セル単位のデータ来歴—データ引用に向けて

朴 柱英† 吉川 正俊†† 加藤 弘之†††

† †† 京都大学 情報学研究科 社会情報学専攻 〒 606-8501 京都市左京区吉田本町
††† 国立情報学研究所 コンテンツ科学研究系 〒 101-8430 東京都千代田区一ツ橋 2-1-2 学術総合センタービル内
E-mail: †pjy953@db.soc.i.kyoto-u.ac.jp, ††yoshikawa@i.kyoto-u.ac.jp, †††kato@nii.ac.jp

あらまし 最近，データの公開，共有，再利用の重要性が高くなっている．データの公開，共有，再利用を通じて我々は研究過程の透明性を高めることと同時に，複数のデータの統合やデータの再解釈による新しい研究結果を得ることもできる．データの公開，共有，再利用が重要になることによって，データ引用の重要性も高くなっている．データ引用は自分が使ったデータを正しく引用するために必要なものである．データ引用は，従来の論文引用とは引用方法が異なる．従来の論文引用では一つの論文が引用の単位であったのに対して，データ引用では，様々な部分データが引用の単位となっている．また，引用をすることによってデータを提供した人に対して正しく貢献を与えなければならないが，その方法はまだ提案されていない．本研究では，データ引用に問合せを用い，データ作成者への貢献にデータ来歴 (Data Provenance) を使うことを提案する．そのために，関係データベースのタプル単位のデータ来歴をセル単位のデータ来歴に変換し，セル単位のデータ来歴を用いて貢献評価まで拡張する．また，セル単位のデータ来歴には表れない作成者にも貢献を与えるために，潜在関係 (Latent Relation) を定義し，潜在データ来歴を計算する．そして，求められた潜在データ来歴からセル単位データ来歴の計算，貢献評価ができることも示す．本研究の結果，引用から正しい貢献を計算できるようになり，データ作成者は正しい貢献を受けると同時に，高品質のデータを提供する動機付けを与えると期待する．

キーワード データ引用，セル単位のデータ来歴

1. 序　　論

最近，データの公開，共有，再利用の重要性が高くなっている．データの公開，共有，再利用を通じて我々は研究過程の透明性を高めることと同時に，複数のデータの統合やデータの再解釈による新しい研究結果を得ることもできる．2016 年度に開かれた G7 の科学部長官会議の声明書 (G7 Science and Technology Ministers ’ Meeting) ではデータ，結果の共有に関するオープンサイエンス (Open Science) の話題もあった [1]．また，ホワイトハウス科学技術政策局でも研究開発支出が年間 1 億ドル以上の政府機関に対しては，その研究成果と研究成果に使われたデータを公開させようとしている [9]．また，RDA(Research Data Alliance) のような研究データの共有，再利用の研究を行うための研究団体も登場した (https://rd-alliance.org/)．

このようなデータの公開，共有，再利用が行われる時，論文の場合と同様に，研究者が自らが利用したデータを正しく引用することが重要である．従って，したデータを正しく引用する方法が活発に研究されており，代表的な研究として [3], [8] 等がある．自分が使ったデータを正しく引用することにより得られることの一つとしてデータを提供したデータ作成者に対して正しく貢献を与えられることがある．貢献を与えるために利用できる引用の例として論文の引用がある．論文引用では自分の論文で参考にした論文を引用し，それぞれの論文は “被引用” という形で評価される．従って，高い被引用を持つ論文はその分野で重要な論文として認められる．しかし，データの場合は論文の被引用に該当するものがまだ存在しない．その原因はデータの粒度 (granularity) にある．このデータの粒度が原因になる理由は論文引用とデータ引用の引用対象が違うからである．論文引用の引用対象は論文自体である．すなわち，論文の一部を使っても，すべてを使っても，引用対象は論文自体になる．しかし，データ引用の引用対象はすべてのデータではなく，実際に使ったデータである．従って，データの一部を使ったか，すべてを使ったによって，引用対象が変わる．例えば，図 1 は京都と大阪の 11 月と 12 月の気温と降水量に関する情報であり，インタネット上で公開されて誰でも使えるとする．その時，(1) ある人は京都のデータのみを使い，(2) ある人は京都の気温のデータのみを使い，(3) ある人は京都と大阪の降水量のデータのみを使った場合，(1)，(2)，(3) のデータ使用者が同じ引用方法を使ってしまうとデータ作成者に対して正しく貢献を与えることができなくなるはずだ．なぜかというと，気温と降水量のデータの作成者が異なる可能性があるからである．すなわち，作成者 A が気温のデータを計り，作成者 B が降水量のデータ

都市名	日付	平均気温 (°C)	降水量の月合計値 (mm)	タプルタグ
京都 (c_{11})	2016.11(c_{12})	12.5(c_{13})	82.0(c_{14})	t_1
京都 (c_{21})	2016.12(c_{22})	8.2(c_{23})	92.0(c_{24})	t_2
大阪 (c_{31})	2016.12(c_{32})	9.4(c_{33})	104.0(c_{34})	t_3

図 1: 都市ごとの気温，降水量 (S)

を計った場合は，(2) は A のみに貢献を与え，(3) は B のみに貢献を与えるべきである．

本研究ではこのように公開されている任意のデータベースから問合せにより求められたデータだけを使う場合，自分が使ったデータの作成者に正しく貢献を与えることを目標にしている．貢献を正しく与えるための方法として本研究ではデータ来歴を使う．データ来歴は結果がどのようなデータから起因してきたかに関する情報である．従って，問合せにより求められたデータを使う場合は，そのデータのデータ来歴を計算することで，自分の使ったデータを提供した作成者を正しく探すことができる．その結果，データ来歴を計算することで結果を出すために使われたソースデータのデータ作成者が分かるようになる．また，データ来歴の情報の詳しさによってデータ来歴は幾つかの種類が存在する．それぞれ Why Provenance [4], How Provenance [7]，Where Provenance [4] 等である．また，このすべてを一つの代数構造として統合したものが，Buneman ら [7] の来歴半環 (Provenance Semirings) である．来歴半環は数学の半環構造を使ってデータ来歴を表したものである．

例えば，ある人が図 1 のようなデータベースに対して

Q “SELECT 都市名, 平均気温 FROM S WHERE 都市名=‘京都’’

により求められたデータを使った場合，貢献を受ける対象は京都の 11 月，12 月の気温データの作成者である．しかし，従来のデータ来歴ではタプル単位のデータ来歴であったから，気温と降水量のデータを作った人がそれぞれ異なる場合は，正しく貢献を与えることができなかった．従って，正しく貢献を与えるための方法として本研究では，従来のデータ来歴を細分化したセル単位のデータ来歴を提案する．ここでのセルはタプルより小さい単位で，タプルと属性で決まるものである．

タプル単位とセル単位のデータ来歴の違いを例 1 で説明する．Q で得られる結果レコード毎のデータ来歴は図 2 に表れている通りである．Q は都市名でデータをフィルタリングし，図 1 の都市名と平均気温のみを出力する．すなわち，元の図 1 のデータの中で降水量に関する情報はどこでも使われていない．それにもかかわらず，従来のデータ来歴のリニエッジ (Lineage) による結果は降水量の情報を含むタプル t_1 で結果レコードのデータ来歴を表現している (他の種類のデータ来歴も同様)．それに比べて本研究で提案するセル単位データ来歴はタプルのどの属性が関与しているをピンポイントで表す．その結果，気温データの作成者のみに正しく貢献を与えることができる．

