学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　　氏名　花田　博幸

審査担当者　主　査　教　授　工藤 峰一
副　査　教　授　有村 博紀
副　査　准教授　喜田 拓也

学位論文題名

The q-gram Distance as an Approximation of the Edit Distance
（編集距離の近似としての q-gram 距離の利用）

審査要旨

近年, インターネット上では文字列データが日々増加し続けている. これらの多くは, 辞書や教科書にみられる「専門家による知」ではなく個々人の生活の知恵や経験あるいは日記を集めた「民衆による知」の集積であり, 新しい文化と価値観を生み出している. さらに, 時系列として観測される脳波などのバイタルデータや遺伝子情報の解析結果なども近年異常な速度で蓄積される文字列型のデータである. 文字列データの活用においては, 所望の情報を得るための「検索」や類似性の発見・解析のための「照合」が主に行われる. また, 実用上は “ 完全一致 ” での検索や照合は望めないことが多いため “ 近似文字列照合 ” は重要な技術的要請となる. 中でも, 文字の削除や挿入, 置換の回数に基づく “ 編集距離 ” は応用面での整合性の良さからよく使われる距離である. 本研究は編集距離の各種近似手法に関して精度面での実用的評価および高速化を行ったものである. 特に,q-gram 距離を中心に検討している.

編集距離は比べる二つの文字列の長さの積の計算量を必要とするため, 繰り返して何度も計算する場合やどちらかがひどく長い場合などに計算量の削減が求められる. そこで, 編集距離そのものではないもののその近似値を与える距離がこれまでに多く提案されており, 中でも二つの文字列長の和の計算量で済む q-gram 距離が多く検討されている. 一方で, それらの近似精度は漸近的にのみ評価されている場合が多く, 実用的な効果が判然としていないという問題があった, 本論文の前半部分 (第 4 章) では,q-gram 距離を含む幾つかの編集距離の代理 (近似) 距離に関して, 統一的かつ有限での評価を行うことでこの問題の解決を計っている. 本論文の後半部分 (第 5 章) では, もう一つの問題として, 長大なテキスト文字列の中から所与のパターン文字列に所与の距離以下で近い部分文字列をすべて数え挙げる問題を扱っている. ここでは q-gram 距離を用いたときの速度に関してこれまで知られている最良のアルゴリズムを改良している.

編集距離の代理距離に関する実用的な近似精度の評価 (第 4 章) については二つの貢献を行っている. 一つ目は評価方式の統一であり, もう一つはアルゴリズムの再 (詳細) 評価である. これまでの各種代理距離に関する精度評価は不等式での上下限評価とラージ・オーの記法による漸近的な評価の二つであった. 本論文ではこの二つの評価を統一する方法を与えている. 具体的には,distortion と呼ばれる編集距離の定数倍の上下限の比を用いている. これまでの不等式評価式に対しては閾値 θ 以上の編集距離のみ評価することで distortion を求め, 漸近評価式についてはアルゴリズムの詳細評価により定数を含めたラージ・オーに依らない distortion の値を求めている. これにより, 従来行

えなかった統一的かつ具体的な代理距離の精度比較を行っている. さらに,q-gram 距離型の各種代理距離の実用性を明らかにしている.

テキスト中の近似文字列数え上げ問題における速度の改善 (第 5 章) についてはこれまで知られている最良の q-gram 距離計算アルゴリズムを改良している. このアルゴリズムは,O(n log k + p)(n はテキスト長,p はパターン長,k は近似精度の制約) の平均および最悪計算量を必要としていた. 通常,k は p に比例するため, 結局 O(n log p) となる. 本研究では最悪計算量は変わらないものの平均計算量を O(n+p) に改善している. これまでは探索範囲の限定ならびに探索開始位置を一つ進めたときの距離情報の逐次更新が効率化のアイディアであった. 本研究ではさらに, ベースラインとオフセットによる距離情報更新の効率化と必要時のみの距離情報更新というアイディアを提案している. 実現のために, これまでのデータ構造である「連結リスト」を「木と配列」の混合データ構造に取替えている. 更新必要事象がそれほど頻繁には起きないことに着目して確率的な評価を行うことで平均計算量を大幅に削減している. 実際に計算機シミュレーションにおいて効果を確認している.

これを要するに, 著者は, 近似文字列照合においてよく用いられる編集距離に着目し, 編集距離を近似する q-gram 距離の実用範囲における理論的な精度評価を世界で初めて与えるとともに, 文字列検索問題に q-gram 距離を適用する際の最良のアルゴリズムをさらに改良することで, 文字列検索および文字列照合技術を大きく向上させた. よって著者は, 北海道大学博士 (情報科学) の学位を授与される資格あるものと認める.