6AA-1　最適パタン発見に基づくテキストデータマイニング

有村博紀, 渡木厚, 藤野亮一, 有川節夫
九州大学大学院システム情報科学研究科, 情報理学専攻 *

概要：本研究では，大量の文書の集積から，分類精度を最適化するパタンを見つける問題を考察する．二語相関パタンとよばれる単純なパタンを仮説としたとき，分類精度を最大化する最適パタンを$O(n^2)$ 時間および領域 $O(kn)$ 領域で計算するアルゴリズムを与える．

1　はじめに

データマイニング (Data mining) とは，データベースに蓄積された大量のデータから，自明でない規則性やパタンを半自動的にとりだす方法についての科学的研究である．データマイニングは，現在，ビジネス分野や科学技術分野をはじめとするさまざまな対象分野で，その適用が盛んにおこなわれている．現在のデータマイニングは，明示的な構造をもつ関係データベースが中心である．テキストデータベースに関しては，

1. 明示的な構造をもたない，
2. 多様な内容をもつ電子化文書の，
3. 数ギガバイトから数テラバイトにおよぶ膨大なデータの集積である

などの理由から，従来の方法は適用できず，研究がほとんどおこなわれていない．そこで本研究ではテキストデータからのデータマイニングについて研究する．

2　二語相関パタン

二語相関パタン (two words association pattern) とは，2つの定数文字列がワイルドカードをはさんだ形 $\langle \alpha, k, \beta \rangle$ の単純なパターンである．例えば，以下は2語相関パタンの例である．

$\langle$TATA, 30, AGGAGGT$\rangle$.
$\langle$knowledge, 50, databases$\rangle$.

ここに，ワイルドカードは2つの文字列が文字数 k 以下の距離で連続して出現するという制約を表わしており，単純な2語の論理積とことなり，文脈情報を表現できる．情報検索では，このような2語相関パタンは，"followed by" パタンとして知られ，ウェブ検索や，ゲノム情報学で有用なパタンである．形式的には，二語相関パタンの意味はつぎのように定める．

Def. 1 二語相関パタンを $P = \langle \alpha, k, \beta \rangle$ のテキスト T における出現位置とは，T中の位置の組 $\langle p, q \rangle$ で以下をみたすものをいう．

(i) 語 α と βは，それぞれ，位置 p と q に出現する T の部分語である．

(ii) 位置 p, q は$0 \leq q - p \leq k$ をみたす．

与えられた二語相関パタン P とエントリ T に対して，つぎの適合度を定義する．

- 重み付き頻度： $c(P,T) \in \mathsf{N}$ は，パタンP のT中での異なる出現位置 $\langle p, q \rangle$ の総数である．これは，ランク質問に対応する．

3　最適パタン発見問題

テキストデータマイニングを，情報検索の逆問題として定式化する．分類例集合 (sample) とは，有限集合 $S \subseteq \Sigma^* \times \mathsf{Z}$ である．各要素 $\langle s, b \rangle \in S$ を分類例 (labeled example) という．語 sを例 (example) といい，整数 b を分類値 (label) という．この定義は，実際のデータでは同じ例が何回も出現したり，一つの例が矛盾する正負の分類値をもち得ることを反映している．

重み付き分類精度最大化問題 (Maximizing Weighted Descrepancy Problem)
入力: 分類例集合 $S \subseteq \Sigma^* \times \mathsf{Z}$ および非負整数 K.
問題: 距離パラメータが k の二語相関パタン $\langle \alpha, k, \beta \rangle$ 全体から，S に関する分類精度

$$C(P,S) = \sum_{\langle s,d \rangle \in S} c(P,s) \times d$$

を最大化するパタン P を見つけよ．

この問題は，すべての二語相関パタンを探索する自明な方法を用いて，$O(n^5)$ 時間でとける．しかし，本研究では大規模な入力に対しても働く高速なアルゴリズムを開発する．

*Text data mining with optimal string patterns, Hiroki Arimura, Atsushi Wataki, Ryoichi Fujino, Setsuo Arikawa, Department of Informatics, Kyushu University, Kasuga Koen 6-1, Kasuga, 816 Japan, TEL: 092-583-7632, FAX: 092-583-7635
e-mail:{arim,wataki,fujino,arikawa}@i.kyushu-u.ac.jp

4 アルゴリズム

4.1 接尾語木

テキスト $A = a_1 \cdots a_{m-1}\$$に対して，位置 p からはじまる A の接尾語をA_p で表す．テキスト A の接尾語木 (suffix tree) T_A とは，A の空でない接尾語全体 $\{A_1, \ldots, A_n\}$ を表す圧縮トライ (compacted trie) である．ここで，圧縮トライとは，通常のトライ (trie) から，子を一つしかもたない内部節点を取り除き，辺のラベルを合併することを繰り返して得られる木である．$W(v)$ で，根から節点 v にいたるパス上のラベルを連結して得られる語を表す．接尾語木は，$O(n)$時間で計算可能であり，$O(n)$領域を使用する (McCreight [2])．

