部分文字列増幅法による共通パタン発見アルゴリズム

池田, 大輔
九州大学情報基盤センター

山田, 泰寛
九州大学大学院システム情報科学府

廣川, 佐千男
九州大学情報基盤センター

出版情報：情報処理学会研究報告 ： 数理モデル化と問題解決. 2003 (122), pp.45-48, 2003-12. 情報処理学会
バージョン：

部分文字列増幅法による共通パタン発見アルゴリズム

池田 大輔、山田 泰寛*、廣川佐千男

九州大学情報基盤センター

*九州大学システム情報科学府

1本の文字列　例：HTML/XMLのファイル　2003.12.12

蛋白質配列

問題：共通な部分を探す

例：情報抽出のテンプレート発見

生物学的に意味のある部分配列

モデル化の例：最長共通部分列問題

2003.12.12

文字列の共通部分列

acbbaccbabaa

ccbbcbabbabca

cbaabbcbcab

cbbbab

同じ機能を持つ
部分配列?

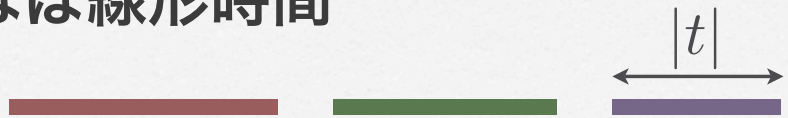
最長共通部分列問題

NP完全(Maier'78)

主な結果

- 入力：未知のパタンからベキ分布に従って生成された語の集合
- すべての語に共通な部分文字列の列を探す問題は平均的に $O(nm|t|^2 \log |t|)$ 時間で解ける

