D-007

誤差の保証がある近似的問合せ処理に関する研究

Approximate Query Processing with Error Guarantees based on Real Data

倪 天嘉 Tianjia Ni　杉浦 健人 Kento Sugiura　石川 佳治 Yoshiharu Ishikawa　陸 可鏡 Kejing Lu

1 はじめに

近年，データ量の増加と分析の要求の高度化に伴い，大量のデータに基づくデータベース問合せを効率的に実行するための技術として，近似的問合せ（approximate query processing, AQP）が着目されている [1–3]．データベース全体ではなく，データベースの一部分のデータや要約データを用いて正確ではないものの効率的に問合せ処理を行う技術として，AQP ではシノプシスを用いた問合せが行われる．シノプシス（synopsis）とは，対象となるデータをコンパクトに集約したデータを指す一般的な概念であり，さまざまなアプローチがある [4–6]．

Bounded Approximate Query（BAQ）はリレーショナルデータベース上のシノプシス構築とそれを用いた問合せに基づく近似処理フレームワークである [1]．BAQ では，ユーザが設定する誤差の閾値と典型的なワークロードを示す問合せ集合を用いてオフライン処理でデータベースからシノプシスを生成し，そのシノプシスを使用して効率的にオンライン問合せに回答する．他の要約データを使用するアプローチ (サンプリング [7,8]，分位数 [9] など) と比べて，BAQ は一般的な集計関数（COUNT，SUM，AVG，MIN，MAX）に対して効率的に誤差を保証できる特徴がある．

ただし，BAQ では COUNT，SUM，AVG 集約について，対象となる集約列のドメインが実数全体である場合に厳密な誤差保証を提供できないという問題がある．また，BAQ で対応できるのは構造が簡単な問合せのみであり，例えば結合などの操作を含む問合せへの対応方法は明示されていない．つまり，BAQ には結合や自己結合などの操作を含む問合せに対するシノプシスの生成手法など，検討と改善の余地がある．

そこで，本研究では以上の問題を解決するために，BAQ のアプローチを拡張した誤差の保証がある近似的問合せ処理手法を提案する．加えて，TPC-H ベンチマークの 22 個の問合せを対象に，それらの問合せに対応可能なシノプシスの生成と問合せの処理のアプローチを検討する．

2 準備

本章では本稿の議論に必要な概念について説明する．以下では，対象とする問合せ，および誤差の指標である相対誤差について述べる．

2.1 対象問合せ

本研究では以下の演算を用いた OLAP 問合せを想定する．

- テーブル間の結合（自己結合を含む）
- COUNT，MIN，MAX，SUM，および AVG での集約
- カテゴリ属性に対する =，≠，及び数値属性に対する =，≠，>，≥，<，および ≤ を用いた条件での選択
- カテゴリ属性に対する存在 EXIST を用いた選択
- グルーピングとランキング

例えば，以下の問合せ 1 を処理対象として扱う．

```
SELECT c_custkey, c_name,
 SUM(l_extendedprice * (1 - l_discount)) AS revenue,
 c_acctbal, n_name, c_address, c_phone
FROM customer, orders, lineitem, nation
WHERE o_orderdate >= '1993-10-01'
 AND o_orderdate < '1994-01-01'
 AND l_returnflag = 'R'
 AND c_custkey = o_custkey
 AND l_orderkey = o_orderkey
 AND c_nationkey = n_nationkey
GROUP BY  c_custkey, c_name, c_acctbal, c_phone,
n_name, c_address
ORDER BY revenue DESC
```

2.2 相対誤差

本研究では与えられた OLAP 問合せに対し，集約結果の真値と近似値との誤差 err として相対誤差を用いる．2 つの値 $x, y \in \mathbb{R}$ の相対誤差を以下の式で定める．

$$err(x,y) = \begin{cases} \frac{|x-y|}{\epsilon} & (x = 0) \\ \left|\frac{x-y}{x}\right| & (\text{otherwise}) \end{cases} \quad (1)$$

ただし，x を真値，y を近似値とし，$\epsilon \in \mathbb{R}$ を非常に小さな正の値とする．シノプシスを生成する時，バケット集合 B_i はある数値属性 $A_i \in R$ を重複なしで分割した範囲の集合であり，各バケット中における平均値と任意の値の相対誤差が δ 以内となるよう生成する．つまり，各バケット $b \in B_i$ の平均値 p を代表値として考える，各バケットは，以下の式が成り立つように定める．

$$p = \frac{\sum_{x \in b} x}{|b|} \quad \forall x \in b, err(x,p) \le \delta \quad (2)$$

本研究では，事前にユーザから相対誤差のしきい値 δ と問合せのワークロードが与えられると想定する．この想定のもとで，誤差以内で式 (2) に基づきバケットを生成し，シノプシスを計算する．オンラインにはユーザがシノプシスを使用し，

名古屋大学大学院情報学研究科　Graduate School of Informatics, Nagoya University

TPC-H の 22 個の問合せのような複雑な問合せを高精度かつ効率的な近似的処理が実行できる．

3 提案手法

本章では結合処理と複雑な判断条件を持つ問合せを高効率的に処理するために，[10] を元にシノプシスの生成と問合せの処理について提案手法を述べる．

3.1 シノプシスの生成

複雑な問合せを処理する場合に，2.1 の問合せ 1 を例として，指定された時間帯において発注された部品の情報を問合せする際，4 つのテーブルの結合と複数の条件に関する選択操作を処理する必要がある．

