関係データ交換フレームワークにおける
弱適切なデータ公開ポリシの存在を判定するアルゴリズム

M2020SC001 相崎聖也

指導教員：石原靖哲

1　はじめに

データ交換とは，ある構造をもつデータを異なる構造のデータに変換する問題である．様々な形式のウェブデータの普及や e-ビジネスアプリケーションの出現によって，異なる構造を持つデータを交換する必要性が高まっている．一方で，データを共有する際には，プライバシやセキュリティにも考慮する必要がある．

通常のデータ交換フレームワーク [1] では，参加者はソース側とターゲット側の 2 種類である．しかしながら，実際の状況では，それらが交換したデータを使用する他の参加者が存在する．本要旨では，以上の 3 種類の参加者がデータのやり取りを行う状況について議論を行う．

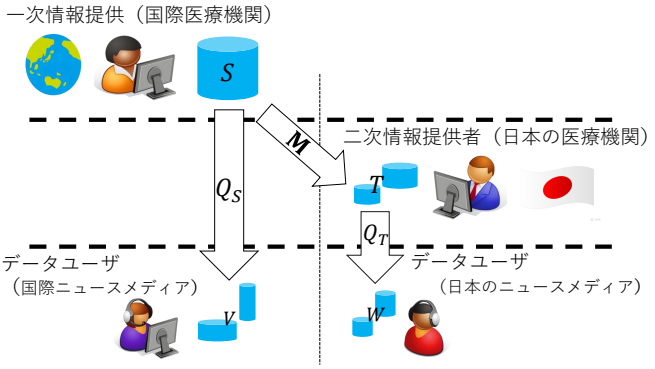

図 1　データ交換フレームワーク

2　問題設定と方針

[2] では，一次情報提供者がソースデータを，二次情報提供者がターゲットデータを所持すると仮定している．また，各情報提供者は所持するデータの一部をデータユーザに公開する（図 1）．一般的に，一次情報提供者はデータユーザに対するデータ公開ポリシ Q_S を持ち，二次情報提供者もまた，ポリシ Q_T を持つ．[2] では，Q_T が Q_S に従った**適切なポリシ**であるための条件として，次の二つを定義している．

1. 秘匿性：Q_T は Q_S が公開する情報のみ公開する．
2. 可用性：Q_T は一つ目の条件を満たす範囲の中で，極大の情報を公開する．

例 1　世界中の傷病者に関するデータを収集する医療機関（以降，WMI と略記する）と，WMI から提供された日本の傷病者に関するデータを管理する医療機関（以降，JMI と略記する）が存在する状況を考える．WMI はその一部のデータを国際的なニュースメディアに公開し，JMI はその一部のデータを日本のニュースメディアに公開する．この例において，WMI は一次情報提供者，JMI は二次情報提供者，ニュースメディアはデータユーザに当たる．WMI は関係スキーマ $S(\mathrm{ID}, \mathrm{Age}, \mathrm{Gender}, \mathrm{Nationality})$ 上のデータを保持しているとする．これは，S という名前のテーブルに，ID，年齢，性別，国籍のデータを保持していることを表す．このとき，WMI と JMI の間のデータ交換メカニズムは次の連言問合せで表現される．

$$\mathbf{M}: \quad T(ID, Age, Gen) \text{ :- } S(ID, Age, Gen, \text{“JPN”}).$$

これは，テーブル S から国籍が JPN である任意のデータを取り出し，テーブル T に ID，年齢，性別のデータを格納することを表す．WMI がニュースメディアに対して ID と国籍の情報を秘匿したいと考え，Q_S として次のような問合せを与えたとする．

$$Q_S: \quad V(Age, Gen) \text{ :- } S(ID, Age, Gen, Nat).$$

ここで，JMI がデータ公開ポリシとして，次のような連言問合せ Q_T を採用した場合を考える．

$$Q_T: \quad W(Age, Gen) \text{ :- } T(ID, Age, Gen).$$

一見すると，Q_S が公開するデータ，つまり，年齢と性別のみを Q_T は公開しているため，Q_T は「適切」なポリシであるように思われる．しかし，日本のニュースメディアは日本の傷病者に関するデータのみを得ているため，国際的なニュースメディアが得ることのできない，傷病者の国籍の情報を得ていることになる．つまり，Q_S が公開していない情報を Q_T は公開しているため，Q_T は秘匿性を満たしていない．この例において，秘匿性を満たす Q_T は何も情報を公開しないポリシのみである．■

