文脈を考慮した文書分類
- 大規模データを深く解析するためのライブラリ群

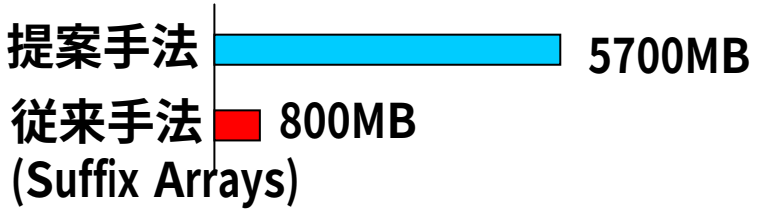
岡野原　大輔　(東京大学情報理工学系コンピュータ科学専攻)

文書分類、情報抽出、機械翻訳、ゲノム解析など多くの場面で必須となる部分列計数、文書頻度を初めとした多くの文字列処理を数GBから数100GBといった大規模なデータに適用可能なライブラリ群を開発した。

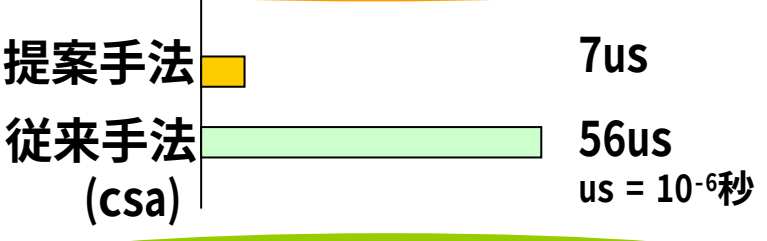
従来手法と比較し、符号法、データ構造等を最新の研究結果を基に理論的、工学的な視点から改良することで約10〜100倍のデータを処理可能となった。

これにより、従来のPC上でゲノム比較や大規模データを基にした機械学習等、今までは大規模クラスタ上でのみ可能だった処理が実行可能となる。

メモリー4GBで処理可能なデータサイズ



パターン(10文字)検索の所要時間



他手法との特徴比較

	単語検索	部分列検索	高度なアプリケーション	必要領域量
転置ファイル	○	×	△	○
Suffix Trees	○	○	○	×
提案手法	○	○	○	○