General Algorithms for Mining Closed Flexible Patterns under Various Equivalence Relations

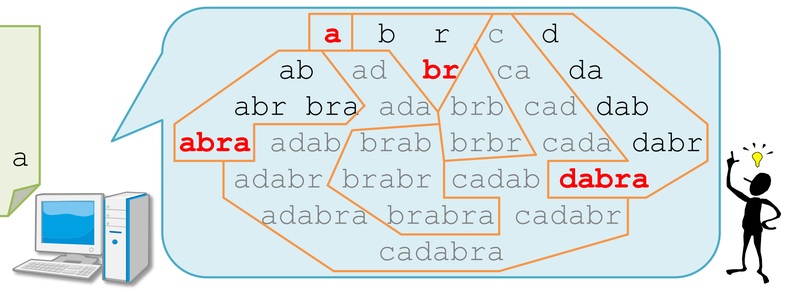
○井 智弘, 荏隈 勇樹, 坂内 英夫, 竹田 正幸　九州大学

頻出飽和文字列パターン列挙問題

入力：テキスト集合 S, 頻度の閾値 σ, 同値関係 $\equiv_S$
出力：S 中の σ 個以上の文字列に出現する飽和パターンを列挙

例　$\sigma = 2$ のときの頻出飽和部分文字列パターン列挙

S
- ✓ a b r a
- ✓ b r a b r a
- ✓ d a b r a
- ✓ c a d a b r a
- ✓ b r b r



■ 同値関係を導入するメリット
1. 簡潔で見やすい出力
2. うまく計算すると探索・出力コスト減

Flexible パターン

$*_1\ w_1\ *_2\ w_2\ *_3\ \cdots\ *_k\ w_k\ *_{k+1}$

◆ $k \geq 1$　◆ $* \notin \Sigma$: ギャップ文字
◆ $w_i \in \Sigma^+$: i 番目のセグメント

パターン P : * a b * e * r a *

埋め込み $\theta = \langle$cd, r, r, bra$\rangle$ を適用 → $P\theta$: c d a b r e r r a b r a

埋め込み $\theta = \langle$c*, r, ε, b*a$\rangle$ を適用 → $P\theta$: c * a b r e r a b * a

■ パターンの出現区間, 極小出現区間

パターン P : * a b * e * r a *

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
テキスト T	y	a	b	x	a	b	e	x	r	a	x	r	a	y

$\theta = \langle$x, xab, x, xrax$\rangle$
$\theta = \langle$x, xab, xrax, x$\rangle$
$\theta = \langle$xabx, ε, x, xrax$\rangle$
$\theta = \langle$xabx, ε, xrax, x$\rangle$

極小

出現区間(P, T) = {[2, 10], [2, 13], [5, 10], [5, 13]}
極小出現区間(P, T) = {[5, 10]}

■ パターン上の半順序

◆ パターン $P \neq Q$ に対して,
$P\theta = Q$ なる埋め込みが存在するとき
Q は P より具体的であるという

P : * a b * e * r a *

埋め込み $\theta = \langle$*, x*b, x, xr*$\rangle$ を適用

Q : * a b x * b e x r a x r *

Flexible パターン上の様々な同値関係に対して統一的な多項式時間遅延アルゴリズムを提案

同値関係	説明	遅延計算量
$\overset{\mathrm{I}}{\equiv}_S$	出現区間の一致	$O(\lvert\Sigma\rvert\ \lVert S\rVert\ d)$
$\overset{\mathrm{IX}}{\equiv}_S$	$\overset{\mathrm{I}}{\equiv}_S$ の拡張	$O(\lvert\Sigma\rvert\ \lVert S\rVert\ d)$
$\overset{\mathrm{M}}{\equiv}_S$	極小出現区間の一致	$O(\lvert\Sigma\rvert\ \lVert S\rVert\ d)$
$\overset{\mathrm{MX}}{\equiv}_S$	$\overset{\mathrm{M}}{\equiv}_S$ の拡張1	$O(\lvert\Sigma\rvert\ \lVert S\rVert\ d)$
$\overset{\mathrm{MXG}}{\equiv}_S$	$\overset{\mathrm{M}}{\equiv}_S$ の拡張2	$O(\lvert\Sigma\rvert\ \lVert S\rVert^2\ d)$
$\overset{\mathrm{E}}{\equiv}_S$	出現の終了位置集合の一致	$O(\lvert\Sigma\rvert\ \lVert S\rVert)$
$\overset{\mathrm{EX}}{\equiv}_S$	$\overset{\mathrm{E}}{\equiv}_S$ の拡張	$O(\lvert\Sigma\rvert\ \lVert S\rVert)$
$\overset{\mathrm{B}}{\equiv}_S$	出現の開始位置集合の一致（有村&宇野）	$O(\lvert\Sigma\rvert\ \lVert S\rVert)$

