圧縮・復元技術の最前線

センシングデータのスマート化に不可欠な技術とは

センシングデバイスから収集される膨大なデータを効率的に蓄積、かつ、高速で処理するためには、圧縮・復元技術が不可欠となる。だが、従来の圧縮・復元技術は、その処理を行うに際して、さまざまな課題を有していた。そうした課題を解決するため、より効率的な圧縮・復元技術の実現に向け、NIIの定兼邦彦准教授と北海道大学大学院の喜田拓也准教授らによる共同研究が進められている。

センシングデータの効率的な圧縮・復元が課題に

近年、ビッグデータの中でも、GPS（全地球測位網）やカメラ、携帯電話、スマートフォンなどのデバイスから得られる「センシングデータ」に注目が集まっている。CPS（サイバー・フィジカル・システム）の実現には、それらの多様なデバイスから連続的に収集される膨大なデータを効率的に蓄積するとともに、自由に取り出してコンピューター上で活用できる仕組みが不可欠だ。そして、そのためにも、より効果的な圧縮・復元技術が求められている。

しかし、大量のデータを効率的に処理したり、高速に検索したりするためには、従来の圧縮・復元技術には限界があった。NIIの定兼邦彦准教授は、「大量のデータを高速処理するにあたり、これまでの圧縮技術では、圧縮データをコンピューターのメモリ上でそのまま読み書きできないという課題がありました。また、特定のデータへのランダムアクセスについても、非圧縮データであれば容易に行えますが、圧縮データでは不可能なため、先頭から1つずつ復元しなければなりません。対して復元の時間を短くするには、データを複数のブロックに分割し、個々に圧縮しなければならず、結果、圧縮率の低下を招いていました」と説明する。

北海道大学大学院の喜田拓也准教授も、「データ量が少なければ、圧縮や復元を行うことなくそのままメモリ上で処理できます。しかし、センシングデータのような膨大な量の情報を圧縮・復元せずに高速に処理するためには、大容量のメモリを搭載した大規模なコンピューター環境が必要となります。また、データを保管しておくためのハードディスクの容量もおのずと増大してしまいます。そうしたことから、従来はある一定の近似値でデータをまとめたり、平均値だけを保存したりするといった手法がとられていましたが、その過程で間引かれたデータに存在する知見を見過ごしてしまうということになってしまいます」と補足する。

「このような課題の解決に向け、喜田先生、九州大学の竹田先生とともに効率的な圧縮と復元を可能とする基盤技術の研究を、2012年より共同で進めてきました」（定兼准教授）

喜田拓也 Takuya Kida
北海道大学　大学院情報科学研究科
コンピュータサイエンス専攻　准教授

VF符号とRe-pairアルゴリズムによる圧縮を推進

では、効率的な圧縮・復元の実現に向け、現在、どのような共同研究が進められているのか。喜田准教授が手掛けているのが、「VF符号（Variable-length-to-Fixed-length code）」と「Re-pairアルゴリズム」を用いた圧縮・復元である。VF符号とは、元のデータにおける長さの異なる部分系列に対して、固定長の符号を割り当てる符号化方式だ。

「例えば、あるデータの1つの塊が5bitで表現されている場合、圧縮されているデータの中から5番目の符号語で表現されている情報だけを抽出したい場合、21番目～25番目を取り出

図1：従来のデータ構造の問題点

データのサイズが大きい	索引サイズが大きい
●**圧縮するとランダムアクセスが**できない（遅い） ●検索を行うには索引を追加する必要がある	●メモリに載らない ●ディスクに置く→低速処理 ●索引の情報量を落としてサイズ削減→精度低下