例 1 で言及したような何も情報を公開しない Q_T は，すべての Q_S と $\mathbf{M}$ に対して，秘匿性を満たすことが容易に示せる．しかし，そのような Q_T が可用性を満たすとは限らない．

[2] では，与えられた Q_S と $\mathbf{M}$ に対して，適切な Q_T を見つけるためのヒューリスティックアルゴリズムが提案されている．ヒューリスティックアルゴリズムから出力されるポリシは秘匿性を満たすことが示されているが，可用性を満たすかどうかは未解決であった．

本要旨では，まず，[2] のヒューリスティックアルゴリズムにより導出されるポリシが一般には可用性を満たさないことを示す．次に，適切なポリシの存在を判定する問題を解決する前段階として，適切なポリシであるための条件

を次のように弱め，それらを満たすポリシ（**弱適切なポリシ**と呼ぶ）が存在するかどうかを判定する問題に注目する．

1. Q_T は秘匿性を満たす．
2. Q_T は何らかの情報を公開する．

ヒューリスティックアルゴリズムが導出するターゲットポリシが弱適切であるための二つ目の条件を満たすかどうかをチェックすることにより，弱適切なポリシが存在するための十分条件を判定できる．本要旨では，その判定が必要条件の判定にもなるための Q_S や $\mathbf{M}$ に関する制約を与える．さらに，本要旨では，ヒューリスティックアルゴリズムとは別の，弱適切なポリシが存在するための十分条件を判定するアルゴリズムを提案し，その判定が必要条件の判定にもなるための制約を与える．

3 諸定義

3.1 関係データベース

関係スキーマ $R[n]$ は関係データの構造を表すものであり，形式的には関係名 R とアリティと呼ばれる非負整数 n から成る．関係スキーマ上の実データは**タプル** t の集合で与えられる．$R[n]$ 上のタプルは n 個のエントリをもつ．各エントリは**ドメイン**と呼ばれる，定数の可算集合 Dom の元をとる．**関係インスタンス** I はタプルの有限集合である．

データベーススキーマ $\mathbf{R}$ は互いに異なる関係名を持つ関係スキーマの有限なリスト $\mathbf{R} = \langle R_1[n_1], \ldots, R_k[n_k] \rangle$ であり，関係データベースの構造を表す．**データベースインスタンス** $\mathbf{I}$ はリスト $\langle I_1, \ldots, I_k \rangle$ である．ただし，各 I_i は $R_i[n_i]$ の関係インスタンスである．$\mathbf{I}$ の i 番目の関係インスタンスがタプル $(a_i, \ldots, a_{n_i})$ を含むとき，$R_i(a_i, \ldots, a_{n_i}) \in \mathbf{I}$ と記す．

$\mathbf{R} = \langle R_1[n_1], \ldots, R_k[n_k] \rangle$ をデータベーススキーマとする．変数の可算集合 Var を固定する．$\mathbf{R}$ における**原子式**は $R_i(x_1, \ldots, x_{n_i})$ と表される．ただし，各 x_i は Dom∪Var に属する．$\mathbf{R}$ における**連言式**は原子式のリストであり，空の場合 $\top$ と記す．$\mathbf{R}$ における**連言問合せ**（以降，CQ と略記する）Q は次の形式で与えられる規則である:

$$Q: \quad V(y_1, \ldots, y_m) \text{ :- } R_1, \ldots, R_\ell.$$

ただし，V はアリティ m の，$\mathbf{R}$ に属さない関係名であり，各 y_i は Dom∪Var に属する．$R_1, \ldots, R_\ell$ は $\mathbf{R}$ における連言式である．Q の体部に現れる変数の集合を $\mathbf{const}(Q)$ と $V(y_1, \ldots, y_m)$ を Q の**頭部**，$R_1, \ldots, R_\ell$ を Q の**体部**と呼ぶ．全ての $y_1, \ldots, y_m$ が Var に属し，かつ体部に現れるとき，Q は**安全**であるという．Q の体部の原子式の数が一つのとき，Q は *join-free* であるという．Q の体部に同じ関係名が 2 回以上現れないとき，Q は *self-join-free* であるという．Q の体部にある全ての変数が頭部にも現れるとき，Q は *projection-free* であるという．

