

簡潔データ構造による 高速ビッグデータ処理

情報学プリンシプル研究系 准教授

定兼 邦彦



研究背景・目的

ビッグデータ処理での問題点は、データの読み書き速度と処理に必要な、計算機のメモリ量です。これらを改善する方法として考えられるのはデータを圧縮することですが、古典的な圧縮法ではうまくいきません。なぜなら、圧縮されたデータは自由に読み書きできないため、目的の処理を行えなくなるからです。また、複雑なデータ処理を行う場合にはデータに索引情報を追加して高速化することがよく行われますが、ビッグデータに対してはその索引のサイズも問題となります。例えば100GBの文書データに全文検索のための索引を追加するとサイズは合計で600GBになり、計算機のメモリ上に置けずディスクに格納する必要があり、速度が大幅に低下してしまいます。もしデータや索引がうまく圧縮できればメモリに格納でき、高速処理が行えます。そのような圧縮法を開発するのが研究の目的です。

研究内容

本研究ではビッグデータを圧縮したまま高速に処理できる新しい圧縮法を開発しています。これは簡潔データ構造と呼ばれます。文書の全文検索のための簡潔データ構造として圧縮接尾辞配列を開発しました。上述の文書

データに適用すると600GBの文書とその索引が22GBに圧縮でき、任意の語句の検索と文書の任意箇所の部分復元を高速に行えます。DNA配列に対しては、配列のアセンブリ(短い部分配列を繋いで全体を復元する処理)で用いられるグラフ構造に対する簡潔データ構造を開発しました。これを用いることで、これまで複数の計算機を用いて行っていたアセンブリ処理を1台の計算機で行うことができます。また、様々なデータ検索で用いられる木構造の簡潔データ構造では、サイズを50分の1以下にできます。木構造の動的な変更も可能です。さらに、汎用的な圧縮法として、データを圧縮したまま高速に読み書きできるものを開発しました。1バイトの読み書きにかかる時間は定数で、データ量には依存しません。これは透過的データ圧縮法と呼ばれ、世界初の技術です。

産業応用の可能性

- ビッグデータ処理を安価な計算機で行える
- モバイル端末、組み込み機器などでの省メモリデータ処理
- ゲノムアセンブリへの適用

