NEWS RELEASE

Inter-University Research Institute Corporation / Research Organization of Information and Systems
National Institute of Informatics

東京大学
THE UNIVERSITY OF TOKYO

September 12, 2017

Toward big-data clustering on a personal computer: New algorithm achieves high processing speeds with small memory

The operation of clustering is a fundamental task in data processing. Now, four researchers—Yusuke Matsui, Project Researcher in the Digital Content and Media Sciences Research Division at Japan's National Institute for Informatics (NII), one of four organizations that constitute the inter-university research institute corporation, the Research Organization of Information and Systems; Keisuke Ogaki, head of the R&D Group at Dwango Media Village Inc.; and Professor Kiyoharu Aizawa and Associate Professor Toshihiko Yamasaki of the Department of Information and Communication Engineering at the University of Tokyo—have developed a high-speed clustering technique, using only small amounts of memory and applicable to a wide variety of problems, that can handle big data sets containing as many as 1 billion elements. The new algorithm allows enormous volumes of data, such as collections of images from social-media sites, to be processed easily even on a typical personal computer. By allowing any engineer or researcher to handle big data with ease, the new technique is expected to find applications to a wide variety of problems, particularly those involving the development of artificial-intelligence (AI) methods based on deep learning.

Because the algorithm performs clustering on compressed data, it uses less memory than previous methods. Moreover, the algorithm accelerates the clustering process by introducing a new technique for efficiently computing *means* of groups of similar data items. Compared to the *k*-means method, a standard approach to clustering, the new algorithm achieves 10–1000-fold acceleration with 100–4000-fold less memory requirement, though with reduced accuracy.

The results of this research were announced at the ACM International Conference on Multimedia 2017, the most prestigious conference in the field of multimedia, held October 23-27 in Mountain View, California. The algorithm was described in a preprint titled "PQk-means: Billion-Scale Clustering for Product-Quantized Codes," uploaded on September 14, 2017 to the arXiv (https://arxiv.org/), a public site for preserving and sharing papers in computer science and other fields.

Key features of the new algorithm

1. The operation is carried out on compressed data, reducing memory requirements.
2. A new technique for efficiently computing the *mean* of a group of similar data items accelerates the process.
3. The algorithm allows clustering of big data on a typical personal computer.

Some of the research results reported here were obtained as part of the following project, research domain, and research topic:
Japan Science and Technology Agency (JST), Strategic Basic Research Program, Advanced information & Communication Technology for Innovation
Research domain: "Information and the future" (Prime Senior Researcher: Masataka Goto, National Institute of Advanced Industrial Science and Technology)
Research topic: "Quantized Linear Algebra: Fast and Memory-efficient Matrix Computation by Data Quantization" (Grant number: JPMJPR16UO)
Principal Investigator: Yusuke Matsui
Operated as part of the AIP project of Ministry of Education, Culture, Sports and Technology, a comprehensive project involving artificial intelligence, big data, IoT, and cybersecurity.

Research Organization of Information and Systems
National Institute for Informatics
The University of Tokyo

Dwango Co., Ltd.
Japan Science and Technology Agency

Background

Research in artificial intelligence and similar fields frequently requires the ability to process gargantuan quantities of complex information, that is, big data. The operation of *clustering*, which simplifies huge data sets by categorizing similar items into groups, is one of the most fundamental procedures in data processing. One example of a clustering operation would be a procedure for separating huge numbers of images, perhaps uploaded to a social-media site, into groups such as *images containing animals* and *images showing urban landscapes*. However, for truly enormous data sets, containing 1 billion or more elements, existing algorithms for clustering run too slowly and require too much memory to be executed on a standard personal computer(*1) of the sort available to typical PC users. For this reason, clustering of large-scale data sets has generally required distributed parallel-processing operations carried out on a large number of servers.

The new technique and what it has accomplished

The new algorithm developed by the research team begins by using a method known as product quantization(*2) to compress the data. This allows the data to be expressed with less memory than existing methods require. Next, the compressed data are subjected repeatedly to a two-step process: collect similar data items to form groups and compute *mean* values of groups. For this purpose, the team invented a new technique for computing group means efficiently, allowing high-speed clustering. The method used by the algorithm to collect similar data items was previously proposed by Matsui(*3).

To illustrate the power of the algorithm, the researchers applied it to the Yahoo Flickr Creative Commons 100M (YFCC100M) data set of 100 million images, grouping images into 100,000 categories such as *sporting competitions on ice*, *European-style churches*, and *palm trees*. Running on a high-performance personal computer of the sort readily available to individuals, with 32 GB of memory and 4 CPU cores, the algorithm completed in approximately 1 hour. To execute the same task in the same time using existing clustering methods would require approximately 300 personal computers of the same specifications. A second test, in which 1 billion images were categorized into 100,000 groups, required approximately 12 hours to execute.