$\mu : \text{Var} \to \text{Dom}$ を変数割当てとする．$\mathbf{R}$ 上の $\mathbf{I}$ における Q の**答え** $Q(\mathbf{I})$ は次のように表される．

$$Q(\mathbf{I}) = \{(\mu(y_1), \ldots, \mu(y_m)) \mid \text{各 } R_i(x_1, \ldots, x_{n_i}) \text{ に対し，} R_i(\mu(x_1), \ldots, \mu(x_{n_i})) \in \mathbf{I} \text{ である.}\}$$

Q が安全であるならば $Q(\mathbf{I})$ は有限であり，したがって $Q(\mathbf{I})$ は $V[m]$ 上の関係インスタンスである．

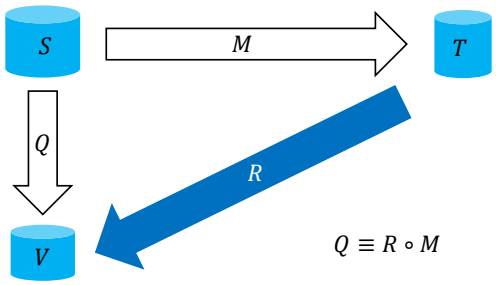

図 2 $\mathbf{M}$ を使った Q の CQ-rewriting R

3.2 問合せの等価性と書き換え

$\mathcal{I}(\mathbf{R})$ は $\mathbf{R}$ 上の全てのデータベースインスタンスの集合，Q_1 と Q_2 は $\mathbf{R}$ 上の CQ とする．任意の $\mathbf{I} \in \mathcal{I}(\mathbf{R})$ に対し $Q_1(\mathbf{I}) \subseteq Q_2(\mathbf{I})$ であるとき，Q_1 は Q_2 に含まれるといい，$Q_1 \subseteq Q_2$ と記す．$Q_1 \subseteq Q_2$ であることと Q_2 から Q_1 への準同型写像が存在することは同値である [3]．$Q_1 \subseteq Q_2$ かつ $Q_2 \subseteq Q_1$ であるとき，Q_1 と Q_2 は**等価**であるといい，$Q_1 \equiv Q_2$ と記す．

information-collapsing な CQ とは，$Q_\perp : W() \text{ :- } \top$ の形式で与えられる CQ である．これは，何も情報を公開しないポリシに相当する．安全な CQ Q について，$Q \not\equiv Q_\perp$ であるとき，Q は *information-revealing* であるという．これは，何らかの情報を公開するポリシに相当する．

Q を CQ，$\mathbf{M}$ を CQ のリストとする．$\mathbf{M}$ を使った Q の *CQ-rewriting* は次のような CQ R である（図 2）．

- R の体部は $\mathbf{M}$ の頭部に存在する関係名のみを含む．
- $R \circ \mathbf{M} \equiv Q$ が成り立つ．ただし，$\circ$ は問合せの合成を表す．

例 2 次の CQ Q と CQ のリスト $\mathbf{M}$ が与えられたとする．以降では，特に断らない限り，変数を大文字で，定数を小文字で表す．

$$\begin{aligned} Q: &\quad V(A, D) \text{ :- } S_1(A, b), S_2(b, C, D). \\ \mathbf{M}: &\quad T_1(X, Y) \text{ :- } S_1(X, Y), \\ &\quad T_2(Z, W) \text{ :- } S_2(Z, V, W). \end{aligned}$$

$\mathbf{M}$ の頭部の関係名が体部に現れる，安全な CQ R を次のように定める．

$$R: \quad N(X, W) \text{ :- } T_1(X, b), T_2(b, W).$$

R を $\mathbf{M}$ で展開すると，次の CQ $R \circ \mathbf{M}$ を得る．

$$R \circ \mathbf{M}: \quad N(X, W) \text{ :- } S_1(X, b), S_2(b, V, W).$$

このとき，Q から $R \circ \mathbf{M}$ への準同型写像が存在し，かつ $R \circ \mathbf{M}$ から Q への準同型写像が存在するため，$Q \equiv R \circ \mathbf{M}$ が成立する．よって，R は $\mathbf{M}$ を使った Q の CQ-rewriting である． ■

