2ZB-3

最小汎化された曖昧な頻出配列パターンの抽出

荒木　康太郎†　　田村　慶一‡　　加藤　智之†　　黒木　進‡　　北上　始‡

広島市立大学大学院†　　広島市立大学‡

1. はじめに

アミノ酸配列やテキスト情報などを含むデータベースに対する配列データマイニングでは，正規表現として表現された頻出配列パターンを抽出する方法が注目されている[1].

本論文では，サフィックス木を用いて，基準となる検索キーPに対して，最小支持数 Mini_sup を満たし，かつ許容誤差半径 r 内にある部分文字列の集合を高速に検索し，検索結果として得られた集合から最小汎化された曖昧パターンの正規表現を効果的に導出する方法を提案する．ここで扱う問題は，以下の 2 つの処理を行う必要がある．

(1) 基準となる k-部分文字列に対して，曖昧検索を行い，許容誤差半径 r 内にある k-部分文字列集合を求める(集合内の各要素は互いに誤差直径 d=2r 内にある).

(2) 検索結果として得られた k-部分文字列の集合に対して，最小汎化された曖昧パターン表現を導出する．

2. 汎化パターン

k−部分文字列集合の部分集合を正規表現によって，1 つのパターンで表現したものを汎化パターンと呼ぶ．ここで，長さ k の部分文字列上の各文字位置に配置される文字の種類を$\sum_1,\sum_2,\cdots,\sum_k$とする．ただし，$\sum_i$は正の k−部分文字列集合の i 番目の文字位置から見つけ出された文字の集合とする(1≦i≦k)．汎化パターンの例として<[ABC]-x(1)-C>が挙げられる．この場合，$\sum_1$=[ABC]であり, A,B,C のどれかの文字が対応することを示す．x(1)はワイルドカード領域である．また，正の部分文字列集合全体に対する汎化パターン$<\sum_1\sum_2\cdots\sum_k>$を最汎パターンと呼ぶ.

3. 高速な曖昧検索

検索キーに対して，最小支持数を満たし，許容誤差半径内にある部分文字列の集合を求めるため，ディスク上のサフィックス木を用いている．DynaCluster アルゴリズム[2]に改良を加え，ディスク上に一般化サフィックス木を構築した．

曖昧検索は構築したサフィックス木を深さ優先にスキャンすることで達成している.

4. 最小汎化された正規表現の導出

ここで，α_iを$\sum_i$の要素とする．長さ k の部分文字列$<\alpha_1\alpha_2\cdots\alpha_k>$をある汎化パターン$<\sum_1\sum_2\cdots\sum_k>$のインスタンス(k-部分文字列)とするとき，$<\sum_1\sum_2\cdots\sum_k>$は以下の 2 要素の汎化パターンとして分解可能である(1≦i≦k).

$$<\sum_1\cdots\sum_{i-1}(\sum_i-\{\alpha_i\})\sum_{i+1}\cdots\sum_k>+<\sum_1\cdots\sum_{i-1}\alpha_i\sum_{i+1}\cdots\sum_k>\} \quad (1)$$

式(1)の右辺第二項を左辺に移項すると，以下のようになる.

$$<\sum_1\sum_2\cdots\sum_{k-1}\sum_k>-<\sum_1\sum_2\cdots\sum_{i-1}\alpha_i\sum_{i+1}\cdots\sum_{k-1}\sum_k>=\sum_1\sum_2\cdots\sum_{i-1}(\sum_i-\{\alpha_i\})\sum_{i+1}\cdots\sum_{k-1}\sum_k> \quad (2)$$

式(1)の分解の組み合わせは，k 通り存在する.

