NEWS RELEASE

平成 29 年（2017 年）9 月 12 日

ビッグデータのクラスタリングがパソコンで可能に
少ないメモリー容量でも高速に処理できる手法を開発

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所（NII）コンテンツ科学研究系 特任研究員 松井 勇佑（まつい・ゆうすけ）、株式会社ドワンゴ メディアヴィレッジ研究開発グループ グループリーダー 大垣 慶介（おおがき・けいすけ）、国立大学法人 東京大学 大学院情報理工学系研究科 電子情報学専攻 教授 相澤 清晴（あいざわ・きよはる）、同 准教授 山崎 俊彦（やまさき・としひこ）の研究グループは、データ処理の基本操作であるクラスタリングを、10 億個程度のビッグデータに対して、高速で、かつ、少ないメモリー容量で実行できる実用性の高い手法を開発しました。これにより、例えばソーシャルメディアの膨大な画像データを一般的なパソコンでも手軽に処理することが可能となります。一般の技術者や研究者にもビッグデータの扱いが容易になるため、深層学習を応用した人工知能（AI）の開発をはじめとする広い分野での活用が期待されます。

本手法では、データを圧縮した状態でクラスタリングを行うため、従来手法よりも少ないメモリー容量で処理が可能になりました。さらに、似たデータを集めたグループの「平均」を効率良く計算する新技術を考案したことで、処理の高速化を実現しました。クラスタリングの基本的手法の一つである k 平均法に対して、精度は劣るものの、10～1000 倍程度高速化し、100～4000 倍程度の省メモリーとなります。

本研究成果は、マルチメディア分野のトップ国際会議「ACM International Conference on Multimedia 2017」（10 月 23 日～27 日、米カリフォルニア州マウンテンビュー）で発表されます。また、論文（PQk-means: Billion-scale Clustering for Product-quantized Codes）は 9 月 14 日に計算機科学などの論文を保存・公開するウェブサイト「arXiv（アーカイブ）」（https://arxiv.org/）に先行掲載されます。

《本手法のポイント》

① データを圧縮して処理することで省メモリーを実現
② 似たデータを集めたグループの「平均」を効率良く計算する新技術を考案して処理を高速化
③ これにより、一般的な能力のパソコンでもビッグデータのクラスタリング処理が可能に

本研究成果の一部は、以下の事業・研究領域・研究課題によって得られました。
国立研究開発法人 科学技術振興機構（JST）戦略的創造研究推進事業 ACT-I
研究領域：「情報と未来」※（研究統括：後藤真孝 産業技術総合研究所 首席研究員）
研究課題：「圧縮線形代数：データ圧縮による省メモリ高速大規模行列演算」（グラント番号：JPMJPR16UO）
研 究 者：松井勇佑
※文部科学省の人工知能/ビッグデータ/IoT/サイバーセキュリティ統合プロジェクト（AIP プロジェクト）の一環として運営