CQ-rewriting の存在判定およびその導出をするためのアルゴリズム [4] が知られている．

3.3 適切なターゲットポリシ

図 1 のような状況で CQ Q_S と CQ のリスト $\mathbf{M}$ が与えられているとする．ターゲットポリシ Q_T が以下で定義される秘匿性と可用性を満たすとき，Q_T は Q_S と $\mathbf{M}$ に関して適切なポリシであるという．

- 秘匿性：Q_S を使った $Q_T \circ \mathbf{M}$ の CQ-rewriting R が存在する．
- 可用性：秘匿性を満たすすべての Q'_T について，Q'_T を使った Q_T の CQ-rewriting が存在するならば，Q_T を使った Q'_T の CQ-rewriting が存在する．

また，Q_T が秘匿性を満たし，かつ information-revealing であるとき，Q_T は Q_S と $\mathbf{M}$ に関して弱適切なポリシであるという．

4 ヒューリスティックアルゴリズムが正しく動作しない例

本節では，[2] で提案されたヒューリスティックアルゴリズムが，与えられた Q_S と $\mathbf{M}$ に関して適切でないターゲットポリシ Q_T を導出する例を示す．

ヒューリスティックアルゴリズムは以下のように動作する．与えられた Q_S と $\mathbf{M}$ に対し，まず，Q_S の頭部にあるいくつかの変数に Q_S または $\mathbf{M}$ に現れる定数を代入した原子式が体部に 1 つだけ現れる，安全かつ projection-free な CQ の集合を $\mathcal{N}$ とする．次に，$\mathbf{M}$ と $N \circ Q_S (N \in \mathcal{N})$ に対して，[4] のアルゴリズムから求まる有限個のポリシ Q_T を計算し，これが弱適切であるかどうかを判定する．最後に，弱適切な Q_T の中で極大の情報を公開するポリシを出力する．

例 3 次のような Q_S と $\mathbf{M}$ が与えられたとする．

$$Q_S: \quad V_S(A,B) \text{ :- } S(A,B),$$
$$\mathbf{M}: \quad T(X) \text{ :- } S(X,y), S(X,z).$$

このとき，$\mathcal{N}$ の元は同型なものを除いて次の 9 つである．

$$N_1 : V_T(A,B) \text{ :- } V_S(A,B), \quad N_2 : V_T(A) \text{ :- } V_S(A,y),$$
$$N_3 : V_T(A) \text{ :- } V_S(A,z), \quad N_4 : V_T(B) \text{ :- } V_S(y,B),$$
$$N_5 : V_T(B) \text{ :- } V_S(z,B), \quad N_6 : V_T() \text{ :- } V_S(y,y),$$
$$N_7 : V_T() \text{ :- } V_S(y,z), \quad N_8 : V_T() \text{ :- } V_S(z,y),$$
$$N_9 : V_T() \text{ :- } V_S(z,z).$$

よって，ヒューリスティックアルゴリズムによって計算される Q_T は information-collapsing な CQ のみである．ここで，次のような，Q_S を使った CQ N' を考える．

$$N': \quad V_T(A) \text{ :- } V_S(A,y), V_S(A,z).$$

このとき，$N' \circ Q_S \equiv Q'_T \circ \mathbf{M}$ が成立する，次のような弱適切な CQ Q'_T を得る．

$$Q'_T: \quad V_T(X) \text{ :- } T(X,y), T(X,z).$$

Q'_T は弱適切なポリシであるから，ヒューリスティックアルゴリズムから出力されるポリシ Q_T は明らかに可用性を満たしていない． ■

以上より，適切なポリシを導出するためには Q_S を 2 回以上使った CQ を考える必要があるとわかる．また，この例は与えられた Q_S と $\mathbf{M}$ に対して弱適切なポリシが存在するにも関わらず information-collapsing なポリシが導出されてしまう例にもなっていることから，一般的には，弱適切なポリシが存在するための必要条件を判定しないことがわかる．

