

データベース問合せ結果の差分に基づくセルの価値付けに関する研究

鍋谷 優斗[†] 吉川 正俊^{††}

[†] 京都大学大学院情報学研究科 〒606-8501 京都府京都市左京区吉田本町

^{††} 京都大学大学院情報学研究科 〒606-8501 京都府京都市左京区吉田本町

E-mail: [†]ynabetani@db.soc.i.kyoto-u.ac.jp, ^{††}yoshikawa@i.kyoto-u.ac.jp

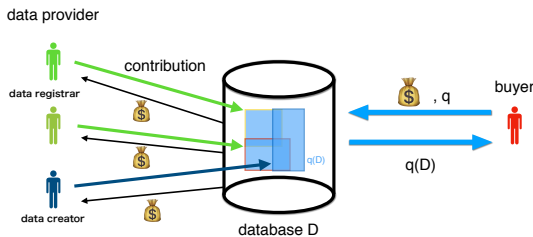


図1 利益分配のイメージ

1. はじめに

日本は海外に比べてデータを活用したビジネス展開があまり進んでいない状況であり、情報銀行[9]という事業が考案されている。情報銀行とは、個人とのデータ活用に関する契約等に基づき、PDS(Personal Data Store)と呼ばれる、他社保有データの集約を含め、個人が自らの意思で自らのデータを蓄積・管理するための仕組みを活用して、個人のデータを管理すると共に、個人の支持またはあらかじめ指定した条件に基づき個人に変わり妥当性を判断の上、データを第三者（他の事業者）に提供する事業のことである。このような、個人情報を含めた多種多様で大量のデータの円滑な流通を実現するために個人の関与のもとでデータ流通・活用を進める仕組みが有効だと考えられている[7]。

情報銀行を実現に近づけるため、パーソナルデータの価値付けを行うことは必要不可欠である。パーソナルデータの価値を考える研究は[6], [8]でも行われている。本研究は[10]に続いて、データの価値づけにおける尺度および価値付けの手法の開発を目指す。そもそも情報は他人に利用されて初めて価値があると思われるもので、実際に使われたデータから起源をたどり提供者に報酬を与える、という形が適していると考えられる。図1は、問合せを行う者 data buyer が、その問合せ q とデータに対する報酬 $\$$ をデータベース D に送り、データベース D は受け取った問合せに対する結果 $q(D)$ を data buyer に返し、その結果のデータを提供した人物である、データ登録者やデータ作成者のようなデータ提供者に報酬を分配する流れを示す。

データの起源を考えるという研究は[2]すでに多くのものがあり、

Cheney らはサーベイ[2]で三つの起源 “Where-Provenance”, “How-Provenance”, “Why-Provenance” を紹介している。

そこで本研究では、2. で上記の研究について説明した後、4. でその研究を拡張し、各セルごとにデータの起源を示す直接来歴注釈と、そのデータがどのデータの貢献で存在しているかを示す間接来歴注釈を定義し、直接来歴注釈及び間接来歴注釈の与えられた関係データベースに対して演算（選択、射影、和、自然結合、集約関数）の実行結果の関係における各セルの直接来歴注釈と間接来歴注釈の定義を与える。これを定義することで結果のデータを提供した人物および結果のデータを得るために貢献した人物を特定することができると考える。次に5. で、データベースの差分に基づいたデータの価値の定義を行い、6. ではさらに、4. で定義した注釈付きの問合せ演算を利用して、5. での差分計算の効率化が可能となる差分等価性について証明を行い、最後に、7. で、注釈に分配した報酬を、さらにデータ提供者へと還元する手法について考える。以上の流れで、データに与えられた報酬を、データ提供者に適切に分配する考えを提案する。

2. 関連研究

文献[2]では、データ起源について “Where-Provenance”, “How-Provenance”, “Why-Provenance” の三つの考え方を紹介している。

“Where-Provenance” は、ある関係に対する問合せの結果得られたデータのセルが元の関係のどのセルからきたものなのかを考えるもので、問合せ前の各セルそれぞれにラベル付けを行なっておくことで、問合せ結果の関係のセルが問合せ前の関係のどのセルから得られたものかを考えるものである。例えば図2のように二つの関係 R_1, R_2 を用意し、この二つの関係を自然結合した結果 $R_1 \bowtie R_2$ について “Where-Provenance” を考えると、まずはじめに自然結合を行う前の二つの関係のセルにそれぞれ $a_1, a_2, \dots$ のようにラベル付けを行なっている。そして自然結合の結果の関係のセルが、元の二つの関係のどのセルから得られたものなのかを見て、そのセルのラベルを結果のセルに付与している。結果の関係の $(1, 1)$ 成分のセル1は、 R_1 の $(1, 1)$ 成分1と R_2 の $(1, 1)$ 成分1が結合してできたものなので、元の二つのセルのラベル a_1 と a_5 を結果のセルの1に付与している。“Where-Provenance” は等価な問合せに対して違う結果を返してしまう場合もあり、[2]では等価な問合せの結果