$\lVert S\rVert$: テキスト長の総和
d : セグメント数の最大値
計算領域はいずれも $O(\lVert S\rVert d)$

同値関係の粗化 $\overset{\mathrm{M}}{\equiv}_S \subseteq \overset{\mathrm{MX}}{\equiv}_S \subseteq \overset{\mathrm{MXG}}{\equiv}_S$

■ $\overset{\mathrm{MX}}{\equiv}_S$
◆ 全ての極小出現に関して左右に共通の文字列で伸長できるとき, それらのパターンも同値とする

テキスト T : xxaberaxxxxaboraxxxxxx

パターン P : *ab*ra*
パターン P' : *ab*rax*
パターン P'' : *ab*raxx*
⋮
*xxab*raxx*
*xxab*raxxx*
*xxab*raxxxx*

$\overset{\mathrm{M}}{\equiv}_S$ では, 極小出現区間が異なるため各々異なる同値類に属する

■ $\overset{\mathrm{MXG}}{\equiv}_S$
◆ さらに, ギャップ付でも左右に伸ばせるようにする（ただし, 極小出現区間の数が変わらない範囲で）

テキスト T : xzaberaxxxxaborayxyxxx

パターン P : *ab*ra*
パターン P' : *x*ab*ra*
パターン P'' : *ab*ra*x*x*
⋮
*x*ab*ra*x*xxx*

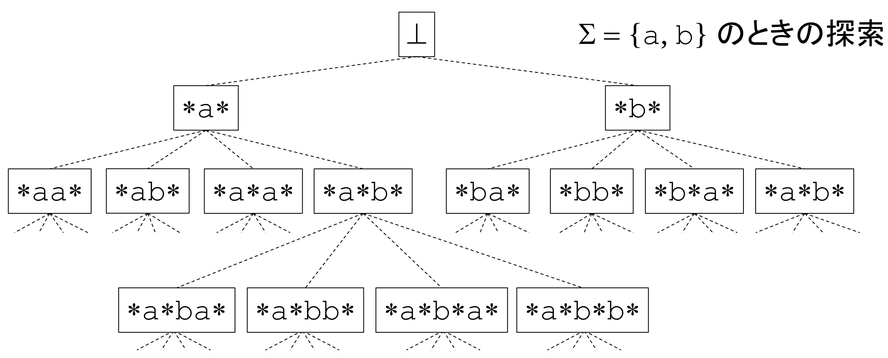
これらも同値とする

*ab*ra*xy* ← これは同値ではない

■ 次の親子関係で定義される探索木を深さ優先探索する

◆ 任意の x ∈ Σ に対して, *x* の親は ⊥
◆ 任意の x ∈ Σ, $w_1, \ldots, w_k \in \Sigma^+$ に対して,
$*w_1*\cdots*w_k$x* と $*w_1*\cdots*w_k*$x* の親は $*w_1*\cdots*w_k*$

Σ = {a, b} のときの探索木



■ 任意のパターン P に対して以下が成り立つ

◆ 頻度チェック
P が頻出でないなら, P の子孫も頻出でない（枝刈り）
◆ 左端飽和チェック
左端のギャップ文字を具体的にしてP と同値なパターンがあれば
P と P の子孫は飽和パターンではない（枝刈り）
◆ 内部飽和チェック
左右端以外のギャップ文字を具体的にしてP と同値なパターンがあれば
P と P の子孫は飽和パターンではない（枝刈り）
◆ 右端飽和チェック
右端のギャップ文字を具体的にしてP と同値なパターンがあるとき
P は飽和パターンでは無いが
P の子孫に頻度が等しく飽和なパターンが必ず存在する