例 1. 従来のデータ来歴 *(例 : リニエッジ [6])* とセル単位データ来歴の比較

t_1 は $c_{11}, c_{12}, c_{13}, c_{14}$ を全部含む．従って，従来のデータ来歴をによると Q では c_{12}, c_{14} を使っていないのにその情報を含む t_1 が結果になっている．それに比べて，セル単位データ来歴では使ったセルである c_{11}, c_{13} のみを正しい結果として出す．

都市名	平均気温 (°C)	リニエッジ	セル単位データ来歴
京都	12.5	$\{t_1\}$	$c_{11} \cdot c_{13}$
京都	8.2	$\{t_2\}$	$c_{21} \cdot c_{23}$

図 2: 図の関係に対する Q の結果

以降，３節では問合せを用いたデータ引用において，使われたデータに関するより詳細なセル単位の情報を得るために，データ来歴をセル単位で定義する．これは，既存のタプル単位のデータ来歴の定義 [7] をセル単位に拡張したものとなっている．４節では，３節で定義したセル単位のデータ来歴だけでは表現できない貢献度を計算するために，各関係演算子のファクタを考慮した潜在関係を定義する．これにより，直接結果に出現するデータだけでなく，結果を得るために用いられたデータをセル単位で特定可能となり，よりきめ細かい貢献度の計算が可能となる．

2. 関 連 研 究

本研究の目標はデータの作成者に正しく貢献を与えることであるが，本研究は二つの研究分野，“引用分野” と “データ来歴” の融合である．引用に関する従来の研究は主に引用文の生成，引用文からデータを特定，引用文の参照などであり，データ作成者への貢献評価はまだできていない．本研究では貢献評価のためにデータ来歴を使っている．その理由は，データ来歴の本来の目標と貢献評価の目標が一致するところがあるからである．データ来歴の目標は自分が使ったデータの来歴を探すことであり，貢献評価での目標は自分が使ったデータの作成者を探すことである．従って，データ来歴の方法を貢献評価のために拡張できる．本研究で仮定している引用方法は，問合せによる引用である．問合せによる引用方法はデータベースからある問合せで求められたデータを使い，その情報を引用することを意味する．これに関しては, [3] で定義している引用方法をそのまま使うことにする. [3] ではデータ引用を問合せとデータベースから一意に決まるものとして定義している．すなわち，データベースに対して問合せが計算されたことが分かっているならその計算結果によるデータ引用文の自動生成ができる．また，その自動生成の具体的な方法まで提案されている．

また，本研究で貢献評価のために使っている方法論である “データ来歴” は最近活発に研究されている分野である．データの来歴はデータの起源とデータの生涯中 (Life Span) の歴史を記述するためのものである．伝統的にデータベースのデータはどんな疑問もなく真 (True) と信じられてきた．しかし，最近，オンラインてデータが収集されたり，いろいろなソースからデータが収集されたりするので，データベースのデータをそのまま信じることはリスクが高い．従って，自分が使ったデータがどこから起因してきたのかが分かることは最近のデータベースを扱う上で大事なことである．これを解決するために来歴に関する研究がいろいろなところで行なわれている．来歴の研究成果のサーベイ [5] によると来歴研究は大きく三つの分野に分かれる．各々Why，How，Where Provenance である．Why Provenance は “結果がなぜ出たのか？” [4]，How Provenance は “結果がどうして出たのか？” [7]，Where Provenance は “結果がどこから出たのか？” [4] を表現するためのものである．そ

の中で本研究では [7] の来歴半環 (provenance semirings) をセル単位に変換したものをデータ来歴として使う．

3. セル単位データ来歴

c この節では従来のデータ来歴を貢献評価で使えない理由と，その問題点を克服するためにデータ来歴をより細分化したセル単位データ来歴を定義し，セル単位データ来歴による貢献評価について述べる．セル単位データ来歴を使って貢献評価することで我々はより正解な貢献評価ができるようになる．

従来のデータ来歴はデータ来歴がタプルに付けられているタグの組み合わせによって表現された．しかし，最近のデータは複数の作成者が作ったデータを一つのタプルに統合して提供する場合も多く，複数の観測装置から集まったデータを一つのタプルとして統合して提供する場合も少なくない．この時，従来のデータ来歴を使うと，タプル単位の演算になり，属性に関する情報がなくなってしまう．従って，本研究ではこの問題を解決するためにタプル単位のデータ来歴をより細分化したセル単位のデータ来歴を使う．セル単位のデータ来歴でのセルは直観的に関係データベースがタプルの行と属性の列により構成されていると見なした時，一つの行の一つの列に該当するものである (例：図 1 の c_{11} 等)．セルを定義するためにまずタプルを定義する．本研究ではタプルの定義として関係モデルの named perspective [2] を使う．named perspective は属性の有限集合 U と値のドメイン $\mathbb{D}$ を用いてタプルを写像 $t : U \to \mathbb{D}$ で表現する．また，ある時刻で $\mathbb{D}$ を固定し，その時可能なすべてのタプルの集合を U-Tup と表現する．U に対する関係は U-Tup の部分集合である．

定義 1. *U-Cell とセル*
属性の有限集合 U が与えられたとき，*U-Cell* を次のように定義する．

$$U\text{-}Cell = \{(t,A) \mid t \in U\text{-}Tup\ ,\ A \in U\}$$

セルは *U-Cell* の一つの要素であり，c で表現する．

本研究では同じ値を持つセルでも違うものとして区別して使っている．従って，各々のセルにはタグが付けられていると仮定する．そのために，U-Cell に含まれているすべてのセルにタグを付ける必要がある．元のデータベースから求められる U-Cell にタグ付ける関数を本研究では **TAG** とする．

定義 2. *C は区別される要素 0 を持つ集合である．U-Cell を C に写像する関数 **TAG** は次のように定義される．*

$$\boldsymbol{TAG} : U\text{-}Cell \to C$$

また，*U-Cell* のすべての要素に ***TAG*** を計算し，その結果の集合を ***TAG**(U-Cell)* と表現し，次のように定義される．

$$\boldsymbol{TAG}(U\text{-}Cell) = \{\boldsymbol{TAG}((t,A)) \mid (t,A) \in U\text{-}Cell\}$$

また，集合 C に対して，単位元 *0* を持つ $+$演算と，単位元 *1* を持つ $\cdot$ 演算が定義されていることにする．

ただし，これ以降曖昧性が生じない場合は，TAG((t,A)) を TAG(t,A) のように表記する．関数 **TAG** は U-Cell の各々の要素を C の要素に写像する．本研究では C の値をタグとも呼ぶ．すなわち，(t,A) で表現されるある c は **TAG** により C の要素である値を持つことになる．U-Cell と同様に U-Tup の各々の要素も C に写像する．また，U-Tup の有限部分集合である関係上の関係代数演算に対応して，U-Tup から写像された C の集合上の代数操作を定義する．