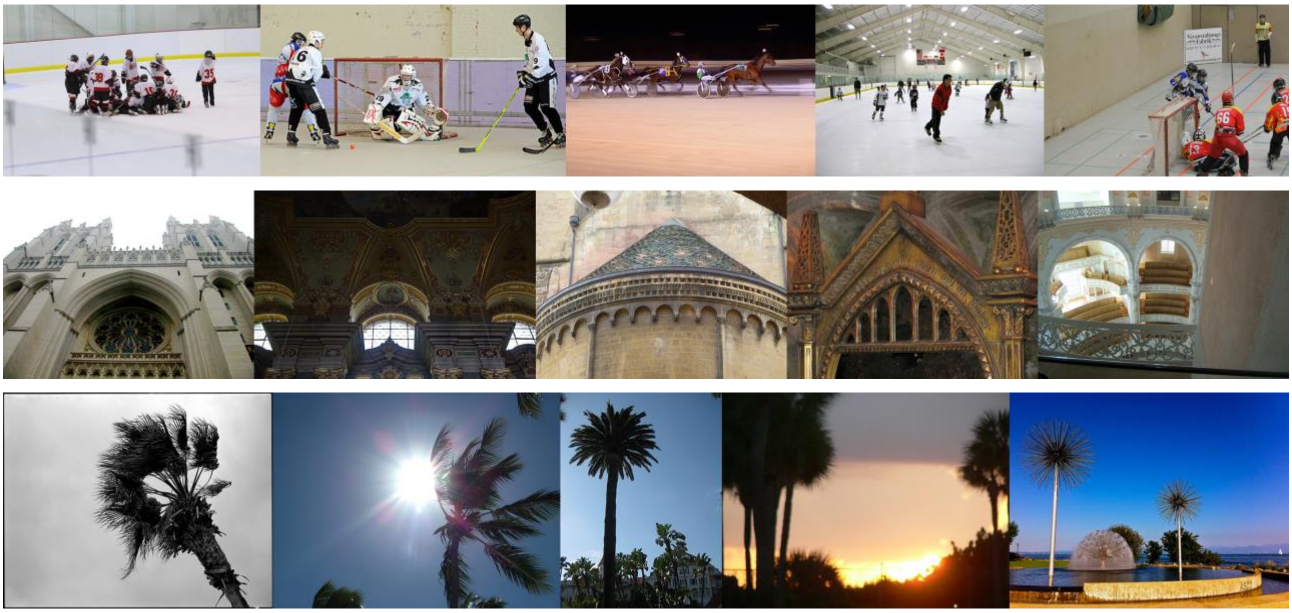

Figure: Sample results obtained by clustering 100 million images. The algorithm constructs groups containing similar images, in this case, *sporting competitions on ice* (top), *European-style churches* (center), and *palm trees* (bottom). (Each row here shows a subset of the images identified by the algorithm as belonging to the corresponding group.)

Compared to the binary *k*-means method(*4), one of the most recent existing methods, the new algorithm offers several advantages:

(1) The new method allows data to be approximately recovered after clustering. Many existing methods transform the original data in major, irreversible ways to increase processing speeds; this has the drawback that the original data cannot be recovered after clustering, preventing users from interpreting the results of the clustering operation or using them for further processing steps. The new algorithm solves this problem.

(2) The new method is simple and requires no complicated configuration of parameter settings. Many existing techniques require tuning depending on the type of data to be processed; in contrast, the new algorithm requires no parameter settings, making it particularly easy to use.

The research team set its sights on tackling data sets with 1 billion entries, considered an upper limit on the size of data sets used to date in research on nearest neighbor search(*5), and the successful development of an algorithm capable of working at this scale may be seen as a milestone in the evolution of clustering algorithms.

Future directions

Clustering is the first and most basic operation used to process large-scale data sets. The proposal of a new clustering algorithm that allows data sets with a billion elements to be handled easily on a standard PC should be of great value to all engineers and researchers who work regularly with large data sets. The new algorithm may also enable embedded microcontroller systems with minimal memory capacity, such as edge devices in the Internet of Things (IoT), to perform clustering-based preprocessing, making it a valuable data-processing tool for the IoT era.

To make the algorithm available to all engineers and researchers involved in big-data processing, on September 14, 2017, the research team released the source code of its implementation via the following URL.

https://github.com/DwangoMediaVillage/pqkmeans

(End of document)

Media Contacts:

National Institute for Informatics
Communications
TEL: +81(0)3-4212-2164 FAX: 4212-2150
E-mail：media@nii.ac.jp

The University of Tokyo
PR office
Graduate School of Information Science and Technology
TEL:+81(0)3-5841-8981
E-mail：ist_pr@adm.i.u-tokyo.ac.jp

Dwango Co., Ltd.
PR department
E-mail：dwango-pr@dwango.co.jp

Japan Science and Technology Agency
PR Division
Department of General Affairs
TEL: +81(0)3-5214-8404 FAX: 5214-8432
E-mail：jstkoho@jst go.jp

(*1) A standard personal computer of the sort available to typical PC users: According to the April 2017 Avast PC Trends Report (http://files.avast.com/files/marketing/materials/pctrendsreportjan2017.pdf) prepared by Czech cybersecurity firm Avast Software**,** fewer than 10,000 of 9.6 million PC users surveyed (0.1%) used machines with 64 GB or more memory, but 92% of users had dual or quad-core (2- or 4-core) machines.

(*2) Product quantization: Quantization is a coding technique in which multiple *candidate data* items are prepared and numbered in advance; then, for each element in the data set to be compressed, only the *number* of the closest candidate item is stored. This allows data items to be represented solely by their *numbers*. The method of product quantization extends this idea by subdividing data items into multiple dimensions and applying quantization to each dimension. Reference: H. Jégou, M. Douze, and C. Schmid, “Product Quantization for Nearest Neighbor Search”, IEEE TPAMI 2011.

(*3) A technique previously proposed by Matsui: This method uses a data structure known as a hash table for rapid detection of related items in data sets compressed by product quantization. Reference: 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*-means method: An accelerated version of the *k*-means method, a common clustering algorithm. 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) Considered an upper limit on the size of data sets used to date in research on nearest neighbor search: See, for example, ANN_SIFT1B (http://corpus-texmex.irisa.fr/) and Deep1B (http://sites.skoltech.ru/compvision/noimi/).