D-002

機械学習を用いた近似的問合せ処理
Approximate Query Processing Based on Machine Learning

倪　天嘉
Tianjia Ni

石川　佳治
Yoshiharu Ishikawa

杉浦　健人
Kento Sugiura

1　はじめに

近年，データの量の増加と分析などの要求の高度化に伴い，データベースおける問合せ処理がより重要となってきている．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については，これまでも多くの研究があるが，近年さらに着目されている．特にサンプリングについては，30年以上に渡る研究が存在する．

ただし，複雑な述語，結合処理，GROUPBY，副問合せなどを含む問合せについては，効果的な近似的問合せの手法について，未だに研究が進んでいる．さらに，近年発展が著しい機械学習の技術を近似的問合せにおいてどのように活用するかは大きな課題となっている．

本研究では，データベースへの問合せを効率的に実行するため，サンプリングと機械学習を組み合わせることについて検討を行う．複雑な述語と集計関数を含む問合せに対し，サンプルの活用により効率的な問合せ処理の実現を図る．

2　近似的問合せ処理

大量のデータを対象とした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，近年，**近似的問合せ処理**（approximate query processing, AQP）が着目されている [1–3]．本章では提案手法における要素技術として，サンプリングベースの近似的問合せ処理と，関連の深い分野であるデータクリーニングについて述べる．

2.1　サンプリングに基づく近似的問合せ処理

サンプリングを用いた近似的問合せ処理は，古くから継続的に研究されている．サンプリングに基づく近似的問合せ処理（sampling-based AQP, SAQP）では，データのランダムサンプルを作成し，そのサンプルを用いて問合せ結果を推定する図 1 に，SAQP の概念図を示す．例えば，10TB のデータベースから10GBのデータをサンプリングしておき，近似的な問合せ処理ではそのサンプルを用いて集計関数を計算し，近似的な結果として返す．

ランダムサンプリングは，データベースの近似的問合せ処理において頻繁に使用される主要な技術の一つであり，他に層化（stratified）サンプリングもしばしば用いられている．最近のAQP システムの事例としては，一般的なアドホック問合せの近似的処理をサポートすることを目的とした BlinkDB [4]や

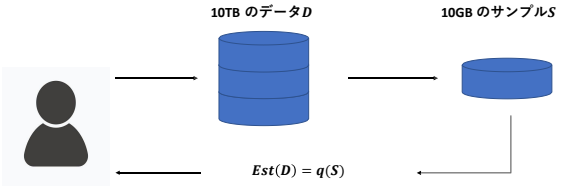

図1　サンプリングベースの近似的問合せ処理の例

VerdictDB [5] がある．これらのシステムは，集約の事前計算（aggregate precomputation）と並列計算などを組み合わせることにより非常に高速な応答時間を実現できるが，複雑な述語の処理能力は十分ではない．

サンプリングにより複雑な問合せを支援することは長年の課題である．また，サンプリングに基づく近似的問合せ処理の性能を向上させるために，多くの手法が提案されている．特に，層化サンプリングの活用は注目されており，また，サンプリングのために索引を活用する研究も見られる．

2.2　サンプリングに基づくデータクリーニング

ビッグデータは一般に多様で誤りやノイズを含んでいることから，近年**データクリーニング**（data cleaning）が重要となっており，さまざまな研究が行われている [6]．サンプリングを用いる近似的問合せでは抽出したサンプルの質が重要となるが，データクリーニングにおいても，着目しているデータの品質をどう評価するかが問題となり，目的は異なるものの技術的に近いアプローチを取ることがある．

特に関連するのが，データクリーニングのためにサンプリングを活用するSampleClean である [7]． SampleClean のフレームワークでは，最初に元データのサンプルが生成され，オラクル（oracle）として扱われるデータクリーニング関数を呼び出し，サンプルをクリーニングする．問合せ処理では，クリーニングされたサンプルを使用して集約問合せに回答するが，信頼区間を伴う推定値を提供する． [7]では，このフレームワークのもとで，集約問合せの処理において，誤りや重複を含むデータベースに対してその問題を軽減できることが示された．

3　機械学習の応用

近年，データベースシステムでの機械学習の使用に関する研究が急速に進展している [8]．本研究に関連が深いものとしては，問合せ最適化で用いられる選択率（selectivity）を推定するために機械学習を用いるものが挙げられる（例： [9]）．選択率の推定は近似的問合せと関連が深く，機械学習の利用により，複雑な述語や問合せへ対応できる可能性がある．

先に述べた，データクリーニングのためのSampleCleanのア

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

Copyright © 2020 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

イデアを近似的問合せに用いたものとして，[10] の “Learning to Sample” のアプローチがある．この研究では，複雑な述語を含むカウント問合せを，サンプリングを用いて近似的に支援する問題について考えている．近年，量化学習（quantification learning) [11]の研究が進展しているが，その中に対象オブジェクトが条件を満たすか判定するための分類器の学習を行い，その分類器を用いてカウントを推定する手法がある．そこでは，対象となる複雑な述語を満たすか否かを判定するために，機械学習を適用し分類器を構築する．[10]では，SampleClean で用いられたアイデアを適用し，サンプルに対するカウント値の推定をもとに，データベースに含まれるサンプル以外のオブジェクトについて条件を満たすカウント値の推定を行っている．

