データベースシステム 120－9
（2000．1．24）

タプル間の依存関係を表現できる確率的データベースモデルの提案

財部 倫孝† 清水 將吾† 石原 靖哲‡ 伊藤 実†

† 奈良先端科学技術大学院大学 情報科学研究科
〒 630-0101 奈良県生駒市高山町 8916-5
‡ 大阪大学大学院基礎工学研究科 情報数理系
〒 560-8531 大阪府豊中市待兼山町 1-3
E-mail: †{tomota-t,shougo-s,ito}@is.aist-nara.ac.jp, ‡ishihara@ics.es.osaka-u.ac.jp

あらまし 従来の確率的データベースモデルは，データベース中の各タプルに対して実際にそのタプルが存在する確率を付加するというものが一般的であった．しかしこのようなモデルではタプル間の依存関係，例えばある二つのタプルの少なくとも一方がデータベース中に存在するという関係にあることを表現できない．そこで本稿では，Imielinski らによって提案された conditional table を基本的枠組として採用し，conditional table 中に現れる各変数を確率変数として扱うことによって，タプル間の依存関係を表現できるような確率的データベースモデルを提案する．

キーワード 確率的データベース，条件付きテーブル，依存関係，確率的質問，関係代数演算

A Probabilistic Database Model with Representability of Dependency among Tuples

Tomotaka Takarabe† Shougo Shimizu† Yasunori Ishihara‡ Minoru Ito†

† Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara 630-0101, Japan
‡ Department of Informatics and Mathematical Science
Graduate School of Engineering Science, Osaka University
1-3, Machikaneyama, Toyonaka, Osaka 560-8531, Japan
E-mail: †{tomota-t,shougo-s,ito}@is.aist-nara.ac.jp, ‡ishihara@ics.es.osaka-u.ac.jp

Abstract In the conventional models of probabilistic databases, the probability of existence of each tuple is attached to the tuple. However, these models are not able to represent dependency among tuples, e.g., there is at least one tuple in a database. In this paper, we propose a probabilistic database model which is able to represent dependency among tuples. For that purpose, we adopt conditional tables proposed by Imielinski et al. as a standard framework, and treat each variable which appears in a conditional table as a random variable.

key words: probabilistic database, conditional table, dependency, probabilistic query, relational operation

NAME	TYPE	pS
Maria	Frigate	0.4
Maria	Tugboat	0.6
Seawolf	Frigate	0.5
Seawolf	Submarine	0.4

図 1: 各タプルに確率を付加するモデル [3]

NAME	TYPE	MAX-SPEED
Maria	0.4[Frigate] 0.5[Tugboat] 0.1[Submarine]	0.3[20-knots] 0.7[*]
Seawolf	0.5[Submarine] 0.5[*]	0.2[20-knots] 0.8[30-knots]

図 2: 不明な属性値も扱えるモデル [2]

NAME	TYPE	LB	UB	PATH
Maria	Frigate	0.4	0.7	w_1
Maria	Tugboat	0.3	0.9	w_2
Seawolf	Frigate	0.6	0.8	w_3
Seawolf	Submarine	0.5	0.8	w_4

図 3: 確率の値に上限，下限をもつモデル [5]

1 まえがき

従来のデータベースにおいて確率的な事象等を扱うには，各ユーザが確率の付加の方法や質問に応じた確率計算等の定義や計算を行う必要があった．そこで，確率的な事象や知識，情報を表すことができるデータベースモデルが幾つか提案されている．例えば文献 [3] のモデルにおいて，各タプルには図 1 のようにそれぞれの存在確率を与える属性 pS が付加されている．図 1 では，Maria が Frigate である確率は 0.4 である．キーとなる属性 NAME の各値に関連付けられた pS を合計すると，キーが示すオブジェクトが確実に存在する場合には 1，存在が不確実ならば 1 未満となる．図 1 では，Maria が存在する確率は 1 であるが，Seawolf が存在する確率は 0.9 であり，その存在は不確実である．文献 [6] のモデルでも確率を表す属性を新たにもうけ，タプルごとに確率を付加する方法を採用している．

また，文献 [2] のモデルでは，各属性が各値をとる確率の下限が付加されている．例えば，図 2 において，Maria の MAX-SPEED が 20 ノットである確率の下限は 0.3 である．また，割り当ての決まっていない残りの確率は不明な値を表す $*$ という記号に付加され，各属性の値ごとの確率の合計は必ず 1 になる．これにより属性値に不確実性をもつタプルも取り扱うことができる．