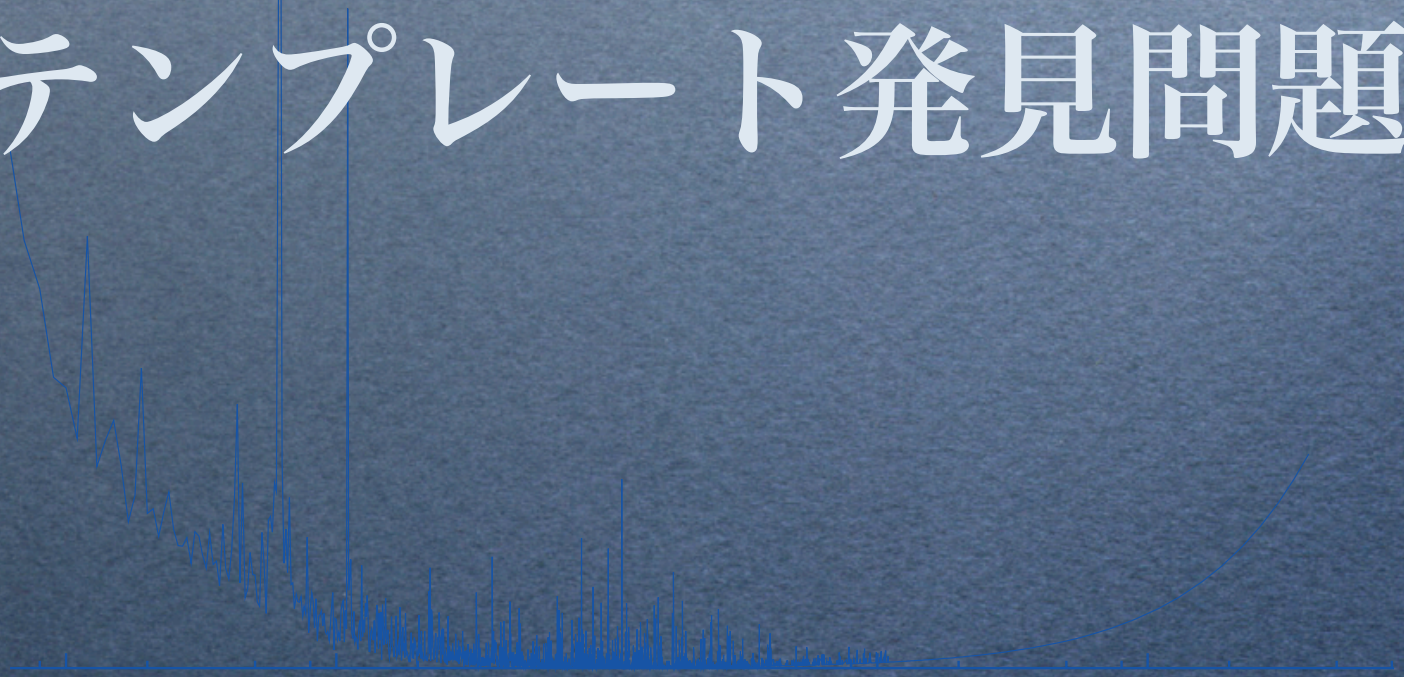
nは入力長で、ほぼ線形時間



目次

- テンプレート発見問題
  - 主に情報抽出の観点からの問題の見直し
  - 共通部分の表現方法
- 部分文字列増幅法
  - 共通部分の高速な特定方法
- 実験
  - Web上のデータからの情報抽出

