統計情報制限下における RDF 問合せ最適化のための結合選択率の間接的見積り手法

的野 晃整[1,a)]　小川 宏高[1]

概要：オープンデータの流行に後押しされ，昨今 RDF データ が急増しており，SPARQL エンドポイントの性能向上が急務となっている．解決策の一つとしてキャッシュサーバの導入があるが，エンドポイントの負荷やネットワーク転送量のなどを最小化するためには，キャッシュサーバ上でのコストベースの問合せ最適化の実施が必要となる．コストベース最適化で行われる選択率見積りには，十分な統計情報を事前に取得しておくことが不可欠であるが，キャッシュサーバではそれが困難であるという問題がある．そこで本稿では，統計情報が制限された環境において適用可能な SPARQL 問合せのための結合選択率見積り手法を提案する．提案する手法は，RDF 問合せで頻出する結合演算において，結合する 2 集合の結合選択率の統計情報が存在しなくても，その 2 集合と共通に結合する第三の集合との結合選択率を利用して，間接的に結合選択率を見積る手法である．実データを用いた実験を通じて，提案手法と従来の手法を比べた結果，提案した手法は改善の余地は残されているものの，従来手法と比べ大幅な (最大 11 桁) 精度改善を確認した．

1. はじめに

近年のオープンデータの潮流に促されて，Linked Data に注目が集まっている．Linked Data とは RDF (Resource Description Framework) [1] に基づいて，記述された外部データへのリンクを含むデータセットのことである．RDF データの公開方法の一つに SPARQL エンドポイント (以後，エンドポイントと言う) がある．エンドポイントとは，RDF データベース (Triplestore と言う) の Web インタフェースで，RDF の問合せ言語である SPARQL [2] を用いて RDF を検索できる Web サービスである．

エンドポイントを用いたアプリケーションの提案や開発は各所で進められてはいるものの，現実的にはまだブレークスルーには至っていない．その最大の理由の一つにエンドポイントの性能が不十分であることが挙げられる．現在のエンドポイントでは，問合せを発行してもタイムアウトが発生して解が得られなかったり，解のサイズを制限したりなど，アプリケーションに耐え得る安定性・信頼性が確保されていない．

エンドポイントの性能を向上させる解決策として，幾つか手法が提案されている．最もナイーブな解決策として，エンドポイントのバックエンドである Triplestore の性能を向上させる研究があるが，これらはこれまでに非常に多く行われ，多様な手法が提案されており，著者らもいくつか提案している [3,4]．また，他の解決策の一つとしてキャッシュを利用した手法が幾つか提案されている．例えば，エンドポイントの直前にプロキシを配置してキャッシュするシンプルな手法 [5] や HTTP レベルでキャッシュできるようにクライアントを拡張する手法 [6] などがある．これらは問合せとその解をそのままキャッシュするアプローチであるため，同一クエリの場合でしか利益を享受できない問題がある．その解決策として，SPARQL を構文解析して実行木を構築し，その部分木で一致する場合にもキャッシュを利用する手法も提案されている [7,8]．また，後続の問合せで利用するための付加情報を取得しておくために，与えられた問合せの構造を拡大変更する手法も提案されている [9]．Verborgh ら [10] は Linked Data Fragment と呼ぶ，負荷が集中しやすいサーバは単純な検索のみを行い，複雑な問合せはクライアントが解決する包括的なモデルを提案している．

このように，キャッシュサーバにおいて，問合せの分割や拡張によって問合せをリライトする手法，すなわち問合せ最適化を行うことが現在主流となっている．前述した RDF キャッシュのための問合せ最適化 [7–9] は，主にルールベースで行われているが，関係データベース管理システ

[1] 国立研究開発法人産業技術総合研究所 (産総研)
National Institute of Advanced Industrial Science and Technology (AIST)
[a)] a.matono@aist.go.jp

ムにおける現在の大多数はコストベースの問合せ最適化を採用している．キャッシュサーバ上でコストベースの問合せ最適化ができれば，サーバの負荷コストやネットワーク転送量を最小にすることなどが可能になるという利点がある．データベース処理では主にディスクやネットワークなどの I/O コストが支配的であることから，一般的にコストベース最適化でのコスト見積りには，結果集合の要素の個数 (以下濃度と言う) の見積りが不可欠である．結果濃度の見積りは，一般的に選択率と濃度の積によって求めるため，濃度見積りには選択率見積りが必須である．

本稿では，SPARQL エンドポイントのキャッシュサーバで利用することを前提に，SPARQL 問合せ最適化のための選択率見積り手法を提案する．6 章 にて後述するが，SPARQL 問合せの問合せ最適化のための濃度見積り手法や選択率見積り手法はこれまで幾つか提案されているが，これらはすべて Triplestore の内部で利用することを前提とした手法であるため，データ格納時に統計情報を取得したり，索引等の内部データ構造を利用できるなど，豊富な統計情報を利用できる前提となっている．しかしながら，キャッシュサーバは Triplestore と同一計算機上に存在するとは限らないどころか，管理ドメインが異なることも十分ありえるため，既存手法の適用は困難である．そのため，本稿では事前に入手できる統計情報は限られているという制約を前提とした選択率見積り手法を提案する．

本稿では，統計情報が限られている制約下において，問合せの結合選択率を見積る手法を提案する．結合演算は RDF 問合せにおいて最も高コストかつ頻発するため，最も重要な演算の一つである．また，RDF における結合は暗黙的に発生するため，関係データベースの外部キーなどのように，事前に結合する集合が決定しておらず，あらゆる結合選択率を取得しておくことは困難である．

