2N-02

正解スキーマを必要としないプロパティグラフのスキーマ評価手法の提案

湯川 楓祐 †　　塩川 浩昭 ‡
筑波大学 情報学群 情報科学類 †　　筑波大学 計算科学研究センター ‡

1 導入

プロパティグラフ（PG）は，ノードとエッジに属性を持たせたグラフ構造のデータモデルであり，ソーシャルネットワークのモデル化などで利用される．またスキーマとは，データが従うべき制約を定義したもので，データの一貫性を保つための仕組みである．PG は必ずしもスキーマを明示的に定義する必要がなく，いわゆるスキーマレスの運用も可能である．PG をスキーマレスで運用する場合，データの属性を自由に追加・削除できるためデータの柔軟性が高まる一方，データの一貫性が保証されないという課題がある．

この課題を解決するために，スキーマレスで運用を開始した PG に対し，後からスキーマを定義するためのスキーマ抽出手法が提案されている [1] [2]．これらの研究では，データセットの設計者や専門家が作成した正解スキーマ（ground truth schema）との比較を通じて，抽出したスキーマの品質を評価する．そのため，正解スキーマが存在しないデータからスキーマを抽出した場合，従来の評価方法ではそのスキーマの品質を評価できない．

そこで本稿では，正解スキーマを必要とせずに，抽出したスキーマを PG のデータそのものと直接比較して，スキーマの品質を評価する新たな手法を提案する．本稿の主な貢献は，この評価手法を提案することと，複数のデータセットを用いた実験によりその有効性を実証することである．

2 事前準備

本稿では (a) 単体のノードまたはエッジを「オブジェクト」，(b) ラベルとプロパティを「属性」，(c) データ構造の制約を定義した「スキーマ」に対し，実際のデータが格納された PG を「インスタンス」と呼ぶ．

また，プロパティグラフ $\mathcal{G}$ が与えられたとき，表 1 に示す記号を用いる．例えば図 1 に示す PG[*1] を $\mathcal{G}_1$ とおくと，$V(\mathcal{G}_1) = \{v_1, v_2, v_3\}$，$\mathcal{L}_{V(\mathcal{G}_1)} = \{\{\text{Message, Post}\}, \{\text{Person}\}\}$，$V(\mathcal{G}_1, \{\text{Message, Post}\}) = \{v_1, v_2\}$，$V(\mathcal{G}_1, \{\text{Person}\}) = \{v_3\}$，$E(\mathcal{G}_1) = \{e_1, e_2\}$，$\mathcal{L}_{E(\mathcal{G}_1)} = \{\{\text{Likes}\}\}$，$E(\mathcal{G}_1, \{\text{LIKES}\}) = \{e_1, e_2\}$ である．

3 提案手法

本稿では，2 つの観点からスキーマを評価する指標を提案する．1 つ目はスキーマがインスタンスをどの程度網羅的に表現できているかを示す「カバレッジ」，2 つ目はスキーマがどの程度冗長性を排除しているかを示す「簡潔性」である．

表 1: 本稿で用いる主な記号一覧

記号	説明
$X \in \{V, E\}$	ノードまたはエッジを表す添字
$X(\mathcal{G})$	$\mathcal{G}$ のノード（エッジ）集合
$\mathcal{L}_{X(\mathcal{G})}$	$\mathcal{G}$ のノード（エッジ）ラベルの集合
$X(\mathcal{G}, L)$	ラベル $L \in \mathcal{L}_{X(\mathcal{G})}$ を持つノード（エッジ）の集合

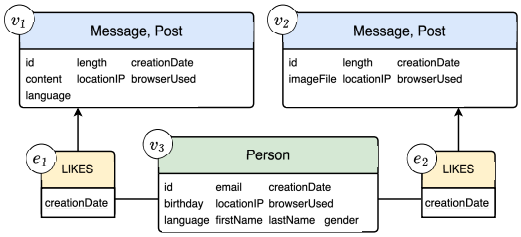

図 1: LDBC データセットから手法 1 で抽出したスキーマの例

3.1 カバレッジ

カバレッジは，スキーマがインスタンスのデータをどの程度網羅的に表現できているかを示す指標である．カバレッジを計算するために，「被表現率（expressed rate）」を導入する．被表現率は，インスタンスがスキーマによってどの程度表現されているかを示す指標であり，インスタンスの単一のオブジェクト，オブジェクトの集合，およびインスタンス全体に対して計算できる．カバレッジの定義を以下に示す．