本研究では，まず，選択条件に基づく相関属性のデータ要約と結合処理をオフラインに行って，[10] で提案した手法でシノプシスを生成する．問合せ 1 を例として，4 つのテーブルを結合し， (l_extendedprice, l_discount,c_acctbal,c_phone) のバケットを計算し，シノプシスを生成し，残るカテゴリ属性 (c_custkey, c_name, n_name, c_address, o_orderdate) に関するシノプシスと結合し，元テーブルから最後のシノプシスを生成し，保存する．問合せが与えらえたときには，シノプシスに基づく具体的な条件からデータを選択し，近似計算を行う．

3.2 問合せの処理

■選択条件に数値属性を含む場合 カテゴリ属性による選択・グルーピングに加え，選択条件に数値属性が用いられる場合について述べる．つまり，以下のような問合せ 2 が例となる．

```
SELECT l_orderkey,o_orderdate, COUNT(*)
FROM lineitem, orders
WHERE l_extendedprice > 11000
  AND o_orderkey   =l_orderkey
GROUP BY l_orderkey,o_orderdate
ORDER BY l_extendedprice
```

この問合せに対し，提案手法による対応するシノプシスのスキーマを (orderkey, orderdate, price_min, price_max, SF) とし，lineitem と orders テーブルを用いてインスタンスを生成する．SF は l_extendedprice における [price_min, price_max] 範囲以内のレコードの数である．シノプシス中には SF = 0 となるレコードは含まない．つまり，選択・グルーピング条件で指定されたカテゴリ属性に加え，選択条件で指定された数値属性のバケットをカテゴリ属性とみなしてグルーピングを行い，各グループに属するタプル数をシノプシスのレコードとする．

ここで，バケットに分割された数値属性に対して選択条件が与えられたとき，各バケットは 1) バケット全体が条件を満たす，2) バケット全体が条件を満たさない，3) バケットの一部が条件を満たすの 3 つの場合に分けられる．

提案手法では，1) の場合，バケットについて，元の問合せを以下の問合せに書き換えて近似的に計算する．

```
SELECT l_orderkey,o_orderdate, sum(SF)
FROM synopsis
WHERE price_min > 11000
GROUP BY l_orderkey,o_orderdate
ORDER BY l_extendedprice
```

3) の場合，以下の問合せで正確に計算する

```
SELECT l_orderkey,o_orderdate, count(*)
FROM lineitem, orders
WHERE l_extendedprice > 11000
  AND l_extendedprice < synopsis.price_max
  AND o_orderkey = l_orderkey
GROUP BY l_orderkey,o_orderdate
ORDER BY l_extendedprice
```

ここで, バケットについて COUNT 関数を計算するとき, シノプシスによる部分的に条件を満たすグループの最大値 price_max を返し，元データから以上の問合せで選択条件を満たすレコードの数を返す．元データに関して問合せを処理するので，3) の集計計算の誤差は 0 となる．

2) の場合，COUNT 集約の計算の誤差は 0 なので，提案手法では誤差 0 で処理できる．つまり，[10] のアプローチにより SUM と AVG の集計計算について誤差 δ 以下の保証ができる．

4 まとめと今後の課題

本稿では，誤差の保証がある近似的問合せ処理に対して，データを要約する提案手法について議論した．今後の課題としては，今回議論した解決策について具体的な手法の考案とその実現が実現する．

謝辞

本研究は，JSPS 科研費（16H01722，20K19804）の助成による．

参考文献

[1] K. Li, Y. Zhang, G. Li, W. Tao, and Y. Yan, “Bounded approximate query processing,” *IEEE TKDE*, vol. 31, no. 12, pp. 2262–2276, 2019.
[2] S. Chaudhuri, B. Ding, and S. Kandula, “Approximate query processing: No silver bullet,” in *Proc. SIGMOD*, pp. 511–519, 2017.
[3] B. Mozafari and N. Niu, “A handbook for building an approximate query engine,” *IEEE Data Engineering Bulletin*, vol. 38, no. 3, pp. 3–29, 2015.
[4] B. Walenz, S. Sintos, S. Roy, and J. Yang, “Learning to sample: Counting with complex queries,” in *PVLDB*, vol. 13, pp. 389–401, 2019.
[5] M. Halford, P. Saint-Pierre, and F. Morvan, “An approach based on bayesian networks for query selectivity estimation,” in *Proc. DASFAA*, pp. 3–19, 2019.
[6] Q. Ma and P. Triantafillou, “DBEST: Revisiting approximate query processing engines with machine learning models,” in *Proc. SIGMOD*, pp. 1553–1570, 2019.
[7] S. Agarwal, A. Panda, B. Mozafari, S. Madden, and I. Stoica, “BlinkDB: Queries with bounded errors and bounded response times on very large data,” in *Proc. EuroSys*, pp. 29–42, 2013.
[8] B. Ding, S. Huang, S. Chaudhuri, K. Chakrabarti, and C. Wang, “Sample + Seek: Approximating aggregates with distribution precision guarantee,” in *Proc. SIGMOD*, pp. 679–694, 2016.
[9] N. Potti and J. M. Patel, “DAQ: a new paradigm for approximate query processing,” *PVLDB*, vol. 8, no. 9, pp. 898–909, 2015.
[10] 倪 天嘉, 杉浦 健人, 石川 佳治, “誤差を保証する近似的問合せについて,” in **第 *13* 回データ工学と情報マネジメントに関するフォーラム**, 2021.