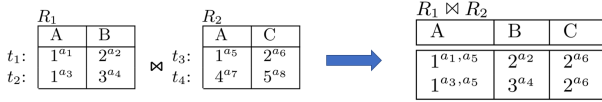


図2 “Where-Provenance” におけるラベルづけ

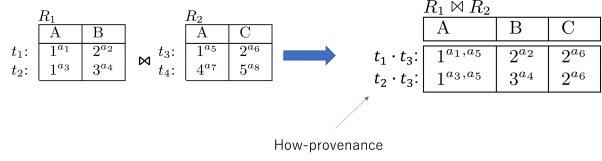


図3 “How-Provenance” におけるタグづけ

果のラベルを全て含んだ結果を返すことで等価性を保つことができるということを紹介している。

次に, “How-Provenance” はある関係に対する問合せの結果得られたデータの組が, どのようにして得られたのかを考えている。結果の組がどのような演算の結果により得られたのかを “provenance semiring” と呼ばれる, 可換半環 $(K, 0, 1, +, \cdot)$ による多項式で表現する。ここで可換半環 K は, 区別された値 0 を含む, 要素の集合である。これを利用して, 関係代数演算を可換半環の演算に写像する。写像モデルを “K -relation” としている。“K -relation” は, 任意の組に対する関数 R として定義する。ここで関数 R とは, 関係における組を K の 0 以外の要素へ写像するものである。先ほどの関係 R_1, R_2 を自然結合した例で “How-Provenance” について考えると, 図3で表される通り, まず二つの関係 R_1, R_2 の各組に t_1, t_2, t_3, t_4 のようにタグづけをして自然結合を行う。結果の関係の各組のタグは, その組の元となった組のタグ同士の積 (AND) で表現される。これはその二つのタグの両方が存在して初めて結果の関係が得られるということを意味している。“Why-Provenance” では, “How-Provenance” よりも簡略的に起源を考えるもので, 演算式ではなく集合でそれを表現する。図3において一目の組の “Why-Provenance” は $\{t_1, t_3\}$, 二目の組は $\{t_2, t_3\}$ となる。

Green ら [4] はデータ起源に関する種々の定義の関係についてまとめている。また, 起源の情報自体を対象にする問合せ言語 [5] に関しても考えられている。

3. 準備

本研究で利用する用語についてここで定義しておく。

ある関係 R における組 t の属性 A のセルを (t, A) で表し, 関係 R が含む属性集合を $att(R)$ と表すものとする。

本研究で対象とする問合せは, 関係代数演算として選択 (selection), 射影 (projection), 自然結合 (natural join), 和 (union) を許す SPJU 問合せと, 平均 (average), 総数 (sum), 最大 (max), 最小 (minimum), カウント (count) の五つの集約関数から成り立つ任意の問合せとする。ただし, 選択条件に集約関数は含まないものとする。またある関係 R, S において, 条件式を c としてこの条件式を満たす組の集合を関係として返す演算を選択と言い, 関係 R についてこの演算を行う場合, $\sigma_c R$ と

	A	B
t_1	1	2
t_2	2	3

図4 R

	A
t_1	1
t_2	2

図5 $\pi_A R$

	B
	2
	3

図6 $\pi_B(R_1 \bowtie R_2)$

表す。ここで, 条件式 c については, 最小の条件式を二項演算子 $\wedge$ (AND), $\vee$ (OR) で繋いだ条件式として捉える。最小の条件式とは,

- (1) 属性と属性
- (2) 属性と値

のどちらかの間に等号または不等号のみが入る条件式のことである。最小の条件式に $\neg$ (NOT) が付いている場合はその条件式の余事象の条件式に置き換える。例えば, 条件式 $c = (A = B) \wedge \neg(C = D)$ は, $\neg(C = D)$ を $(C = D)$ の余事象 $(C \neq D)$ に置き換えると, 最小の条件式 $(A = B), (C \neq D)$ を二項演算子 $\wedge$ で繋いだ条件式と捉えることができる。

関係 R を属性 A 上に射影する演算を $\pi_A R$ と表す。関係 R, S の和演算を $R \cup S$ と表す。関係 R, S の自然結合演算を $R \bowtie S$ と表す。関係 R の属性 A について, 平均, 総数, 最大, 最小を取る演算をそれぞれ, $avr_A R, sum_A R, max_A R, min_A R$ と表す。関係 R の属性 A について, カウントを取る演算を, DISTINCT を指定しない場合は $count_A R$, 指定する場合は $count_distinct_A R$ と表す。

4. 直接来歴注釈, 間接来歴注釈の定義と計算

本研究では, 問合せ結果のデータの起源を辿り, 元となっているセルに価値を与えたい。そのため Green ら [3] の, 可換半環による多項式で結果を表現する “How-Provenance” の考えをセル単位で適用する。組単位ではなくセル単位で考える理由としては, 組ごとでは最終結果に残っておらず, 演算にも使用されていないデータにまで価値を与える可能性があるからである。また, セル単位で考えると組単位で考えるよりもより詳細にデータに価値を与えることができる。例えば, 図4のように組にタグ t_1, t_2 が与えられた関係 R の属性 A の上に射影演算を行った結果は図5のようになるが, 組ごとに価値を与えると, t_1, t_2 が結果に残っていない属性 B の値 $2, 3$ も含んでいるため, これらの値にも価値を与えることになる。

ここで, 問合せによって得られたそのセルが元のどのセルから得られたものなのかを考えるために, 問合せによって演算が行われる前のデータのセルにそれぞれデータの起源を示す直接来歴注釈を与える。さらに, 問合せによって得られたそのセルは, どのセルの貢献によって得られたものなのかも考えるため, 間接来歴注釈も与える。間接来歴注釈を考える理由としては, 起源そのものを考えるだけでは, そのデータを得るために貢献したデータには価値を与えることができない可能性があるからである。例えば, 図2の結果の関係を属性 B の上に射影する演算を行なった場合, 図6のような結果が得られるが, この二つのセルの値 $2, 3$ が存在するのは図2で関係 R_1 の属性 A における二つの値 1 と, 関係 R_2 の属性 A における値 1 が同じ値を取っ