文献 [5] のモデルでは，本質的に，各タプルには図 3 のようにその存在確率の上限値と下限値が与えられている．従って，正確な確率値が不明であるタプルでも扱うことができる．図 3 の場合，Maria が Frigate である確率の下限は 0.4，上限は 0.7 である．また，各タプルには，それぞれを唯一に区別するために，真偽値をもつ変数を PATH という属性で与えている．PATH の値は質問時に評価され，結果の PATH の値に反映される．

これらのモデルはいずれも各タプルに確率を付加するという方法を採用しており，異なるタプルの存在確率はそれぞれ独立したものとして扱われている．そのため，これらのモデルではタプル間に何らかの依存関係がある場合にはその情報を表現することができない．例えば文献 [3] のモデル（図 1）において「Frigate は必ず存在する」，つまり「Maria と Seawolf がともに Frigate ではない確率は 0 である」という情報が与えられたとしても，それをこのモデルで表現することは不可能である．実際，このモデルでは，Maria と Seawolf がともに Frigate ではない確率は $(1-0.4)\times(1-0.5)=0.3$ と計算され，その確率が 0 にはならない．

本稿ではこのような問題点を解決するため，タプル間の依存関係を表現できるような確率的データベースモデルの提案を行う．我々のモデルは，Imielinski らによって提案された条件付きテーブル（conditional table）[4] を基本的枠組とし，条件付きテーブル中に現れる各変数を確率変数として扱うことにより，確率的な情報を表現する．

本稿の構成は以下の通りである．2 節では，我々が提案する確率的データベースの定義を示す．3 節では，タプルの存在確率に関する質問について考察し，データベースがどのような条件を満たしていればその質問を効率良く処理できるかを示す．最後に，4 節でまとめと今後の課題を述べる．

2 確率的データベースモデル

2.1 確率的データベースの定義

U を属性の有限集合とする．U 中のすべての属性に対して値の定義域は同じ可算集合であると仮定し，これを **dom** とする．**dom** 中の要素を定数と呼ぶ．V を変数の可算集合（但し，$V \cap \mathbf{dom} = \phi$）とする．属性集合 $X \subseteq U$ 上のタプルとは，X から $\mathbf{dom} \cup V$ への写像 u である．u の属性 $A \in X$ の値を $u(A)$ と書く．$Y \subseteq X$ に対して，$u[Y]$ は各 $A \in Y$ に対して $v(A) = u(A)$ であるような Y 上のタプル v を表す．X 上のテーブルとは，X 上のタプルの有限集合である．

条件式とは，$x = a$，$x = y$，$a = a'$（ここで，x，$y \in V$，a，$a' \in \mathbf{dom}$）という形の原子式から $\neg$，$\vee$，$\wedge$ を有限回用いて構成される式である．属性集合 X 上の条件付きタプルとは，$t[X]$ が X 上のタプルであり，$t(con)$ が条件式であるような，$X \cup \{con\}$ 上で定義された写像 t である．X 上の条件付きテーブル（conditional table）とは，X 上の条件付きタプルの有限集合である．型 $\langle X_1, \ldots, X_n \rangle$ の条件付き多重テーブルとは，各 $i \in [1, n]$ に対して，T_i が X_i 上の条件付きテーブルであるような列 $\mathbf{T} = \langle T_1, \ldots, T_n \rangle$ である．

付値関数とは，V から **dom** への写像 ν である．付値関数は以下のように拡張できる．

- 任意の $a \in \mathbf{dom}$ に対して，$\nu(a) = a$ とする．
- X 上のタプル t に対して，$\nu(t)$ を以下の式を満たす X 上のタプルとする．