4 研究のアイデア

本節では，機械学習を用いた近似的問合せ処理問題について，研究のアイデアについて述べる．

問合せ対象のデータベースを D とし，そこから抽出されたサンプル集合を S とする．また，データベースに問合せとして適用される述語 $q : D \rightarrow \{0, 1\}$ を考える．実際には，述語は SQL の WHERE 句における複雑な条件や，ユーザ定義関数に基づく述語などを想定する． SQL の単純な検索条件に対しては従来からの手法が適用できることから，対象とはしない．

具体的な問合せの例として，[10]で用いられた，点オブジェクトのデータベースに対する以下の問合せが考えられる．

```
SELECT COUNT (*) FROM
 (SELECT o1.id FROM D o1, D o2 WHERE o2.x >= o1.x
 AND o2.y >= o1.y AND (o2.x > o1.x OR o2.y > o1.y)
GROUP BY o1.id HAVING COUNT (*) < k)
```

この問合せは，スカイライン問合せを一般化したk-skyband 問合せ [12] であり，その点を支配するような他の点の数が k 未満であるような点を求める．ここでは最終的にはカウントを行っているため，k-skyband オブジェクトの個数を求める集約問合せとなっている． k-skyband 問合せを処理するアルゴリズムも存在してはいるが，処理に時間がかかることから，近似的でよいのでカウント値を高速に求めたいという要求がある．WHERE 句に含まれる条件が上記の述語 q に該当するが，q を満たすオブジェクト数を推定することは自明ではない．サンプル S を用いた機械学習によりその値を推定する関数 $f()$ を求め，それを活用することにより近似的問合せを実現する．

ここまでの内容については，[10]の方針に従ったものであるが，本研究では以下のような拡張を目指す．

1. カウント問合せ以外の集約問合せへの拡張： [10]ではカウント問合せのみに限定して議論を行っていたが，集約問合せには SUM, AVG など，他の問合せも存在する．それらの問合せに対して一般化を行い，より汎用性を高めることを目標とする．その際，単に精度のよい推定値を求めるだけでなく，値の信頼区間を合せて提示することができれば，より有効であると考えられる．
2. 述語のクラスに対する一般化： [10]のアプローチでは指定された述語に対して学習を行うが，わずかでも異なる述語については学習結果が適用できず，別の学習を行う必要があった．データベースではさまざまな問合せ条件が与えられることから，想定する述語を限定してしまうことは，手法の適用範囲を大きく狭めることになってしまう．ある程度形式を限定した述語のクラスを想定し，その述語に対して学習を行うことができれば，より一般的に近似的問合せを行うことが可能となる．
3. データクリーニングとの連携：誤りやノイズを含んだ大規模なデータベースに対してサンプリングをもとに近似問合せを行うことも，現実的なシナリオとして存在する．サンプリングとデータクリーニングの連携については，SampleClean [7] のようなアプローチがあるが，その近似的問合せへの拡張も考えられる．
4. ユーザの知識の活用：対象のデータベースと想定する問合せ条件によっては，人手により条件を評価する，クラウドソーシングのアプローチも有効であると考えられる．述語の学習においてユーザの知識を導入することについて検討の余地がある．

5 まとめと今後の課題

本稿では，サンプリングに基づく近似的問合せ処理に対する新たなアプローチについて，そのアイデアを述べた．今回挙げた項目について詳細な検討を行い，具体的な手法を提案する．

謝辞

本研究は JSPS 科研費（16H01722, JP19K21530）による

参考文献

[1] S. Chaudhuri, B. Ding, and S. Kandula, “Approximate query processing: No silver bullet,” in *Proc. SIGMOD*, pp. 511–519, 2017.
[2] B. Mozafari and N. Niu, “A handbook for building an approximate query engine,” *IEEE Data Eng. Bull.*, vol. 38, no. 3, pp. 3–29, 2015.
[3] K. Li and G. Li, “Approximate query processing: What is new and where to go? A survey on approximate query processing,” *Data Science and Engineering*, vol. 3, pp. 379–397, 2018.
[4] 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.
[5] Y. Park, B. Mozafari, J. Sorenson, and J. Wang, “VerdictDB: Universalizing approximate query processing,” in *Proc. SIGMOD*, pp. 1461–1476, 2018.
[6] V. Ganti and A. Das Sarma, *Data Cleaning: A Practical Perspective*. Morgan & Claypool, 2013.
[7] J. Wang, S. Krishnan, M. J. Franklin, K. Goldberg, T. Kraskay, and T. Milo, “A sample-and-clean framework for fast and accurate query processing on dirty data,” in *Proc. SIGMOD*, pp. 469–480, 2014.
[8] M. Boehm, A. Kumar, and J. Yang, *Data Management in Machine Learning Systems*. Morgan & Claypool, 2019.
[9] M. Halford, P. Saint-Pierre, and F. Morvan, “An approach based on Bayesian networks for query selectivity estimation,” in *Proc. DASFAA*, vol. 11447, pp. 3–19, 2019.
[10] B. Walenz, S. Sintos, S. Roy, and J. Yang, “Learning to sample: Counting with complex queries,” in *Proc. VLDB Endowment*, vol. 13, pp. 389–401, 2019.
[11] P. González, A. Castaño, N. V. Chawla, and J. J. D. Coz, “A review on quantification learning,” *ACM Comput. Surv.*, vol. 50, no. 5, 2017.
[12] D. Papadias, Y. Tao, G. Fu, and B. Seeger, “Progressive skyline computation in database systems,” *ACM TODS*, vol. 30, no. 1, pp. 41–82, 2005.

Copyright © 2020 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.