シャープレイ値に基づく関係データベース問合せ結果販売利益分配に関する研究

鍋谷　優斗†　　吉川　正俊††

† 京都大学大学院情報学研究科　〒 606–8303 京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8303 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

あらまし　今日情報銀行という事業が社会に出始めていたり，データの価値に関する注目が高まっているということがあり，データに価値が与えられた際に，その価値をデータを提供したデータ提供者に分配する手法の提案に取り組んでいる．具体的には，データベースに対してデータを買うデータバイヤーが欲しいデータの問合せと報酬をデータベースに与え，データベースは受け取った問合せの結果をデータバイヤーに返すが，この時受け取った報酬は，結果となったデータを提供してくれた人物に分配するべきだと考え，その分配手法を考えている．そのために，データの起源を表す provenance という概念をセル単位に拡張し，問合せ結果が誰のデータのものなのか，誰のデータの貢献によるものなのかを特定する手法を提案した．さらに，特定したデータ提供者に得られた報酬を分配する手法を，ゲーム理論において複数人が協力して利得を得た際に各人に公正に利得を計算する shapley 値を利用して定義した．

キーワード　情報銀行，データベース，provenance, shapley 値

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 [16] 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 [12].

情報銀行の普及に向けて，パーソナルデータの価値付けおよびその分配方法を開発することが有効だと考える．パーソナルデータの価値を考える研究は [15], [10], [11], [13] でも行われている．上記に加えて，問合せを元にデータの価値付けを行う研究も [9] で行われている．これらの研究は問合せ一つ一つに対して具体的に価値付けを行なっていない．そこで本研究は [17] に続いて，関連研究よりもさらに細かく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わかるものなので，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る．データの起源を考えるという研究は [5] すでに多くのものがあり，Cheney らはサーベイ [5] で三つの起源 “Where-Provenance”, “How-Provenance”, “Why-Provenance” を紹介している．

図 1 は，問合せを行うデータ購入者が，その問合せ q とデータ

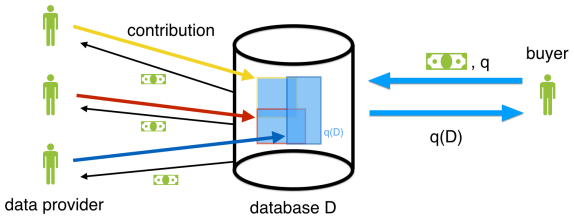

図 1　利益分配のイメージ

に対する価値をデータベース 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 $q(D)$ をデータ購入者に返し，その結果のデータを提供した人物である，データ提供者に報酬を分配する流れを示す．

さらに，データベースの各データが結果に対してどの程度貢献したかについてを測る尺度として Shapley 値 [14] を用いる．Shapley 値はゲーム理論において協力によって得られた利得を各プレイヤーへ公正に分配する方法の一案であり，これを本研究に応用することが可能だと考える．そこで本研究では，2. で，本研究で使用する用語や問合せの定義を行い，Shapley 値の一般的な定義を紹介する．次に 3. でデータベースにおける Shapley 値の定義と，Shapley 値を用いたデータの価値付けを，具体的な例を用いて行う．また，来歴を用いた価値付け方法についても提案し，二つの価値付け方法が等しい結果になるという仮説を立てる．そして，2., 3. をもとに，4. ではセル単位での Shapley 値の計算方法を演算ごとに定義する．これによってセル単位での価値付けが可能になると考える．

2. 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．

name	based_ in	phone
BayTours	San Francisco	415-1200
HarborCruz	Santa Cruz	831-3000

図 2　*Agencies*

2.1　セルの表現と問合せ

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 問合せまたは，さらに平均 (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，$\sigma_c R$ と表す．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．

2.2　Shapley 値

本研究では想定している報酬の分配の実現のために Shapley 値 [14] を利用する．Shapley 値とは，ゲーム理論において協力によって得られた利得を各プレイヤーへ公正に分配する方法の一案である．Shapley 値の形式的な定義を以下に示す．

プレイヤーの集合 N および関数 $v : \mathcal{P}(N) \to \mathfrak{R}$ を考える．ここで，$\mathfrak{R}$ は実数の集合である．あるプレイヤーの集合を $P \subset N$ とすると，関数 $v(P)$ は P に含まれるプレイヤーの協力によって得られる利得を表す．関数 v は以下の性質を持つ．

（1）　$v(\varnothing) = 0$

（2）　$v(S \cup T) \geqq v(S) + v(T)$

ここで S, T は N の任意の非交の部分集合．また，関数 v を特性関数と呼ぶ．

この時プレイヤー i は関数 v の元で以下の配分 $\phi_i(v)$ を得る．

$$\phi_i(v) = \sum_{S \in N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

ここで，n はプレイヤーの総数であり，足し合わせの範囲は N の部分集合 S のうちプレイヤー i を除いたもの．この式はプレイヤーが加わる全順列に対する，注目しているプレイヤー i が加わった際の貢献度増加の平均値を計算している．