4.1 残差汎化パターン

式(2)の右辺第一項を k 通り集めたものを$<\Sigma_1\Sigma_2\cdots\Sigma_k>$に対する$<\alpha_1\alpha_2\cdots\alpha_k>$の除去により得られる残差汎化パターン集合 RPS($<\sum_1\sum_2\cdots\sum_k>$, $<\alpha_1\alpha_2\cdots\alpha_k>$)と呼ぶ.

$$RPS(<\sum_1\sum_2\cdots\sum_k>,<\alpha_1\alpha_2\cdots\alpha_k>)=<(\sum_1-\{\alpha_1\})\sum_2\cdots\sum_{k-1}\sum_k>+\cdots+<\sum_1\sum_2\cdots\sum_{i-1}(\sum_i-\{\alpha_i\})\sum_{i+1}\cdots\sum_{k-1}\sum_k>+\cdots+<\sum_1\sum_2\cdots\sum_{k-1}(\sum_k-\{\alpha_k\})> \quad (3)$$

ただし，右辺の中で冗長なパターンは除去する.

式(1),(2),(3)をさらに一般化してみよう．ある汎化パターン$<\Gamma_1\Gamma_2\cdots\Gamma_k>$に対して，$\Delta_i=\Gamma\cap\sum_i\neq\varphi$ が成立するとき，他の汎化パターン$<\sum_1\sum_2\cdots\sum_k>$を以下のように分解可能である(1≦i≦k).

$$<\sum_1\sum_2\cdots\sum_{k-1}\sum_k>=<\sum_1\sum_2\cdots\sum_{i-1}(\sum_i-\Gamma_i)\sum_{i+1}\cdots\sum_{k-1}\sum_k>+<\sum_1\sum_2\cdots\sum_{i-1}\Delta_i\sum_{i+1}\cdots\sum_{k-1}\sum_k> \quad (4)$$

式(4)の右辺第二項を左辺に移項すると，以下のようになる.

$$<\sum_1\sum_2\cdots\sum_{k-1}\sum_k>-<\sum_1\sum_2\cdots\sum_{i-1}\Delta_i\sum_{i+1}\cdots\sum_{k-1}\sum_k>=<\sum_1\sum_2\cdots(\sum_i-\Gamma_i)\sum_{i+1}\cdots\sum_{k-1}\sum_k> \quad (5)$$

また，$<\sum_1\sum_2\cdots\sum_k>$に対する$<\Gamma_1\Gamma_2\cdots\Gamma_k>$の除去によって得られる残差汎化パターン集合 RPS($<\sum_1\sum_2\cdots\sum_k>$, $<\Gamma_1\Gamma_2\cdots\Gamma_k>$)は以下のとおりで

Extraction of sequential patterns with least minimum generalization based on ambiguous retrival.

†Kotaro ARAKI・Graduate School of Hiroshima City University

‡Keiichi TAMURA・Hiroshima City University

†Tomoyuki KATO・Graduate School of Hiroshima City University

‡Susumu KUROKI・Hiroshima City University

‡Hajime KITAKAMI・Hiroshima City University

ある．

$$RPS(<\sum_1\sum_2\cdots\sum_k>, <\Gamma_1\Gamma_2\cdots\Gamma_k>,)=<(\sum_1-\Gamma_1)\sum_2\cdots\sum_k>+\cdots+<\sum_1\cdots\sum_{i-1}(\sum_i-\Gamma_i)\sum_{i+1}\cdots\sum_k>+\cdots+<\sum_1\sum_2\cdots\sum_{k-1}(\sum_k-\Gamma_k)> \quad (6)$$

ただし，右辺の中で冗長なパターンは除去する．

4.2 最小汎化された正規表現の導出

ある k-部分文字列の集合を被覆する汎化パターンの中で最小の汎化パターンを最小汎化パターンと呼ぶ．正(あるいは負)の k-部分文字列の集合から得られる最小汎化パターンをそれぞれ正(あるいは負)の最小汎化パターンと呼ぶ．以下に，最小汎化パターンの導出手順を示す．

(1) k-部分文字列の集合 PS から$<\sum_1,\sum_2,\cdots,\sum_k>$それぞれを見つけ出し，最汎パターン$<\sum_1\sum_2\cdots\sum_k>$を構成する．

(2) 式(5)を用いて，$<\sum_1\sum_2\cdots\sum_k>$に対する PS の除去により得られる残差汎化パターン集合 RPS を計算する．ただし，冗長なパターンは除去する．