  任意の $A \in X$ に対して，

  $$(\nu(t))(A) = \nu(t(A)).$$

- 原子式 $l = r$ に対して，

  $$\nu(l = r) = (\nu(l) = \nu(r))$$

  と定義する．条件式 c に対して，$\nu(c)$ は c 中の各原子式 $l = r$ を $\nu(l = r)$ で置き換えて得られる条件式である．
- X 上の条件付きテーブル T に対して，

  $$\nu(T) = \{\nu(t[X]) \mid t \in T, \nu(t(con)) = \mathbf{true}\}.$$

- 型 $\langle X_1, \ldots, X_n \rangle$ の条件付き多重テーブル

  $$\mathbf{T} = \langle T_1, \ldots, T_n \rangle$$

  に対して，

  $$\nu(\mathbf{T}) = \langle \nu(T_1), \ldots, \nu(T_n) \rangle.$$

$\mathbf{T}$ が表すテーブルの集合は以下のように定義される．

$$rep(\mathbf{T}) = \{\nu(\mathbf{T}) \mid \nu \text{ は付値関数 }\}$$

条件付きテーブルは Imielinski らによって提案された不完全な情報を表すための枠組である [4]．本稿では，付値関数のとる値に対して確率を付加することにより，条件付きテーブルの拡張を行う．

定義 1 確率的データベースは二字組 $\mathbf{T}_\mu = \langle \mathbf{T}, \mu \rangle$ で表される．ここで，

- $\mathbf{T}$ は条件付き多重テーブル；
- μ は $V \times \mathbf{dom}$ から $[0, 1]$ への写像である．但し，任意の $x \in V$ に対して，

  $$\sum_{a \in \mathbf{dom}} \mu(x, a) = 1$$

  が成り立つものとする．$\mu(x, a)$ は変数 x の値が a である確率を表す． □

確率的多重テーブルとは，二字組 $\langle T, p \rangle$（ここで，T はテーブルの列，p は $0 \leq p \leq 1$ を満たす実数）である．確率的データベース $\mathbf{T}_\mu = \langle \mathbf{T}, \mu \rangle$ が表す確率的多重テーブルの集合は以下のように定義される．

$$rep(\mathbf{T}_\mu) = \{\langle \nu(\mathbf{T}), \prod_{x \in V} \mu(x, \nu(x)) \rangle \mid \nu \text{ は付値関数 }\}$$

例 1 $\mathbf{T}_\mu = \langle \langle T_1 \rangle, \mu_1 \rangle$ の例を図 4 に示す．ここで，Maria, Seawolf, Frigate, Tugboat, Submarine $\in \mathbf{dom}$，x，$y \in V$ である．$\mathbf{T}_\mu$ が表す確率的テーブルの集合 $rep(\mathbf{T}_\mu)$ を図 5 に示す． □

T_1

NAME	TYPE	con
Maria	x	$(x = \text{Frigate}) \vee (y = \text{Frigate})$
Seawolf	Frigate	$(y = \text{Submarine})$

μ_1

V	dom	p
x	Frigate	0.3
x	Tugboat	0.7
y	Frigate	0.8
y	Submarine	0.2

図 4: 確率的データベース $\mathbf{T}_\mu = \langle\langle T_1\rangle, \mu_1\rangle$

$\nu(x) = \text{Frigate}$
$\nu(y) = \text{Frigate}$

$\langle T_1, 0.24\rangle$

NAME	TYPE
Maria	Frigate

$\nu(x) = \text{Frigate}$
$\nu(y) = \text{Submarine}$

$\langle T_1, 0.06\rangle$

NAME	TYPE
Maria	Frigate
Seawolf	Frigate

$\nu(x) = \text{Tugboat}$
$\nu(y) = \text{Frigate}$

$\langle T_1, 0.56\rangle$

NAME	TYPE
Maria	Tugboat

$\nu(x) = \text{Tugboat}$
$\nu(y) = \text{Submarine}$

$\langle T_1, 0.14\rangle$

NAME	TYPE
Seawolf	Frigate

図 5: $rep(\mathbf{T}_\mu)$

条件式 c の大きさ $|c|$ を c 中に現れる原子式の個数と定義する．

$\mathbf{T}_\mu = \langle \mathbf{T}, \mu\rangle$ の大きさ $|\mathbf{T}_\mu|$ を以下のように定義する．

$$|\mathbf{T}_\mu| = \sum_{T_i \in \mathbf{T}} \left(|X_i| \cdot |T_i| + \sum_{t \in T_i} |t(con)| \right) + |\mathcal{P}|$$

ここで，$\mathcal{P} = \{\langle x, a\rangle \mid \mu(x, a) > 0\}$ であり，$|T|$ は T 中のタプルの総数を表す．