ており、自然結合することができたからである。そのため問合せ結果には残っていないが関係 R_1 の属性 A における二つの値 1 と、関係 R_2 の属性 A における値 1 にも報酬を与えるべきである。

つまり、演算において注目しているセルが存在しているのはどのデータの貢献によるのかを知るために間接来歴注釈を考えている。

Green ら [3] の “How-Provenance” は, “provenance semiring” と呼ばれる, 可換半環 $(K, 0, 1, +, \cdot)$ による多項式で表すと定義している。そのため “How-Provenance” をセル単位に拡張した直接来歴注釈に関しては, 直接来歴注釈のみの集合 K_s とし, 可換半環 $(K_s, 0, 1, +, \cdot)$ による多項式と定義する。また, 間接来歴注釈に関しても, 間接来歴注釈のみの集合 K_l とし, 可換半環 $(K_l, 0, 1, +, \cdot)$ による多項式と定義する。[3] では関係代数演算を可換半環の演算に写像するモデルを “K-relation” として定義し, 関係における組を可換半環の要素 K へ写像する関数を R としていた。本研究では, このモデルに関してもセル単位に拡張し, “ K^n -relation” とする。

K^n -relation : 関係における列数を n とする。関係におけるセルごとの直接来歴注釈の要素の集合を K_s とする。関係における写像モデル K_s^n -relation とは, 任意の組に対する関数 R_s^n のことである。ここで関数 R_s^n とは, 関係における組の, n 個のセルそれぞれを K_s の 0 以外の要素へ写像するもので, ここのセルは特殊な要素 0 への関係をもたないものである。次に, 間接来歴注釈の集合を K_l とし, 関係における写像モデル K_l^n -relation についても同様に定義する。写像モデル K_s^n -relation および, K_l^n -relation をまとめて, 写像モデル K^n -relation とする。

K^n -relation を持つ K を定義し, 可換半環 $(K, 0, 1, +, \cdot)$ による多項式で表した直接来歴注釈および, 間接来歴注釈を問合せ前の関係の各セルに追加し, 問合せによる演算の結果得られたデータの各セルの直接来歴注釈, 間接来歴注釈はどのようになるかを考える。Buneman ら [1] もデータを木構造として捉え, 独自の問合せ言語を用いて広い範囲で起源を考える方法を示しているが, 本研究では関係代数演算を前提とする。選択, 射影, 和, 自然結合および平均と総数, 最大と最小, カウントから成り立つ問合せ演算を行った場合に, 結果の関係の各セルの直接来歴注釈, 間接来歴注釈成分の計算方法を以下で定義する。セル (t, A) の直接来歴注釈, 間接来歴注釈をそれぞれ $source(t, A)$, $lineage(t, A)$ で表すものとする。また, 具体例として図 7,8,9 のデータベースを用いることとする。図 7,8,9 の各セルの成分は第一成分が直接来歴注釈, 第二成分が間接来歴注釈を表す。例では予め上記の直接来歴注釈, 間接来歴注釈を持っているものとする。

4.1 演算ごとの注釈演算の定義

選択, 射影, 和, 自然結合および, 平均と総数, 最大と最小, カウントの, それぞれの演算について注釈演算を定義する。任

name	based_in	phone
BayTours (s_{11}, l_{11})	San Francisco (s_{12}, l_{12})	415-1200 (s_{13}, l_{13})
HarborCruz (s_{21}, l_{21})	Santa Cruz (s_{22}, l_{22})	831-3000 (s_{23}, l_{23})

図 7 Agencies

name	destination	type
BayTours (s'_{11}, l'_{11})	San Francisco (s'_{12}, l'_{12})	cable car (s'_{13}, l'_{13})
BayTours (s'_{21}, l'_{21})	Santa Cruz (s'_{22}, l'_{22})	bus (s'_{23}, l'_{23})
BayTours (s'_{31}, l'_{31})	Santa Cruz (s'_{32}, l'_{32})	boat (s'_{33}, l'_{33})
BayTours (s'_{41}, l'_{41})	Monterey (s'_{42}, l'_{42})	boat (s'_{43}, l'_{43})

図 8 ExternalTours

name	destination	type
BayTours (s''_{11}, l''_{11})	San Francisco (s''_{12}, l''_{12})	cable car (s''_{13}, l''_{13})
HarborCruz (s''_{21}, l''_{21})	Santa Cruz (s''_{22}, l''_{22})	bus (s''_{23}, l''_{23})

図 9 OriExternalTours

意の問合せについては, 定義した注釈演算を再帰的に利用すれば良い。

4.1.1 選択

関係 R に対してある選択条件 c による選択を行った結果の関係 $\sigma_c R$ の中のあるセル (t_a, A) の直接来歴注釈, 間接来歴注釈を考える。ここで, t_a, A は関係 R のある組, および属性を表す。

ここで, 選択条件 c は最小の条件式集合と $\wedge$ (AND), $\vee$ (OR), $\neg$ (NOT) の 3 つの記号で構成されているものとする。最小の条件式に $\neg$ が付いている場合は 3. の $\neg$ に対する操作をあらかじめ行っておく。