提案手法の特徴は，結合選択率が統計情報として保持されていない場合でも，他の統計情報を用いて間接的にそれを見積ることができる点である．提案手法のアイデアは，本来の検索条件同士は従属関係にあるが，第三の検索条件を介在することで独立になるという仮説を立てた点である．例えば，$A \bowtie B$ を求める場合を考えたとき，A と B の結合選択率 $\mathrm{sel}(A \bowtie B)$ を求めたいとする．そのとき，A と B が独立であれば，単に積でもとめることができる，すなわち，$\mathrm{sel}(A \bowtie B) = \mathrm{sel}(A) \times \mathrm{sel}(B)$ が成り立つが，実際には従属関係にあるため，この式に適用しても差が大きいため精度が悪い [11]．先の仮説に基づいて，第三の検索条件 O を介在することで独立になるのであれば，すなわち $A \bowtie O$ と $B \bowtie O$ が互いに独立であるなら，$\mathrm{sel}(A \bowtie B \bowtie O) = \mathrm{sel}(A \bowtie O) \times \mathrm{sel}(B \bowtie O)$ で求めることができる．

本稿の構成は次の通りである．2 章 で前提知識として定義等を行い，3 章 にて提案する仮説および定理の詳細を述べる．4 章 では事前に取得しておく統計情報について簡単に説明する．5 章 で実験を通じて仮説の確からしさを検証する．6 章 で既存の関連研究について紹介し，7 章 にてまとめる．

2. 準備

2.1 RDF と SPARQL

RDF と SPARQL について説明する．本稿で扱う RDF のモデルはおよび SPARQL は主要部のみに単純化して定義する．なお，本稿の定義は [12] に基づいている．

まずは RDF について定義する．互いに素な無限集合 I と B，L が存在する．I は資源を示す識別子 IRI の集合，B は識別されない無名ノードを意味する空白ノード (Blank node) の集合，L は文字列や数字等の値を意味するリテラル (Literal) の集合である．本稿では，識別子は rdf:type のように，リテラルは “23” にように表記する．*RDF* **トリプル** *(RDF triple)* は**主語** *(subject)* および**述語** *(predicate)*，**目的語** *(object)* の三つ組で構成され，主語を s，述語を p，目的語を o とすると $(\mathsf{s}, \mathsf{p}, \mathsf{o})$ と表記され，$(\mathsf{s}, \mathsf{p}, \mathsf{o}) \in (I \cup B) \times I \times (I \cup B \cup L)$ と定義される．RDF データは RDF トリプルの有限集合で，同一識別子は同一資源とみなして，主語と目的語を頂点，述語を辺とした有向グラフ構造を形成する．

次に RDF データへの問合せ言語である SPARQL について定義する．SPARQL は本質的にはグラフのパターンマッチングを用いた，RDF のための問合せ言語である．I と B，L に対して互いに素である無限集合を V とする．V は変数の集合である．本稿では変数は $?x$ や $?name$ と表記する．本稿で対象とするグラフパターンは次のように再帰的に定義する．(1) トリプルパターン $t \in ?s \times I \times V$ はグラフパターンである $(?s \in V)$．(2) もし P_1 と P_2 がグラフパターンであれば，$(P_1 \textsc{ And } P_2)$ はグラフパターンである．本来であれば，トリプルパターンは $t \in (I \cup V) \times (I \cup V) \times (I \cup L \cup V)$ と定義されるが，本稿では主語は同一の変数に，述語は IRI に，目的語は任意の変数に束縛されたトリプルパターンのみを扱うものとし，これを述語トリプルパターンと呼ぶ．また，グラフパターンについても本来なら `OPTIONAL` や `FILTER`，`UNION` 等が利用でき，より複雑なパターンを定義できるが，本稿では And のみを利用するものとする．さらに，本来の SPARQL では `ASK` や `CONSTRUCT` 等複雑な処理が可能であるが，本稿では `SELECT *` のみを対象とする．結果的に，本稿で対象とする SPARQL 問合せは，述語トリプルパターンが主語同士で結合するスター型の構造を持つ．なお，SPARQL の構文やセマンティクスの定義は本稿では割愛する．

本稿では説明のために，グラフパターン P の選択率を $\mathrm{sel}(P)$ と表し，P を RDF データセット D に対して評価した結果を $[\![P]\!]_D$ と表す．また，集合 X の濃度を $|X|$ と表記

する．したがって，$\mathrm{sel}(P) = \frac{|[\![P]\!]_D|}{|D|}$ である．

2.2 RDF 結合選択率見積りの困難さ

まず，本稿での重要な概念である独立事象および従属事象についての定義を示す．本稿では，事象 X が発生する確率を $\Pr(X)$ として表記し，事象 X が発生したことが判明した後に事象 Y が発生する確率 (条件付き確率) を $\Pr(Y \mid X)$ と表す．

定義 1 (事象の独立と従属)**:** 独立とは，二つの事象 X, Y のいずれの事象が発生する確率も，他方の事象の発生に依存しない．すなわち，

$$\Pr(X \mid Y) = \Pr(X \mid Y^c)$$

であるとき，この X と Y は**互いに独立である**といい，X, Y を**独立事象**という．また $X \perp\!\!\!\perp Y$ と表わす．したがって，

$$\begin{aligned} X \perp\!\!\!\perp Y &\iff \Pr(X \mid Y) = \Pr(X) \\ &\iff \Pr(Y \mid X) = \Pr(Y) \\ &\iff \Pr(X \cap Y) = \Pr(X)\Pr(Y) \end{aligned}$$

また事象 X と Y が互いに独立でないとき，**従属である**といい，X, Y を**従属事象**という．また $X \not\perp\!\!\!\perp Y$ と表わす．