2.2 確率的データベース上の関係代数演算

通常の（確率に関する質問を含まない）関係代数演算は，条件付きテーブルに対する演算と同様に定義できる．但し，演算は同じ μ の定義をもつ条件付きテーブルについてのみ定義され，μ の定義は演算適用後もそのまま保持される．以下に条件付きテーブル上の関係代数演算の定義を示す．詳細は文献 [4] を参照されたい．

T, T' をそれぞれ X, X' 上の条件付きテーブルとする．

射影：$\pi_Y(T) = \{t[Y \cup \{con\}] \mid t \in T\}$．但し，$Y \subseteq X$．

選択：$\sigma_E(T) = \{\sigma_E(t) \mid t \in T\}$．ここで，$\sigma_E(t)$ は以下の式を満たす X 上のタプルである．

$$\begin{aligned} \sigma_E(t)[X] &= t[X] \\ \sigma_E(t)(con) &= t(con) \wedge E(t) \end{aligned}$$

$E(t)$ は E 中の各属性 $A \in X$ を $t(A)$ で置き換えることにより得られる条件式である．

結合：$T \bowtie T' = \{t \bowtie t' \mid t \in T, t' \in T'\}$．ここで，$t \bowtie t'$ は以下の式を満たす $X \cup X'$ 上のタプルである．

$$\begin{aligned} (t \bowtie t')(A) &= \begin{cases} t(A) & \text{if } A \in X \\ t'(A) & \text{if } A \in X' - X \end{cases} \\ (t \bowtie t')(con) &= t(con) \wedge t'(con) \wedge \bigwedge_{A \in X \cap X'} (t(A) = t'(A)) \end{aligned}$$

和：$X = X'$ のとき定義され，$T \cup T' = \{t \mid t \in T \text{ または } t \in T'\}$．

差：$X = X'$ のとき定義され，$T - T' = \{t_u \mid t \in T\}$．ここで，$t_u$ は以下の式を満たす X 上のタプルである．

$$\begin{aligned} t_u[X] &= t[X] \\ t_u(con) &= t(con) \wedge \bigwedge_{t' \in T'} \psi(t, t') \end{aligned}$$

ここで，

$$\psi(t, t') = \left(\bigvee_{A \in X} (t(A) \neq t'(A)) \right) \vee \neg t'(con).$$

属性名の変更：$\theta^B_A(T) = \{\theta^B_A(t) \mid t \in T\}$．ここで，$\theta^B_A(t)$ は以下の式を満たす $(X - \{A\}) \cup \{B\}$ 上のタプルである．

$$\begin{aligned} (\theta^B_A(t))(C) &= \begin{cases} t(C) & \text{if } C \in X - \{A\} \\ t(A) & \text{if } C = B \end{cases} \\ (\theta^B_A(t))(con) &= t(con) \end{aligned}$$

3 タプルの存在確率に関する質問 Q_μ

本節では，$\mathbf{T} = \langle T \rangle$ と仮定し，$\mathbf{T}_\mu = \langle \langle T \rangle, \mu \rangle$ を単に $\mathbf{T}_\mu = \langle T, \mu \rangle$ と書く．

3.1 Q_μ の定義と NP 困難性

確率的データベースでは，通常の関係代数質問の他に，確率を考慮した質問も当然要求される．確率に関する質問は幾つか考えられるが，本稿ではその中でも基礎となる質問について検討する．それは，与えられたタプルが，条件付きテーブルが表すテーブルの中に存在する確率を求めるというもので，その定義を次に示す．

定義 2 T を X 上の条件付きテーブルとし，$\mathbf{T}_\mu = \langle T, \mu \rangle$ を確率的データベースとする．このとき，$t \in rep(T)$ である確率 $Q_\mu(t, T)$ は以下のように定義される．

$$Q_\mu(t,T) = \sum_{\substack{\langle T, p \rangle \in rep(\mathbf{T}_\mu) \\ \text{s.t. } t \in T}} p$$

□

$Q_\mu(t, T)$ は以下のようにして求められる．

$$Q_\mu(t,T) = p(c)$$

ここで，

$$\begin{aligned} c &= \bigvee_{t_i \in T} (t_i(con) \wedge \varphi(t, t_i)), \\ \varphi(t, t_i) &= \bigwedge_{A \in X} (t(A) = t_i(A)) \end{aligned}$$