定義 3. *属性の有限集合 U に対する C-relation は関数 $R : U\text{-}Tup \to C$ である．*

C-relation では U-Tup の要素を C に写像している．しかし，本研究では U-Tup から得られる C の値は ***TAG***(U-Cell) の要素から決まるものとする．正確には，U-Tup のある要素を t とした時，C-relation 上での R(t) は ***TAG***(U-Cell) の要素の $\cdot$ 演算でできるものである．また，その具体的な値は次のように定義される．

定義 4. *タプルのタグ*
あるタプル $t(\in U\text{-}Tup)$ の C-relation 上での $R(t)$ の値は次のように定義される．

$$R(t) = \prod_{A \in U} \boldsymbol{TAG}(t,A)$$

定義 4 により得られる結果は，タプルのタグとも呼び，そのタプルのタグ上での代数演算を行う．タプルのタグ上での代数演算のために本研究では C-relation に対して関係代数演算を定義する．すなわち，タプルのタグ上での代数演算は C-relation に対する関係代数演算で計算される．また，C-relation 上での関係代数演算は [7] の来歴半環をベースにしている．従って，C-relation での関係代数演算は K-relation [7] での関係代数演算のように，選択，射影，集合和，結合を表すために “足し算” と呼ばれる 2 項演算 “+” と “掛け算” と呼ばれる 2 項演算 “$\cdot$” を使う．このように，C-relation の演算は基本的に K-relation と同じ計算方法を使うが，射影だけは違う．属性情報を表すために，C-relation の射影は射影に使われなかった属性のセルのタグを 1 に変換する．C-relation に対する関係代数演算の定義は以下のようである．

定義 5. *$(C,+,\cdot,0,1)$ を二つの演算と二つの区別される原子を持つ代数構造だとすると，集合差を除いた関係代数演算は次のように定義される*
空関係 ***(empry relation)*** *任意の属性集合 U に対して，$\emptyset(t) = 0$ を満足する $\emptyset : U\text{-}Tup \to C$ がある．*
集合和 ***(union)*** *もし，$R_1, R_2 : U\text{-}Tup \to C$ であると，$R_1 \cup R_2 : U\text{-}Tup \to C$ は次のように定義される．*

$$(R_1 \cup R_2)(t) = R_1(t) + R_2(t)$$

射影 ***(projection)*** *もし，$R : U\text{-}Tup \to C$ であり，$V \subseteqq U$ であると，$\pi_V R : V\text{-}Tup \to C$ は次のように定義される．*

$$(\pi_V R)(t) = \sum_{V \text{ の上で } t=t' \wedge R(t') \neq 0} (\boldsymbol{PROJ}(R(t'))$$

$\boldsymbol{PROJ}(R(t'))$ は C から C' への写像で，セルのタグを変換するための関数である．C' は以下のようになる．

$C' = \{\boldsymbol{TAG}(c) \mid c \in C\}$

$$\boldsymbol{TAG}(c) = \begin{cases} \boldsymbol{TAG}(c), & \boldsymbol{att}(c) \in V. \\ 1, & \text{それ以外の場合.} \end{cases}$$

この C' を用いて，$\boldsymbol{PROJ}$ は次のように表現される．$\boldsymbol{PROJ} : C \to C'$.

ただし，V の上で t=t' は V により制限され属性 V 上でのタプル t になる属性 U 上でのタプル t' を意味する．

選択 ***(selection)*** もし，$R : U\text{-}Tup \to C$ であり，選択述語 $\boldsymbol{P}$ が各々の U-*tuple* を *0* か *1* に写像すると，$\sigma_{\boldsymbol{P}} R : U\text{-}Tup \to C$ は次のように定義される．

$$(\sigma_{\boldsymbol{P}} R)(t) = R(t) \cdot \boldsymbol{P}(t)$$

$$P(t) = \begin{cases} 1, & t \in \sigma_{\boldsymbol{P}} R. \\ 0, & \text{その以外の場合.} \end{cases}$$

自然結合 ***(natural join)*** もし， $R_i : U_i\text{-}Tup \to C, i = 1,2$ であると，$U_1 \cup U_2$ に対する $R_1 \bowtie R_2$ は次のように定義される．

$$(R_1 \bowtie R_2)(t) = R_1(t_1) \cdot R_2(t_2)$$

t_1は U_1の上で $t_1 = t$ 同様に t_2は U_2の上で $t_2 = t$

セル単位データ来歴の値は可換半環 (commutative semirings) である．定義４により得られるタプルのタグは $\cdot$ 演算の上で可換，結合である．また，集合和，自然結合，選択の場合は [7] の定義と同様なのでそれらの演算による結果は可換半環に従う．定義５の射影による結果も可換半環になる．なぜかというと，可換半環である入力タグの一部を $\cdot$ 演算の単位元である 1 に変更しても可換半環であり，それらの足し算も可換半環であるからである．従って，定義 5 により求められるセル単位データ来歴は可換半環構造である．その結果，我々はタグに具体的な代入することで，信頼度計算等ができるようになる [7].

例２では定義５に従って，図１のデータと問合せからセル単位データ来歴がどう求められるかを示す．例 2 の結果から我々は “問合せ Q と Q’ により求められたデータを使った場合は，c_{11}, c_{13}, c_{21}, c_{23} の作成者に貢献を与えるべきである” ということが分かる．また，セル c_{11}，c_{13} がセル c_{21}, c_{23} より多く使われているので，c_{11}，c_{13} の作成者は c_{21}, c_{23} の作成者より多い貢献を受けると思われる．

例 2. *Q*, *Q’* と図 *1* から結果のレコードに対するセル単位データ来歴を計算する過程

Q’ : SELECT* 都市名, 平均気温 *FROM S WHERE* 都市名＝‘京都’ *AND* 降水量 $>$ *90

1. 問合せを関係代数で表す．

$\boldsymbol{Q}$ *UNION* $\boldsymbol{Q'} = \pi_{\text{都市名, 平均気温}}(\sigma_{\text{都市名=‘京都’}}(S)) \cup \pi_{\text{都市名, 平均気温}}(\sigma_{\text{都市名=‘京都’} \wedge \text{降水量}>90}(S))$

2. タプル毎のタグ

定義 *4* に従って，タプルのタグを U-*Cell* のタグを用いて計算する．

都市名	日付	平均気温 (°C)	降水量の月合計値 (mm)	セル単位データ来歴
京都	2016.11	12.5	82.0	$c_{11} \cdot c_{12} \cdot c_{13} \cdot c_{14}$
京都	2016.12	8.2	92.0	$c_{21} \cdot c_{22} \cdot c_{23} \cdot c_{24}$
大阪	2016.12	9.4	104.0	$c_{31} \cdot c_{32} \cdot c_{33} \cdot c_{34}$

3. 選択を計算する．

元のテーブルから選択演算を計算する．$\sigma_{\text{都市名=‘京都’}}(S)$ の結果，‘京都’の値を持つものは残るが，大阪のデータはなくなる．定義５に従うと選択演算は，$(\sigma_{\boldsymbol{P}} R)(t) = R(t) \cdot \boldsymbol{P}(t)$ なので，Q によるセル単位データ来歴は次のようになる．しかし，セル単位データ来歴が *0* であることはレコードが存在しないことを意味するので，選択された結果残るのは１行目と２行目の ‘京都’ のデータのみである．*Q’* も同様に計算すると，*Q’* の場合は１行目が $c_{11} \cdot c_{12} \cdot c_{13} \cdot c_{14} \cdot 0$ になる．他は Q と同じデータ来歴を持つ．