2.3　具　体　例

本研究で具体例として用いる関係および問合せの例と問合せ結果を用意しておく． 問合せの例として，文献 [5] の第 3 章 “How-Provenance” で使用されている例を用いる．[5] では図 2,3 に対して図 4 の問合せを実行している．この問合せを行った結果が図 5 のようになる．この具体例を 3.，4. で用いる．問合せの過程は 3. の具体例で述べるため，ここでは省略する．

name	destination	type
BayTours	San Francisco	cable car
BayTours	Santa Cruz	bus
BayTours	Santa Cruz	boat
BayTours	Monterey	boat

図 3　*ExternalTours*

```
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 4　[5] 第 3 章 “How-Provenance” p418 の例

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

図 5　図 4 の問合せによる結果

3.　データベースにおける Shapley 値

2.2 節で Shapley 値について紹介した．Shapley 値とは，ゲーム理論において協力によって得られた利得を各プレイヤーへ公正に分配する方法の一案である．この考え方は，データ提供者が提供したデータに対して問合せがあり，問合せ結果に支払われた報酬を各データ提供者に分配するというケースに適用できる．そのため本研究では Shapley 値をデータベースに応用するということを考える．Shapley 値を本研究に応用すると，関係の各組を提供したデータ提供者がプレイヤーであり，協力によってできたものが問合せ結果，得られた利得がデータ購入者によって与えられた報酬と考えることができる．そこで，データベースにおける Shapley 値について，その考え方と定義を行なった後，四つの演算（選択，射影，集合和，自然結合）における Shapley 値の計算方法を示す．なお，本章以降では，データの起源の概念として [5] で紹介されている “How-Provenance” や [17] で提案した直接来歴注釈および間接来歴注釈といった来歴を利用する．

3.1　データベース問合せにおける Shapley 値の定義

データベースにおける Shapley 値を考える．データベースにおけるプレイヤーをデータベースの各組とし，問合せ結果に対して与えられた報酬を，各組に分配するという状況を想定する．データベースにおけるプレイヤーの貢献としては主に二つが考えられる．

- 問合せ結果の来歴に対する貢献（Ⓟと表す)
- 問合せ結果の値に対する貢献 (Ⓥと表す)

よって，データベース問合せにおける関数 v は上記の二通りとする．したがって，選択，射影，集合和，自然結合の 4 つの演算について上で述べた前提において Shapley 値を計算する．

3.1.1 選　　択

選択における shapley 値を考える．ある関係 R の k 行目の組を T_k とする．また，1 行目から順に “How-Provenance” を $t_1, t_2, ...$ のようにつけておく．t_k は T_k の “How-Provenance” である．関係は n 行存在するとする．

問合せとしてある条件式 e を持つ選択 $\sigma_e R$ を考える．この結果を $Q(R)$ として，$Q(R)$ の行数を m とする．このとき組 T_k が条件式 e を満たす場合を $True$，満たさない場合を $False$ とすると，組 T_k の問合せ結果に対する二つの貢献度ⓟ,ⓞⓟの Shapley 値は以下のように表せる．

$$ⓟ : \begin{cases} \frac{n\cdot(n-1)!}{n!}t_k = t_k & (True) \\ 0 & (False) \end{cases}$$

$$ⓞⓟ : \begin{cases} \frac{n\cdot(n-1)!}{n!}T_k = T_k & (True) \\ 0 & (False) \end{cases}$$

$True$ の場合について，選択では任意の組の並べ替えについて，組 T_k が追加された時点で結果には新しい組 T_k が現れるので，を関係 R の全組の集合 N としたとき，集合 S を N の部分集合 S のうち組 T_k を除いたものとすると，二つの貢献度に対する，組 T_k が加わった際の貢献の増加分 $v(S \cup \{T_k\}) - v(S)$ は以下のように表すことができる．

$$ⓟ : v(S \cup \{T_k\}) - v(S) = t_k$$

$$ⓞⓟ : v(S \cup \{T_k\}) - v(S) = T_k$$

以上により，上記の Shapley 値となる．

3.1.2 射　　影

射影における Shapley 値を考える．ある関係 R は属性 A を持っており，関係 R の k 行目の組 T_k は属性 A に値 a を持つとする．また，関係 R の属性 A に値 a を持つ組は組 T_k を含めて m 個存在するとする．問合せとして，属性 A への射影 $\pi_A R$ を考える．この時組 T_k における二つの貢献度ⓟ,ⓞⓟに対する Shapley 値はそれぞれ以下のように表せる．

$$ⓟ : \frac{n\cdot(n-1)!}{n!}t_k = t_k$$

$$\begin{aligned} ⓞⓟ &: \frac{{}_n\mathrm{C}_m(n-m)!(m-1)!}{n!}T_k \\ &= \frac{n!}{(n-m)!m!}\cdot\frac{(n-m)!(m-1)!}{n!}T_k \\ &= \frac{1}{m}T_k \end{aligned}$$