テンプレート発見問題



2003.12.12

入力文字列の性質

正例が与えられる

には自然な代入

テンプレート

7

パタン言語

Σ : 有限アルファベット

V : 変数の集合　　$(\Sigma \cap V = \phi)$

パタン $\Leftrightarrow$ $\Sigma \cup V$ 上の文字列

例： $axbyax (\Sigma = \{a, b\}, V = \{x, y\})$

代入 $\Leftrightarrow$ $\theta : V \to \Sigma^*$

パタン p の言語

$$L(p) = \{w \in \Sigma^* \mid \exists \theta; p\theta = w\}$$

テンプレート

正則パタン⇔各変数の出現は高々1回

例： $axby$　　$axbx$

正則パタンの言語族は正規言語族の真の部分クラス

$a.*b.*$

$w_0x_1w_1\ldots w_{m-1}x_mw_m$：正則パタン $(w_i \in \Sigma^*, x_j \in V)$

テンプレート⇔定数文字列の列 $(w_0, \ldots, w_m)$

2003.12.12

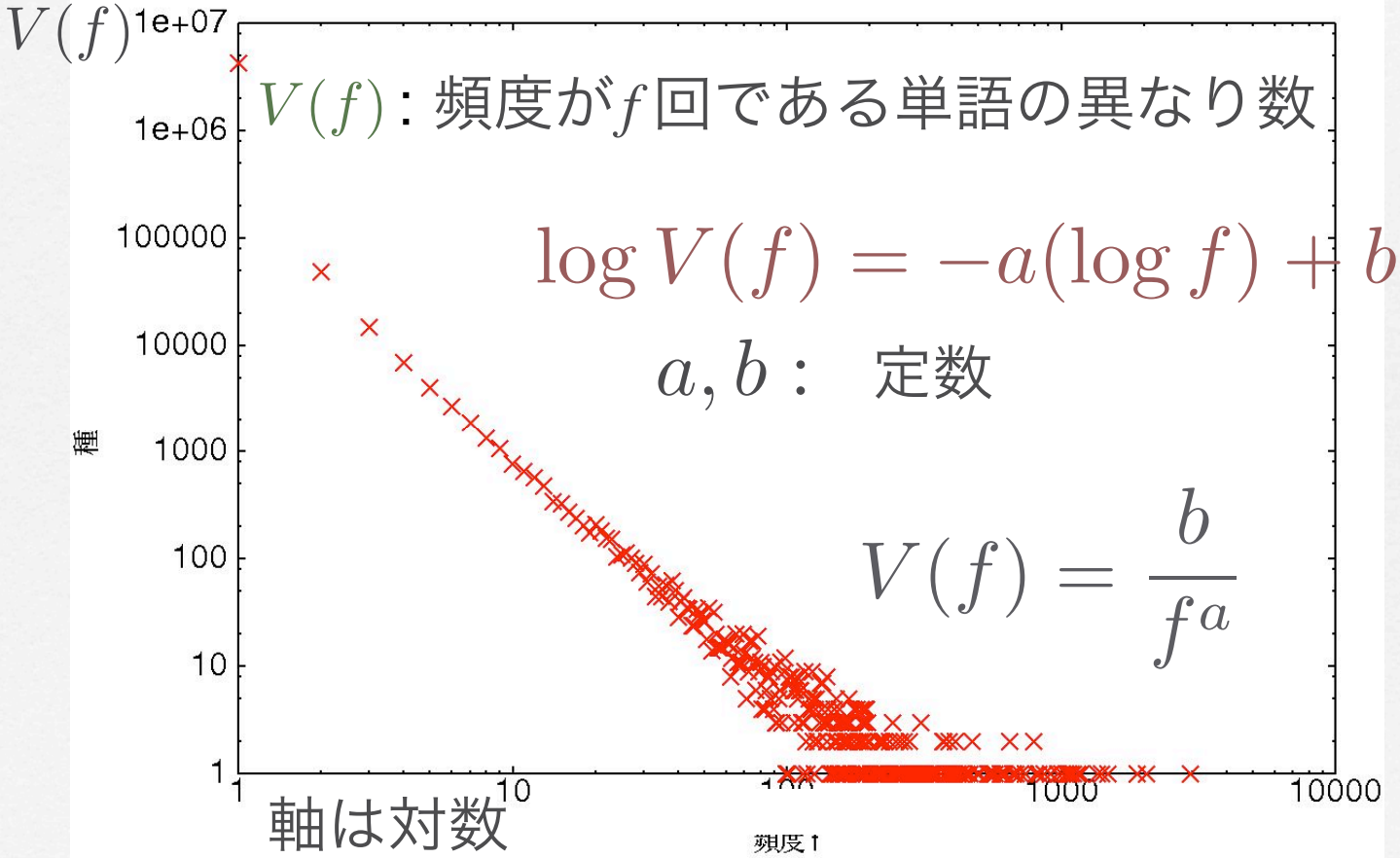
テンプレートと代入

仮定：代入される文字列の頻度は
ベキ分布に従う

$$p = x_1 A x_2 B x_3 C x_4$$

$$(A, B, C \in \Sigma^*, x_i \in V)$$

べキ分布：Zipfの(第二)法則



テンプレート発見問題 (池田'03)

