編集距離空間における1-挿入被覆符号のサイズ限界式

田中 俊介　　泉 泰介

1　導入

類似ジョインとは入力集合から類似ペアを全列挙する問題である．本研究は等長文字列からなる編集距離空間上の類似ジョイン，すなわちアルファベット Σ 上の長さ n の文字列の集合 $S \subseteq \Sigma^n$ を入力とし，$s_1, s_2 \in S$，$ed(s_1, s_2) \leq \epsilon$ となる対 (s_1, s_2) を全出力する問題を，並列処理フレームワークの代表的なプラットフォームである MapReduce 上で処理することを考える [1]．ここで ϵ は閾値パラメータであり，$ed(s_1, s_2)$ は s_1 と s_2 の間の編集距離とする。

MapReduce は mapper と reducer と呼ばれる 2 種類のプロセスにより構成される。mapper は入力データより複数の key-value ペアを生成する。生成された key-value に従って入力データを各 reducer に送信し，reducer は並列処理を行い出力を返す。この処理において複数の mapper および reducer が複平行動作するが，通常，同一キー値を持つ key-value ペアは単一の reducer に集約されることが保障される。本研究では MapReduce 上で類似ジョインを計算するためのアルゴリズムの一つであるアンカーポイントアルゴリズムに注目する。アンカーポイントアルゴリズムでは入力文字列集合 Σ^n を被覆する被覆符号 C（アンカーポイント集合）を事前に用意する．ここで，被覆符号 C とは文字列の集合（長さ n とは限らない）で，任意の文字列 $s \in \Sigma^n$ に対して $ed(c, s) \leq \delta$ を満たす文字列 $c \in C$ が存在するようなものである．アンカーポイントアルゴリズムでは，各入力文字列 $s \in S$ について，自身からの距離が $\delta + \epsilon/2$ 以内に存在する（すなわち，自身の被覆している）C 中のすべての語 $c_1, c_2, \cdots, c_k$ に対して，$(c_1, s), (c_2, s), \cdots (c_k, s)$ を生成する．その後，共通のキー値を持つすべての文字列対にのみ全対比較を行うことで類似ペアを列挙する．アンカーポイントアルゴリズムにおいて，生成される key-value ペアの個数を出来るだけ小さくするためには，被覆符号 C が重複して被覆する領域を出来るだけ小さくすることが望ましい．そのため，被覆効率のよい符号 C の構成可能性，およびその限界を知ることは重要な問題である．しかしながら，被覆の重複量を正確に評価することは容易ではない．そこで，本研究では単純に可能な限り語数の少ない被覆符号を構成することに注目し，特に $\Sigma = 0, 1$ における 1-挿入被覆符号のサイズ限界式について考察する．

2　1-挿入被覆集合

Σ^n 上の 1-挿入被覆集合 C とは，Σ^{n+1} の部分集合であり，任意の $s \in \Sigma^n$ に対して適切な位置に Σ 中の 1 文字を挿入することで C 中にある語に変形可能であるものである．1-挿入被覆符号は Σ^n 上の編集距離空間における $\epsilon = 2$ の被覆符号となる．また，符号語 $c \in C$ が被覆する Σ^n 中の領域は，c から任意の 1 文字を削除して得られる Σ^n 中の語の数に等しいため，その領域サイズは高々$(n+1)$ である．このことから，1-挿入被覆符号のサイズに対する自明な下界は $|\Sigma|^n/(n+1)$ となる．一方で，既知の結果として，サイズ $|\Sigma|^{n+1}/(n+1)$ の 1-挿入被覆符号が存在することが知られている [2].

2.1　1-挿入被覆集合のサイズ

以下に，今回新たに得られた 1-挿入被覆符号のサイズ限界式を示す.

Theorem 2.1 k を以下の不等式を満たす最大の自然数とする．

$$\sum_{i=0}^{n-k} 2\binom{n}{n-i}(n+1-i) \leq 2^n$$

また，上記の条件を満たす k の値に対して決まる左辺の値を v とする．このとき，任意の *1*-挿入被覆符号 C に対して，

$$|C| \geq \sum_{i=0}^{n-k} 2\binom{n}{n-i} + \lceil \frac{2^n - v}{k} \rceil$$

が成立する．

2.2 既存の結果との比較

既存の上下界と今回の限界式を数値計算を用いて比較した．結果を図 1 に示す．(横軸：文字列の長さ n，縦軸：被覆符号の個数)

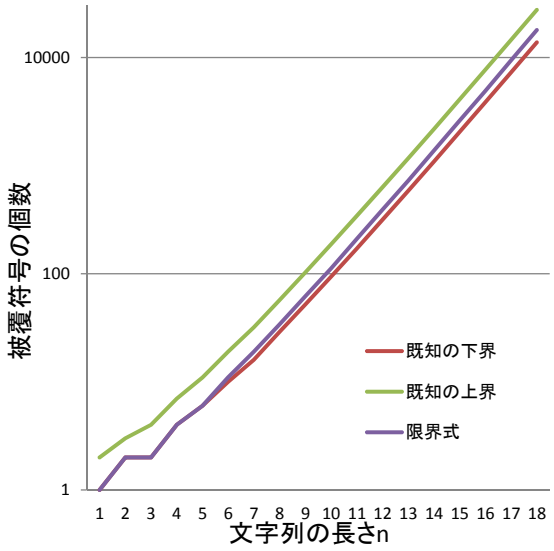

図 1: 既存の上下界と今回の限界式比較

参考文献

[1] F .N. Afrati, A. D. Sarma, D. Menestrina, A. G. parameswaran, and J. D. Ullman. Fuzzy Joins Using MapReduce. In ICDE,2012.

[2] F. Afrati, A. Das Sarma, A. Rajaraman, P. Rule, S. Salihoglu, and J. Ullman. Anchor points algorithms for hamming and edit distance.In Proceedings of ICDT, 2014.