演算の結果の関係のセルが元の関係のセル (t, A) から得られたものとする。この時, セル (t_a, A) の直接来歴注釈 $source(t_a, A)$ および間接来歴注釈 $lineage(t, A)$ を以下のように定義する。

$$source(t_a, A) = source(t, A)$$

$$lineage(t_a, A) = lineage(t, A) \cdot l_{sub}$$

ここで, l_{sub} は選択条件 c によって決定される, 間接来歴注釈の多項式である。

選択条件 c が, その全体を二分する記号によって, $c = C_1 \wedge C_2$ または $c = C_1 \vee C_2$ で構成されている場合, l_{sub} は C_1 及び C_2 の間接来歴注釈成分をそれぞれ l_{C_1}, l_{C_2} とすると, 以下のよう

(1) $c = C_1 \wedge C_2$ の場合

$$l_{sub} = l_{C_1} \cdot l_{C_2}$$

(2) $c = C_1 \vee C_2$ の場合

$$l_{sub} = l_{C_1} + l_{C_2}$$

条件式 C_1, C_2 の間接来歴注釈成分 l_{C_1}, l_{C_2} に対して上記の操作を、条件式が最小の条件式になるまで再帰的に利用する。その後、各最小の条件式についての間接来歴注釈成分は以下で求める。ある最小の条件式を c_{sub} とし、 c_{sub} を満たす属性集合を B とすると、それぞれの間接来歴注釈成分 l_c は以下である。

$$l_c = \prod_{b \in B} lineage(t, b)$$

以上の手順によって得られた l_{sub} を代入することで求めたい間接来歴注釈が得られる。なお $\neg$ (NOT) に関しては、ド・モルガンの法則によって最小の条件式の否定にできるため、上記の手順で計算可能である。

4.1.2 射 影

関係 R を属性集合 X 上に射影した結果の関係 $\pi_X R$ の中のあるセル (t_a, A) の直接来歴注釈、間接来歴注釈を考える。これらは、それぞれ、元の関係 R の組で X 上に射影すると t_a になるものを t とすると t の直接来歴注釈、間接来歴注釈の和とする。すなわち、

$$source(t_a, A) = \sum_{t \in R, \pi_X t = t_a} source(t, A)$$

$$lineage(t_a, A) = \sum_{t \in R, \pi_X t = t_a} lineage(t, A)$$

4.1.3 和

関係 R, S の和演算による結果の関係 $R \cup S$ の中のあるセル (t_a, A) の直接来歴注釈、間接来歴注釈を考える。結果の関係の組 t_a が、それぞれもとの関係 R と関係 S に存在する同一の組 t_R, t_S に対応する場合、 t_a のすべての属性 A に対し、

$$source(t_a, A) = source(t_R, A) + source(t_S, A)$$

$$lineage(t_a, A) = lineage(t_R, A) + lineage(t_S, A)$$

結果の関係の組 t_a がもとの関係 R の組 t_R のみに対応する場合、 t_a のすべての属性 A に対し、

$$source(t_a, A) = source(t_R, A)$$

$$lineage(t_a, A) = lineage(t_R, A)$$

結果の関係の組 t_a がもとの関係 S の組 t_S のみに対応する場合、 t_a のすべての属性 A に対し、

$$source(t_a, A) = source(t_S, A)$$

$$lineage(t_a, A) = lineage(t_S, A)$$

4.1.4 自然結合

関係 R, S の自然結合による結果の関係 $R \bowtie S$ の中のあるセル (t_a, A) の直接来歴注釈、間接来歴注釈を考える。結果の関係の組 t_a が元の関係 R と関係 S の組 t_R, t_S の自然結合に対応する場合、 t_a の属性のうち

- (1) $A_J \in att(R) \cap att(S)$
- (2) $A_R \in att(R) - att(A_J)$
- (3) $A_S \in att(S) - att(A_J)$

とすると、

直接来歴注釈：

$$source(t_a, A_J) = source(t_R, A_J) \cdot source(t_S, A_J)$$

$$source(t_a, A_R) = source(t_R, A_R)$$

$$source(t_a, A_S) = source(t_S, A_S)$$

間接来歴注釈：

$$\begin{aligned} lineage(t_a, A_J) \\ = \prod_{A_J} (lineage(t_R, A_J) \cdot lineage(t_S, A_J)) \end{aligned}$$

$$\begin{aligned} lineage(t_a, A_R) \\ = lineage(t_R, A_R) \\ \cdot \prod_{A_J} (lineage(t_R, A_J) \cdot lineage(t_S, A_J)) \end{aligned}$$

$$\begin{aligned} lineage(t_a, A_S) \\ = lineage(t_S, A_S) \\ \cdot \prod_{A_J} (lineage(t_R, A_J) \cdot lineage(t_S, A_J)) \end{aligned}$$

4.1.5 平均、総数

平均と総数の二つの演算については、結果として得られたその値を出すのに関わった全ての値が、その値を出すために必要だったと考える。そのため直接来歴注釈については結果として得られたその値を出すのに関わった全ての値の直接来歴注釈の積で表される。間接来歴注釈についても、同様に結果として得られたその値を出すのに関わった全ての値の間接来歴注釈の積で表される。