2003.12.12

入力：文字列の有限集合S

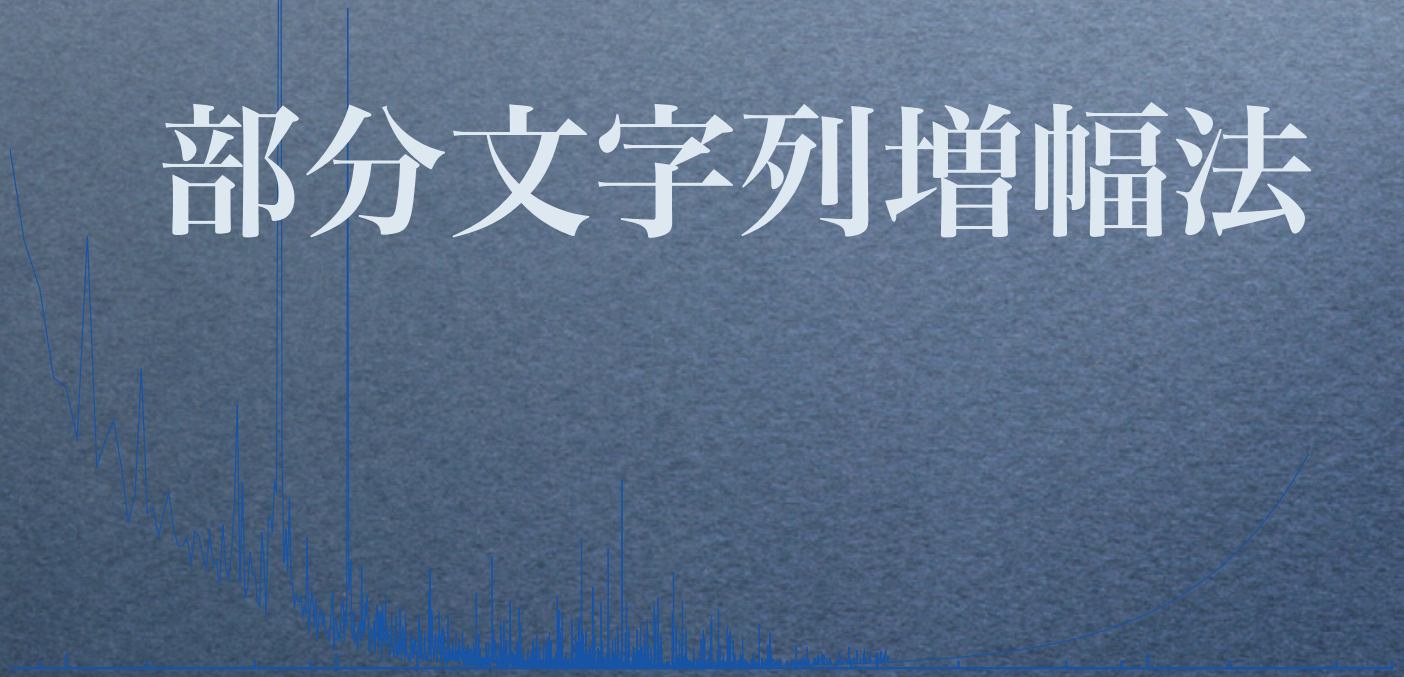
Sは未知のパタンからベキ分布に従って生成された

問題：以下を満すテンプレートtを見つけよ

tは正規パタンpのテンプレートである
$S \subseteq L(p)$ かつ $\nexists q; S \subseteq L(q) \subset L(p)$