射影では重複削除が行われるため，組 T_k が加わることによる貢献度ⓞⓟの変化は，組の任意の並べ替えのうち，組 T_k の前に属性 A が m である組が現れない場合 T_k，一つ以上現れる場合 0 となる．そのため全通り $n!$ のうち，T_k 以外の属性が a である $m-1$ 個の組の場所とその並べ替え ${}_n\mathrm{C}_m(m-1)!$ と，T_k を除いた残りの $n-m$ 個の並べ替え $(n-m)!$ の積で計算することができる．一方で組 T_k が加わることによる貢献度ⓟの変化は，組みの任意の並べ替えにおいて変わらず t_k である．以上により，上記の Shapley 値となる．

t_1^R	T_1^R
⋮	⋮
t_k^R	T_k^R
⋮	⋮
t_n^R	T_n^R

図 6　table　R

t_1^S	T_1^S
⋮	⋮
t_n^S	T_n^S

図 7　table　S

3.1.3 集　合　和

次のような関係 R, S(図 6, 図 7) を用意する．それぞれ n_1 行，n_2 行存在するとする．問合せとして，二つの関係の集合和 $R \cup S$ を考える．関係 R 中の k 番目の組 T_k が関係 S 中のある組と重複している場合と，いずれの組とも重複していない場合の二通りについて T_k の Shapley 値を計算する．

重複無しの場合

$$ⓟ : \frac{(n_1+n_2)\cdot(n_1+n_2-1)!}{(n_1+n_2)!}t_k = t_k$$

$$ⓞⓟ : \frac{(n_1+n_2)\cdot(n_1+n_2-1)!}{(n_1+n_2)!}T_k = T_k$$

重複ありの場合

$$ⓟ : \frac{(n_1+n_2)\cdot(n_1+n_2-1)!}{(n_1+n_2)!}t_k = t_k$$

$$ⓞⓟ : \frac{\frac{1}{2}(n_1+n_2)!}{(n_1+n_2)!}T_k = \frac{1}{2}T_k$$

重複無しの場合，任意の組の並べ替えについて，組 T_k を加えることにより二つの貢献度ⓟ,ⓞⓟはそれぞれ t_k, T_k だけ変化する．重複ありの場合，貢献度ⓟは任意の並べ替えについて組 T_k を加えることによる貢献度の増加は t_k である．一方で，貢献度ⓞⓟは任意の並べ替えのうち，重複するもう一つの組の前に組 T_k を加えた場合 T_k，重複するもう一つの組の後に加えた場合 0 となる．そしてその確率は $\frac{1}{2}$ のため，上記の Shapley 値となる．

3.1.4 自 然 結 合

自然結合における Shapley 値を考える．次のような関係 R, S(図 8, 図 9) を用意する．問合せとして二つの関係の自然結合 $R \bowtie S$ を考える．関係 R のある組 T_k^R が関係 S の m 個の組と結合するとする．この m 個の組の集合および来歴の集合をそれぞれ M，$\mathcal{M}$ とする．この時組 T_k^R に対する二つの貢献度ⓟ,ⓞⓟに対する Shapley 値は以下のように表せる．

t_1^R	T_1^R
$\vdots$	$\vdots$
t_k^R	T_k^R
$\vdots$	$\vdots$
t_n^R	T_n^R

図 8　table　R

t_1^S	T_1^S
$\vdots$	$\vdots$
$t^S(\in \mathcal{M})$	$T^S(\in M)$
$\vdots$	$\vdots$
t_n^S	T_n^S

図 9　table　S

$$\text{ⓟ}: \frac{\sum_{i=1}^{m}(m-1)!i(\sum_{t_j^S\in\mathcal{M}} t_j^S)t_k^R}{(m+1)!}$$
$$= \frac{\frac{1}{2}m(m+1)(m-1)!(\sum_{t_j^S\in\mathcal{M}} t_j^S)t_k^R}{(m+1)!}$$
$$= \frac{1}{2}(\sum_{t_j^S\in\mathcal{M}} t_j^S)t_k^R$$

$$\text{ⓞⓟ}: \frac{\sum_{i=1}^{m}(m-1)!i(\sum_{T_j^S\in M} T_j^S)T_k^R}{(m+1)!}$$
$$= \frac{\frac{1}{2}m(m+1)(m-1)!(\sum_{t_j^S\in M} T_j^S)T_k^R}{(m+1)!}$$
$$= \frac{1}{2}(\sum_{T_j^S\in M} T_j^S)T_k^R$$

ただしⓞⓟにおける結合は積の形で表現している．
まず，組 T_k^R に対する自然結合の Shapley 値を考える場合において，組 T_k^R と，組 T_k^R と結合する関係 S の m 個の組 $T^S(\in M)$ の計 $(m+1)$ 個の順列を全通りとして考えれば良い．

（略証）． 全ての組の数を n とし，考えたい組の数を $(m+1)$ とする．このとき考えたい $(m+1)$ 個の組の出現確率は以下のようになる．

$$\frac{{}_n\mathrm{C}_{m+1}(\text{m+1 行の並び替え})(n-m-1)!}{n!} = \frac{(\text{m+1 行の並び替え})}{(m+1)!}$$

したがって，考えたい $(m+1)$ 個の組の全ての並び替えを分母として考えて良い．　□

次にⓟについて，その Shapley 値の計算方法を説明する．

t_k^R が先頭に来る場合　　t_k^R が加わることによる貢献度増加は，結合できる組がこの時点では存在しないため任意の場合において 0 となる．

t_k^R が $i+1$ 番目に来る場合（$i: 1 \leqq i \leqq m$　の整数）　ある $t_j^s(\in \mathcal{M})$ に注目する．組 T_k^R が加わることによる貢献度増加 $v(S\cup\{i\})-v(S)$ のうち $t_j^S\cdot t_k^R$ について考える．