もし二つの条件 (A と B) が独立であるなら，それらの条件の選択率は，各条件の選択率の積で求めることができ，すなわち $\mathrm{sel}(A \text{ And } B) = \mathrm{sel}(A) \times \mathrm{sel}(B)$ が成り立つ．しかしながら，RDF データでは，Neumann ら [11] も主張しているように，検索条件同士が従属事象であるために，複数条件の積，すなわち結合演算を，各条件の選択率の乗算で求めることは難しい．

RDF において従属事象であること多いかどうかを確認するため，予備実験を行った．予備実験に用いた RDF データは 100,091,676 トリプルからなる DBpedia Japanese [*1] である．実験の方針は，先の式が成り立つかどうか評価することで，独立か従属かを調べる．具体的には，実測値に対する選択率の乗算による見積値の比率を調査する．

問合せは，表 1 に示すプロパティ用いた述語トリプルパターンの組合せで構成されるものとした．クラスが限定されると見込まれる特殊プロパティと，任意のクラスで利用されることが見込まれる汎用プロパティをそれぞれ 5 つずつ用いた．これらを用いた理由としては，特殊プロパティ同士は従属である傾向が高く，汎用プロパティ同士が独立性が高いと考えたためである．問合せは，特殊と汎用の各 5 つのプロパティで構成される述語トリプルパターンの集合 $S = \{(?s, \mathsf{p}_i, ?o) \mid 1 \le i \le 5\}$ の冪集合のうち，濃度 1 である集合を除く集合 $\mathfrak{P}(S) = \{A \mid A \subseteq S \wedge |A| \neq 1\}$ を問合せ集合とした．すなわち，述語トリプルパターンの全組合せで

[*1] DBpedia Japanese: http://ja.dbpedia.org/

表 1: 実験で用いた述語一覧

		プロパティ	トリプル数
特殊	s1	dbo:postalCode	42,035
	s2	geo:lat	31,107
	s3	dbo:locationCity	20,884
	s4	dbo:owner	14,178
	s5	dbo:numberOfEmployees	17,008
汎用	g1	dct:subject	3,616,298
	g2	rdfs:label	1,681,458
	g3	foaf:name	421,578
	g4	foaf:homepage	128,988
	g5	dbo:genre	99,160

ある ${}_5C_2 + {}_5C_3 + {}_5C_4 + {}_5C_5 = 26$ 個の問合せ集合を用いた．

選択率の見積りは，プロパティ p の述語トリプルパターン $P = (?s, \mathsf{p}, ?o)$ に一致するトリプル数 $|[\![P]\!]_D|$ を統計情報として取得しておき (表 1)，それを総トリプル数 $|D|$ (100,091,676) で割った値を各トリプルパターンの選択率 $\mathrm{sel}(P) = |[\![P]\!]_D|/|D|$ として用いる．問合せ $Q \in \mathfrak{P}(S)$ の選択率見積りは，問合せを構成する述語トリプルパターン P の選択率の総乗 $\mathrm{sel}(Q) = \Pi_{P \in Q}\mathrm{sel}(P)$ によって求める．

予備実験の結果を図 1 に示す．図 1 は横軸が問合せの種類 (26 通り) で縦軸が濃度の見積値と実測値の比率 (見積値/実測値) である．したがって，1 に近ければ見積りが正確で，独立であることを意味し，1 から遠い場合は見積りが実測からはずれており，特に 1 より小さい場合に従属事象であることを意味している．図 1 から判断できるように実際の濃度に比べると，選択率の総乗による見積りはかなりのずれがあることが判断でき，最大比では 10^{-13}，最小比でもおよそ 0.02，すなわち 50 倍もの違いがある．すべての問合せにおいて見積値が少く見積っていることから，RDF では検索条件同士が従属事象にあることを示している．また，特殊プロパティだけでなく，独立性が高いと見込まれる汎用プロパティであっても従属事象であることがわかる．

このように RDF では二つの事象が従属である場合が多いため，個別の選択率の積で代替すると，予備実験からは少なくとも 50 倍の誤差が発生することがわかった．したがって，個別条件の確率の積による見積りではなく，複数条件が同時に満足される確率，言い換えると結合選択率を求めることがより正確な選択率のためには不可欠である．もっとも単純かつ正確な方法としては，結合選択率を事前に統計情報として取得しておく手法である．しかしながら，前述したように RDF は 関係データベースとは異なり，結合選択率を事前に取得しておくことは非常に困難である．その理由として，1) 結合が頻出することと，2) 何と何が結合するか不明であることの 2 点がある．1) について，ある資源に関連する情報を取得するためには，関係データベースではある関係表の 1 行を選択するだけで済むが，RDF で

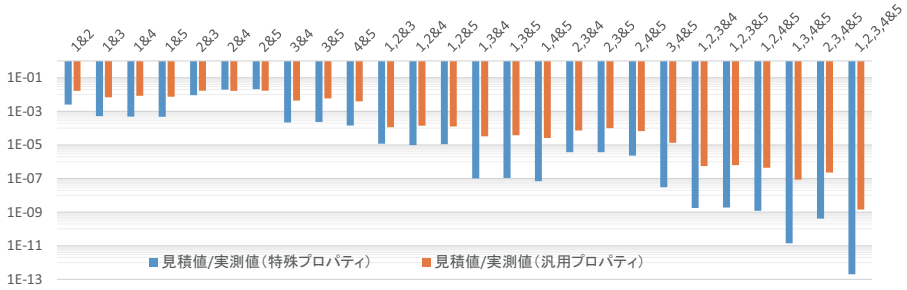

図 1: 総乗による選択率見積りの性能

はあらゆる情報をトリプル単位に細ぎれにして表現するため，その資源に関連するトリプル数に応じた結合処理が必要になってしまう．2) について，関係データベースでは主キーや外部キーを用いることが一般的で，結合する関係表とそのカラムを前もって予測できるのに対し，RDF では IRI が同一であれば同一資源であり，主キーや外部キーの考え方がないため，結合する可能性がある組合せは膨大となる．