pはS に対し極小

部分文字列増幅法



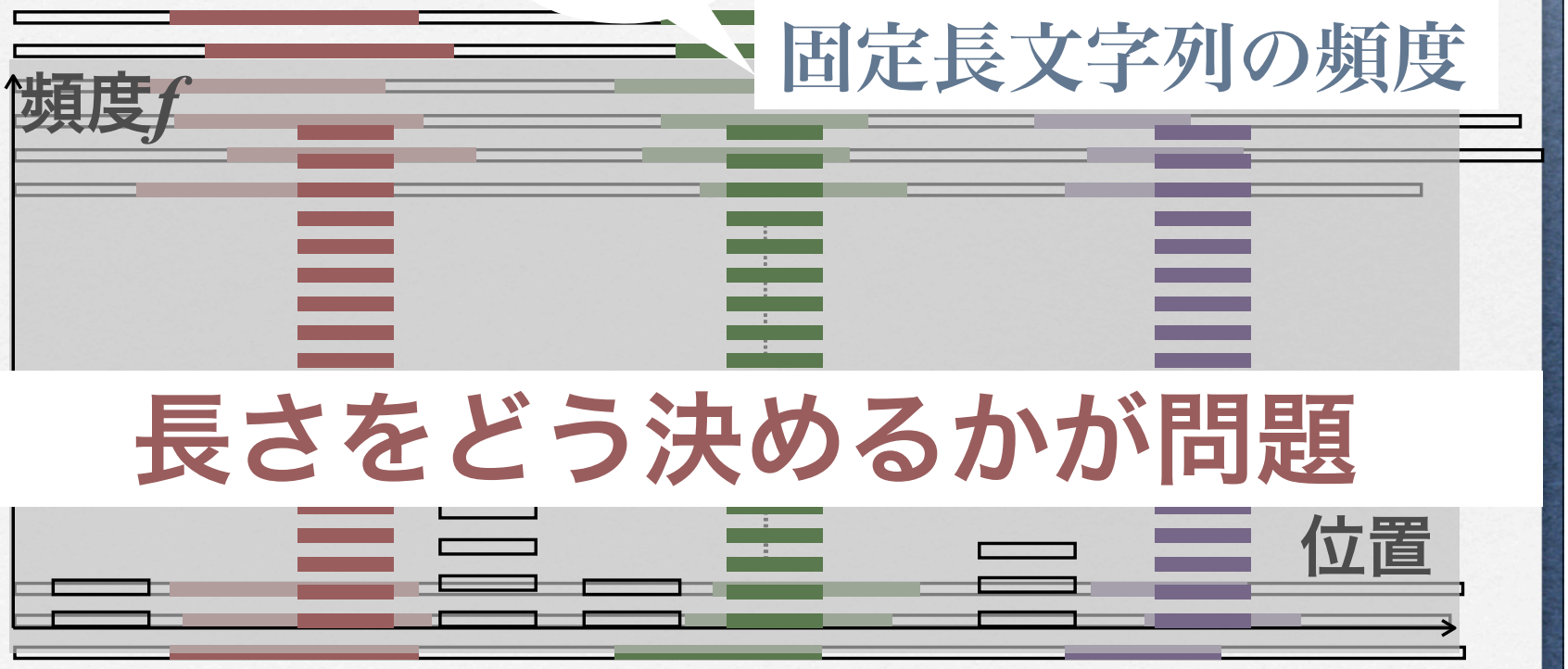
2003.12.12

共通パタンと頻度

例えば長さ10

1ファイルでみる

固定長文字列の頻度



長さをどう決めるかが問題

文字列の総出現数

$F(f)$：頻度(出現回数)がf回であるような文字列の総出現数

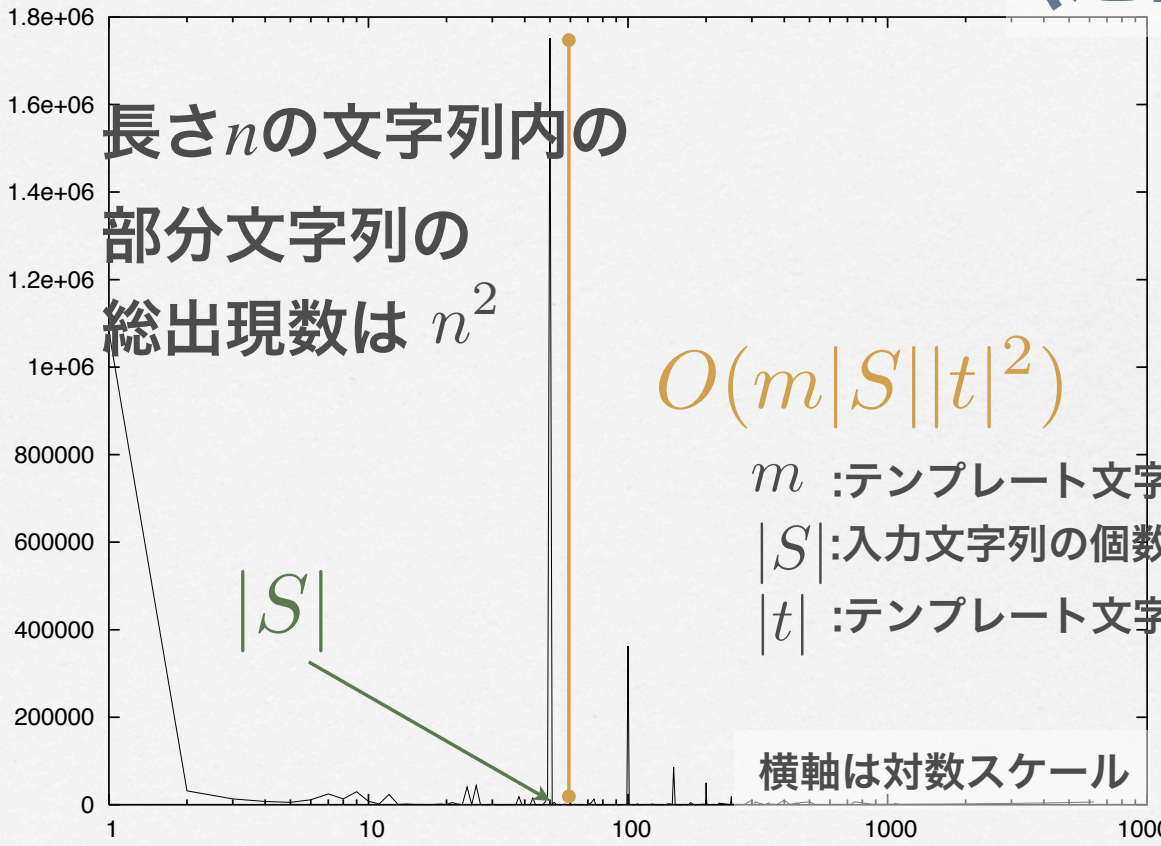
$$F(f) = fV(f)$$

$$\log V(f) = -a(\log f) + b$$
$$\log f + \log V(f) = (1-a)(\log f) + b$$
$$\log fV(f) = (1-a)\log f + b$$
$$\log F(f) = (1-a)\log f + b$$

2003.12.12

部分文字列増幅法 (池田'03)