都市名	日付	平均気温 (°C)	降水量の月合計値 (mm)	セル単位データ来歴
京都	2016.11	12.5	82.0	$c_{11} \cdot c_{12} \cdot c_{13} \cdot c_{14} \cdot 1$
京都	2016.12	8.2	92.0	$c_{21} \cdot c_{22} \cdot c_{23} \cdot c_{24} \cdot 1$
大阪	2016.12	9.4	104.0	$c_{31} \cdot c_{32} \cdot c_{33} \cdot c_{34} \cdot 0$

4. 射影を計算する．

選択された結果から射影演算を計算する．射影演算は $(\pi_V R)(t) = \sum_{V \text{ の上で } t=t' \wedge R(t') \neq 0} (\boldsymbol{PROJ}(R(t'))$ であり，$\boldsymbol{PROJ}$ から日付，降水量の属性のセルのタグを *1* に変更した関係代数式が得られる．従って，セル単位データ来歴は日付，降水量のタグを１に変更したものになる．*Q’* も同様に計算すると，*Q’* の場合は *2* 行だけが結果になる．

都市名	平均気温 (°C)	セル単位データ来歴
京都	12.5	$c_{11} \cdot 1 \cdot c_{13} \cdot 1$
京都	8.2	$c_{21} \cdot 1 \cdot c_{23} \cdot 1$

5. 集合和を計算する．

集合和は同じ値を持つものを $+$演算で結合して表現する．

都市名	平均気温 (°C)	セル単位データ来歴
京都	12.5	$c_{11} \cdot c_{13} + c_{11} \cdot c_{13}$ ($= 2\ c_{11} \cdot c_{13}$)
京都	8.2	$c_{21} \cdot c_{23}$

本研究ではセル単位データ来歴をタプルのタグとも呼ぶ．その理由は，関係代数演算の計算で得られる結果タプルのタグがセル単位データ来歴で表現されるからである．

4. データ作成者への貢献評価

この節ではセル単位データ来歴だけには表現できない貢献を与えるための方法を扱う．セル単位データ来歴は基本的に結果の来歴であるので，結果には現れないが演算などで使われたデータを取れない．しかし，結果には現れていなくても，演算などで使われたデータの作成者は貢献を受けるべきであると考

える．従って，本研究では隠れている作成者にも貢献を与えるために潜在関係を定義し，潜在関係の関係代数演算から計算される潜在データ来歴からの貢献評価方法を提案する．

データ作成者が貢献を受けるためには，次のような過程が要求される．まず，データ使用者は公開されているデータを使い論文等を書く．その次，論文等が評価されて，論文の中で使われたデータの作成者にも論文の評価に従って貢献が与えられる．データ使用者がどのデータを使ったかは分かるために，データ使用者は自分の使ったデータを正しく引用すべきである．本研究では引用方法として [3] の引用方法と同じものを使う．データベースと問合せにより求められた結果を自分が使ったデータとして見なし，引用の対象だと仮定する．例えば，ある人が問合せにより求められた結果データを使うと，結果に含まれているデータの生成に影響を与えているソースデータの作成者に貢献を与える．しかし，セル単位データ来歴だけには正しく貢献を与えることができない．なぜかというと，結果には現れていないが問合せの計算には使われる情報もあるからである．例えば，例 2 では c_{11}, c_{13}, c_{21}, c_{23} の作成者に貢献を与えている．しかし，実際は c_{31} の作成者にも貢献を与えなければならない．なぜなら，c_{31} の作成者が都市名が ‘大阪’ だと正しく値を提供したからである．すなわち，c_{31} の作成者には ‘NOT 京都’ という値を提供したことに対して貢献を与える．

従って，本研究ではセル単位データ来歴をベースにして，貢献評価，また観測データベースの貢献評価までの方法を提案する．その方法としてセル単位データ来歴と貢献評価を含む潜在データ来歴を定義する．潜在はセル毎に付けられる変数で，潜在に値を代入することでセル単位データ来歴になったり，貢献評価になったりする．

図 3 では，セル単位データ来歴，貢献評価，観測データベースの貢献評価で必要なセルをベンダイアグラムで表現している．セル単位データ来歴によると貢献を受ける作成者は c_{11}, c_{13}, c_{21}, c_{23} の作成者である．しかし，c_{31} が ‘NOT 京都’ という情報を提供しているので，貢献評価ではセル単位データ来歴に c_{31} を加えてその作成者に貢献を与える．また，観測データベースの貢献評価ではキーの情報 (都市名と日付がキー) を除いた c_{13}, c_{23} の作成者に貢献を与える．観測データベースでキーの情報を除く理由は，観測データベースが持つ特徴に起因する．観測データベースで重要なものは観測値のみであり，他の属性は観測値を説明するためなものである．すなわち，観測値を説明するためのものはその属性の値が重複していたり，観測値から自然に与えられたりする場合が多い．従って，観測値データベースでは観測値の作成者のみに貢献を与える．

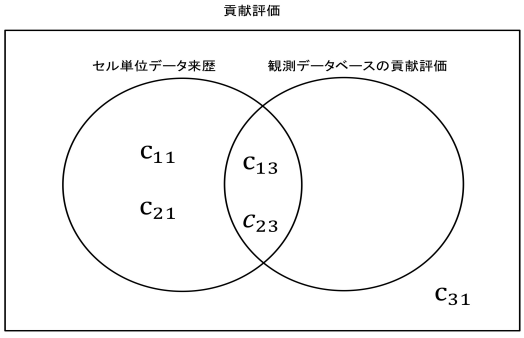

図 3: セル単位データ来歴，貢献評価，観測データベースの貢献評価

本研究での貢献評価の対象は大きく二つに分類できる．一つは結果に現れているものへの貢献評価と，もう一つは結果に現れていないものへの貢献評価である．前者の場合はセル単位データ来歴を用いて計算し，後者の場合は潜在データ来歴を用いて計算する．

4.1 セル単位データ来歴による貢献評価

セル単位データ来歴による貢献評価は基本的には結果に現れているものだけである．しかし，セル単位データ来歴を計算するために使われた問合せを一部変更した問合せを用い，隠れている作成者まで貢献を与えることもできる．セル単位データ来歴を用いた貢献評価は次のようである．

定義 6. セル単位データ来歴からの貢献評価
問合せ Q，関係 $R : U\text{-}Tup \rightarrow C$ があり，$Q(R)$ のレコードの集合を T，属性集合を $\boldsymbol{att}(Q(R))$ とする．
$U\text{-}Cell(=\{(t,A) \mid t \in T\ ,\ A \in \boldsymbol{att}(Q(R))\})$ の要素 c に貢献しているセルを $\boldsymbol{CNTRB}(c)$ とし，$\boldsymbol{CNTRB}(c)$ は次のように計算される．

$$\boldsymbol{CNTRB}(c) = \sigma_{c[t]}(\pi_{c[a]}(Q(R)))$$

ただし，$\sigma_{c[t]}$ は c のタプルを選択し，$\pi_{c[a]}$ は c の属性で射影する．

定義 6 はセル単位データ来歴に選択と射影をすることで，結果に含まれているセル毎の貢献評価を行っている．このようにセル毎に貢献評価することで我々はセル毎に違う貢献評価ができるようになる．例 3 は，セル単位データ来歴を用いた貢献評価を計算する過程である．