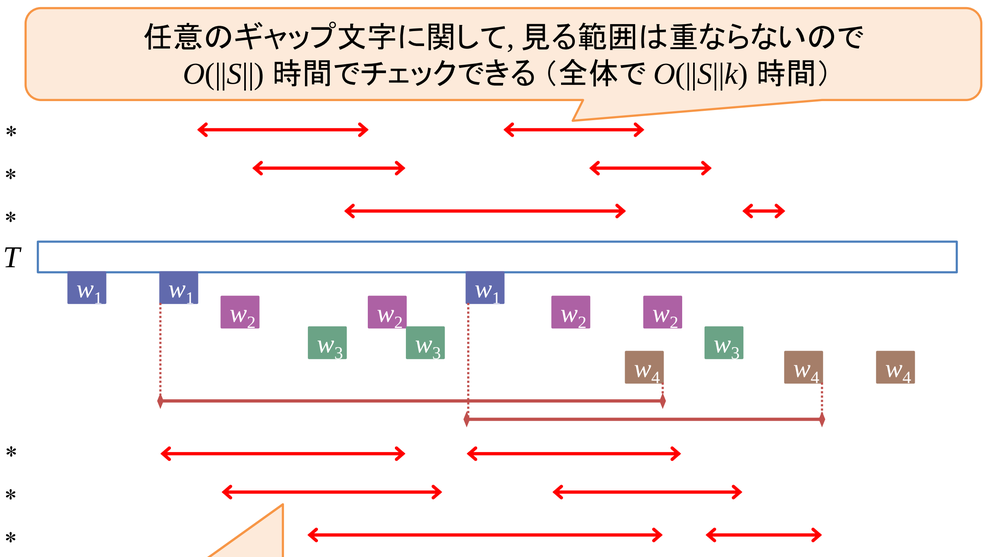
■ $\overset{\mathrm{M}}{\equiv}_S\ \overset{\mathrm{MX}}{\equiv}_S\ \overset{\mathrm{MXG}}{\equiv}_S$ における内部飽和チェック〈$O(\lVert S\rVert k)$ 時間〉

$P = * w_1 * w_2 * w_3 * w_4 *$

任意のギャップ文字に関して, 見る範囲は重ならないので
$O(\lVert S\rVert)$ 時間でチェックできる（全体で $O(\lVert S\rVert k)$ 時間）

◆ non-erasing
$* w_1 *x* w_2 * w_3 * w_4 *$
$* w_1 * w_2 *x* w_3 * w_4 *$
$* w_1 * w_2 * w_3 *x* w_4 *$

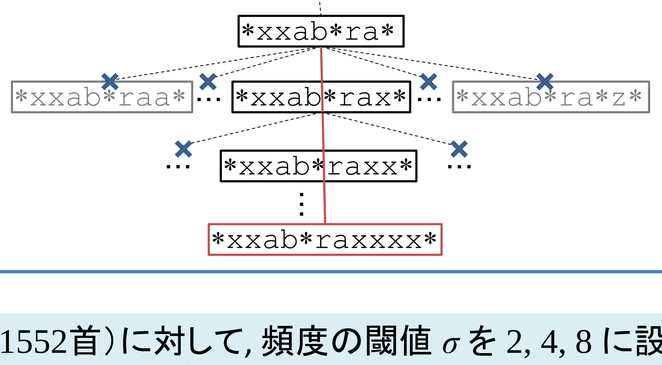
◆ erasing
$* w_1 w_2 * w_3 * w_4 *$
$* w_1 * w_2 w_3 * w_4 *$
$* w_1 * w_2 * w_3 w_4 *$



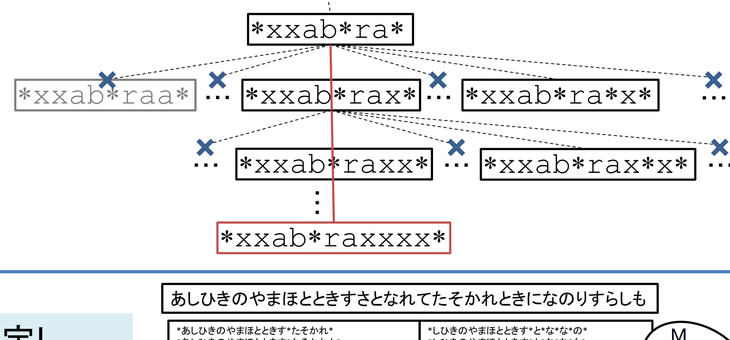
テキストの各位置に対して, w_2 w_3 が次に出現する位置を記録しておけば
$O(\lVert S\rVert)$ 時間でチェックできる（全体で $O(\lVert S\rVert k)$ 時間）