$v(S\cup\{i\})-v(S)$ が $t_j^S\cdot t_k^R$ を含むには t_j^R の前の i 個のうち一つが t_j^S となる必要がある．したがってその条件において場合の数を計算すれば，貢献度増加に $t_j^S\cdot t_k^R$ を含む総数が求められる．t_j^S のはじめの i 箇所の並べ替えと，その各々について t_j^S と t_k^R を除いた残りの $m-1$ 個の並べ替えの計算なので，$t_j^S\cdot t_k^R$ の総数は以下のようになる．

$$\sum_{i=1}^{m} i\cdot(m-1)!t_j^S\cdot t_k^R = \frac{1}{2}(m+1)!t_j^S\cdot t_k^R$$

m 個の任意の組について同様の計算ができるので，その総和を全通り $(m+1)!$ で割ったものが求める Shapley 値となる．よって，以下のような結果となる．

$$\frac{\sum_{t_j^S\in\mathcal{M}}\frac{1}{2}(m+1)!t_j^S\cdot t_k^R}{(m+1)!} = \frac{1}{2}(\sum_{t_j^S\in\mathcal{M}} t_j^S)t_k^R$$

また，ⓞⓟについても同様の計算で Shapley 値を求めることができる．ただし，ここでは組同士の結合結果を例えば $T_k^R\cdot T_j^S$ のように表している．

3.2　価 値 付 け

データ購入者が問合せ Q をデータベース D に与え，問合せ結果 $Q(D)$ を受け取り，報酬として V をデータベースに与えたとする．この報酬 V を各データに分配する場合を考える．報酬分配の方法として，Shapley 値を用いる方法と来歴を用いる方法の二通りが存在する．それぞれの方法についてその手順を以下に記す．

Shapley 値による分配：
分配は以下の手順で行う．

（1） 問合せ結果 $Q(D)$ の各組に報酬 V を分配する．

（2） 問合せ Q を，選択，射影，集合和，自然結合のいずれか一つを用いた問合せ $q_1, q_2, \cdots, q_n$ に分解する．すなわち，$q_n(\cdots q_2(q_1(D))\cdots) = Q(D)$ となる．また，$q_0(D) = D$ となる問合せ q_0 を定義しておく．これにより，$q_n(\cdots q_2(q_1(q_0(D)))\cdots) = Q(D)$ となる．

（3） 問合せ結果 $q_{n-1}(\cdots q_2(q_1(q_0(D)))\cdots)$ に問合せ q_n を行なった場合の Shapley 値計算から，問合せ結果 $q_{n-1}(\cdots q_2(q_1(q_0(D)))\cdots)$ の各組への報酬を計算する．

（4） 3 を $n=1$ まで繰り返す．

（5） 各組に与えられた報酬の総和がその組に分配される最終報酬となる．

	name	destination	type
t_1 :	BayTours	San Francisco	cable car
t_2 :	BayTours	Santa Cruz	bus
t_3 :	BayTours	Santa Cruz	boat
t_4 :	BayTours	Monterey	boat

図 10　$ExternalTours$

	name	destination	type
t_3 :	BayTours	Santa Cruz	boat
t_4 :	BayTours	Monterey	boat

図 11　$\sigma_{type='boat'}ExternalTours$

なお，上記で前提としている報酬 V が問合せ結果に分配される場合では，各組の貢献度は分配された報酬となる．例えばデータベース D のある組 T がある条件式 e とした場合の問合せ $\sigma_e D$ において，条件式 e を満たし結果に表れたとする．結果のその組に報酬 v が分配されたとすると，その問合せにおける元データベース D の組 T の貢献度は v となる．3.1.1 節の定義では条件式を満たす場合の$\oplus$をそのまま出力される組を使い T_k と表していたが，上記の前提の場合出力された組 T_k に対して分配された報酬 v が$\oplus$となる．

来歴による分配：

分配は以下の手順で行う．

（1） 問合せ結果 $Q(D)$ の各組に報酬 V を分配する．

（2） $Q(D)$ の各組に分配された報酬を，来歴を元に元のデータベースの組に分配する．来歴に対する分配方法は，ある来歴集合 A, B に対して，二項演算 $A \circ B$ が報酬 v が与えられた組に付与された来歴であったとすると，A, B に分配される報酬はそれぞれ $\frac{1}{2}v$ とする．括弧の外から再帰的に分配する．

（3） それぞれの来歴に対して与えられた報酬の総和が求める最終報酬となる．

以上の二通りの分配方法について，四つの演算（選択，射影，集合和，自然結合）における報酬分配の例を示す．なお例に用いる関係として，図 3 の関係を用いる．この関係に “How-Provenance” を付与したものを図 10 とする．問合せ前の組を一行目から T_1, T_2, T_3, T_4 とする．

3.2.1 選　　択

図 10 の関係に対して問合せ $\sigma_{type='boat'}ExternalTours$ を行う．この問合せ結果が図 11 となる．問合せ結果に価値 $V = 100$ が与えられたとし，Shapley 値による分配と来歴による分配それぞれの分配結果を手順に沿って考える．

Shapley 値による分配：