例 3. セル単位データ来歴からの貢献評価

都市名	平均気温 (°C)	セル単位データ来歴
京都 ($c_{11} \cdot 1 \cdot 1 \cdot 1$)	12.5($1 \cdot 1 \cdot c_{13} \cdot 1$)	$c_{11} \cdot 1 \cdot c_{13} \cdot 1$
京都 ($c_{21} \cdot 1 \cdot 1 \cdot 1$)	8.2($1 \cdot 1 \cdot c_{23} \cdot 1$)	$c_{21} \cdot 1 \cdot c_{23} \cdot 1$

セル単位データ来歴に，選択と射影演算をすることでセル毎の貢献評価ができる．平均気温 *12.5* のセルを c だとすると，$\boldsymbol{CNTRB}(c) = \sigma_{(\text{京都},\ 12.5)}(\pi_{\text{平均気温}}(Q(S)))$ である．ただし，$Q = \pi_{\text{都市名},\ \text{平均気温}}(\sigma_{\text{都市名}=\text{‘京都’}}(S))$

セル単位データ来歴を使って貢献評価を行うことで，我々はセル毎に貢献している作成者を求められる．その結果，属性毎に違う貢献を与えることができるようになる．例えば，‘都市名’ より ‘平均気温’ の方に貢献を与えるときは，c_{13}, c_{23} の作成者が c_{11}, c_{21} の作成者より高い貢献を受けることになる．

また，セル単位データ来歴を求める過程で使った問合せから射影演算を変更した新しい問合せを作る．その問合せからでき

るセル単位データ来歴の貢献評価は隠れている情報の一部を取ることができるものである．新しい問合せを作る方法は，元の問合せに含まれている射影演算をより大きい属性による射影演算に書き換えることである．また，射影演算を書き換えてできた問合せを Q_{bigP} で表し，次のように定義する．

定義 7. 元の問合せ Q があり，Q のすべての選択で一回でも使われた属性の集合を $\mathbb{S}$，すべての自然結合で一回でも使われた属性の集合を $\mathbb{J}$，すべての射影で一回でも使われた属性の集合を $\mathbb{V}$ だとし，$\boldsymbol{att}(Q) = \mathbb{S} \cup \mathbb{J} \cup \mathbb{V}$ にする．
Q_{bigP} は元の問合せ Q にあるすべての射影演算をそれぞれ $\pi_V(R) \rightarrow \pi_{\boldsymbol{att}(R) \cap \boldsymbol{att}(Q)}(R)$ に書き換えた問合せである．

例 4 は問合せ Q' と ${Q'}_{bigP}$ によるセル単位データ来歴であり，それぞれの違いを表している．

例 4. Q'と ${Q'}_{bigP}$ からの貢献評価
$Q' = \pi_{\text{都市名, 平均気温}}(\sigma_{\text{都市名='京都'} \wedge \text{降水量}>90}(S))$
${Q'}_{bigP} = \pi_{\text{都市名, 平均気温, 降水量}}(\sigma_{\text{都市名='京都'} \wedge \text{降水量}>90}(S))$

都市名	平均気温 (°C)	セル単位データ来歴
京都 (c_{21})	8.2(c_{23})	$c_{21} \cdot c_{23}$

図 4: Q' のセル単位データ来歴

都市名	平均気温 (°C)	降水量の月合計値 (mm)	セル単位データ来歴
京都 (c_{21})	8.2(c_{23})	92.0(c_{24})	$c_{21} \cdot c_{23} \cdot c_{24}$

図 5: Q'$_{bigP}$ のセル単位データ来歴

ある問合せ Q を Q_{bigP} に書き換えることで，結合や，選択で使われたか結果には現れていない情報の一部を取ることができる．しかし，それだけでは図 3 の c_{31} のような情報を取ることはできない．従って，次の節ではすべての隠れている情報を取るために潜在関係を定義し，潜在データ来歴を求める．

4.2 潜在関係と関係代数演算

潜在関係はタプルを構成するすべての値を U-Cell のタグで表現したタプルでできる関係である．潜在関係の上で関係代数演算を計算することで，潜在データ来歴を得ることができ，セル毎に違う貢献を与えることができるようになる．

潜在データ来歴ではセル毎に潜在が付けられている．例えば，例 3 で '都市名' と '平均気温' が同じ貢献を受けるなら，それぞれのセルに付けられている潜在は一緒である．しかし，'都市名' より '平均気温' の方が価値が高い（より貢献を受ける）なら，'平均気温' のセルが '都市名' のセルより高い潜在を持つことになる．また，結果には含まれていないが選択，結合などで使われたものは，結果に現れているセルよりは小さく貢献しているので小さい貢献を持ち，どこでも使われなかったものはそれよりも小さい貢献を持つ．

潜在データ来歴を求めるための関係代数演算は従来の関係代数演算の計算方法とは違う．その違う関係代数演算が成立する関係が潜在関係である．まず，潜在関係を定義するために，U_C-Tup を定義する．属性の有限集合 U とタグ集合 C の有限集合である $\mathbb{C}$ を用いてタプルを関数 $t_c : U \rightarrow \mathbb{C}$ で表現する．その時，可能なすべてのタプルの集合を U_C-Tup と表現する．潜在関係は元のドメインからできるタプルをタグによるタプルに写像したものであり，次のように定義される．

定義 8. 潜在関係は関数 $R_{\mathscr{L}} : U\text{-}Tup \rightarrow U_C\text{-}Tup$ である．ただし，$t_c \in U_C\text{-}Tup$ に対して，$R_{\mathscr{L}}(t) = t_c$ である．

我々は潜在関係上でタグを計算する関係代数演算の定義ができるようになる．例 5 は図 1 のセルのタグ付けと潜在関係を表している．潜在関係の各々のセルの値がタグであることを確認できる．

例 5. セルのタグ付けと潜在関係
1. 定義 *2* によるセルのタグ付け
各々のセル毎に c_{ij} が付けられているとする．ただし，c_{ij} はセルを区別するためのタグでもある．

都市名	日付	平均気温 (°C)	降水量の月合計値 (mm)
京都 (c_{11})	2016.11(c_{12})	12.5(c_{13})	82.0(c_{14})
京都 (c_{21})	2016.12(c_{22})	8.2(c_{23})	92.0(c_{24})
大阪 (c_{31})	2016.12(c_{32})	9.4(c_{33})	104.0(c_{34})

2. 潜在関係
潜在関係は元の関係の値をすべてタグに入れ替えたものであるので，同じ値を持つもの *(*例 *:* c_{11} と c_{21}*)* も違うものとして扱われる．

都市名	日付	平均気温 (°C)	降水量の月合計値 (mm)
c_{11}	c_{12}	c_{13}	c_{14}
c_{21}	c_{22}	c_{23}	c_{24}
c_{31}	c_{32}	c_{33}	c_{34}

潜在関係は関係代数演算により変化していく．そのために，一般的な関係代数演算とは多少違う演算方法を使う．潜在関係での集合和は重複を許している．重複を許すことで同じ値を持つものも違うものとして扱うことができるようになり，セル毎の潜在が計算ができる．また，潜在関係での自然結合は直積で計算する．その結果潜在関係の上での自然結合でできる関係は同じ名前の属性を持つことになり，それぞれを区別して扱える．