(3) 上記の残差汎化パターン RPS から負の最小汎化パターンの集合を選び出す．それぞれの最小汎化パターンは，部分 PS のどんな集合(ただし，空集合は除く)も被覆しないパターンである．

(4) 式(6)を用いて，$<\sum_1\sum_2\cdots\sum_k>$に対する負の最小汎化パターン集合の除去により得られる残差汎化パターン集合 RPS'を計算する．ただし，冗長なパターンが出現した場合は，ただちに取り除く．

(5) 上記の残差汎化パターン集合 RPS'から正の最小汎化パターンの集合を選び出す．各パターンはそれぞれの最小汎化パターン，$NG=\sum_1\times\sum_2\times\cdots\times\sum_k$－PS のどんな部分集合(ただし，空集合は除く)も被覆しないが，PS のある部分集合を被覆する汎化パターンである．

5. 実験

ここでは，Cytochrome C に対してサフィックス木を用いて曖昧検索を行い，出力として得られた k-部分文字列集合に対して最小汎化を行なう．これにより，提案手法の有効性を確認する．データセットの詳細は以下の表 1 の通りである．

表 1 データセット詳細

データセット	データ件数	総長(bytes)	最大長(bytes)	最小長(bytes)
Cytochrome C	29	4235	631	116

Cytochrome C を用いて曖昧検索を行い，得られた部分文字列集合に対して最小汎化を行なった実験結果について示す．Cytochrome C モチーフは，

<C-x(2)-[STAQ]-x-[STAMV]-C-[STA]-T-C-[HR]>

の形式で知られている．検索キーとしてモチーフの一部分である<C-x(2)-A-x-A-C-A-T-C-R>を与えた．曖昧検索結果を表 2 に示す．

表 2 Cytochrome C を用いた曖昧検索結果

検索条件		曖昧検索結果		
許容誤差半径	最小支持数	出力文字列数(個)	支持数	検索時間(sec)
4	8	15	29	0.117

曖昧検索によって得られた部分文字列は 15 個であった．最小汎化パターンを抽出するために，まず，最汎パターンを計算し，そこから曖昧検索によって得られた 15 個の全ての部分文字列を除去することで 25 の負の汎化パターン集合を抽出した．次に，負の汎化パターン集合を最汎パターンから削除した．この汎化処理によって，7 個の汎化パターンが抽出できた．また，汎化処理にかかった時間は 0.132(sec)であった．

曖昧検索で得られた部分文字列集合を少数の汎化パターンで表現することに成功した．これにより，少数の汎化パターンを参照するだけで全体像を容易に把握できる．また，実際に抽出された汎化パターンは，<C-x(2)-[AS]-x-A-C-[AS]-T-C-H>などのパターンであり，全てがモチーフに含まれるものであった．このことから，汎化処理を行うことで，データセットに含まれるモチーフの正確なパターンを知ることができる可能性が高まる．また，他のデータセットを用いた実験でも同様の結果が得られた．

6. おわりに

本論文では，サフィックス木を用いて曖昧検索を行い，誤差半径内にある k-部分文字列集合を全て求め，汎化処理によって曖昧な表現を含む配列パターンの正規表現化を行なった．Cytochrome C データセットを用いて提案手法の有効性を確認した．その結果，部分文字列集合を少数の汎化パターンで表現することに成功した．また，データセットに含まれるモチーフの正確なパターンを知ることができた．

今後は，大規模なデータセットでの実験，さらなる正規表現化の研究を行う予定である．

参考文献

[1] 加藤 智之，北上 始，森 康真，田村 慶一，黒木 進：極小かつ非冗長な可変長ワイルドカード領域を持つ頻出配列パターンの抽出，電子情報通信学会論文誌 D，データ工学特集号，Vol. J90-D，No. 2，2006 年 2 月．

[2] Ching-Fung Cheung, Jeffrey Xu Yu, Hongjun Lu: Constructing Suffix Tree for Gigabyte Sequences with Megabyte Memory, IEEE Transaction Knowledge Data Engineering, Vo.17, No.1, pp.90-105, 2005.