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所

株式会社ドワンゴ

国立大学法人 東京大学

国立研究開発法人 科学技術振興機構

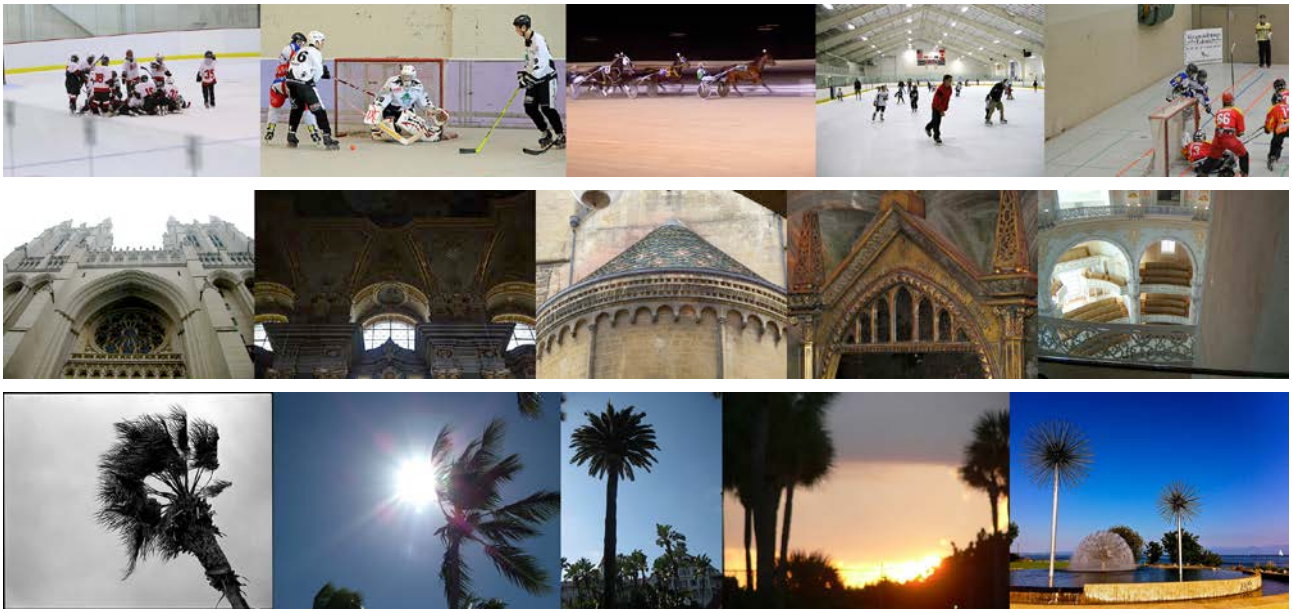
【背景】

AI などの研究においては、巨大で複雑なデータ（ビッグデータ）を処理する必要があります。こうした大量のデータのうち似たものをまとめてグループに分ける「クラスタリング」は、データ処理の最も基本的な作業の一つです。例えば、ソーシャルメディアにアップロードされている膨大な数の画像データを対象に、動物のようなものが写っている写真、街の風景が写っている写真といったグループに分ける処理がクラスタリングです。しかし、枚数が 1 億以上の巨大なデータに対しては、従来の手法では、処理速度が遅すぎたり、必要なメモリー容量が大き過ぎたりして、個人パソコンユーザーが入手・利用できる仕様のパソコン (*1) 1 台ではクラスタリングを実行することは難しいという問題がありました。このため、大規模なクラスタリングを行うためには、多数のサーバーを用いた分散並列処理が必要でした。

【今回開発した技術と成果】

研究チームが開発した手法では、まず、直積量子化 (*2) という技術を用いてデータを圧縮します。これにより、データを従来手法よりも少ないメモリーで表現することができます。次に、圧縮されたデータに対して、「似ているデータを集めてグループを作る」「グループの『平均』を計算する」という処理を繰り返します。今回は、グループの平均を効率的に計算する技術を新たに考案し、これにより、高速なクラスタリングが可能になりました。似ているデータを集める手法については、松井が過去に提案した技術 (*3) を利用しています。

この手法を用いることで、例えば、画像データセット「Yahoo Flickr Creative Commons 100M（YFCC100M）」の 1 億枚の画像を対象に、「氷上のスポーツ試合」や「欧風の教会」、「ヤシの木」など 10 万種類のグループに分類する処理を、個人が一般に入手できる高性能機種の仕様であるメモリー容量 32GB、CPU のコア数 4 のパソコン 1 台でも約 1 時間で実行できました。これを一般的なクラスタリング手法を用いて同じ所要時間で実行しようとすると、同じ仕様のパソコンが約 300 台必要になります。また、10 億の画像データを 10 万種類のグループに分ける処理も約 12 時間で実行できました。



〈図〉1 億枚の画像をクラスタリング処理した結果の例。似た画像がまとめられている（上から、「氷上のスポーツ試合」グループ、「欧風の教会」グループ、「ヤシの木」グループそれぞれの画像の一部）

さらに、本手法を最新の既存手法である「Binary k 平均法」(*4) などと比較した場合、以下のような利点があります。

(1) 本手法は、クラスタリング終了後にデータを近似的に復元できます。多くの既存手法は高速化を実現するために元データを不可逆的に大きく変形してしまうため、クラスタリングが終わった後に元データを復元できず、クラスタリング結果を解釈したり、別の処理に利用したりすることができないという弱点がありました。本手法はこの問題を解決しています。