（1） 問合せ結果に与えられた報酬 $V = 100$ を各組に分配する．ここでは単純に各組に均等に分配するとする．組は T_3, T_4 の二組が結果に残っているのでこの二組に報酬を 50 ずつ分配する．

	name	based_ in	phone
t_1 :	BayTours	San Francisco	415-1200
t_2 :	HarborCruz	Santa Cruz	831-3000

図 12　$Agencies$

	name	destination	type
t_3 :	BayTours	San Francisco	cable car
t_4 :	BayTours	Santa Cruz	bus
t_5 :	BayTours	Santa Cruz	boat
t_6 :	BayTours	Monterey	boat

図 13　$ExternalTours$

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

図 14　図 4 の問合せによる結果

（2） 今回の問合せは $\sigma_{type='boat'}ExternalTours$ より，q_1 は $\sigma_{type='boat'}$ となる．

（3） 問合せ結果 $q_0(D) = D = ExternalTours$ に問合せ q_1 を行なった結果 $q_1(D) = \sigma_{type='boat'}ExternalTours$ に対する，問合せ結果 $q_0(D) = D = ExternalTours$ の Shapley 値計算から，各組への報酬を計算する．3.1.1 節の定義より，条件式を満たす組に対して，結果の組に与えられた報酬がそのまま分配されるので，報酬分配結果は，$T_3 = 50, T_4 = 50$ となる．

（4） すでに $n = 1$ のためスキップ.

（5） 以上より各組に分配される最終報酬は，$T_1 = 0, T_2 = 0, T_3 = 50, T_4 = 50$ となる．

来歴による分配：

（1） 問合せ結果に与えられた報酬 $V = 100$ を各組に分配する．ここでは単純に各組に均等に分配するとする．組は T_3, T_4 の二組が結果に残っているのでこの二組に報酬を 50 ずつ分配する．

（2） 各組に分配された報酬を，来歴を元に元のデータベースの組に分配する．今回は来歴が二項演算をもたないため t_3 が付与された組 T_3 と t_4 が付与された組 T_4 にそれぞれ 50 の報酬が与えられる．

（3） 以上より各組に分配される最終報酬は，$T_1 = 0, T_2 = 0, T_3 = 50, T_4 = 50$ となる．

3.3 具 体 例

2.3 節の例において Shapley 値の計算を行なう．図 2，3 に “How-Provenance” をつけたものを図 12，13 で表す．二つの関係の組をそれぞれ一行目から T_1, T_2 および T_3, T_4, T_5, T_6 とする．また，データベース $D = Agencies, ExternalTours$ とする．この関係に対して図 4 の問合せを行う．問合せ結果は 14 で表される．問合せに対して報酬 $V = 2000$ が与えられたとする．行なった問合せを元に，Shapley 値による分配と来歴によ

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

図 15　D_5

る分配の二通りの方法で報酬分配を行う．

Shapley 値による分配：

（1） 問合せ結果の図 14 に与えられた報酬 $V = 2000$ を各組に分配する．ここでは単純に各組に均等に分配するとする．四つの組にそれぞれ 500 ずつ分配する．

（2） 今回の問合せは図 4 である．これを関係代数式で表すと，

$$\begin{aligned}&\pi_{destination,phone}(\\&\quad Agencies\\&\quad \bowtie\\&\quad ((\delta_{based\_in \to destination}(\pi_{name,based\_in} Agencies))\\&\qquad \cup\\&\qquad (\pi_{name,destination} ExternalTours)))\end{aligned}$$

これを分解すると，

$$\begin{aligned}q_1 &= \pi_{name,based\_in} Agencies\\q_2 &= \delta_{based\_in \to destination} D_1\\q_3 &= \pi_{name,destination} ExternalTours\\q_4 &= D_2 \cup D_3\\q_5 &= Agencies \bowtie D_4\\q_6 &= \pi_{destination,name} D_5\end{aligned}$$

となる．ここで，D_1, D_2, D_3, D_4, D_5 はそれぞれ問合せ q_1, q_2, q_3, q_4, q_5 の問合せ結果である．

（3） 問合せ結果 $D_5 = q_5(q_4(q_3(q_2(q_1(q_0(D))))))$ に問合せ q_6 を行なった結果 $q_6(q_5(q_4(q_3(q_2(q_1(q_0(D)))))))$ に対する，問合せ結果 D_5 の Shapley 値計算から，各組への報酬を計算する．D_5 は図 15 であり，これに問合せ q_6 を行なった場合の D_5 の各組に分配される報酬は，全て 500 となる．

（4） 手順 3 を $n = 1$ まで繰り返す．

（5） 以上より各組に分配される最終報酬は，$T_1 = 125 + 750 = 875, T_2 = 250 + 250 = 500, T_3 = 125, T_4 = 125, T_5 = 125, T_6 = 250$ となる．