本稿では，問合せを述語トリプルパターンのみを前提としているが，実際には，述語が束縛されていない上，多重に結合する場合も考慮すると，事前に結合選択率の全組合せを取得しておくことは困難で，特にキャッシュサーバ上でこれを行うことは現実的には不可能である．代替案として，スキーマ情報を元に結合する可能性のある述語を特定することはできる [13] が，述語定義の定義域 (domain) クラスや値域 (range) クラスに rdfs:Resource などの高次のクラスを指定されると，結局組合せが爆発してしまう．

3. 間接独立定理の提案

本節では，RDF の結合選択率の見積りに利用できる仮説と定理について述べる．

3.1 述語とクラスの独立性に関する仮説

我々は RDF が興味深い特性を持つことを観察した．

仮説 1 (述語とクラスの独立性に関する仮説)**:** RDF において，ある資源が任意の二つの述語を持つ確率について，データセット全体では従属事象である傾向が高いが，クラスのインスンタンス集合に限定すると独立事象になる傾向がある．

仮説前半の「データセット全体では従属事象である」という特性は，前述した通り予備実験によって確からしいと評価した通りで，後半部分の「クラスのインスンタンス集合に限定すると独立事象になる」という特性を生かした手法が本稿での提案である．

例 1 (従属事象の例)**:** 本と著者に関する RDF データセット D から， name と email を同時に持つ資源集合を求める場合を考える．D は本 10 冊および著者 20 人の合計 30 の資源で構成されるとし，著者のうち 15 人の名前 (name) が特定しており，10 人がメール (email) も特定しているとする．D 全体からランダムに取得した資源が name を持つ確率は 15/30 = 1/2 で，同様に email を持つ確率は 10/30 = 1/3 である．次に条件付き確率を考える．name が判っている 15 人からランダムに選んだ人が email を持つ確率は，1/3 ではなくもっと高い確率となる．すなわち，「D 全体から取得した資源」と「name を持つ資源集合から取得した資源」では email を持つ確率が異なり，name を持つかどうかが email を持つ確率に影響を与えている．つまり，name と email を持つ資源を選択することは従属事象である．

次に，仮説 1 の本質である後半部分の例として，D 全体からではなく，著者集合から選択した場合を考える．

例 2 (クラス限定によって独立事象になる例)**:** 著者集合からランダムに取得した資源が name を持つ確率は 15/20 = 2/3 で，同様に email を持つ確率は 10/20 = 1/2 である．次に同様に条件付き確率を考える．name が判っている 15 人から選んだ人が email を持つ確率は，以前と同様に 1/2 である．また逆に，email が判っている 10 人から選んだ人が name を持つ確率は，以前と同様に 2/3 である．すなわち，「著者集合から取得した資源」と「name 持つ資源集合から取得した資源」が email を持つ確率が等しく，逆に「著者集合から取得した資源」と「email 持つ資源集合から取得した資源」が name を持つ確率も等しい．つまり，name と email を持つ資源を選択することは独立事象である．

3.2 間接独立に関する定義

本節では 仮説 1 の関係を形式的に定義する．定義 1 を拡張して，間接独立と呼ぶ関係を次のように定義する．

定義 2 (事象の間接独立)**:** 事象 X， Y および O が，

$$\Pr((X \cap Y) \mid O) = \Pr(X \mid O) \Pr(Y \mid O)$$

となることを，この X と Y は O に対して，**互いに間接独立であるといい**，X, Y を O に対する**間接独立事象**という．また，$X \perp\!\!\!\perp_O Y$ と表わす．さらに，事象 O を X, Y の**間接独立空間**と言う．

この定義に基づくと， 仮説 1 の後半部は「RDF では述

語トリプルパターンはクラスに対して互いに間接独立である」と言い変えることができる．

3.3 確率の間接独立に関する定理

本節では間接独立に関する定理を提案する．

定理 1 (間接独立の基礎定理)**:** 事象 A と B が 事象 O に対して互いに間接独立 ($A \perp\!\!\!\perp_O B$) であるとき，次の式が成り立つ．

$$\Pr(A \cap B \cap O) = \Pr(A \mid O) \Pr(B \mid O) \Pr(O) \tag{1}$$

証明: 確率の乗法定理 $\Pr(X \cap Y) = \Pr(Y \mid X) \Pr(X)$ より，式 (1) は次に展開できる．

$$\Pr(A \cap B \cap O) = \Pr((A \cap B) \mid O) \Pr(O)$$

前提条件 ($A \perp\!\!\!\perp_O B$) および定義 2 より，

$$\Pr((A \cap B) \mid O) \Pr(O) = \Pr(A \mid O) \Pr(B \mid O) \Pr(O)$$

したがって，式 (1) が成り立つ． □

式 (1) を展開すると以下を得ることができる．

$$\begin{aligned} Pr(A \cap B \cap O) &= \Pr(A \mid O) \Pr(B \mid O) \Pr(O) \\ &= \frac{\Pr(A \cap O)}{\Pr(O)} \frac{\Pr(B \cap O)}{\Pr(O)} \Pr(O) \\ &= \frac{|A \cap O|/|D|}{|O|/|D|} \frac{|B \cap O|/|D|}{|O|/|D|} |O|/|D| \\ &= \frac{|A \cap O||B \cap O|}{|O||D|} \end{aligned} \tag{2}$$