(2) 本手法はシンプルであり、煩雑な設定が必要ありません。多くの既存手法は使うデータに応じてチューニングする必要がありましたが、本手法は何も設定することなく簡単に使うことができます。

研究チームでは、従来の最近傍探索分野の研究で使われているデータセットの中で最大規模と言われている 10 億個のデータ (*5) を目標として研究に取り組み、今回、その規模のクラスタリングを実現したことを一つのマイルストーンと考えています。

【今後の展望】

クラスタリングは大規模なデータを扱う際に最も基本的で最初に行う処理です。10 億個規模の大規模データを一般的な能力のパソコンでも手軽に扱えるクラスタリング手法を新たに提案できたことは、大規模なデータ処理に日常的に取り組むエンジニアや研究者にとって有益であると考えられます。また、マイコンなどメモリー容量があまりない IoT（モノのインターネット）のエッジデバイスにおいてもクラスタリング前処理を行えるようになるなど、IoT 時代のデータ処理にも有効だと考えられます。

研究チームでは、ビッグデータ処理に関わるすべてのエンジニアや研究者の利用に供するため、本手法のコードを 9 月 14 日より、以下の URL で一般公開する予定です。

https://github.com/DwangoMediaVillage/pqkmeans

以上

本件は NII、ドワンゴ、東京大学、JST が共同で発表するものです。文部科学記者会、科学記者会、東京大学記者会を通じて各加盟メディアの皆様に資料提供しているほか、各機関・企業から関係するメディアの方々に個別に本リリースをお送りしています。重複して配信される場合がありますことをご了承お願いいたします。

〈メディアの皆様からのお問い合わせ先〉

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
総務部企画課 広報チーム
TEL:03-4212-2164 FAX:03-4212-2150
E-mail：media@nii.ac.jp

株式会社ドワンゴ
広報部
E-mail：dwango-pr@dwango.co.jp

国立大学法人 東京大学
大学院情報理工学系研究科 広報室
TEL:03-5841-8981
E-mail：ist_pr@adm.i.u-tokyo.ac.jp

国立研究開発法人 科学技術振興機構
総務部広報課
TEL:03-5214-8404 FAX:03-5214-8432
E-mail：jstkoho@jst go.jp

（*1）個人パソコンユーザーが入手・利用できる仕様のパソコン：　チェコのサイバーセキュリティー会社「Avast Software」が今年 4 月に発表した、同社の製品から得られたデータに基づく「Avast PC Trends Report」（http://files.avast.com/files/marketing/materials/pctrendsreportjan2017.pdf）によると、メモリー容量 64GB 以上のパソコンを使っているのは調査対象 960 万人中 1 万人以下（0.1%）で、CPU のコア数はデュアルコア（2 個）とクアッドコア（4 個）を合わせて全体の 92%を占めている。

（*2）直積量子化：　事前に、「候補データ」を複数用意しておく。圧縮したいデータに対し、最も近い候補データの「番号」だけを記録する。これにより、データを「番号」だけで表すことが出来る。この考え方を発展させ、データを次元方向に複数個に分割し、上記の番号表現を行ったものが直積量子化である。H. Jégou, M. Douze, and C. Schmid, “Product Quantization for Nearest Neighbor Search”, IEEE TPAMI 2011

（*3）松井が過去に提案した技術：　直積量子化により圧縮されたデータに対し、ハッシュテーブルというデータ構造を用いて高速に似ているデータを探す手法。Y. Matsui, T. Yamasaki, and K. Aizawa, “PQTable: Fast Exact Asymmetric Distance Neighbor Search for Product Quantization using Hash Tables”, ICCV 2015

（*4）Binary k 平均法：　一般的なクラスタリング手法である k 平均法を高速化した手法。Y. Gong, M. Pawlowski, F. Yang, L. Brandy, L. Bourdev, and Rob Fergus, “Web Scale Photo Hash Clustering on A Single Machine”, CVPR 2015

（*5）最大規模と言われている 10 億個のデータ：　「ANN_SIFT1B」データセット（http://corpus-texmex.irisa.fr/）、「Deep1B」データセット（http://sites.skoltech.ru/compvision/noimi/）。