来歴による分配：

（1） 問合せ結果に与えられた報酬 $V = 2000$ を各組に分配する．ここでは単純に各組に均等に分配するとする．四つの組にそれぞれ 500 ずつ分配する．

（2） 各組に分配された報酬を，来歴を元に元のデータベースの組に分配する．問合せ結果の一行目の組は $t_1 \cdot (t_1 + t_3)$ より，まず括弧の外の $\cdot$ に対してその左右に $\frac{500}{2} = 250$ ずつ分配する．さらに括弧内の $+$ に対してその左右に $\frac{250}{2} = 125$ ずつ分配する．したがって結果的に t_1 が付与された組 T_1 には $250 + 125 = 375$ を分配し，t_3 が付与された組 T_3 には 125 を分配する．二行目以降の組にも同様に分配を行う．

（3） 以上より各組に分配される最終報酬は，$T_1 = 250 + 125 + 250 + 250 = 875, T_2 = 500, T_3 = 125, T_4 = 125, T_5 = 125, T_6 = 250$ となる．

3.4　価値付けの仮説

3.2 で，問合せに対する報酬の分配から，その問合せにおける元データの価値を組単位で付けることができた．そしてその報酬分配の方法は二通り存在し，どちらの方法でも同じ価値が付けられた．このことから，価値付けに関する次の仮説を立てることができる．

価値付けの仮説：

Shapley 値による分配を使用した報酬分配結果と，来歴による分配を使用した報酬分配結果は等しい．

4.　セル単位の Shapley 値

3. では組単位の来歴を利用してデータベース問合せにおける Shapley 値を考えた．そして, [17] ではセル単位での来歴として，直接来歴注釈および間接来歴注釈を定義した．4. ではさらに，エル単位の来歴を利用してデータベース問合せにおける Shapley 値を考える．

4.1　セル単位の Shapley 値の定義

3. と同様に，データベースにおけるプレイヤーをデータベースの各セルとし，問合せにおける結果に対して与えられた報酬を，各セルに分配するという状況を想定する．選択，射影，集合和，自然結合の 4 つの演算について上で述べた前提においてShapley 値を計算する．なお，ここでは “あらかじめ元データが組単位で揃っている場合にのみ問合せにそのデータが使用される” と言う前提のもと計算方法を定義するとする．

4.1.1　射　　影

問合せとして，ある属性集合 A の上に射影を行うとする．A の属性数を n とする．関係の i 行目の組を T_i とし，組 T_i の j 列目のセルを C_{ij} とする．また，j 列目の属性は $X \in A$ とする．セル C_{ij} の直接来歴注釈および間接来歴注釈を (s_{ij}, l_{ij}) とする．セル C_{ij} を含めて m 個のセルが属性 X に値 C_{ij} を持つとする．このとき，属性集合 A の上から射影を行なった場合のセル C_{ij} における二つの貢献度ⓟ,ⓞⓡに対する Shapley 値はそれぞれ以下のように表せる．

$$ⓟ : \frac{(nm-1)!}{(nm)!}(s_{ij}, l_{ij}) = \frac{1}{nm}(s_{ij}, l_{ij})$$

$$\textcircled{op} : \frac{(nm-1)!}{(nm)!}C_{ij} = \frac{1}{nm}C_{ij}$$

4.1.2 集 合 和

二つの関係 R, S を用意する．それぞれ n_1 行 m 列，n_2 行 m 列存在するとする．問合せとして，二つの関係の集合和を考える．関係代数式では $R \cup S$ と表せる．関係 R 中の i 行 j 列目の組 T_i^R が関係 S 中のある組と重複している場合と，いずれの組とも重複していない場合の二通りについて C_{ij} の Shapley 値を計算する．

重複無しの場合

$$\textcircled{p} : \frac{1}{m}(s_{ij}, l_{ij})$$

$$\textcircled{op} : \frac{1}{m}T_i^R$$

重複ありの場合

$$\textcircled{p} : \frac{1}{m}(s_{ij}, l_{ij})$$

$$\textcircled{op} : \frac{{}_{2m-1}\mathrm{C}_m(m-1)! \cdot m!}{(2m)!}T_i^R = \frac{1}{2m}T_i^R$$

重複無しの場合，任意のセルの並べ替えについて，m 個のカラムのうち，いずれか一つのセルが貢献度増加をもつ．そして対称性からこの場合のセル C_{ij} が加わることによる貢献度増加 ⓟ および ⓞⓟ は以上のようになる．重複ありの場合，セル C_{ij} が加わることによる貢献度増加 ⓟ は，関係 S の重複する組が先に追加されるかどうかにかかわらず，関係 R の組 T_i^R が揃った時点で貢献度増加が行われる．したがって重複無しの場合と同様．ⓞⓟ は，関係 S の重複する組よりも先に関係 R の組 T_i^R が揃う必要がある．したがって，それら二つの組の合計セル数 $2m$ 個の並べ替えのうち，先頭から $2m-1$ 個のいずれか m 箇所に組 T_i^R の m 個のセルが入るような並べ替えを考えればよく，以上の計算方法となる．

4.1.3 自 然 結 合

二つの関係 R, S を用意する．それぞれ n_1 行 m_1 列，n_2 行 m_2 列存在するとする．問合せとして二つの関係の自然結合を考える．関係 R の i 行 j 列のセル C_{ij} が属する組 T_i^R が，関係 S の a 個の組と結合するとする．この a 個の組の集合および来歴の集合をそれぞれ M，$\mathcal{M}$ とする．この時組 T_i^R に属する i 行 j 列のセル C_{ij}^R 対する二つの貢献度ⓟ,ⓞⓟに対する Shapley 値は以下のように表せる．