これから言えることは，定理 1 の特徴は右辺では $A \cap B$ が存在するが，左辺では登場しないことである．言い換えると， $A \bowtie B$ の結合選択率は，$A \bowtie O$ と $B \bowtie O$ の結合選択率の積で代替できると言える．

例 3: 著者集合を O とし，name を持つ人の集合を A，email を持つ人の集合を B とする．name と email を同時に持つ人が選択される確率 $\Pr(A \cap B)$ を求める．この例では，$A \cap B \cap O = A \cap B$ であるため，式 (2) に代入すると，

$$\Pr(A \cap B) = \Pr(A \cap B \cap O) = \frac{15 \cdot 10}{20 \cdot 30} = 1/4$$

となる．したがって，name と email を同時に持つ人の数は

$$\Pr(A \cap B) \times |D| = 1/4 \times 30 = 7.5$$

と見積ることができる．

3.4 間接独立定理の拡張

本節では，前節で述べた定理 1 をより複雑な状況で利用できるよう拡張する．定理 1 は，二つの事象において第三の事象を経由することで，間接的な独立性が生じる場合を一般化したものである．しかし現実の RDF データではより複雑で，実際には，間接独立である事象が三つ以上存在する場合 (3.4.1 節) や，間接独立空間が複数存在する場合 (3.4.2 節)，およびこれらの複合 (3.4.3 節) がある．次節以降では，これらの二つの側面から定理を拡張する．

3.4.1 3 以上の間接独立事象への拡張

本節では間接独立事象が 3 以上存在する場合について 定理 1 を拡張する．SPARQL 問合せに含まれるグラフパターンは，三つ以上の述語で構成される場合があり，この拡張はそのような場合に使うことができる．特に，3 以上の辺からなるグラフを分解することなく，大きなグラフのまま直接計算できる利点がある．

補助定理 2 (複数間接独立事象の定理)**:** 複数の事象 $A_t (1 \leq t \leq m)$ について，それらのどの二つも，事象 O に対して，互いに間接独立 ($\forall x \in \{1, 2, \ldots, m\} \forall y \in \{1, 2, \ldots, m\}, A_x \perp\!\!\!\perp_O A_y \wedge x \neq y$) であるとき，次の式が成り立つ．

$$\Pr\left(\bigcap_{t=1}^{m} A_t \cap O\right) = \Pr(O) \prod_{t=1}^{m} \Pr(A_t \mid O) \tag{3}$$

証明: 定理 1 の証明と同様であるため割愛する． □

3.4.2 複数の間接独立空間への拡張

前節とは異なる側面，すなわち間接独立空間が複数存在する場合を考える．実際の RDF では この状況が頻出する．例えば name という述語が出現するクラスは，Person クラスもあれば，Book クラスでも利用される可能性が高い．このような汎用的な述語は，複数のクラス，すなわち複数の間接独立空間において用いられる場合が一般的である．

複数の間接独立空間がある場合，それらの空間の和集合の濃度を求める必要がある．和集合の濃度を求めるには以下の包除原理が適用できる．

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{\emptyset \neq J \subseteq \{1,2,\ldots,n\}} (-1)^{|J|-1} \left|\bigcap_{j \in J} A_j\right| \tag{4}$$

これを 定理 1 の左辺を拡張したものに適用して，展開すると式 (6) のようになる．

$$\Pr\left(\bigcup_{i=1}^{n} A \cap B \cap O_i\right) \tag{5}$$

$$\begin{aligned} &= \sum_{i=1}^{n} \Pr(A \cap B \cap O_i) \\ &\quad - \sum_{1 \leq i < j \leq n} \Pr(A \cap B \cap O_i \cap O_j) \\ &\quad + \sum_{1 \leq i < j < k \leq n} \Pr(A \cap B \cap O_i \cap O_j \cap O_k) \\ &\quad - \ldots \\ &\quad + (-1)^{n-1} \Pr(A \cap B \cap O_1 \cap \cdots \cap O_n) \end{aligned} \tag{6}$$

式 (6) のように，包除原理では n が大きいとき，+ と − が交互に出現する長大な式となるため計算が困難である．ま

た，間接独立の定理を利用することで，$A \cap B$ を排除できたとしても，第 2 項以降に $O_1 \cap \cdots \cap O_n$ が新たに出現してしまい，本末転倒となってしまう．そのため，本稿では，複数の事象 $O_i (1 \leq i \leq n)$ について，それらのどの二つも互いに共通部分をもたない排反事象 $(\bigcap_{i=1}^{n} O_i = \emptyset)$ であるという仮定を置くことで，式 (6) の第 2 項以降をすべて排除した次の式を用いることとした．実データが，この仮定を満足することが困難であるかどうかは，後述の実験で評価する．この仮定を簡単のために**空間排反仮定**と呼ぶ．

$$\Pr\left(\bigcup_{i=1}^{n} A \cap B \cap O_i\right) = \sum_{i=1}^{n} \Pr(A \cap B \cap O_i) \qquad (7)$$

これらを纏めると，次の補助定理を導くことができる．

補助定理 3 (複数間接独立空間の定理)**:** 複数の事象 $O_i (1 \leq i \leq n)$ について，それらのどの二つも互いに共通部分をもたない排反事象 $(\bigcap_{i=1}^{n} O_i = \emptyset)$ であり，かつすべての O_i に対して，事象 A と B が互いに間接独立 $(\forall O_i, A \perp\!\!\!\perp_{O_i} B)$ であるとき，次の式が成り立つ．

$$\Pr\left(\bigcup_{i=1}^{n} A \cap B \cap O_i\right) = \sum_{i=1}^{n} \Pr(A \mid O_i)\Pr(B \mid O_i)\Pr(O_i)$$