簡潔データ構造

従来の索引（接尾辞配列）＋データ	680.4G
簡潔データ構造（索引＋データ）	21.6 G

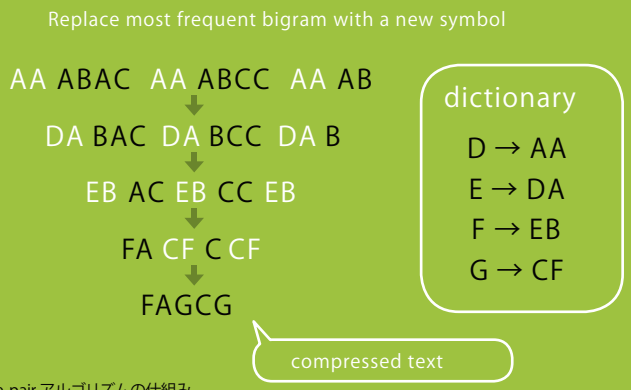

図2：Re-pair アルゴリズムの仕組み

せば済むわけです。これが可変長の符号語で圧縮されたデータでは、前から順番に見ていかなければなりません」（喜田准教授）

このように、符号語の境界が明確であることから、圧縮されたデータへのアクセスが容易というメリットをもつが、VF符号は可変長符号に比べて圧縮率を向上させることが難しい、という側面があった。喜田准教授は「そこで、Re-pairアルゴリズムを用いた文法変換に基づく圧縮方法を組み合わせることで、圧縮の効率を保ちながら、データを復元することなしにキーワード検索等の処理を高速に行える仕組みを実現しています（図2）」と説明する。この組み合わせにより、一般的な圧縮ツールである「gzip」を上回る圧縮率が実現されている（図3）。

簡潔データ構造による新しい圧縮・復元法

一方、定兼准教授は、「簡潔データ構造」に基づいた圧縮・復元技術の研究を推進している。

「私は、簡潔データ構造と呼ばれる新しい圧縮法を研究していますが、その簡潔データ構造において、文字列を圧縮したまま高速に検索できる圧縮接尾辞配列を開発しました」と、定兼准教授は話す。

「文字列の検索用の索引をコンピューター上に実装しようとした場合、余分なデータを付加してしまうことが多く、結果、索引が元データよりも大きくなってしまうことが多々あります。この問題を解決するために圧縮接尾辞配列が提案されていますが、検索にはテキスト自体が必要となるため、索引のサイズがテキストよりも小さくならないという課題がありました。そこで、私は圧縮接尾辞配列を用いた検索アルゴリズムにおいてテキスト自体が不要になるように変更を加えたり、テキスト全体やその一部を圧縮接尾辞配列から復元するアルゴリズムを用いることを勘案しました。これにより、テキストや索引の圧縮に加え、任意の語句の検索や文書の任意箇所の部分復元を実現しました」（定兼准教授）

喜田准教授は、「定兼先生が研究している簡潔データ構造を活用した圧縮・復元技術を用いることで、メモリ上に乗せられるほどデータをコンパクト化できるため、データアクセスが容易になります。すなわち、大容量メモリを搭載した大規模なコンピューティング環境を用意せずとも、膨大なデータの処理や分析が可能となるわけです」と強調する。

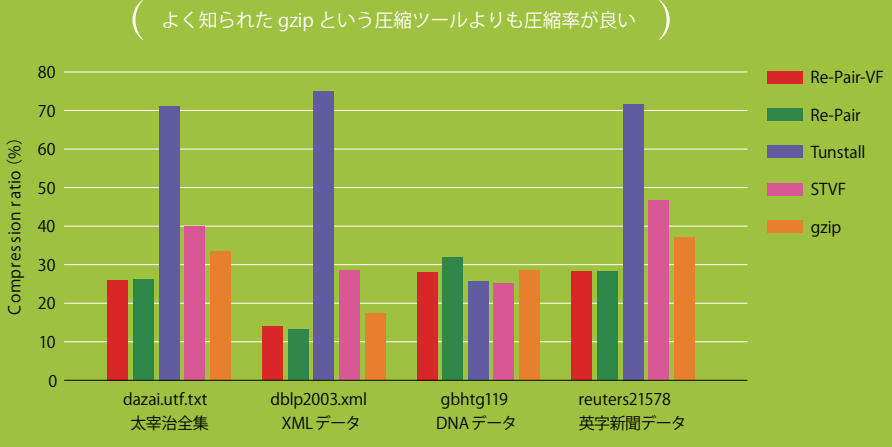

図3：Re-Pairアルゴリズムと VF符号化の組み合わせにより、圧縮率が向上

現在、テキストデータを対象とした圧縮・復元の研究が共同で進められており、期待される実用例の1つとして、DNAやヒトゲノムのデータ処理などが挙げられている。また、今後はセンサーデータへの適用も進めていき、まずは、車載位置情報システムから収集されるデータの圧縮・復元に着手していくことを予定しているという。

新しい圧縮・復元技術により、センシングデータの処理がより高速に行えるようになれば、膨大なデータの中に埋もれ、これまで見過ごされてきたようなさまざまな知見が得られるようになるだろう。その実現に期待が寄せられる。

（取材・文＝伊藤秀樹）

定兼邦彦 Kunihiko Sadakane
国立情報学研究所　情報学プリンシプル研究系　准教授
総合研究大学院大学　複合科学研究科　情報学専攻　准教授