学　位　論　文　内　容　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

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）

コンピュータ上で扱われる文字列データの量は増大の一途を辿っており, それにより文字列データを分析する技術の重要性も増大している. 例えば文書データはもちろんのこと音声や動画といったデータも文字列の一種と考えることができる. このようなデータがインターネット上で急速に増加していることを受けて, 文字列データを検索・分析するための研究もますます重要性を増している. また生物学分野においては, ゲノム配列をゲノムの分子から取得するシーケンシング技術が向上したことで, 次々に生成される配列データに対して機能解析や分析が追いつかない状況にある.

文字列を分析する有用な技術の一つとして, データベース中の文字列を完全一致のみならず類似するものまで含めて検出する技術が挙げられる. その類似度指標の代表例が「編集距離」である. 編集距離は, 二つの文字列を文字の挿入・削除・置換という編集操作をもって一致させるための最小の操作回数と定義される. 編集距離の概念は 1960 年代に登場して以来, 多くの研究がなされている. その理由の一つとして, 文書データにおける誤字・脱字, 音声データにおけるノイズ, ゲノム配列における突然変異による変化などを表すのに親和性が高く, それらの変化を考慮した検索において編集距離を基準とすることが有効であることが挙げられる. 編集距離に対するもう一つの研究の視点として, 時間計算量の削減が挙げられる. 二つの文字列 x, y に対する編集距離の時間計算量は $O(|x||y|)$($|\cdot|$ は文字列長) である. データベースからの検索など多くの応用においては距離計算の回数が多大になるため計算量を削減する需要は大きい. 本研究は主にこれを目的とする.

編集距離の計算を高速化する従来の代表的な研究としては, suffix tree や空間分割木などの索引を用いる方法, 編集距離の上限値を設けて計算を高速化する方法, そして文字列中の「文字とその出現位置の組」や「q-gram の出現数」などをもって編集距離を近似計算する方法がある. 本論文ではそのうち, 編集距離を近似計算する方法に着目し, その中でも特に計算量が小さい「q-gram 距離」について様々な角度から検討を行った.

q-gram 距離は, 二つの文字列に含まれる異なる q-gram(長さ q の部分文字列) の個数を測るものである. 二つの文字列 x, y に対する q-gram 距離は時間計算量 $O(|x| + |y|)$ で計算でき (q には依存しない), 編集距離に比べて非常に高速である. また編集距離を q-gram 距離で上下界付ける式もこれまでにいくつか得られている. しかしながらこれらの従来研究には, 実際の応用における有用性評価という観点で二つの課題がある. 一つ目は, q-gram 距離を利用して編集距離を近似する手法が他の手法に比べて精度がよいのか比較できていないこと, 二つ目は q-gram 距離を基準とした類似文字列検索のアルゴリズムに改良の余地があることである. 本論文ではこれら二つの課題に取り組んでいる.

第一の課題は第四章にて扱っている. 従来研究においては, q-gram 距離を利用して編集距離を近似する 2 手法の精度は上記の通り上下界の不等式により与えられている一方, これ以外の編集距離近似手法は近似精度を distortion(近似精度の指標で, 近似による変動範囲を倍率により表す) の文字列長 n に対する漸近評価 (例:$O(n \log n)$) で表しているものが多い. これらを統一的に比較するため, 前者は不等式を変換し, 後者は漸近評価の式を辿り直すことで, これらを漸近評価を含まない distortion に統一した. さらに, これらの手法を実装して実験的にも distortion を評価した. これらの結果, q-gram 距離を利用する方法は n が比較的小さい (300 以下) 場合に distortion が他手法より小さくなる場合が多いことが, 理論評価ならびに実験評価の両方で明らかになった. また実験結果により, アルファベット数が比較的大きい場合にも distortion が他手法より小さくなる場合が多いことが示された.

第二の課題は第五章にて扱っている. 長大なテキスト文字列 t と比較的短いパターン文字列 p に対し, 「q-gram 距離がしきい値 $k(= O(|p|))$ 以下である文字列を検索する」という問題を考察する. この研究では, 従来の最良のアルゴリズムの時間計算量が平均・最悪ともに $O(|t| \log k + |p|) = O(|t| \log |p| + |p|)$ であったのに対し, 従来のアルゴリズムにて利用されていた探索木を索引付き連結リストに置き換えることで, 最悪時間計算量を増加させずに平均時間計算量 $O(|t| + |p|)$ を達成した. また実験により, 計算時間が理論評価の通りの挙動を示すことを確かめるとともに, 計算時間を文字列長 100 の場合で 2 分の 1 程度, 文字列長 500 の場合で 3 分の 1 程度に削減できることが示された.

本論文の貢献は以下にまとめられる. 第一には, q-gram 距離による編集距離の近似精度について, 他の手法との統一的な比較を試みたとともに, 各手法の近似精度が高くなる条件を n により表現できるようにしたことである. 第二には, 類似文字列検索の高速化を行い, これまで示されていた最良の平均時間計算量よりも小さい $O(|t| + |p|)$ 時間を達成したことである.