5 ヒューリスティックアルゴリズムが弱適切なポリシの存在を判定できるための制約

本節では，Q_S が self-join-free な CQ であり，かつ $\mathbf{M}$ が join-free な CQ のリストであるとき，ヒューリスティックアルゴリズムは弱適切なポリシが存在するための必要条件を判定することを示す．

定理 1 self-join-free な CQ Q_S と join-free な CQ のリスト $\mathbf{M}$ が与えられたとする．弱適切なターゲットポリシ Q_T が存在するならば，Q_S を 1 回使った CQ N が存在して，$N \circ Q_S \equiv Q_T^N \circ \mathbf{M}$ が成立する弱適切なターゲットポリシ Q_T^N が存在する．

ヒューリスティックアルゴリズムにおいて，Q_S の脳部に現れる変数に代入する値の集合を Q_S または $\mathbf{M}$ に現れる定数の集合，Q_S に現れる変数の集合と定数 ♣ の和集合に拡張することで，次の定理が成り立つ．ただし，♣ は Q_S または $\mathbf{M}$ に現れる定数の集合に属さない定数である．

定理 2 Q_S を self-join-free な CQ とし，$\mathbf{M}$ を join-free な CQ のリストとする．Q_S を 1 回使った CQ N に対して弱適切なポリシが存在するとき，ヒューリスティックアルゴリズムは弱適切なポリシを導出する．

定理 1 と定理 2 より，Q_S が self-join-free な CQ，$\mathbf{M}$ が join-free な CQ のリストであるとき，弱適切なポリシが存在するならば，ヒューリスティックアルゴリズムは弱適切なポリシを導出することが示される．しかしながら，導出されたポリシが一般に可用性を満たすかどうかは未解決である．

6 ヒューリスティックアルゴリズムとは別の十分条件を判定するアルゴリズム

本節では，インスタンス化関数 [2] を定義し，これを使った，ヒューリスティックアルゴリズムとは別の，弱適切なターゲットポリシが存在するための十分条件を判定するアルゴリズムを提案する．また，提案アルゴリズムが必要条件を判定するための制約を与える．

Q の **アクティブドメイン** （Q の体部に現れる全ての定

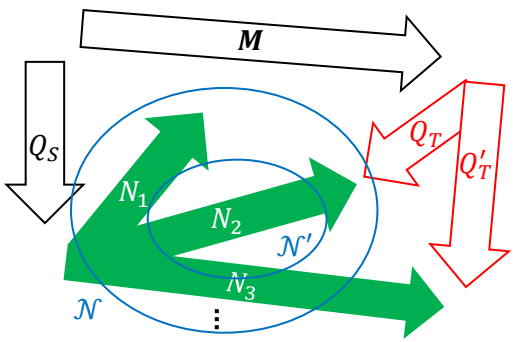

図 3　アルゴリズムの基本方針

数の集合）を $adom(Q)$ と記す．$adom$ の定義域は CQ のリストに対して自然に拡張される．CQ のリスト $\mathbf{M}$ に関する Q の**インスタンス化関数** θ は，Q の頭部にある変数から $adom(\mathbf{M}) \cup adom(Q) \cup \{\clubsuit\}$ の定数への写像である．ただし，$\clubsuit$ は $adom(\mathbf{M}) \cup adom(Q)$ に存在しない定数である．θ が全域写像であるとき，θ は**全域**であるという．Q に θ を適用して得られる CQ を $Q\theta$ と記す．

提案アルゴリズムはヒューリスティックアルゴリズムと基本方針は同じである．Q_S と $\mathbf{M}$ が与えられたとき，以下のように動作する（図 3）．

1. Q_S を使った CQ の全体集合 $\mathcal{N}$ を考える．
2. Q_S，$\mathbf{M}$ と各 $N \in \mathcal{N}$ に対して [4] のアルゴリズムを用いることでターゲットポリシの候補を求める．

$\mathcal{N}$ は無限集合である．したがって，ヒューリスティックアルゴリズムでは Q_S を使った join-free な CQ N の有限集合 $\mathcal{N}' \subseteq \mathcal{N}$ に対して手順 2 を実行し，ターゲットポリシの候補を有限に抑えている．しかしながら，join が存在する $N \notin \mathcal{N}'$ に対して求められるターゲットポリシが導出されず，例 3 のように適切なポリシが導出されない場合がある．提案アルゴリズムでは，$\mathcal{N}'$ を join が存在する CQ の集合とする．一方で，$\mathcal{N}'$ を有限集合とするため，Q_S の頭部にあるすべての変数に $adom(\mathbf{M}) \cup adom(Q) \cup \{\clubsuit\}$ の定数を代入した Q_S を使った CQ のみを考える．