セルの貢献評価はセルのタグに付けられる潜在により行われる．従って，本研究では潜在の値を先に計算し，最後にその潜在の値をタグに付けることにする．そのためには，元の潜在関係のタプルを潜在とタグの対で表現する必要がある．

定義 9. 潜在とタグの対によるタプル
潜在関係上でのあるタプル $t_c(=(c_1, \ldots, c_n))$ は以下のようにも表現される．

$$t_c = ((\boldsymbol{L}_1, \ldots, \boldsymbol{L}_n) \circ_{\cdot_c} (c_1, \ldots, c_n))$$

また，潜在とタグの対によるタプルを明示的に表す時は，タプルを $(\boldsymbol{L} \circ_{\cdot_c} c)$ で表現し，$\boldsymbol{L}_i$ をセル c_i の潜在とも呼ぶ．

定義9のようにタプルを表現できる理由は，セルのタグがCの要素であるからである．Cの要素は半環構造に従うので，$\cdot$演算が定義されている．また，例5の潜在関係はすべての潜在が1である特殊な場合である．

定義 10. 潜在関係での関係代数演算

潜在関係での自然結合は直積と同じ役割をし，$\bowtie_{\mathscr{L}}$で表す．

集合和 ***(union)*** 潜在関係上での集合和は重複を許す多重集合での演算である．潜在関係上での関係RとSがあり，$R \cup S$は，次のように定義される関係である．

$$R \cup S = R \uplus S$$

ただし，$\uplus$は多重集合での足し算を意味する．例えば，$\{1\} \uplus \{1\} = \{1,1\}$

自然結合 ***(natural join)*** 潜在関係上での関係RとSがあり，$R \bowtie_{\mathscr{L}} S$は，次のように定義される関係である．

$$R \bowtie_{\mathscr{L}} S = R \times S$$

$$\begin{aligned} R \times S = \{\ & \mathscr{L}\circ_{\cdot_{\mathscr{L}}}(\boldsymbol{L}_{r_1}, \boldsymbol{L}_{r_2}, \ldots, \boldsymbol{L}_{r_n}, \boldsymbol{L}_{s_1}, \boldsymbol{L}_{s_2}, \ldots, \boldsymbol{L}_{s_m})\circ_{\cdot_c} \\ & (r_1, r_2, \ldots, r_n, s_1, s_2, \ldots, s_m)(=w) \mid \\ & ((\boldsymbol{L}_{r_1}, \boldsymbol{L}_{r_2}, \ldots, \boldsymbol{L}_{r_n})\circ_{\cdot_c}(r_1, r_2, \ldots, r_n)(=r)) \in R \wedge \\ & ((\boldsymbol{L}_{s_1}, \boldsymbol{L}_{s_2}, \ldots, \boldsymbol{L}_{s_m})\circ_{\cdot_c}(s_1, s_2, \ldots, s_m)(=s)) \in S\ \} \end{aligned}$$

ただし，$\circ_{\cdot_{\mathscr{L}}}$は潜在同士の演算，$\mathscr{L}=(\mathscr{L}_1, \mathscr{L}_2, \ldots, \mathscr{L}_{n+m})$，$\boldsymbol{ATT} = \boldsymbol{att}(R) \cup \boldsymbol{att}(S)$である．

$$\mathscr{L}_i = \begin{cases} \mathscr{L}_{JP}, & t_c^{-1}(w_i) \in \boldsymbol{ATT}\ AND\ R_{\mathscr{L}}^{-1}(r) \bowtie R_{\mathscr{L}}^{-1}(s) \in R_{\mathscr{L}}^{-1}(R) \bowtie R_{\mathscr{L}}^{-1}(S). \\ \mathscr{L}_{JP\text{-}}, & t_c^{-1}(w_i) \notin \boldsymbol{ATT}\ AND\ R_{\mathscr{L}}^{-1}(r) \bowtie R_{\mathscr{L}}^{-1}(s) \in R_{\mathscr{L}}^{-1}(R) \bowtie R_{\mathscr{L}}^{-1}(S). \\ \mathscr{L}_{JN}, & t_c^{-1}(w_i) \in \boldsymbol{ATT}\ AND\ R_{\mathscr{L}}^{-1}(r) \bowtie R_{\mathscr{L}}^{-1}(s) \notin R_{\mathscr{L}}^{-1}(R) \bowtie R_{\mathscr{L}}^{-1}(S). \\ \mathscr{L}_{JN\text{-}}, & t_c^{-1}(w_i) \notin \boldsymbol{ATT}\ AND\ R_{\mathscr{L}}^{-1}(r) \bowtie R_{\mathscr{L}}^{-1}(s) \notin R_{\mathscr{L}}^{-1}(R) \bowtie R_{\mathscr{L}}^{-1}(S). \end{cases}$$

また，$t_c^{-1} : \mathbb{C} \to U$で，$\mathbb{C}$に含まれるあるセルは属性が優一なので，そのセルからのt_c^{-1}はいつも一意に決まる．

射影 ***(projection)*** 潜在関係上での関係Rがあり，$\pi_V(R)$は，次のように定義される関係である．

$$\begin{aligned} \pi_V(R) = \{ & \mathscr{L}\circ_{\cdot_{\mathscr{L}}}(\boldsymbol{L}_{r_1}, \boldsymbol{L}_{r_2}, \ldots, \boldsymbol{L}_{r_n})\circ_{\cdot_c}(r_1, r_2, \ldots, r_n) \mid \\ & ((\boldsymbol{L}_{r_1}, \boldsymbol{L}_{r_2}, \ldots, \boldsymbol{L}_{r_n})\circ_{\cdot_c}(r_1, r_2, \ldots, r_n)(=r)) \in R\ \} \end{aligned}$$

ただし，$\mathscr{L} = (\mathscr{L}_1, \mathscr{L}_2, \ldots, \mathscr{L}_n)$である．

$$\mathscr{L}_i = \begin{cases} \mathscr{L}_{PP}, & t_c^{-1}(r_i) \in \boldsymbol{att}(R). \\ \mathscr{L}_{PN}, & t_c^{-1}(r_i) \notin \boldsymbol{att}(R). \end{cases}$$

選択 ***(selection)*** 潜在関係上での関係Rがあり，$\sigma_P(R)$は，次のように定義される関係である．

$$\begin{aligned} \sigma_P(R) = \{ & \mathscr{L}\circ_{\cdot_{\mathscr{L}}}(\boldsymbol{L}_{r_1}, \boldsymbol{L}_{r_2}, \ldots, \boldsymbol{L}_{r_n})\circ_{\cdot_c}(r_1, r_2, \ldots, r_n)(=r) \mid \\ & ((\boldsymbol{L}_{r_1}, \boldsymbol{L}_{r_2}, \ldots, \boldsymbol{L}_{r_n})\circ_{\cdot_c}(r_1, r_2, \ldots, r_n)(=r)) \in R\ \} \end{aligned}$$