関係 R の属性集合 X について平均、または総数をとった場合の結果の関係 $avr_X R, sum_X R$ の中のセル (t_a, X) の直接来歴注釈および間接来歴注釈を以下のように定義する。

$$source(t_a, X) = \prod_{t \in R} source(t, X)$$

$$lineage(t_a, X) = \prod_{t \in R} lineage(t, X)$$

4.1.6 最大、最小

最大と最小の二つの演算については、結果として得られたその値の元となった全セルの直接来歴注釈および間接来歴注釈の和が、それぞれ結果の値の直接来歴注釈および間接来歴注釈として表される。その値が最大または最小となるためには、比べた他の値がその値よりも小さかった、または大きかった必要がある、という考えもあるが、仮に最大、最小以外の値が存在しなかったとしても最大、最小の値は変化しないため、それらの値の間接来歴注釈は無視することとする。

関係 R の属性集合 X について最大、または最小をとった場合の結果の関係 $max_X R, min_X R$ の中のセル (t_a, X) の直接来歴注釈および間接来歴注釈を以下のように定義する。

$max_X R$ ：

$$source(t_a, X) = \sum_{t \in R, (t, X) = MAX(X)} source(t, X)$$

$$lineage(t_a, X) = \sum_{t \in R, (t, X) = MAX(X)} lineage(t, X)$$

$min_X R$:

$$source(t_a, X) = \sum_{t \in R, (t, X) = MIN(X)} source(t, X)$$

$$lineage(t_a, X) = \sum_{t \in R, (t, X) = MIN(X)} lineage(t, X)$$

4.1.7 カウント

カウントについては、レコードの数を数えるので、指定した属性集合の全セルの直接来歴注釈および間接来歴注釈の積を、結果のセルの直接来歴注釈および間接来歴注釈として表す。重複する行はカウントに含まない DISTINCT を指定する場合もあるが、注釈付き演算においてはその場合、予め重複している行の直接来歴注釈および間接来歴注釈の和をとっておき、その注釈と他の行の注釈との積を結果のセルの直接来歴注釈および間接来歴注釈として表す。そのため DISTINCT を指定しない場合と指定する場合に分けて定義する。

DISTINCT を指定しない場合：関係 R の属性集合 X についてカウントをとった場合の結果の関係 $count_X R$ 中のセル (t_a, X) の直接来歴注釈および間接来歴注釈を以下のように定義する。

$$source(t_a, X) = \prod_{t \in R} source(t, X)$$

$$lineage(t_a, X) = \prod_{t \in R} lineage(t, X)$$

DISTINCT を指定する場合：DISTINCT を指定する場合、予め重複する行の注釈についてはそれぞれ和をとっておき、その後行ごとの注釈の積を結果のセルの注釈とする。これは予め射影演算を行い、その結果について積を取る演算と同じである。したがって、集約関数 $count_{distinct_X} R$ を以下のように変換し、注釈演算を行う。

$$count_{distinct_X} R = count_X (\pi_X R)$$

5. 差分に基づくセルの価値付け

データベースのあるセルが価値を持つかどうかを考える手法を定義する。考え方として、注目しているセルがある場合の問合せ結果と、注目しているセルが存在していない場合の問合せ結果との差を取り、その差を注目しているセルの価値とする方法がある。

あるデータベース D のあるセルの値を c とする。このデータベースに対してある問合せ q を行なった際、その問合せ結果を $q(D)$ と表し、 $q(D)$ におけるセル c の価値を $contribution(q(D), c)$ と表す。また、セル c の直接来歴注釈および間接来歴注釈を (s, l) とした時、この (s, l) を $(0, 0)$ で置き換えたデータベースを

Result of Q :

destination	phone	
San Francisco	415-1200	$t_1 \cdot (t_1 + t_3)$
Santa Cruz	831-3000	t_2^2
Santa Cruz	415-1200	$t_1 \cdot (t_4 + t_5)$
Monterey	415-1200	$t_1 \cdot t_6$

図 11 図 10 の問合せによる結果

destination	phone
San Francisco	415-1200
Santa Cruz	831-3000
Santa Cruz	415-1200
Monterey	415-1200

図 12 $q(D)$

destination	phone
Santa Cruz	831-3000

図 13 $q(D_{\perp c_1})$

$D_{\perp c}$ 、問合せ結果を $q(D_{\perp c})$ とすると、 $contribution(q(D), c)$ は以下の式で定義する。

$$contribution(q(D), c) = difference(q(D), q(D_{\perp c}))$$

ここで $difference(q(D), q(D_{\perp c}))$ は、問合せ結果 $q(D)$ と $q(D_{\perp c})$ との差である。

また、注目しているセルが存在しない場合の問合せ結果 $q(D_{\perp c})$ を、問合せ前のセル c の直接来歴注釈および間接来歴注釈 (s, l) を $(0, 0)$ とした場合の問合せ結果としているが、これは [2] にあるように、組単位の provenance が 0 であるとは、その組が存在しないことであると定義されていることから、[2] の “How-Provenance” をセル単位に拡張した本研究の直接来歴注釈および間接来歴注釈が 0 であるとは、そのセルが存在しないことであると定義できるからである。

5.1 例

文献 [2] の第 3 章 “How-Provenance” で使用されている例では、図 7,8 に対して以下の問合せを実行している。

```

Q:
SELECT  e.destination, a.phone
FROM    Agencies a,
        (SELECT name,
                 based_in AS destination
         FROM Agencies a
         UNION
         SELECT name, destination
         FROM ExternalTours ) e
WHERE   a.name = e.name