である．$p(c)$ は c が真となる確率を表し，c を真にするような付値関数の集合を $\mathcal{V}$ としたとき，

$$p(c) = \sum_{\nu \in \mathcal{V}} \prod_{x \in V} \mu(x, \nu(x)).$$

以下，X 上の（変数を含まない）タプル t と確率的データベース $\mathbf{T}_\mu = \langle T, \mu \rangle$ が与えられたとき，$Q_\mu(t, T)$ を求める問題を考える．

例 2 図6に示す確率的データベース $\mathbf{T}_\mu = \langle T_2, \mu_2 \rangle$ を考える．ここで，タプル $t = (\text{Maria}, \text{Submarine})$ が $rep(T_2)$ 中に存在する確率を求める質問 $Q_\mu(t, T_2)$

T_2

	NAME	TYPE	*con*
t_1	Maria	Frigate	$x = 0$
t_2	Maria	y	$y \neq$Frigate
t_3	Seawolf	Frigate	$x = 1$
t_4	Seawolf	Submarine	$x \neq 1$

μ_2

V	dom	p
x	0	0.2
x	1	0.5
x	2	0.3
y	Frigate	0.5
y	Submarine	0.1
y	Tugboat	0.4

図 6: 確率的データベース $\mathbf{T}_\mu = \langle T_2, \mu_2 \rangle$

を考える．このとき，t が $rep(T_2)$ 中に存在するための条件式は

$$\begin{aligned} c &= t_2(con) \wedge \varphi(t, t_2) \\ &= (y \neq \text{Frigate}) \wedge (y = \text{Submarine}) \\ &= (y = \text{Submarine}) \end{aligned}$$

となり，その確率は $\mu_2(y, \text{Submarine}) = 0.1$ である．

一方，$t' = (\text{Seawolf}, \text{Frigate})$ について考えた場合，

$$\begin{aligned} c &= t_3(con) \wedge \varphi(t', t_3) \\ &= (x = 1) \end{aligned}$$

となり，その確率は $\mu_2(x, 1) = 0.5$ である． □

しかし，$Q_\mu(t, T)$ を求める問題は以下に述べるように一般に NP 困難である．

定理 1 条件付きテーブル T とタプル t が与えられたとき，$t \in rep(T)$ であるかどうかを判定する問題は NP 困難である．

証明: SAT からの帰着を行う．SAT のインスタンス F が与えられたとき，$\{A\}$ 上の条件付きテーブル T を以下のように構成する．

$$T = \{t\}$$

但し，t は以下の式を満たす $\{A\}$ 上のタプルである．

$$\begin{aligned} t(A) &= 1 \\ t(con) &= c_F \end{aligned}$$

ここで，c_F は F 中の各変数 x を $((x=0)\vee(x=1))$ で置き換えることにより得られる条件式である．

このとき，$\langle 1\rangle \in rep(T)$ であるときかつそのときのみ，F は充足可能である． □

$Q_\mu(t,T)$ を求める問題の NP 困難性は，上の問題が $Q_\mu(t,T)$ の特別な場合であることから示せる．

3.2 Q_μ を多項式時間で解くための制約条件

$Q_\mu(t,T)$ を多項式時間で求めるためにはデータベースに対する何らかの制約条件が必要となる．本稿では，与えられた確率的データベース $\mathbf{T}_\mu = \langle T,\mu\rangle$ が以下の制約条件のいずれかを満たすときに $Q_\mu(t,T)$ が多項式時間で計算できることを示す．

定義 3 以下の条件を満たす $T_1,\ldots,T_n$ を T の分割と呼ぶ．

$$T = \bigcup_{i=1}^{n} T_i$$

但し，V_{T_i} を T_i 中に現れる変数の集合としたとき，任意の $i,j\in[1,n]$（但し，$i\neq j$）に対して，

$$V_{T_i}\cap V_{T_j} = \phi.$$

従って，異なる i,j に対して

$$T_i\cap T_j = \phi$$

である． □

この分割は $|\mathbf{T}_\mu|$ に対する多項式時間で行える．

条件 1 $T_1,\ldots,T_n$ を T の分割としたとき，各 T_i（$i\in[1,n]$）に対して $Q_\mu(t,T_i)$ が多項式時間で求まる． □