$F(f)$



長さnの文字列内の
部分文字列の
総出現数は n^2

$O(m|S||t|^2)$

m :テンプレート文字列の個数

$|S|$:入力文字列の個数

$|t|$:テンプレート文字列の最小幅

$|S|$

横軸は対数スケール

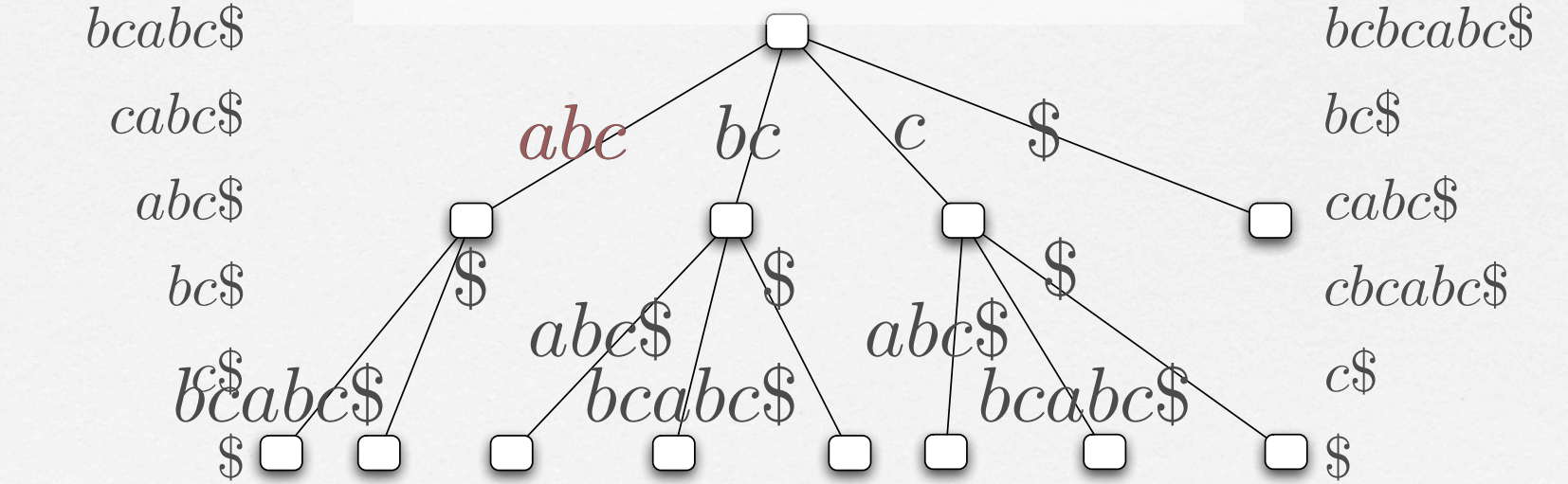
頻度 f

2003.12.12

abcbcabc の接尾辞木

branching word

⇔ 根からあるノードまでの辺のラベルをつなげたもの

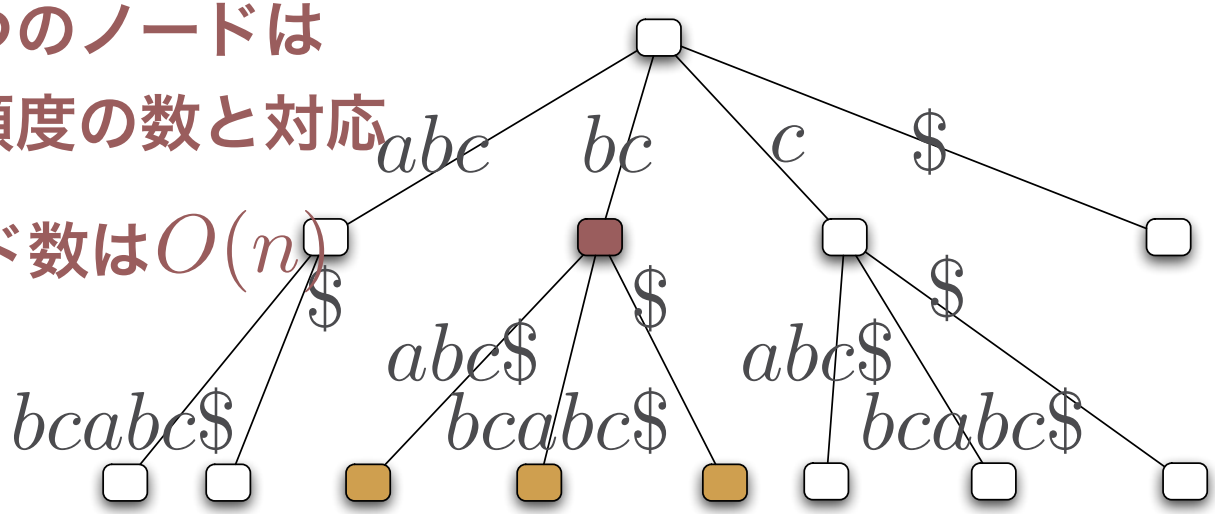


Lemma 1

サンプルS中に現われる部分文字列の異なる頻度の数は高々$O(n)$個である

ただしnはS中の全て文字列の長さの和である

一つのノードは
ある頻度の数と対応
ノード数は$O(n)$



Lemma 2

テンプレート文字列内の全ての部分文字列に対し、
同じ文字列の総出現回数は高々 $O(m \log |t|)$
個である

m :テンプレート文字列の個数

$|t|$:テンプレート文字列の最小幅

```
function Template(S):template
begin
   V:=Count(S);  #V(f)={f回出現するbranching wordの集合}
   for f in keyes(V):
      F(f):=f|V(f)|;
   end
   P:=FindPeaks(F);
   t:=Reconstruct(V, P);
   return(t)
end
```

$O(n)$

$O(n)$

$O(n)$

$|P| \leq O(m \log t)$

```
subroutine Reconstruct(S):template
begin
   for f in P:
      W:=V(f); #V(f)={f回出現するbranching wordの集合}
      for w in W:
         Sにおける出現回数と位置をチェック
      end
      tに追加
   end
   return(t)
end
```