ただし，$\mathscr{L} = (\mathscr{L}_1, \mathscr{L}_2, \ldots, \mathscr{L}_n)$であり，選択条件$P$で使われた属性の集合を$\boldsymbol{att}(P)$とする．

$$\mathscr{L}_i = \begin{cases} \mathscr{L}_{SP}, & t_c^{-1}(r_i) \in \boldsymbol{att}(P)\ AND\ P(r) = TRUE. \\ \mathscr{L}_{SP\text{-}}, & t_c^{-1}(r_i) \notin \boldsymbol{att}(P)\ AND\ P(r) = TRUE. \\ \mathscr{L}_{SN}, & t_c^{-1}(r_i) \in \boldsymbol{att}(P)\ AND\ P(r) = FALSE. \\ \mathscr{L}_{SN\text{-}}, & t_c^{-1}(r_i) \notin \boldsymbol{att}(P)\ AND\ P(r) = FALSE. \end{cases}$$

定義10に従うと，自然結合，射影，選択の場合は計算の段階でそれぞれの$\mathscr{L}$が与えられ，$\mathscr{L}$同士で計算される．また，この$\cdot_{\mathscr{L}}$に適切な演算子とそれぞれの$\mathscr{L}$に値を代入することで，貢献評価ができるようになる．例6は潜在関係による関係代数演算である．

例 6. 潜在関係による関係代数演算

$\boldsymbol{Q"} = \pi_{AC}(\sigma_{A=a}(R) \bowtie_{\mathscr{L}} S)$

R

A	B
a	b
d	e

S

B	C
b	c
f	g

1. U-Cell のすべての要素に対して，***TAG*** を適用することですべての要素のタグを求められる．その結果セル毎にタグが付けられる．

R

A(タグ)	B(タグ)
a(r_{11})	b(r_{12})
d(r_{21})	e(r_{22})

S

B(タグ)	C(タグ)
b(s_{11})	c(s_{12})
f(s_{21})	g(s_{22})

2. 潜在関係に従う関係を作る．その結果セルそれぞれを違うタグで区別することができるようになる．

R

A	B
$1 \cdot_c r_{11}$	$1 \cdot_c r_{12}$
$1 \cdot_c r_{21}$	$1 \cdot_c r_{22}$

S

B	C
$1 \cdot_c s_{11}$	$1 \cdot_c s_{12}$
$1 \cdot_c s_{21}$	$1 \cdot_c s_{22}$

3. $\sigma_{A=a}(R)$ を計算する．

R

A	B
$\mathscr{L}_{SP\cdot c} r_{11}$	$\mathscr{L}_{SP\text{-}\cdot c} r_{12}$
$\mathscr{L}_{SN\cdot c} r_{21}$	$\mathscr{L}_{SN\text{-}\cdot c} r_{22}$

4. $\sigma_{A=a}(R) \bowtie_{\mathscr{L}} S$ を計算する．

A	B	B	C
$\mathscr{L}_{JP\text{-}\cdot\mathscr{L}}\mathscr{L}_{SP\cdot c} r_{11}$	$\mathscr{L}_{JP\cdot\mathscr{L}}\mathscr{L}_{SP\text{-}\cdot c} r_{12}$	$\mathscr{L}_{JP\cdot c} s_{11}$	$\mathscr{L}_{JP\text{-}\cdot c} s_{12}$
$\mathscr{L}_{JN\text{-}\cdot\mathscr{L}}\mathscr{L}_{SP\cdot c} r_{11}$	$\mathscr{L}_{JN\cdot\mathscr{L}}\mathscr{L}_{SP\text{-}\cdot c} r_{12}$	$\mathscr{L}_{JN\cdot c} s_{21}$	$\mathscr{L}_{JN\text{-}\cdot c} s_{22}$
$\mathscr{L}_{JN\text{-}\cdot\mathscr{L}}\mathscr{L}_{SN\cdot c} r_{21}$	$\mathscr{L}_{JN\cdot\mathscr{L}}\mathscr{L}_{SN\text{-}\cdot c} r_{22}$	$\mathscr{L}_{JN\cdot c} s_{11}$	$\mathscr{L}_{JN\text{-}\cdot c} s_{12}$
$\mathscr{L}_{JN\text{-}\cdot\mathscr{L}}\mathscr{L}_{SN\cdot c} r_{21}$	$\mathscr{L}_{JN\cdot\mathscr{L}}\mathscr{L}_{SN\text{-}\cdot c} r_{22}$	$\mathscr{L}_{JN\cdot c} s_{21}$	$\mathscr{L}_{JN\text{-}\cdot c} s_{22}$

5. $\pi_{AC}(\sigma_{A=a}(R) \bowtie_{\mathscr{L}} S)$ を計算する．

A	B	B	C
$\mathscr{L}_{PP\cdot\mathscr{L}}\mathscr{L}_{JP\text{-}\cdot\mathscr{L}}$ $\mathscr{L}_{SP\cdot c} r_{11}$	$\mathscr{L}_{PN\cdot\mathscr{L}}\mathscr{L}_{JP\cdot\mathscr{L}}$ $\mathscr{L}_{SP\text{-}\cdot c} r_{12}$	$\mathscr{L}_{PN\cdot\mathscr{L}}$ $\mathscr{L}_{JP\cdot c} s_{11}$	$\mathscr{L}_{PP\cdot\mathscr{L}}$ $\mathscr{L}_{JP\text{-}\cdot c} s_{12}$
$\mathscr{L}_{PP\cdot\mathscr{L}}\mathscr{L}_{JN\text{-}\cdot\mathscr{L}}$ $\mathscr{L}_{SP\cdot c} r_{11}$	$\mathscr{L}_{PN\cdot\mathscr{L}}\mathscr{L}_{JN\cdot\mathscr{L}}$ $\mathscr{L}_{SP\text{-}\cdot c} r_{12}$	$\mathscr{L}_{PN\cdot\mathscr{L}}$ $\mathscr{L}_{JN\cdot c} s_{21}$	$\mathscr{L}_{PP\cdot\mathscr{L}}$ $\mathscr{L}_{JN\text{-}\cdot c} s_{22}$
$\mathscr{L}_{PP\cdot\mathscr{L}}\mathscr{L}_{JN\text{-}\cdot\mathscr{L}}$ $\mathscr{L}_{SN\cdot c} r_{21}$	$\mathscr{L}_{PN\cdot\mathscr{L}}\mathscr{L}_{JN\cdot\mathscr{L}}$ $\mathscr{L}_{SN\text{-}\cdot c} r_{22}$	$\mathscr{L}_{PN\cdot\mathscr{L}}$ $\mathscr{L}_{JN\cdot c} s_{11}$	$\mathscr{L}_{PP\cdot\mathscr{L}}$ $\mathscr{L}_{JN\text{-}\cdot c} s_{12}$
$\mathscr{L}_{PP\cdot\mathscr{L}}\mathscr{L}_{JN\text{-}\cdot\mathscr{L}}$ $\mathscr{L}_{SN\cdot c} r_{21}$	$\mathscr{L}_{PN\cdot\mathscr{L}}\mathscr{L}_{JN\cdot\mathscr{L}}$ $\mathscr{L}_{SN\text{-}\cdot c} r_{22}$	$\mathscr{L}_{PN\cdot\mathscr{L}}$ $\mathscr{L}_{JN\cdot c} s_{21}$	$\mathscr{L}_{PP\cdot\mathscr{L}}$ $\mathscr{L}_{JN\text{-}\cdot c} s_{22}$

