] Garofalakis M, Suciu D, et al. Probabilistic Data Management [J]. IEEE Data Engineering Bulletin, 2006, 29(1)

[8] Choenni S, Blok H E, Leertouwer E. Handling uncertainty and ignorance in databases: A rule to combine dependent data[C]// Database Systems for Advanced Applications. 2006

[9] Xia Yu-ni, Prabhakar S, Lei Shan, et al. Indexing continuously changing data with mean-variance tree[C]// ACM Symposium on Applied Computing. 2005

(上接第 234 页)

抽取的搭配如表 4 所列。

表 4 名词与形容词搭配(N+A)，中心词为“企业”的搭配

困难	知名	像样	赚钱	单一	雄厚	著名	微小	吃力	正规	不错

名词与量词搭配(N+Q)，以中心词为“企业”为例，共抽取 4 个候选搭配，经人工校对，3 个正确，准确率为 75%。抽取的搭配如表 5 所列。

表 5 名词与形容词搭配(N+A)，中心词为“企业”的搭配

家	户	批	一下子

动词与动词搭配(V+V)，以中心词为“突破”为例，共抽取 16 个候选搭配，经人工校对，15 个正确，准确率为 93%。抽取的搭配如表 6 所列。

表 6 动词与动词搭配(V+V)，心词为“突破”的搭配

可	创新	取得	获	快攻	予以	无	有所	分区
寻求	亩产	加以	求	鉴定	争取	可望		

动词与副词搭配(V+F)，以中心词为“突破”为例，共抽取 8 个候选搭配，经人工校对，8 个正确，准确率为 100%。抽取的搭配如表 7 所列。

表 7 动词与副词搭配(V+F)，中心词为“突破”的搭配

首先	已	轻易	重点	一举全力	难以	由此

形容词与形容词搭配(A+A)，以中心词为“繁荣”为例，共抽取 4 个候选搭配，经人工校对，3 个正确，准确率为 75%。抽取的搭配如表 8 所列。

表 8 形容词与形容词搭配(A+A)，中心词为“繁荣”的搭配

全面	稳定	空前	繁荣

形容词与副词搭配(A+F)，以中心词为“繁荣”为例，共抽取 4 个候选搭配，经人工校对，4 个正确，准确率为 100%。抽取的搭配如表 9 所列。

表 9 形容词与副词搭配(A+F)，中心词为“繁荣”的搭配

更加	长期	共同	真正

将各个类型的准确率统一计算，则：

平均准确率=(71+220+10+3+15+8+3+4)/(73+235+11+4+16+8+4+4)=94.1%

相对于基于框架的提取方法准确率为 83.73%，本文的方法不仅在准确率上有了很大的提高，而且在全面性上也有了很大的突破。通过对以名词、动词、形容词为中心词的主要实词的统计，基本提取了各种搭配类型，这些信息对于后期建立搭配知识库有很大的作用。

结束语 本文提出了一种全新的基于典型句型的词语搭配自动提取模型。引入语言学知识，利用标注词性的大规模语料库，构建基于典型句型的模型，利用统计量共现频率及互信息筛选候选搭配，获得了 94.1%的准确率及较高的召回率。

分析错误提取的词语搭配，主要有以下几个方面：

1)构建典型句型考虑的主要是两个词，对于单个词考虑不够。例如“企业”+“添”及“企业”+“赚”。

2)只考虑了简单句型，对复杂句型引起的局部句型相同从而产生的错误考虑不够。

3)对语言学知识的运用还很粗浅，只利用了词与词之间的搭配限制，还没有利用相关的语义制约关系。

下一步工作：(1) 提高分词与词性标注的性能，为词语搭配研究提供质量更好的语料。(2) 对词性相同的搭配，典型句型的限制要进一步加强，限制其结构。(3) 对两个词以上的搭配以及短语组块间的搭配进行深入研究，寻找更好的解决方案。(4)引入语义信息，加强词语搭配的融合，扩张数据库的实用性。

参 考 文 献

[1] Smadja F. Retrieving Collocations from Text: Xtract[J]. Computational Linguistics, 1993, 19(1): 143-177

[2] Benson M, Benson E, Ilson R. The BBI Combinatory Dictionary of English: A Guide to Word Combinations[M]. John Benjamins, Amsterdam and Philadelphia, 1986

[3] 孙茂松，黄昌宁，方捷. 汉语搭配定量分析初探[J]. 中国语文，1997(1)：29-38

[4] Choueka Y, Klein T, Neuwitz E. Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus [J]. Literacy and Linguistic Computing, 1983, 4(1): 34-38

[5] 曲维光，陈小荷，吉根林. 基于框架的词语搭配自动抽取方法[J]. 计算机工程，2004，30(23)：22-24