$|W| < |t|^2$

$O(nm|t|^2 \log |t|)$

21

Theorem

Templateは平均的に正しくテンプレート発見問題を解き、その計算量は$O(nm|t|^2 \log |t|)$である

m :テンプレート文字列の個数
$|S|$:入力文字列の個数
$|t|$:テンプレートの幅

○示すべきこと

$S \subseteq L(p)$

極小である

2003.12.12

極小性の証明

○示すべきこと

余計なものをテンプレートとしない

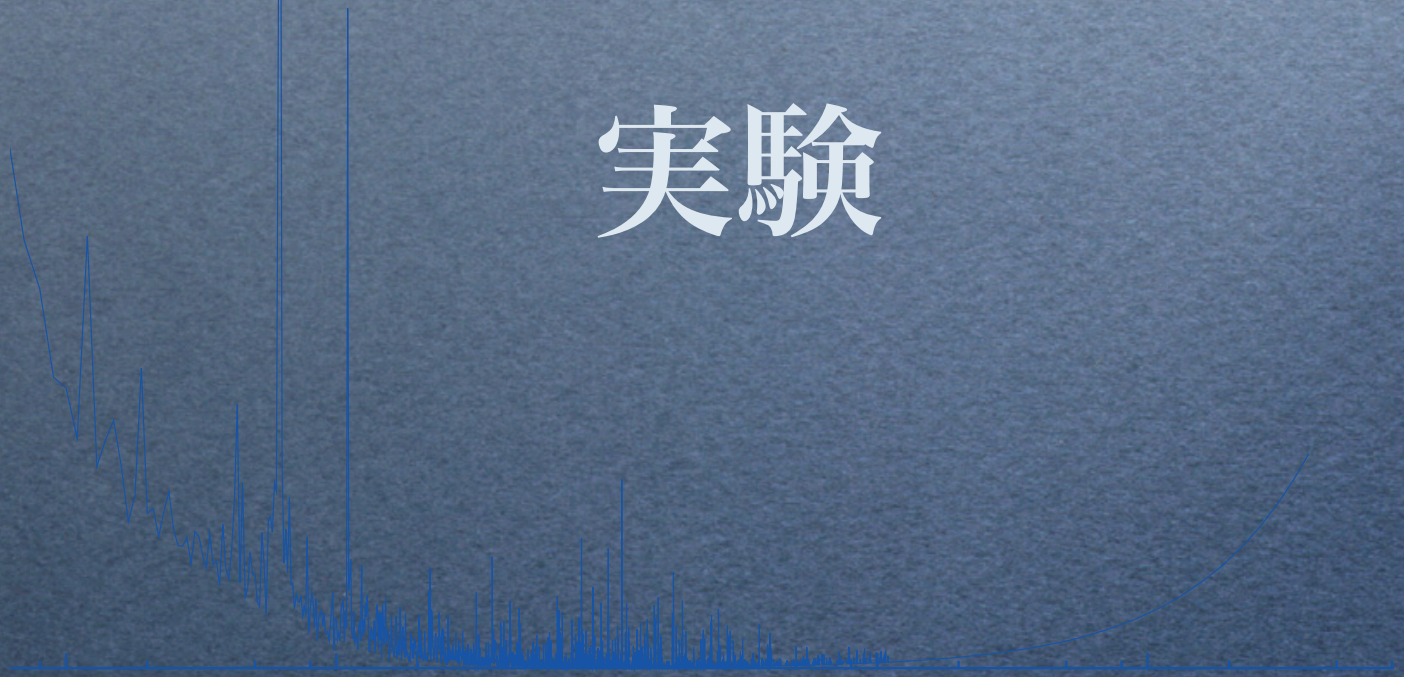
テンプレートの見落としがない

テンプレート文字列の