$$\textcircled{p} : \frac{1}{2m_1}(\sum_{t_j^S \in \mathcal{M}} t_j^S)t_k^R$$

$$\textcircled{op} : \frac{1}{2m_1}(\sum_{T_j^S \in M} T_j^S)T_k^R$$

4.1.4 選 択

選択における Shapley 値を考える．n 行 m 列の関係 R を用意する．問合せとしてある条件式 e を持つ選択を考える．このとき組 T_i が条件式 e を満たす場合を $True$，満たさない場合を $False$ とすると，組 T_i に属する i 行 j 列のセル C_{ij} が加わることによる二つの貢献度ⓟ,ⓞⓟ に対する Shapley 値は以下のように表せる．

$$\textcircled{p} : \begin{cases} \frac{1}{m}t_k & (True) \\ 0 & (False) \end{cases}$$

$$\textcircled{op} : \begin{cases} \frac{1}{m}T_k & (True) \\ 0 & (False) \end{cases}$$

なお，あらかじめ元データが組単位で揃っている場合にのみ問合せにそのデータが使用されるとしていたが，組が揃っていなくても追加された時点で問合せに使用される場合も考えることは可能で，そうすることで例えば選択においては選択条件によって各セルの Shapley 値が変化する．実際はこちらの方が自然だが計算方法は複雑となる．

4.2 価 値 付 け

3.2 節と同様に，データ購入者が問合せ Q をデータベース D に与え，問合せ結果 $Q(D)$ を受け取り，報酬として V をデータベースに与えたとする．この報酬 V を各データに分配する場合を考える．報酬分配の方法として，Shapley 値を用いる方法を考える．Shapley 値による分配の手順は 3.2 節と同様である．ただし，3.2 節では組単位で分配を行なっていたが，ここではセル単位で分配を行う．なお，3.2 節と同様に上記で前提としている報酬 V が問合せ結果に分配される場合では，各セルの貢献度は分配された報酬となる．4.1 節で定義したセル単位の Shapley 値の定義では，セルが属している組が Shapley 値の一部として利用されているが，セル単位の価値付けでは組の価値を，組に属する全セルの貢献度の和とする．また，問合せ結果のセルの貢献度が 0 の場合には以下の二通りの分配方法が考えられる．

(i) 問合せ前のセルの貢献度を定義通り割り振る

(ii) 問合せ結果のセルの貢献度が 0 の場合は問合せ前のセルの貢献度を 0 とする

Shapley 値による分配方法を使用して 3.3 節と同じ例で報酬分配を行なった場合を考える．2.3 節の例において Shapley 値の計算を行なう．二つの関係の組をそれぞれ一行目から T_1, T_2 および T_3, T_4, T_5, T_6 とする．また，二つの関係の i 行 j 列目のセルをそれぞれ C_{ij}^A, C_{ij}^E とする．また，データベース $D = Agencies, ExternalTours$ とする．この関係に対して図 4 の問合せを行う．問合せに対して報酬 $V = 2000$ が与えられたとする．行なった問合せを元に，Shapley 値による分配方法で報酬分配を行う．

Shapley 値による分配：

（1） 問合せ結果に与えられた報酬 $V = 2000$ を各セルに分配する．ここでは単純に各セルに均等に分配するとする．八つのセルにそれぞれ 250 ずつ分配する．

（2） 3.3 節と同様に問合せを分解し，$q_1, ..., q_6$ とする．

（3） 問合せ結果 $D_5 = q_5(q_4(q_3(q_2(q_1(q_0(D))))))$ に問合せ q_6 を行なった結果
$q_6(q_5(q_4(q_3(q_2(q_1(q_0(D)))))))$ に対する，問合せ結果 D_5

の Shapley 値計算から，各セルへの報酬を計算する．D_5 に問合せ q_6 を行なった場合の D_5 の各セルに分配される報酬は，属性が *destination* および *phone* の全セルに 250 ずつとなる．

（4） 手順 3 を $n = 1$ まで繰り返す．

（5） 以上より各組に分配される最終報酬は，i の考え方では，$C^A_{11} = \frac{125}{2} + 250 = \frac{625}{2}, C^A_{12} = \frac{125}{2} + 250 = \frac{625}{2}, C^A_{13} = 250, C^A_{21} = \frac{250}{3} + 125 = \frac{625}{3}, C^A_{22} = \frac{250}{3} + 125 = \frac{625}{3}, C^A_{23} = \frac{250}{3}, C^E_{11} = \frac{125}{2}, C^E_{12} = \frac{125}{2}, C^E_{13} = 0, C^E_{21} = \frac{125}{2}, C^E_{22} = \frac{125}{2}, C^E_{23} = 0, C^E_{31} = \frac{125}{2}, C^E_{32} = \frac{125}{2}, C^E_{33} = 0, C^E_{41} = 125, C^E_{42} = 125, C^E_{43} = 0$ となる．ii の考え方では，$C^A_{11} = 0 + 0 = 0, C^A_{12} = 125 + 0 = 125, C^A_{13} = 750 + 0 = 750, C^A_{21} = 0 + 125 = 125, C^A_{22} = 0 + 250 = 250, C^A_{23} = 250 + 0 = 250, C^E_{11} = 0, C^E_{12} = 125, C^E_{13} = 0, C^E_{21} = 0, C^E_{22} = 125, C^E_{23} = 0, C^E_{31} = 0, C^E_{32} = 125, C^E_{33} = 0, C^E_{41} = 0, C^E_{42} = 250, C^E_{43} = 0$ となる．

