

研究のはじめの一步

藤原 靖宏 日本電信電話(株)



〔受賞論文〕

- ・ 大量データストリームの類似探索手法
- ・ 藤原靖宏, 櫻井保志, 山室雅司(日本電信電話(株))
- ・ 情報処理学会論文誌, Vol.48, No.SIG7(TOD33), pp.1-14 (2007)

標記の論文に対して本会論文賞という名誉ある賞をいただき、非常に感激していると同時に今後も情報処理分野の発展へさらに貢献しなければと身が引き締まる思いである。このたび受賞論文についての紹介の場をいただいたが、主に本研究の提案手法を得るに至った基本的な考え方の経緯について述べることにする。研究の着想を明らかにすることにより、データベースの分野のみならず、広く情報処理技術の分野の方々に対して研究におけるひらめきを与えられればと考えたからである。

受賞論文の最大の特徴はそれまでは近似解を求めることに主眼が置かれていたデータストリームの処理において、正確な解を求めることを目的とし、そのためのアーキテクチャおよびアルゴリズムを理論的背景とともに提案したことにある。データストリームとはネットワークなどから高速に流れてくる大量のデータであるが、本研究にとりかかった当時からデータストリーム処理技術の重要性が高まっていた。研究をはじめるとあたり関連研究の調査を行ったが、調査を進めてみるとデータストリームの処理においては解の正確性が求められていないことが分かった。問合せの解を高速に計算するにはデータストリームをメモリ上で処理しなければならないが、サイズが膨大になるデータストリームのすべてをメモリ上に置くことができないため、一般的にメモリ上に置いたデータストリームの特徴量のみから解を計算していたからである。しかしこのデータストリームの処理の考え方には必ずしも納得できないところがあった。アプリケーションを使うユーザの視点から考えたとき、解の正確性が保証されないシステムは許容できない場合があると考えたからである。そこで本研究においては解の正確性を保証することを目標としたのだが、振り返るとこの発想が本研究における技術的な優位性の芽となり育っていくこととなった。

正確な解を求めるためにまず考えたのはオリジナルのデータを利用することである。特徴量からのみ解を計算するから近似解しか求められないのであれば、求めた近

似解の中からさらにオリジナルのデータによって正確な解を求めればよいと考えたのである。しかしデータストリームのサイズは膨大になるためオリジナルのデータをメモリに格納することはできない。そこで本研究では従来のメモリ上の特徴量のみを用いるスキームを拡張し、ディスクにオリジナルのデータを格納し、必要に応じてディスク内のデータも用いるスキームを用いることとした。

ディスクを利用することによって解の正確性を保証することはできたが、この方法にはまだ改善すべき点があった。近似解からオリジナルのデータを用いて正確な解を求めるには高い計算コストが必要だったのである。理論的な説明は省くが、従来提案されていた特徴量の手法による近似解は解の候補でしかなく、正確な解であるかは保証されていなかったのである。そこで近似解が正確な解でもあることを保証する理論的条件を考察し、それに基づき新たな特徴量の手法を提案した。その結果、特徴量のみから正確な解を求めることができるようになり、計算量を低減することができた。

受賞論文では従来研究における考え方に疑問を持ったことを発端に、解の正確性を保証するためディスクを利用することと、近似解の中から正確な解を求めることができる特徴量を利用することを提案したが、これらはすべて関係者の方々との有益な議論の賜物である。特に本研究はそれまでほとんど研究の経験がなかった私が研究所に入り初めて本格的に行ったものであり、研究の基本のみならず会社業務の一般的な進め方もご指導いただいた指導者の方々に心より感謝している。最後に非常に荒削りな論文でありながらそれを評価していただいた査読者の方々およびデータベースシステム研究会の方々に深く感謝する。

(平成 20 年 4 月 29 日受付)

藤原 靖宏(正会員) fujiwara.yasuhiro@lab.ntt.co.jp

平成 15 年早大大学院理工学研究科電気工学専攻修士課程修了。同年日本電信電話(株)入社。DEWS2006 優秀論文賞、平成 19 年度電子情報通信学会論文賞各受賞。電子情報通信学会、日本データベース学会各会員。