```

図 10 [2] 第 3 章 “How-Provenance” p418 の例

この問合せを行った結果は図 11 のようになる。この例における問合せを q として、 $difference$ を考える。

図 7 の (1,1) 成分 “BayTours” に注目し、これをセル c_1 とする。このとき、 $q(D)$ は、図 12 になり、 $q(D_{\perp c_1})$ は図 13 になる。

destination	phone
San Francisco	415-1200
Santa Cruz	831-3000
Santa Cruz	415-1200
Monterey	415-1200

図 14 $q(D_{\perp_{c_2}})$

destination	phone
San Francisco $(s_{12} + s'_{12}, l_{11} \cdot (l_{11} + l'_{11}) \cdot (l_{12} + l'_{12}))$	415-1200 $(s_{13}, l_{11} \cdot l_{13} \cdot (l_{11} + l'_{11}))$
Santa Cruz $(s_{22}, l_{21} \cdot l_{22})$	831-3000 $(s_{23}, l_{21} \cdot l_{23})$
Santa Cruz $(s'_{22} + s'_{32}, l_{11} \cdot (l'_{22} + l'_{32}) \cdot (l'_{21} + l'_{31}))$	415-1200 $(s_{13}, l_{11} \cdot l_{13} \cdot (l'_{21} + l'_{31}))$
Monterey $(s'_{42}, l_{11} \cdot l'_{41} \cdot l'_{42})$	415-1200 $(s_{13}, l_{11} \cdot l_{13} \cdot l'_{41})$

図 15 $q(D)$ (来歴注釈付き)

destination	phone
San Francisco $(s_{12} + s'_{12}, 0)$	415-1200 $(s_{13}, 0)$
Santa Cruz $(s_{22}, l_{21} \cdot l_{22})$	831-3000 $(s_{23}, l_{21} \cdot l_{23})$
Santa Cruz $(s'_{22} + s'_{32}, 0)$	415-1200 $(s_{13}, 0)$
Monterey $(s'_{42}, 0)$	415-1200 $(s_{13}, 0)$

図 16 $q(D)$ (c_1 の注釈を 0 にした場合)

destination	phone
Santa Cruz $(s_{22}, l_{21} \cdot l_{22})$	831-3000 $(s_{23}, l_{21} \cdot l_{23})$

図 17 $q(D)_{\perp_{c_1}}$

二つの図を見ると、セル c_1 の有無でかなり結果が違っていることがわかる。この差が *difference* になる。

次に、図 8 の (1,3) 成分 “cable car” に注目し、これをセル c_2 とする。このとき、 $q(D_{\perp_{c_2}})$ は図 14 になる。図 12 と図 14 を見ると、セル c_2 の有無で結果は同じになっている。この場合 *difference* は 0 であると考えることができる。

ここで、 $q(D)$ を、来歴注釈付きで演算した結果について考えると、図 15 のようになり、この結果において、セル c_1 の注釈 (s_{11}, l_{11}) を $(0, 0)$ にすると、図 16 になる。そもそも、間接来歴注釈は注釈のついているセルが存在するために必要な要素であるので、間接来歴注釈が 0 であるということは、そのセルは存在できないということになる。そのため間接来歴注釈成分が 0 のセルはデータが存在しないということになり、組全ての間接来歴注釈成分が 0 である組は消去できる。従って、図 16 は図 17 のように書き換えることができ、図 13 と同じであることがわかる。これを $q(D)_{\perp_{c_1}}$ とおくと、

$$q(D_{\perp_{c_1}}) = q(D)_{\perp_{c_1}}$$

となる。任意のセルについてこの式が成り立つことが証明できれば、図 18 のように、一つ一つのセルについてそれを $\perp$ に置き換えて問合せ演算を行い *difference* を考える、という手順を、注目しているセルの注釈成分を $(0, 0)$ にして、問合せ結果に代入するだけででき、効率化を測ることが可能となる。

6. 注釈付き問合せ演算における差分等価性

5. では、データの価値を差分を用いて定義した。ここで、注釈付き問合せ演算においては、はじめに注目しているデータの注釈を 0 に置き換えて演算を行なった場合の問合せ結果と、問合せを行なった後、注目しているデータの注釈を 0 にした場合の問い合わせ結果とが、等価になる。

注釈付き問合せ演算における差分等価性： あるデータベース D のあるセルを c とし、その直接来歴および間接来歴注釈を (s, l) とする。このデータベースに対して、集約関数を含まないある問合せ q を行なった際、その問合せ結果 $q(D)$ とし、セル c の (s, l) を $(0, 0)$ で置き換えたデータベースを $D_{\perp_c}$ 、問合せ結果を $q(D_{\perp_c})$ とする。また、問合せ結果 $q(D)$ において s, l に 0 を代入した結果を $q(D)_{\perp_c}$ とする。このとき、

$$q(D_{\perp_c}) = q(D)_{\perp_c}$$

が成り立つ。

(略証)．左辺について、演算の流れを考えると以下のようなになる。

- (1) セル c の注釈成分 (s, l) に 0 を代入する。
- (2) 問合せ q による注釈付き問合せ演算（この時点で注釈による消去は行わない）を行う。
- (3) 仮の問合せ結果 $q(D_{\perp_c})$ に対して、注釈による消去を行う。
- (4) 消去後の $q(D_{\perp_c})$ が左辺となる。