5. 考察および今後の課題

3. では，データベースにおける Shapley 値を定義したが，集約関数を用いる問合せについては定義していないため，こちらについて考える必要がある．3. 2 節では Shapley 値による組の価値付けの一例を示したが，ここでは Shapley 値の定義の一つとして来歴の貢献度増加を利用しているが価値付けの一例には利用していない．したがって来歴を用いた Shapley 値を使った組の価値付け方法についても考える必要がある．3. 4 節では価値付けの仮説を立てたが，これが正しいことの証明を行う．4. では，セル単位の Shapley 値を定義したが，ここではあらかじめ元データが組単位で揃っている場合にのみ問合せにそのデータが使用されるとしていた．しかし組が揃っていなくても追加された時点で問合せに使用される場合も考えることは可能で，そうすることで例えば選択においては選択条件によって各セルの Shapley 値が変化する．したがって組が揃っていなくても追加された時点で問合せに使用される場合についても考える必要がある．また，4. 2 節ではセル単位の Shapley 値を用いた報酬分配の一例を示したが，3. 2 節と同様に注釈の貢献度増加を利用していないため注釈の貢献度増加を利用した価値付けの方法についても考える．また，注釈による分配についてもまだ考えられていないため，その方法についても考える必要がある．

6. 終 わ り に

本研究では，データベースの価値付けを考えるため，2. で Shapley 値の概念を導入し，3. でデータベースにおける組単位の Shapley 値の定義と，それを使用した報酬分配の方法を具体例を用いて示した．また，3. で得られた，最後に 4. で，3. を利用してセル単位での Shapley 値を考え，具体例を用いて報酬分配の一例を示した．

7. 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] Shapley value. wikipedia: Free encyclopedia.
[2] エコノミックゲームセオリー: 協力ゲームの応用. SGC ライブラリ. サイエンス社, 2001.
[3] J. Berstel and D. Perrin. *Theory of Codes*. ISSN. Elsevier Science, 1985.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] Paraschos Koutris, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu. Query-Based Data Pricing. *Journal of the ACM*, Vol. 62, No. 5, pp. 1–44, November 2015. 00090.
[10] Chao Li, Daniel Yang Li, Gerome Miklau, and Dan Suciu. A Theory of Pricing Private Data. *ACM Trans. Database Syst.*, Vol. 39, No. 4, pp. 34:1–34:28, December 2014. 00095.
[11] 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.
[12] OECD. Exploring the Economics of Personal Data. OECD Digital Economy Papers, Organisation for Economic Cooperation and Development, Paris, April 2013.
[13] Aaron Roth. Technical Perspective: Pricing Information (and Its Implications). *Commun. ACM*, Vol. 60, No. 12, pp. 78–78, November 2017.
[14] L.S. Shapley, A.E. Roth, and Cambridge University Press. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.
[15] Sarah Spiekermann, Rainer Böhme, Alessandro Acquisti, and Kai-Lung Hui. Personal data markets. *Electronic Markets*, Vol. 25, No. 2, pp. 91–93, April 2015. 00066.
[16] 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.
[17] 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.
[18] 日本数学会. 岩波数学辞典　第 4 版. 岩波書店, 2007.