証明: 定理 1 および確率の加法定理 $X \cap Y = \emptyset \iff \Pr(X \cup Y) = \Pr(X) + \Pr(Y)$ より，自明である． □

3.4.3 2 拡張の融合

本節では，3.4.1 節 および 3.4.2 節 で述べた拡張を融合して，間接独立に関する最終的な定理を得る．

定理 4 (間接独立の定理)**:** 複数の事象 $O_i (1 \leq i \leq n)$ について，それらのどの二つも互いに共通部分をもたない排反事象 $(\bigcap_{i=1}^{n} O_i = \emptyset)$ にあり，かつすべての O_i に対して，複数の事象 $A_t (1 \leq t \leq m)$ について，それらのどの二つも互いに間接独立 $(\forall O_i \forall x \in \{1, 2, \ldots, m\} \forall y \in \{1, 2, \ldots, m\}, A_x \perp\!\!\!\perp_{O_i} A_y \wedge x \neq y)$ であるとき，次の式が成り立つ．

$$\Pr\left(\bigcup_{i=1}^{n} \bigcap_{t=1}^{m} A_t \cap O_i\right) = \sum_{i=1}^{n} \left(\Pr(O_i) \prod_{t=1}^{m} \Pr(A_t \mid O_i)\right) \qquad (8)$$

証明: 補助定理 3 と 補助定理 2 より自明である． □

なお，仮説が誤っていたとしても，定理 1 と定理 4 は常に真であることに注意したい．すなわち，RDF の特性が定理の前提条件を満足するかどうかが重要で，満足されすれば定理を利用できる．したがって，仮説 1 と空間排反仮定が正しいかどうかに，性能が左右される．

例 4: 本と著者に関する RDF データセット D から，title と date，label を同時に持つ資源集合を求める．D は，本が 10 冊，著者が 20 人の合計 30 の資源集合で構成される．本クラスのうち，title を持つ資源は 10 冊，date があるのは発行年月と印刷年月の重複を含め 14，label があるのは 8 冊である．また，著者クラスのうち，title を持つのは 15 人，date を持つのは 17 人，label を持つのは 9 人である．本と著者クラスの資源を取得する事象を B, A とし，title と date，label を持つ資源を選ぶ事象をそれぞれ T, D, L とする．

本から選択する確率は

$$\begin{aligned}\Pr(T \cap D \cap L \cap B) &= \frac{|T \cap B||D \cap B||L \cap B|}{|B|^2|D|} \\ &= \frac{10 \cdot 14 \cdot 8}{10^2 \cdot 30} = 0.373\end{aligned}$$

となり，著者から選択する確率は

$$\begin{aligned}\Pr(T \cap D \cap L \cap A) &= \frac{|T \cap A||D \cap A||L \cap A|}{|A|^2|D|} \\ &= \frac{15 \cdot 17 \cdot 9}{20^2 \cdot 30} = 0.191\end{aligned}$$

したがって，D から選択する確率はそれらの和の 0.565 となり，濃度は $|D|$ を掛けた 16.938 と見積ることができる．

4. 事前取得する統計情報

本手法の実施に必要な統計情報について簡潔に説明する．すべてのプロパティ間のあらゆる組合せの結合演算の結果濃度 $\forall \mathsf{x} \forall \mathsf{y}, \mathrm{sel}((?s, \mathsf{x}, ?o1) \textsc{And} (?s, \mathsf{y}, ?o2))$ を求めることができれば，全ての結合選択率を特定できるが，前述したように現実的ではない．そのため，提案手法ではプロパティとクラスとの結合選択率を特定しておく．すなわち，$\forall \mathsf{x} \forall \mathsf{c}, \mathrm{sel}((?s, \mathsf{x}, ?o1) \textsc{And} (?s, \mathsf{rdf{:}type}, \mathsf{c}))$ を求める．具体的には，Query 1 ですべてのプロパティを取得し，Query 2 で各プロパティを持つ資源のクラス毎の資源数を求める．

Query 1: 全プロパティ取得

```
SELECT count(DISTINCT ?p) AS ?property
WHERE {?s ?p ?o .}
```

Query 2: 特定プロパティを持つ資源のクラス毎の資源数

```
SELECT count(DISTINCT ?s) AS ?resource
WHERE { ?s <XXX> ?o . ?s rdf:type ?class . }
GROUP BY ?class
ORDER BY DESC(?resource)
```

5. 評価実験

本節では，実験を通じて仮説 1 が確からしいかどうかを評価し，さらに式 (6) から式 (7) に単純化するために導入した空間排反仮定が自然な条件であるかどうかを評価する．

実験方法としては，2.2 節 にて行った予備実験と同様に，実データを用いて問合せを設定し，事前に取得しておいた統計情報を用いて見積った濃度と，実測の濃度とを比較する．実験に用いたデータおよび問合せについても，2.2 節の予備実験と同様に，DBpedia Japanese を用い，表 1 のプロパティによる述語トリプルパターンの全組合せ (26 通り)

を問合せ集合とした．

まず，仮説 1 の検証のために，クラスを 1 つのみとすることで，間接独立空間が一つのみになるようにした．具体的には，特殊プロパティの問合せには dbo:Company クラスを，汎用プロパティの問合せには foaf:Person クラス，あるいは dbo:Work クラスを用いた．