潜在関係上での関係代数演算の結果，セル毎にあるタグに潜在が付けられる． 例 6 の最終結果に対して，$\cdot_{\mathscr{L}}$ に掛け算を意味する $\cdot$，$\mathscr{L}_{*P}, \mathscr{L}_{*P\text{-}}$ に 1，$\mathscr{L}_{*N}, \mathscr{L}_{*N\text{-}}$ に 0 を代入し，潜在が 1 であるものは残して，0 であるものを消すと，最終に残るのは r_{11} と s_{12} だけである．これは潜在関係ではない関係での $\pi_{AC}(\sigma_{A=a}(R) \ltimes S)$ の結果と一致する．

4.3 潜在データ来歴

潜在データ来歴というのは例 6 の最終結果のようなものである．すなわち，潜在関係上での関係代数演算の結果を潜在データ来歴と呼ぶ．例 6 の最終結果は Q" の潜在データ来歴である．この潜在データ来歴の潜在をどう計算するかによって，潜在データ来歴からセル毎の貢献評価になったり，貢献評価になったり，観測データベースの貢献評価になったりする．

定理 1. 潜在データ来歴と潜在

***1.* セル毎の貢献評価**

潜在データ来歴からセル毎の貢献評価までの計算は次のように行われる．

1) $\cdot_{\mathscr{L}}$ に $\cdot$，$\mathscr{L}_{*P}, \mathscr{L}_{*P\text{-}}$ に *1*，$\mathscr{L}_{*N}, \mathscr{L}_{*N\text{-}}$ に *0* を代入し，*0* の値を持つものは消す．

2) 同じ属性名の要素をお互いに $\cdot$ で結合する．

3) $R^{-1}_{\mathscr{L}}$ により同じ値を持つタプルをお互いに $+$で結合する．

***2.* 貢献評価**

潜在データ来歴から貢献評価までの計算は次のように行われる．

1) $\cdot_{\mathscr{L}}$ に $+$，それぞれの潜在に適当な値を代入する．ただし，潜在はパラメーターである．

2) 属性毎の重み *($\mathscr{L}_w$)* をセルの $\boldsymbol{L}$ に掛ける．

***3.* 観測データベースの貢献評価**

潜在データ来歴から観測データベースの貢献評価までの計算は次のように行われる．

1) $\cdot_{\mathscr{L}}$ に $+$，それぞれの潜在に適当な値を代入する．ただし，潜在はパラメーターである．

2) キー属性のセルの $\boldsymbol{L}$ には *($\mathscr{L}_k$=0)* をキーではに属性のセルには *($\mathscr{L}_{nk}$=1)* を掛ける．

3) 属性毎の重み *($\mathscr{L}_w$)* をセルの $\boldsymbol{L}$ に掛ける．

貢献評価で結果には現れないが貢献を受けなければならない対象は，選択と結合で使われたセルである．その時，潜在の値では $\mathscr{L}_{JP} > \mathscr{L}_{JN} > \mathscr{L}_{JP\text{-}}$，$\mathscr{L}_{SP} > \mathscr{L}_{SN} > \mathscr{L}_{SP\text{-}}$，$\mathscr{L}_{PP} > \mathscr{L}_{PN}$ が成り立つ．この潜在値に従って Q" の潜在データ来歴に $\mathscr{L}_{JP}, \mathscr{L}_{SP}$ に 2 を，$\mathscr{L}_{PP}, \mathscr{L}_{JN}, \mathscr{L}_{SN}$ に 1 を，$\mathscr{L}_{JP\text{-}}, \mathscr{L}_{SP\text{-}}, \mathscr{L}_{PN}$ に 0 を代入すると，図 6 のようになる．図 6 の結果をそれぞれのタグ毎に潜在を足し算すると，$6r_{11}$,$4r_{21}$,$4s_{12}$,$4s_{22}$,$3r_{12}$,$3s_{11}$,$2s_{21}$,$2r_{22}$ になる．この潜在値の順序は，現実の貢献とは異なるが，潜在の値ををよく調整することで現実的な結果が得られると期待する．

A	B	B	C
$3 \cdot r_{11}$	$2 \cdot r_{12}$	$2 \cdot s_{11}$	$2 \cdot s_{12}$
$3 \cdot r_{11}$	$1 \cdot r_{12}$	$1 \cdot s_{21}$	$2 \cdot s_{22}$
$2 \cdot r_{21}$	$1 \cdot r_{22}$	$1 \cdot s_{11}$	$2 \cdot s_{12}$
$2 \cdot r_{21}$	$1 \cdot r_{22}$	$1 \cdot s_{21}$	$2 \cdot s_{22}$

図 6: 潜在による貢献評価

5. 結　　論

本研究ではデータ作成者への貢献を与えるための方法としてセル単位データ来歴と潜在データ来歴を提案した．潜在データ来歴はセル単位データ来歴による貢献評価の一般化で，結果に現れていなくても貢献している対象の貢献まで計算できる．データ作成者の貢献を正しく評価することで，データ作成者の功績を認めることができる．

本研究では結果に現れているセルの貢献を計算するためにセル単位データ来歴を用いた．しかし，セル単位データ来歴は信頼度の計算にも使われる．信頼度は結果の信頼度とセルの信頼度に分かれる．結果の信頼度はタプル単位のタグを持つセル単位データ来歴を用いて計算され，セルの信頼度はセル単位のタグを持つセル単位データ来歴による貢献評価を用いて計算される．この時，我々は潜在データ来歴だけで両方を計算できる．そのためには，セル単位データ来歴による貢献評価からセル単位データ来歴を復元する必要がある．その理由は，潜在データ来歴から計算できるのがセル単位データ来歴による貢献評価だけであるからだ．セル単位データ来歴による貢献評価からセル単位データ来歴への復元ができるなら，潜在データ来歴だけで貢献度と（結果，セル) 信頼度を計算できる．その復元に必要な構造を探すことは今後の課題である．

文　　献

[1] Communique of the g7 science and technology ministers' meeting in tsukuba, ibaraki (http://www8.cao.go.jp/cstp /english/others/20160517communique.pdf), May 2016.

[2] Serge Abiteboul, Richard Hull, and Victor Vianu, editors. *Foundations of Databases: The Logical Level*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1995.

[3] Peter Buneman, Susan Davidson, and James Frew. Why data citation is a computational problem. *Commun. ACM*, 59(9):50–57, August 2016.

[4] Peter Buneman, Sanjeev Khanna, and Wang Chiew Tan. Why and where: A characterization of data provenance. In *Proceedings of the 8th International Conference on Database Theory*, ICDT '01, pages 316–330, London, UK, UK, 2001. Springer-Verlag.

[5] James Cheney, Laura Chiticariu, and Wang-Chiew Tan. Provenance in databases: Why, how, and where. *Found. Trends databases*, 1(4):379–474, April 2009.

[6] Yingwei Cui, Jennifer Widom, and Janet L. Wiener. Tracing the lineage of view data in a warehousing environment. *ACM Trans. Database Syst.*, 25(2):179–227, June 2000.

[7] 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*, PODS '07, pages 31–40, New York, NY, USA, 2007. ACM.

[8] Data Citation Synthesis Group. *Joint Declaration of Data Citation Principles*. Martone M. (ed.) San Diego CA: FORCE11, 2014.

[9] John P. Holdren. *Increasing Access to the Results of Federally Funded Scientific Research*. Office of Science and Technology Policy, February 2013.