手順 (2) の時点では、注釈成分が $(0, 0)$ のセルをヌル値 $\perp$ にしたり、組全ての間接来歴注釈が 0 の組を見つけて消去する操作は行わず、この操作は手順 (3) で行う。さらに、注釈はそれぞれ可換半環 $(K_s, 0, 1, +, \cdot)$ または $(K_l, 0, 1, +, \cdot)$ による多項式であると定義している。これは演算の過程でセル内の注釈がなくなることがないことを保証する。従って、左辺の演算方法は右辺において問合せ演算 q を行なった後に、注釈 s, l に 0 を代入する方法と同じであり、

$$q(D_{\perp_c}) = q(D)_{\perp_c}$$

が成り立つ。 □

7. データ提供者と報酬分配

[10] では、直接来歴注釈および、間接来歴注釈を以下のように定義した。

直接来歴注釈・間接来歴注釈： 問合せによって得られたそのセルが元のどのセルから得られたものなのかを考えるために、問合せによって演算が行われる前のデータのセルにそれぞれデータの起源を示す直接来歴注釈を与える。さらに、問合せによって得られたそのセルは、どのセルの貢献によって得られたものなのかも考えるため、間接来歴注釈も与える。

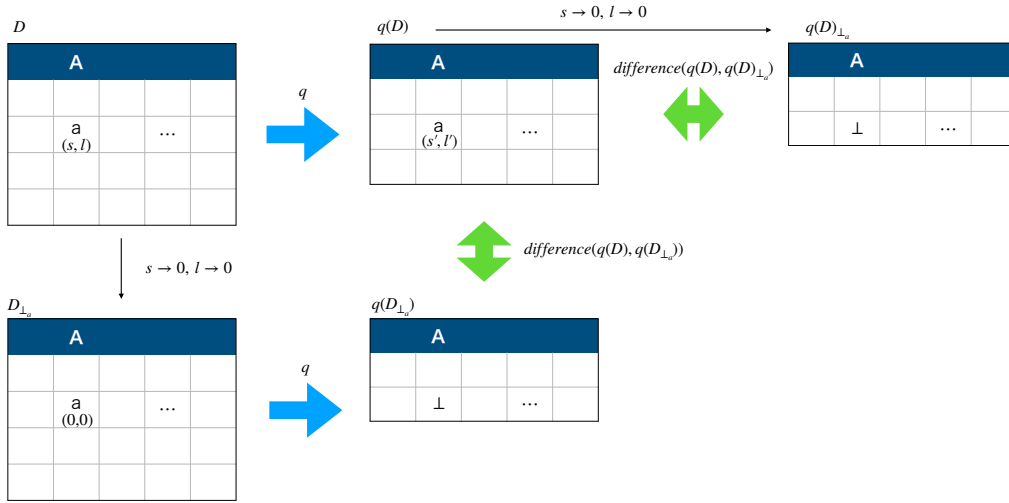


図 18 $q(D_{\perp_a})$ と $q(D)_{\perp_a}$

paperid	papertitle	year
p352 (s_{11}, l_{11})	"OASSIS..." (s_{12}, l_{12})	2014 (s_{13}, l_{13})
p353 (s_{21}, l_{21})	"Monitoring..." (s_{22}, l_{22})	2007 (s_{23}, l_{23})

図 19 *Papers P*

paperid	a.order	c.name	attributes
p352	1	Tova M.	[papertitle]
p352	2	Susan D.	[papertitle]
p352	3	Slava N.	[papertitle]
p353	1	Tova M.	[papertitle]

図 20 *Creators C*

paperid	r.order	r.name	attributes
p352	1	Michael S.	[*]
p353	1	Michael S.	[paperid, papertitle]
p353	2	Alex M.	[year]

図 21 *Registrars R*

実際、元となったセルにはそのデータの提供者が存在する。データの提供者とは、そのデータを記録したデータ記録者および、場合によってはそのデータ自体を生み出したデータ作成者のことだと考えられる。[10] では注釈自体に報酬を分配するところまで話を終えていたが、その先を考える。改めて、データ提供者を以下のように定義する。

データ提供者： データの提供者とは、そのデータを記録したデータ記録者および、場合によってはそのデータ自体を生み出したデータ作成者のことである。

注釈に与えられた報酬は、次にデータ提供者に分配される。例えば図 19 において“OASSIS...”と入力した人物が Michael S. だった場合、データ提供者は、そのデータを記録したデータ記録者である Michael S. および、論文“OASSIS...”そのものを生み出したデータ作成者 Tova M., Susan D., Slava N. があげられる。したがって、“OASSIS...”の直接来歴注釈 s_{12} に与えられた報酬の分配先はこの場合 Michael S. と Tova M., Susan D., Slava N. である。

また、“p352”については、p352 そのものを生み出した人はいないので、この場合はデータ提供者はデータ記録者 Michael S. のみである。

次に、両者に報酬を分配する方法は様々なものが考えられる。決め方の一つとしては、直接来歴注釈の分配先を、問合せで表現しておくと言う方法が考えられる。予めデータ提供者のデータベースを用意しておき、分配手法によって問合せを作