定義 1. カバレッジは，インスタンス全体の被表現率である．

カバレッジの評価アルゴリズムを，Algorithm1 に示す．提案手法では，(1) スキーマをフラット化[*2]する．次に，(2) 各インスタンスオブジェクトに対する被表現率を計算[*3]し，(3) ラベル集合を基準にインスタンスオブジェクトをグルーピング[*4]し，(4) 各グループの被表現率を計算する．最後に，(5) インスタンス全体の被表現率（全グループの被表現率の平均値）を計算し，スキーマのカバレッジとして出力する．

なお，カバレッジのみを用いてスキーマを評価すると，オブジェクトや属性が重複している冗長なスキーマに対しても高い評価を与える可能性がある．そこでスキーマをより総合

Schema Evaluation Methods for Property Graphs without Ground Truth

† Fusuke Yukawa, College of Information Science, University of Tsukuba
‡ Shiokawa Hiroaki, Center for Computational Sciences, University of Tsukuba

*1 PG のスキーマは PG で表現することができる．図 1 はスキーマであり，かつ PG である．

*2 フラット化とは，スキーマの継承関係を展開し，継承親が持つ制約を子孫に分配する処理である．フラット化の詳細については [3] を参照されたい．

*3 インスタンスオブジェクト i の被表現率は，スキーマオブジェクト s との類似度を返す関数 $\text{sim}(i, s)$（Algorithm1 の 4 行目）を用いて決定する．類似度関数の定義については [3] を参照されたい．

*4 グルーピングは，特定のラベルを持つオブジェクトが多数存在する場合に，それらが全体のカバレッジスコアに過度な影響を与えないようにするために行う．

Algorithm 1: カバレッジの評価アルゴリズム

Input: $\mathcal{I}$: インスタンス, $\mathcal{S}$: スキーマ, X
Output: coverage

function `calcCoverage`($\mathcal{I}$, $\mathcal{S}$, X)
  $\mathcal{S} \leftarrow$ `flatten`($\mathcal{S}$);
  foreach $i \in X(\mathcal{I})$ **do**
    $objectExpressedRate[i] \leftarrow \max_{s \in X(\mathcal{S})}$ `sim`(i, s);
  $groups \leftarrow \{X(\mathcal{I}, L) \mid L \in \mathcal{L}_{X(\mathcal{I})}\}$;
  foreach $group \in groups$ **do**
    $groupExpressedRate[group] \leftarrow$ `avg`$(\{objectExpressedRate[i] \mid i \in group\})$;
  return `avg`$(\{groupExpressedRate[g] \mid g \in groups\})$;

Algorithm 2: 簡潔性（Concision）の評価アルゴリズム

Input: $\mathcal{I}$: インスタンス, $\mathcal{S}$: スキーマ, θ: 閾値, X
Output: concision

function `calcConcision`($\mathcal{I}$, $\mathcal{S}$, θ, X)
  originalCoverage $\leftarrow$ `calcCoverage`$(\mathcal{I}, \mathcal{S}, X)$;
  $redundantObjCnt \leftarrow 0$;
  foreach $object \in X(\mathcal{S})$ **do**
    $\mathcal{S}' \leftarrow X(\mathcal{S}) \setminus \{object\}$;
    newCoverage $\leftarrow$ `calcCoverage`$(\mathcal{I}, \mathcal{S}', X)$;
    if originalCoverage − newCoverage $< \theta$ **then**
      $redundantObjCnt \leftarrow redundantObjCnt + 1$;
  return $1 - \dfrac{redundantObjCnt}{|X(\mathcal{S})|}$;

的に評価するため，簡潔性を次節で導入する．

3.2 簡潔性

簡潔性は，スキーマがどの程度重複を避け，効率的にデータを表現しているかを定量化した指標である．スキーマ内のあるオブジェクトを取り除いてもカバレッジがほとんど低下しない場合，そのオブジェクトは冗長とみなされる．冗長なオブジェクトが少ないほど，スキーマは簡潔であると評価する．簡潔性の定義を以下に示す．