$\mathbf{T}_\mu$ が条件 1 を満たすとき，異なる i,j に対して $Q_\mu(t,T_i)$，$Q_\mu(t,T_j)$ は独立であるので，$Q_\mu(t,T)$ は次式により求められる．

$$Q_\mu(t,T) = 1 - \prod_{i=1}^{n}(1-Q_\mu(t,T_i))$$

従って，$Q_\mu(t,T)$ が多項式時間で計算できるためには，T の各分割 T_i に対して $Q_\mu(t,T_i)$ が多項式時間で計算できれば十分である．

条件 2 $|V_T|$ がある定数 k で抑えられる． □

$rep(\mathbf{T}_\mu) = \{\langle T_1,p_1\rangle,\ldots,\langle T_n,p_n\rangle\}$ とする．このとき，n の値はたかだか $|\mathcal{P}|^k$ である．すると，$Q_\mu(t,T)$ は次のように求まる．

$$Q_\mu(t,T) = \sum_{t\in T_i \text{ for all } i\in[1,n]} p_i$$

各 $i\in[1,n]$ に対して，$t\in T_i$ であるかどうかを判定すればよい．p_i を求める計算量は $\mathcal{O}(|\mathbf{T}_\mu|)$ なので，$Q_\mu(t,T)$ は多項式時間で求められる．

条件 3 X 上の条件付きテーブル $T' = \{t'_1,\ldots,t'_n\}$，$T$ が以下のすべての条件を満たす．

- $\{t'_1\},\ldots,\{t'_n\}$ が T' の分割である．
- 各 $i\in[1,n]$ に対して，$Q_\mu(t,t'_i)$ が多項式時間で求まる．
- $T=\{t_1,\ldots,t_n\}$ が以下のように表せる．

$$\begin{aligned} t_i[X] &= t'_i[X] \\ t_i(con) &= t'_i(con)\wedge\Big(w_{k_i}\vee\bigwedge_{j=1}^{m}\neg w_j\Big) \end{aligned}$$

（m は正整数）

各 $j\in[1,m]$ に対して，w_j は $x_j=a_j$（但し，$x_j\in V-V_{T'}$，$a_j\in\mathbf{dom}$）という形の原子式であり，任意の j_1，$j_2\in[1,m]$ に対して，$j_1\neq j_2$ ならば $x_{j_1}\neq x_{j_2}$ である．さらに，各 $i\in[1,n]$ に対して，$k_i\in[1,m]$ である．

□

条件 3 を満たす T'，T の形を図 7 に示す．ここで w_{k_i} による条件を見てみると，$\{w_1,\ldots,w_m\}\subseteq\{w_{k_1},\ldots,w_{k_n}\}$ のとき，各原子式の真偽値がどのような組み合わせになっても，w_{k_i} による各条件部分のうち少なくとも一つは真になることがわかる．従って，少なくとも一つのタプルが存在することを表現したいときにはこの形の条件が有効である．

各 $i\in[1,n]$ に対して，

$$c'_i = t'_i(con)\wedge\varphi(t,t'_i)$$

T'

X	con
X_1	con_1
	$\vdots$
X_i	con_i
	$\vdots$
X_n	con_n

T

X	con
X_1	$con_1 \wedge (w_{k_1} \vee (\neg w_1 \wedge \neg w_2 \wedge \cdots \wedge \neg w_m))$
	$\vdots$
X_i	$con_i \wedge (w_{k_i} \vee (\neg w_1 \wedge \neg w_2 \wedge \cdots \wedge \neg w_m))$
	$\vdots$
X_n	$con_n \wedge (w_{k_n} \vee (\neg w_1 \wedge \neg w_2 \wedge \cdots \wedge \neg w_m))$

図 7: 条件 3 を満たす条件付きテーブル

とする．このとき，$Q_\mu(t,t'_i)$ を求めるために定義 2 に従って生成される条件式は，c'_i と一致する．よって，$p(c'_i)$ は多項式時間で求まる．

$w = \neg w_1 \wedge \cdots \wedge \neg w_m$ とおく．$Q_\mu(t,T)$ を求めるために定義 2 に従って生成される条件式を c とすると，

$$\begin{aligned} c &= \bigvee_{i=1}^{n}((t_i(con) \wedge (w_{k_i} \vee w)) \wedge \varphi(t,t_i)) \\ &= \bigvee_{i=1}^{n}((t'_i(con) \wedge (w_{k_i} \vee w)) \wedge \varphi(t,t'_i)) \\ &= \bigvee_{i=1}^{n}(c'_i \wedge (w_{k_i} \vee w)) \\ &= (\bigvee_{i=1}^{n}(c'_i \wedge w_{k_i})) \vee ((\bigvee_{i=1}^{n} c'_i) \wedge w) \end{aligned}$$