成するという方法をとる。データ提供者データベースは例えば図 20, 21 のような例が考えられる。ここで、attributes は、Creators の場合、c.name に該当する人物が作成したデータの属性の配列が入り、Registrars の場合、r.name に該当する人物が記録したデータの属性の配列が入る。例えば、図 21 において一列目の Michael S. の attributes は [*] となっている。これは、Michael S. が paperid “p352”の全ての属性のデータを記録したということの意味する。また、二列目は、Michael S. が paperid “p353”の属性のうち、paperid と papertitle の属性のデータを記録したということの意味する。

例えば直接来歴注釈 s_{12} の分配先として、記録者および作成者の双方に分配をするのであれば、

$$s_{12} \rightarrow \delta_{c.name \rightarrow provider} (\pi_{c.name} ((\sigma_{pid=p352} P) \bowtie (\sigma_{\text{"papertitle" in attributes} C})) \cup \delta_{r.name \rightarrow provider} (\pi_{r.name} ((\sigma_{pid=p352} P) \bowtie (\sigma_{\text{"papertitle" in attributes} R})))$$

provider
Tova M.
Susan D.
Slava N.
Michael S.

図 22 *providers Pro₁*

provider
Tova M.
Susan D.
Slava N.

図 23 *providers Pro₂*

のように表現できる。図 22 のような結果が得られる。

また、作成者のみに分配をするのであれば、

$$s_{12} \rightarrow \delta_{c.name \rightarrow provider}(\pi_{c.name}((\sigma_{pid=p352}P) \bowtie (\sigma_{"papertitle" \text{ in } attributes}C)))$$

のように表現できる。図 23 のような結果が得られる。問合せとして考えられるのは例えば以上のような例である。

この問合せはさらに以下のように一般化することができる。

問合せ一般化： あるキー *key* を持つテーブル *T* において、(キー) = *ID*, (属性) = *A* であるセルの直接来歴注釈 *s* に与えられた報酬の分配先を考える。属性 “key”, “c.order”, “c.name”, “attributes” を持つテーブル *C* と、属性 “key”, “r.order”, “r.name”, “attributes” を持つテーブル *R* をあらかじめ用意しておく。

直接来歴注釈 *s* に与えられた報酬の分配先は以下のような問合せで表現することができる。

$$s \rightarrow \delta_{c.name \rightarrow provider}(\pi_{c.name}((\sigma_{key=ID}T) \bowtie (\sigma_{A \text{ in } attributes}C)) \cup \delta_{r.name \rightarrow provider}(\pi_{r.name}((\sigma_{key=ID}T) \bowtie (\sigma_{A \text{ in } attributes}R)))$$

8. 考 察

本研究の 5. で提案した差分に基づくセルの価値付けにおいて、差分が具体的にどのようなものかは定義していないため、より具体的に考える必要がある。

また、6. で証明した定理は、集約関数を含まない問合せのみを考えている。将来的に集約関数を含む任意の問い合わせについて定理が成り立つことを証明していく。

そして実データに対する問合せの例を使用して、実際に提案した手法を用いて報酬の分配をする実験をする必要もある。

9. 終わりに

本研究では、[10] に引き続き、データの起源および貢献度に基づいてデータベースに対する問合せ結果に与えられた利益を分配する方法について議論を進めた。[10] で定義した直接来歴注釈および間接来歴注釈を改めて 4. で定義し、5. では新たにセルの価値を決定する尺度として、注目しているセルが存在している場合の問合せ結果と、注目しているセルが存在していない場合の問合せ結果との差分を用いることを提案した。6. では差分の計算において等価であるものを証明し、差分計算の効率化が可能であることを示した。また、7. では、データ提供者を改めて定義し、注釈に分配された報酬を、さらにデータ提供者に分配する方法の一つを述べた。最後に 8. で今後の課題を述べた。

10. 謝 辞

本研究は JSPS 科研費基盤研究 (S) No. 17H06099, (A) No. 18H04093 の助成を受けたものです。

文 献

- [1] Peter Buneman, Sanjeev Khanna, and Tan Wang-Chiew. Why and where: A characterization of data provenance. In *International conference on database theory*, pp. 316–330. Springer, 2001.
- [2] James Cheney, Laura Chiticariu, and Wang-Chiew Tan. Provenance in Databases: Why, How, and Where. *Found. Trends databases*, Vol. 1, No. 4, pp. 379–474, April 2009.
- [3] Todd J Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 31–40. ACM, 2007.
- [4] Todd J. Green and Val Tannen. The Semiring Framework for Database Provenance. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS ’17, pp. 93–99, New York, NY, USA, 2017. ACM.
- [5] Grigoris Karvounarakis, Zachary G Ives, and Val Tannen. Querying data provenance. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pp. 951–962. ACM, 2010.
- [6] Chao Li, Daniel Yang Li, Gerome Miklau, and Dan Suciu. A Theory of Pricing Private Data. *Commun. ACM*, Vol. 60, No. 12, pp. 79–86, November 2017.
- [7] OECD. Exploring the Economics of Personal Data. OECD Digital Economy Papers, Organisation for Economic Co-operation and Development, Paris, April 2013.
- [8] Aaron Roth. Technical Perspective: Pricing Information (and Its Implications). *Commun. ACM*, Vol. 60, No. 12, pp. 78–78, November 2017.
- [9] 内閣官房 IT 総合戦略室. AI、IoT 時代におけるデータ活用ワーキンググループ中間とりまとめの概要, 3 月 2017 年.
- [10] 鍋谷優斗, 吉川正俊. データベース問合せ結果販売利益の分配に関する研究. 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM2018), pp. G1–6, 2018.