提案アルゴリズムは次の 3 ステップから構成される．

1. $\mathbf{M}$ と Q_S に対し，$\mathbf{M}$ に関する Q_S の全ての全域インスタンス化関数の集合 Θ を計算する．$Q_S\Theta = \{Q_S\theta \mid \theta \in \Theta\}$ とする．
2. $Q_S\Theta$ の部分集合 $\mathcal{Q}_S$ を非決定的に選択する．N を，頭部は新しい関係名を持つ原子式であり，体部は $\mathcal{Q}_S$ に属する CQ の頭部から構成される安全な CQ とする．
3. $\mathbf{M}$ と $N \circ Q_S$ の information-revealing な CQ-rewriting Q_T が存在するならば True を出力する．そうでなければ，False を出力する．

次の定理が成立することは容易に証明できる．

定理 3　アルゴリズムが True を出力するならば，弱適切なターゲットポリシが存在する．

提案アルゴリズムは，インスタンス化関数の値域が $adom(\mathbf{M}) \cup adom(Q_S) \cup \{\clubsuit\}$ であるため，アクティブドメインが $adom(\mathbf{M}) \cup adom(Q_S) \cup \{\clubsuit\}$ の部分集合であるようなターゲットポリシのみ導出される．よって，アクティブドメインが $adom(\mathbf{M}) \cup adom(Q_S) \cup \{\clubsuit\}$ に含まれないターゲットポリシのみが弱適切である場合，提案アルゴリズムは正常に動作しない．以下の補題 1 は，アクティブドメインが $adom(\mathbf{M}) \cup adom(Q_S) \cup \{\clubsuit\}$ に含まれない弱適切なターゲットポリシが存在するとき，アクティブドメインが $adom(\mathbf{M}) \cup adom(Q_S) \cup \{\clubsuit\}$ に含まれる弱適切なターゲットポリシもまた存在することを保証する．

補題 1　与えられた $\mathbf{M}$ と Q_S に対して弱適切な CQ Q_T が存在するとき，次を満たす CQ Q'_T が存在する．

- Q'_T は $\mathbf{M}$ と Q_S に関して弱適切なポリシである．
- $adom(Q'_T) \subseteq adom(\mathbf{M}) \cup adom(Q_S) \cup \{\clubsuit\}$ が成立する．

補題 1 を用いて，以下の定理が証明できる．

定理 4　$\mathbf{M}$ は projection-free であるとする．このとき，弱適切かつ projection-free なターゲットポリシが存在するならば，このアルゴリズムは True を出力する．

7　おわりに

本研究では，ソース側とターゲット側，データ利用者がデータの交換を行う関係データ交換フレームワークにおいて，弱適切なポリシが存在するための十分条件を判定するアルゴリズムの提案を行った．また，ソース側またはターゲット側のデータ公開ポリシに対していくつかの制限を加えることで，ヒューリスティックアルゴリズムや提案アルゴリズムが，弱適切なポリシが存在するための必要条件を判定することを証明した．

今後の課題として，制限のないデータ公開ポリシに対して弱適切または適切なポリシが存在するための必要十分条件の判定可能性を考え，判定可能であればそれらを導出するためのアルゴリズムを提案することなどが挙げられる．

参考文献

[1] R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa. Data exchange: semantics and query answering. *Theoretical Computer Science*, Vol. 336, No. 1, pp. 89–124, 2005.

[2] Y. Ishihara. Toward appropriate data publishing in relational data exchange framework. In *Proceedings of SFDI 2020, CCIS 1281*, pp. 131–137, 2020.

[3] A. K. Chandra and P. M. Merlin. Optimal implementation of conjunctive queries in relational data bases. In *Proceedings of STOC '77*, pp. 77–90, 1977.

[4] F. Afrati. Determinacy and query rewriting for conjunctive queries and views. *Theoretical Computer Science*, Vol. 412, No. 11, pp. 1005–1021, 2011.