図 2 に 仮説 1 の検証のための実験結果を示す．図の横軸と縦軸については，図 1 と同様である．図 1 では最大比が 2^{-13} であったが，提案手法では最大比 0.27 程度に収まっており，最小比では 1.001 と非常に 1 に近い場合もある．この 0.27 の比についても，濃度では実測値が 325 で，見積値が 88.1 であるため，それほど差は大きくないと言える．図 2 は図 1 に比べると十分 1 に近いため独立である傾向が高いと言える．すなわち，本実験においては 仮説 1 は確からしいという結果を得ることができた．

つぎに，空間排反仮定が困難な条件かどうかを検証するために，図 2 の実験条件からクラス限定を解除して，実測値と見積値の比較を行った．すなわち，図 2 の実験結果を基準と考えられるため，基本的にはこれより良くなることはない．

図 3 に空間排反仮定の検証のための実験結果を示す．図 2 では 8 割方が小く見積られていたのに対し，図 3 では 1 つ以外大きく見積られるという結果になっている．これは，単純化することために空間排反仮定を置いたことで，第二項に出現するマイナスが排除された影響がでていると予想される．この結果から空間排反仮定は実データで満足することは困難な条件であったことを示している．また，資源が有する平均クラス数を計測するとおよそ 2.06 であった．このことからも空間排反仮定は実データに適用するには不自然な前提であったことを示しているが，これについては今後の課題としたい．

しかしながら，一つの手法としてこの実験結果を見ると，従来手法である図 1 での，実測値と見積値の比は最小で 0.02，最大で 10^{-13} であったのに対し，図 3 では，最小で 1.01，最大で 50 程度に収まっており，従来方法の最良値が提案手法の最悪値と同等程度となっており，最大 11 桁もの大幅な性能改善ができたと言える．

6. 関連研究

RDF のための濃度見積りあるいは選択率見積り手法はすでに幾つか提案されている．Stocker ら [13] は，Triplestore の一つである Jena の SPARQL 処理で用いる問合せ最適化のための選択率見積りを提案している．彼等の手法は，基本的には従来の関係データベースで行われてきた手法を RDF に適用した手法で，トリプルパターンの選択率を各要素の選択率の積 $\mathrm{sel}(s) \times \mathrm{sel}(p) \times \mathrm{sel}(o)$ にて求める手法である．また，結合選択率についても前述したように，スキーマ情報を元に結合する可能性のある述語同士の結合選択率を事前取得する方針で，もしスキーマ情報がない場合は，すべての組合せを求めるとしている．Maduko らの手法 [14] は，RDF グラフから部分グラフを抽出し，それらの出現回数のヒストグラムを統計情報として利用する手法である．与えられた問合せに一致する部分グラフあれば，それの出現回数を利用して濃度を見積る．Neumann らは RDF-3X で利用する問合せ最適のための濃度見積り手法を提案している [11]．彼らは RDF では検索条件同士が従属事象であることを問題視しており，その解決策として，Characteristic Sets [11] と呼ぶ濃度見積り手法を提案している．共通の主語を有する述語の集合を Characteristic Set と呼び，それらの出現頻度を用いる手法で，Maduko らの手法 [14] と類似しているが，著者らは現実のデータにより適した手法であると主張している．

これらの手法はいずれも，基本的には Triplestore 内での問合せ最適化のための濃度見積り手法である．したがって，利用可能な統計情報が豊富に入手できる前提であるのに対し，我々はキャッシュサーバでの利用を想定しているため，統計情報が不十分であるという前提を置いている．また，我々の提案手法は結合選択率そのものを見積ることは従来になかった技術である．

7. おわりに

本稿では，キャッシュサーバ等の統計情報の取得が制限された環境における SPARQL 問合せにおける結果濃度見積り手法を提案した．本稿では，直接的な結合選択率が統計情報として与えられない環境においても，第三者のとの結合選択率を用いることで間接的に結合選択率を求める手法を提案した．本稿では，仮説 1 と独立空間同士が排反であるという仮定を置いて二つの定理を提案した．定理については証明によって真であることを示した．実験によって，仮説 1 については確からしいことを，また空間排反仮定は実データには若干厳しい条件であったことを確認した．しかしながら，一つの結合選択率の見積り手法としてみると，従来の見積り手法に比べると最大 11 桁もの大幅な精度の改善が見られることを確認できた．

今後の課題として， 空間排反仮定の訂正とそれに合せて，補助定理 3 および 定理 4 の修正を行う．また，実際にキャッシュサーバをプロトタイプして，キャッシュにおける問合せ最適化の有用性を評価したい．

謝辞 本研究の一部は科研費 (15H02781) の助成を受けたものである．また，NEDO 次世代ロボット中核技術開発の成果を活用している．加えて，助言を頂いた産業技術総合研究所池上努氏および金京淑氏に深謝する．

参考文献

[1] Cyganiak, R., Wood, D. and Lanthaler, M.: RDF 1.1 Concepts and Abstract Syntax, W3C Recommendation

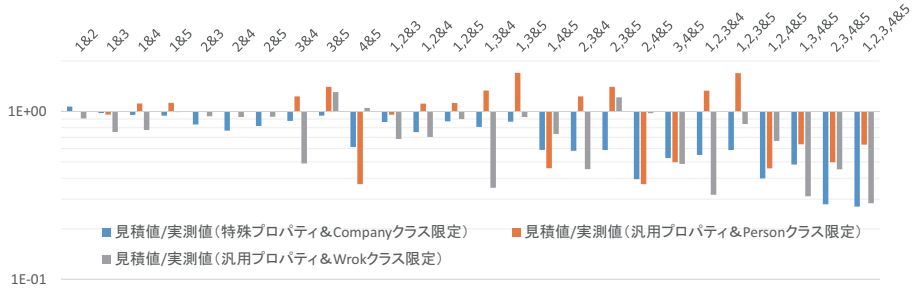

図 2: 1 クラスに限定した場合の間接独立定理に基づく選択率見積りの性能 (仮説 1 の検証実験)