定義 2. 簡潔性は，スキーマ内の冗長でないオブジェクトの割合である．

簡潔性の評価アルゴリズムを，Algorithm2 に示す．提案手法では，まず (1) スキーマのカバレッジを計算する．次に，(2) スキーマから 1 つオブジェクトを取り除いてカバレッジを再計算し，(3) カバレッジの低下度合いが閾値[*5]を下回る場合にそのオブジェクトを冗長なオブジェクトと判定する．最後に，(4) 冗長でないオブジェクトの割合を簡潔性として出力する．

4 評価実験

実験設定 LDBC Social Network Benchmark [4] および NeuPrint MB6 [5] データセットを用いて，提案手法の有効性を評価した．具体的には，データセットの正解スキーマと，スキーマ抽出手法によって抽出したスキーマを，それぞれ提案手法で評価した．評価結果を表 2 に示す．表中の「手法 1」「手法 2」は，それぞれ [1]，[2] により抽出したスキーマを指す．

考察 1 提案手法は正解スキーマに対して非常に高いスコアを与えている．正解スキーマはデータセット作成時に専門家が設計したものであり，データ構造を網羅的かつ簡潔に表現しているため，この評価結果は妥当であると考える．

考察 2 提案手法は [1]，[2] で抽出したスキーマに対し，カバレッジは高いが一部の簡潔性は低い，という評価を与えている．この評価の妥当性を確認するため全ての抽出スキーマを分析した．ここでは一例として，LDBC データセットから手法 1 で抽出したスキーマを取り上げる．このスキーマはインスタンスの全属性を網羅していたが，図 1 の n_1, n_2, e_1, e_2 のように，ほぼ同一の属性を持つノードやエッジが複数存在していた．したがって，カバレッジは高いが簡潔性は低いという評価結果が妥当であると考える．その他のスキーマについても同様の分析を行い，評価が妥当であることを確認している．

結論 提案手法によるスキーマの評価は実際のスキーマの特徴と整合しており，提案手法はスキーマの品質評価において有効であると考える．

表 2: 提案手法による，各スキーマの評価結果

データセット	評価指標	正解スキーマ		手法 1		手法 2	
		ノード	エッジ	ノード	エッジ	ノード	エッジ
LDBC	Coverage	1.00	1.00	1.00	1.00	1.00	1.00
	Concision	1.00	1.00	0.83	0.61	1.00	0.69
MB6	Coverage	0.98	0.98	0.98	0.98	0.98	0.98
	Concision	1.00	1.00	1.00	0.33	1.00	0.40

5 まとめ・今後の課題

本稿では，正解スキーマを必要としない PG のスキーマ評価指標を提案した．また提案手法を用いて複数のスキーマを評価し，その評価の妥当性を示した．今後は，提案した指標がより多様なデータセットやスキーマにも適用可能かを検証し，さらに指標の改善や拡張についても検討を行う予定である．

謝辞

本研究の一部は JST 創発的研究支援事業 JPMJFR232P の支援を受けたものである．

参考文献

[1] Xue Lei. Property graph schema extraction. Master's thesis, Eindhoven University of Technology, 2021.
[2] Hanâ Lbath, Angela Bonifati, and Russ Harmer. Schema inference for property graphs. In *Proceedings of the 24th International Conference on Extending Database Technology (EDBT 2021)*, pp. 499–504, 2021.
[3] 湯川楓祐, 塩川浩昭. プロパティグラフに対する新たなスキーマ評価指標の提案. 第 17 回データ工学と情報マネジメントに関するフォーラム, 発表予定, 2025.
[4] Orri Erling, Alex Averbuch, et al. The ldbc social network benchmark: Interactive workload. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 619–630, 2015.
[5] Shin-ya Takemura, Yoshinori Aso, et al. A connectome of a learning and memory center in the adult *Drosophila* brain. *eLife*, Vol. 6, p. e26975, 2017.

[*5] 閾値 θ は，1 スキーマオブジェクトあたりの平均的なカバレッジ（originalCoverage/$|X(\mathcal{S})|$）に対して，ハイパーパラメータ γ を乗じた値である．γ の設定方法については，[3] を参照されたい．