と表せる．ここで，

$$c_A = \bigvee_{i=1}^{n}(c'_i \wedge w_{k_i}),$$

$$c_B = (\bigvee_{i=1}^{n} c'_i) \wedge w$$

とすると，

$$p(c) = p(c_A) + p(c_B)$$

と求まる．c_Aについて，

$$\begin{aligned} c_A &= (c'_1 \wedge w_{k_1}) \vee \cdots \vee (c'_n \wedge w_{k_n}) \\ &= (c'_{1,1} \vee \cdots \vee c'_{1,q_1}) \wedge w_1 \vee \cdots \\ &\qquad (c'_{j,1} \vee \cdots \vee c'_{j,q_j}) \wedge w_j \vee \cdots \\ &\qquad\qquad (c'_{m,1} \vee \cdots \vee c'_{m,q_m}) \wedge w_m \end{aligned}$$

（但し，$j \in [1,m]$，$i \in [1,q_j]$ に対して，$c'_{j,i} \in \{c'_1,\ldots,c'_n\}$）とすると，$p(c_A)$ は，

$$p(c_A) = 1 - \prod_{j=1}^{m}(1 - \left(1 - \prod_{i=1}^{q_j}(1 - p(c'_{j,i}))\right) \times p(w_j))$$

と求まり，$p(c_B)$ については，

$$p(c_B) = \left(1 - \prod_{i=1}^{n}(1 - p(c'_i))\right) \times p(w)$$

と求まる．また，$p(w)$ は

$$p(w) = p(\neg w_1) \times p(\neg w_2) \times \cdots \times p(\neg w_m)$$

と求まる．

各 $i \in [1,n]$ に対して，$p(c'_i)$ は多項式時間で求まる．また，$x \in V$，$a \in \mathbf{dom}$ に対して，$p(x=a) = \mu(x,a)$ なので，各 $j \in [1,m]$ に対して，$p(w_j)$ は $|\mathcal{P}|$ に対する多項式時間で求まる．よって，$p(c_A)$ は $|\mathbf{T}_\mu|$ に対する多項式時間で求まる．また，$p(\neg(x=a)) = 1 - \mu(x,a)$ なので，$p(w)$ も $m \cdot |\mathcal{P}|$ に対する多項式時間で求まり，$p(c_B)$ も $|\mathbf{T}_\mu|$ に対する多項式時間で求まる．よって，$p(c)$ は $|\mathbf{T}_\mu|$ に対する多項式時間で求まる．

4 あとがき

本稿では，タプル間の依存関係を表現できる確率的データベースモデルを定義するために，Imielinski らによって提案された条件付きテーブルを基本的枠組とし，テーブル中に現れる各変数を確率変数として扱うという方法を提案した．さらに，タプルの存在確率に関する質問について考察し，データベースがどのような条件を満たしていればその質問を効率良く処理できるかを示した．

今後の課題としては，

- Q_μ が多項式時間で解けるようなより緩い条件を提案する

- 他の実用的な問題への応用を検討する

ことなどが挙げられる.

参考文献

[1] S. Abiteboul, R. Hull, and V. Vianu, "Foundations of databases," Addison-Wesley, 1995.

[2] D. Barbará, H. Garcia-Molina, and D. Porter, "The management of probabilistic data," IEEE Trans. Knowledge and Data Engineering, vol.4, no.5, pp.487–502, Oct. 1992.

[3] D. Dey and S. Sarkar, "A probabilistic relational model and algebra," ACM Trans. Database Syst., vol.21, no.3, pp.339–369, Sep. 1996.

[4] T. Imielinski and W. Lipski, "Incomplete information in relational databases," J. ACM, vol.31, no.4, pp.761–791, Oct. 1984.

[5] V. Lakshmanan, N. Leone, R. Ross, and V.S. Subrahmanian, "ProbView: A flexible probabilistic database system," ACM Trans. Database Syst., vol.22, no.3, pp.419–469, Sep. 1997.

[6] E. Zimányi, "Query evaluation in probabilistic relational databases," Theoretical Computer Science, vol.171, no.1–2, pp.179–219, Jan. 1997.