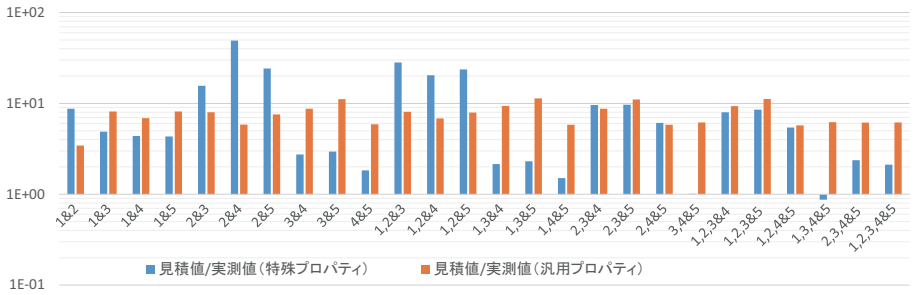

図 3: クラス限定のない間接独立定理に基づく選択率見積りの性能 (空間排反仮定の検証実験)

(25 February 2014). http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

[2] The W3C SPARQL Working Group: SPARQL 1.1 Overview, W3C Recommendation (21 March 2013). http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/.

[3] Matono, A., Pahlevi, S. M. and Kojima, I.: RDFCube: A P2P-Based Three-Dimensional Index for Structural Joins on Distributed Triple Stores, *Databases, Information Systems, and Peer-to-Peer Computing, International Workshops, DBISP2P 2005/2006, Trondheim, Norway, August 28-29, 2005, Seoul, Korea, September 11, 2006, Revised Selected Papers*, pp. 323–330 (online), doi: 10.1007/978-3-540-71661-7_31 (2006).

[4] Matono, A. and Kojima, I.: Paragraph Tables: A Storage Scheme Based on RDF Document Structure, *Database and Expert Systems Applications - 23rd International Conference, DEXA 2012, Vienna, Austria, September 3-6, 2012. Proceedings, Part II*, pp. 231–247 (online), doi: 10.1007/978-3-642-32597-7_21 (2012).

[5] Martin, M., Unbehauen, J. and Auer, S.: Improving the Performance of Semantic Web Applications with SPARQL Query Caching, *Proceedings of the 7th International Conference on The Semantic Web: Research and Applications - Volume Part II*, ESWC'10, Berlin, Heidelberg, Springer-Verlag, pp. 304–318 (online), doi: 10.1007/978-3-642-13489-0_21 (2010).

[6] Williams, G. T. and Weaver, J.: Enabling Fine-grained HTTP Caching of SPARQL Query Results, *Proceedings of the 10th International Conference on The Semantic Web - Volume Part I*, ISWC'11, Berlin, Heidelberg, Springer-Verlag, pp. 762–777 (online), available from ⟨http://dl.acm.org/citation.cfm?id=2063016.2063065⟩ (2011).

[7] Yang, M. and Wu, G.: Caching intermediate result of SPARQL queries, *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011 (Companion Volume)*, pp. 159–160 (online), doi: 10.1145/1963192.1963273 (2011).

[8] Wu, G. and Yang, M.: Improving SPARQL query performance with algebraic expression tree based caching and entity caching, *Journal of Zhejiang University - Science C*, Vol. 13, No. 4, pp. 281–294 (online), doi: 10.1631/jzus.C1101009 (2012).

[9] Lorey, J. and Naumann, F.: Caching and Prefetching Strategies for SPARQL Queries, *The Semantic Web: ESWC 2013 Satellite Events - ESWC 2013 Satellite Events, Montpellier, France, May 26-30, 2013, Revised Selected Papers*, pp. 46–65 (online), doi: 10.1007/978-3-642-41242-4_5 (2013).

[10] Verborgh, R., Sande, M. V., Colpaert, P., Coppens, S., Mannens, E. and de Walle, R. V.: Web-Scale Querying through Linked Data Fragments, *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014.*, (online), available from ⟨http://ceur-ws.org/Vol-1184/ldow2014_paper_04.pdf⟩ (2014).

[11] Neumann, T. and Moerkotte, G.: Characteristic Sets: Accurate Cardinality Estimation for RDF Queries with Multiple Joins, *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, ICDE '11, Washington, DC, USA, IEEE Computer Society, pp. 984–994 (online), doi: 10.1109/ICDE.2011.5767868 (2011).

[12] Pérez, J., Arenas, M. and Gutierrez, C.: Semantics and Complexity of SPARQL, *ACM Trans. Database Syst.*, Vol. 34, No. 3, pp. 16:1–16:45 (online), doi: 10.1145/1567274.1567278 (2009).

[13] Stocker, M., Seaborne, A., Bernstein, A., Kiefer, C. and Reynolds, D.: SPARQL Basic Graph Pattern Optimization Using Selectivity Estimation, *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, New York, NY, USA, ACM, pp. 595–604 (online), doi: 10.1145/1367497.1367578 (2008).

[14] Maduko, A., Anyanwu, K., Sheth, A. P. and Schliekelman, P.: Graph Summaries for Subgraph Frequency Estimation, *The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008, Proceedings*, pp. 508–523 (online), doi: 10.1007/978-3-540-68234-9_38 (2008).