4.2 アルゴリズム

$S = \{\langle s_1, b_1\rangle, \ldots, \langle s_m, b_m\rangle\} \subseteq \Sigma^* \times \mathsf{Z}$ を分類例集合とし，k を非負整数とする．一般性を失うことなく，すべての例 $s_1, \ldots, s_m$ は異なると仮定する．

S の例すべてを連結した語を $A = s_1\$_1 s_2\$_2 \cdots s_m\$_m$ とする $(n = |A|)$．ここに，$\$_1, \ldots, \$_m$ は，$\$_i \notin \Sigma$ をみたす互いに異なる区切り文字である．このアルゴリズムは，テキスト A の接尾木 T_A を作り，これを用いて例の部分語を管理する．

T_A の葉はAの接尾語を表しており，左から右へそれが表す接尾語の辞書式順序に並んでいる．これらの接尾語 A_p の開始位置 p をこの辞書式順序に従っておさめた配列を $suf[1,n]$ とし，その逆関数を表す配列を $pos[1,n]$ とする．

アルゴリズムは，分類例集合 S と非負整数k を受け取ると，つぎのように最適パタンを計算する．

1. $A = s_1\$_1 s_2\$_2 \cdots s_m\$_m$ に対して，接尾語木 T_A を計算する．位置の組 $\langle p, q\rangle$ に対して，もしある $\langle s_i, d_i\rangle$ に対して s_i がp, qを両方含むならば，重みを $w_A(p,q) = d_i$ と定義する．

2. 配列 $pos[1,n]$ を，T_A から計算する．つぎに空の直交領域木 (orthogonal range tree) を D とする．すべての組 $\langle p, q\rangle$ $(0 \le q - p \le k)$ を点 $\langle pos[p], pos[q]\rangle$ に変換し，直交領域木 D に挿入する．各組の重みは，$w_A(\langle pos[p], pos[q]\rangle) = w_A(\langle p, q\rangle)$ とおく．

3. T_A の各節点について，区間 $[L(u), R(v)]$ を計算する．各実節点 v に対して区間 $[L(v), R(v)]$ をつぎのように関連づける．

   - 節点 v がi番目の葉ならば，$L(v) = R(v) = i$とする．
   - 節点 v が子 $v_1, \ldots, v_m$ をもつ内部節点で，すでに区間が計算済みならば，$L(v) = L(v_1)$ および $R(v) = R(v_m)$ とする．

4. T_A のすべての節点の組 $\langle u, v\rangle$ について，以下をくり返し，$C(\langle W(u), k, W(v)\rangle)$ が最大になる二語相関パタン $\langle W(u), k, W(v)\rangle$ を探す．

   直交質問をおこなって長方形 $[L(u), R(v)] \times [L(v), R(v)]$ に含まれる点について，重みの総和を求める．これを分類精度 $C(\langle W(u), k, W(v)\rangle$ とする．

5. 最大分類精度を与えるパタンをすべて出力する．

部分語α が $W(v)$ の接頭語であり，v の親 u に対して，$W(u)$ が α の真の接頭語になるような節点 v が一意に定まる．この節点 u を，語α の実節点 (locus) といい，$locus(\alpha)$ と書く．

Lemma 1 $\langle \alpha, k, \beta\rangle$ を二語相関パタンとし，u, v をそれぞれ α, β の実節点とする．このとき，

$$C(\langle \alpha, k, \beta\rangle, S) = C(\langle W(u), k, W(v)\rangle, S).$$

上の補題から，最適パタンを見つけるには，高々 $O(n^2)$ 個の二語相関パタンを枚挙しながら，分類精度 $C(\langle W(u), k, W(v)\rangle, S)$ の値を比較すればいいことがわかる．アルゴリズムでは，直交質問のための適当なデータ構造を用いて，分類精度の計算を前処理 $O(N)$，領域 $O(N)$，時間 $O(\log^2 N)$ で行っている $(\mathrm{N} = \mathrm{kn})$ [3]．

Lemma 2 二語相関パタンに対する重み付き分類精度最大化問題ha，$O(n^2 \log^2 n)$時間と$O(kn)$領域で計算可能．

副手続きとして毎回新たに直交質問をするかわりに，接尾語木自体を区間木として使い，直交領域木を実現することができる．この場合，接尾語木は一般に平衡木ではないが，探索対象となる区間の数が線形個しかないので，Maass [1] の手法を利用してつぎの定理が示せる．

Lemma 3 二語相関パタンに対する重み付き分類精度最大化問題は，$O(n^2)$時間と$O(kn)$領域で計算可能．

References

[1] W. Maass, Efficient agnostic PAC-learning with simple hypothesis, In Proc. COLT94 (1994), 67–75.

[2] E. M. McCreight, A space-echonomical suffix tree constructiooon algorithm, JACM 23 (1976), 262–272.

[3] F. P. Preparata, M. I. Shamos, Computational Geometry, Springer-Verlag (1985).