■ 最長共通接頭辞を用いた高速化

◆ $\overset{\mathrm{M}}{\equiv}_S\ \overset{\mathrm{MX}}{\equiv}_S\ \overset{\mathrm{MXG}}{\equiv}_S$
極小出現区間の右側に必ず“xxxx”が出現するとき



◆ $\overset{\mathrm{I}}{\equiv}_S\ \overset{\mathrm{IX}}{\equiv}_S\ \overset{\mathrm{E}}{\equiv}_S\ \overset{\mathrm{EX}}{\equiv}_S$
出現区間の右側に必ず“xxxx”が出現するとき



和歌データを用いた実験

	$\sigma = 2$ (0.13%)		$\sigma = 4$ (0.26%)		$\sigma = 8$ (0.52%)	
algorithm/equiv	patterns	seconds	patterns	seconds	patterns	seconds
GenCloFlex $\overset{\mathrm{I}}{\equiv}_S$	85,665,856	4084.4	24,941,537	2088.3	4,634,071	823.54
GenCloFlex+ $\overset{\mathrm{I}}{\equiv}_S$	85,665,856	3983.7	24,941,537	2065.8	4,634,071	817.12
GenCloFlex+ $\overset{\mathrm{IX}}{\equiv}_S$	80,622,811	3932.2	24,661,456	2063.0	4,622,063	816.77
GenCloFlex $\overset{\mathrm{M}}{\equiv}_S$	83,042,793	3993.8	24,596,215	2064.2	4,601,106	819.16
GenCloFlex+ $\overset{\mathrm{M}}{\equiv}_S$	83,042,793	3872.2	24,596,215	2040.3	4,601,106	810.88
GenCloFlex+ $\overset{\mathrm{MX}}{\equiv}_S$	77,392,851	3819.8	24,287,323	2035.1	4,588,408	811.97
GenCloFlex+ $\overset{\mathrm{MXG}}{\equiv}_S$	55,131,507	3509.8	23,508,993	2032.7	4,580,550	814.46
GenCloFlex $\overset{\mathrm{E}}{\equiv}_S$	67,210,837	3567.4	24,253,028	2055.3	4,609,610	825.44
GenCloFlex+ $\overset{\mathrm{E}}{\equiv}_S$	67,210,837	3500.8	24,253,028	2034.5	4,609,610	817.87
GenCloFlex+ $\overset{\mathrm{EX}}{\equiv}_S$	65,242,280	3497.0	24,064,249	2030.3	4,601,551	806.16

■ 山家集（1552首）に対して, 頻度の閾値 σ を 2, 4, 8 に設定し, 各同値関係における頻出飽和パターンの数と計算時間を比較（画面表示はオフ）

■ 「御裳濯和歌集序文」と「八代集の各首」の間で, $\overset{\mathrm{M}}{\equiv}_S, \overset{\mathrm{MX}}{\equiv}_S, \overset{\mathrm{MXG}}{\equiv}_S$ における共通飽和パターンを列挙し結果を比較

御裳濯和歌集序文
やまとうたはあしはらのなかつくによりおこり、いなたひめすかのさとよりひろまれり、あとをたれたまへるしんみやうもこのみちをむねとし、あらはれいてたまへるふつたにのことわさをすてたまはす、おほよそわかくにのふうそくとしてむかしよりいまにたゆることなし、つらつらちよくせんをたつぬれはすてにはちたいになれり、ところところのうちききしふそのかすまたおほし、かれにしたふこころすすみて、……（1241文字）

◆ 最初のセグメントから最後のセグメントの距離が 38 の出現のみを考える
◆ セグメントの長さの総和が 15 以上のパターンのみ表示する

あしひきのやまほととぎすさとなれてたそかれときになのりすらしも