一部が偶然代入された

＝ の分だけピークが低くなる!

実験

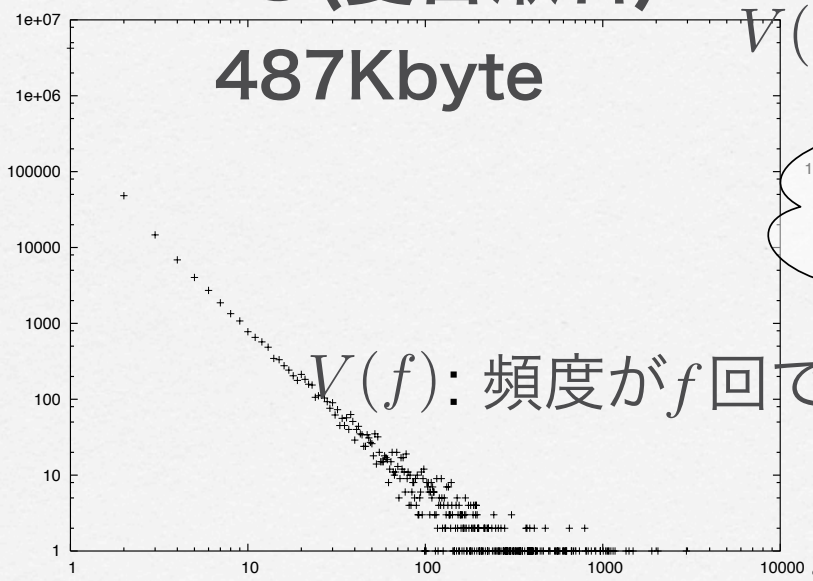


2003.12.12

小説中の部分文字列の分布

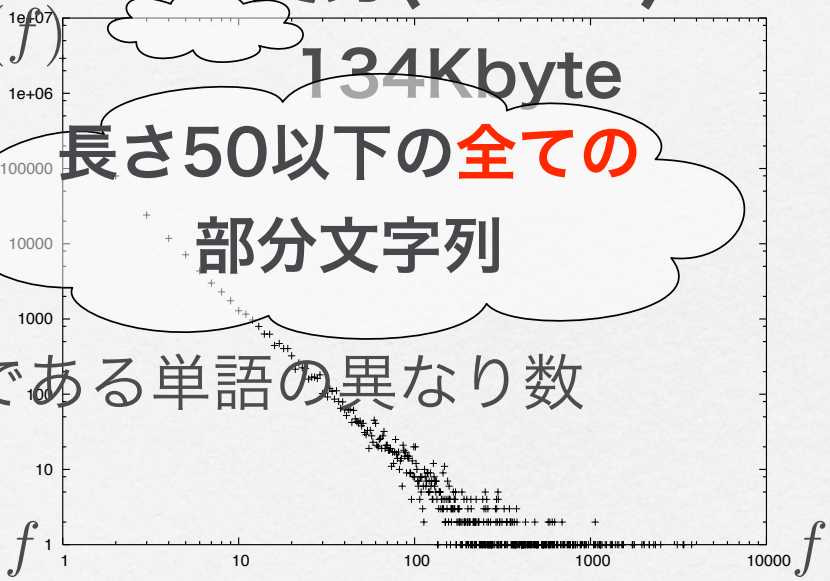
こころ(夏目漱石)

487Kbyte



変身(カフカ)

134Kbyte



長さ50以下の**全ての**部分文字列

$V(f)$: 頻度がf回である単語の異なり数

青空文庫

http://www.aozora.gr.jp/

Project Gutenberg

http://www.promo.net/pg/

25

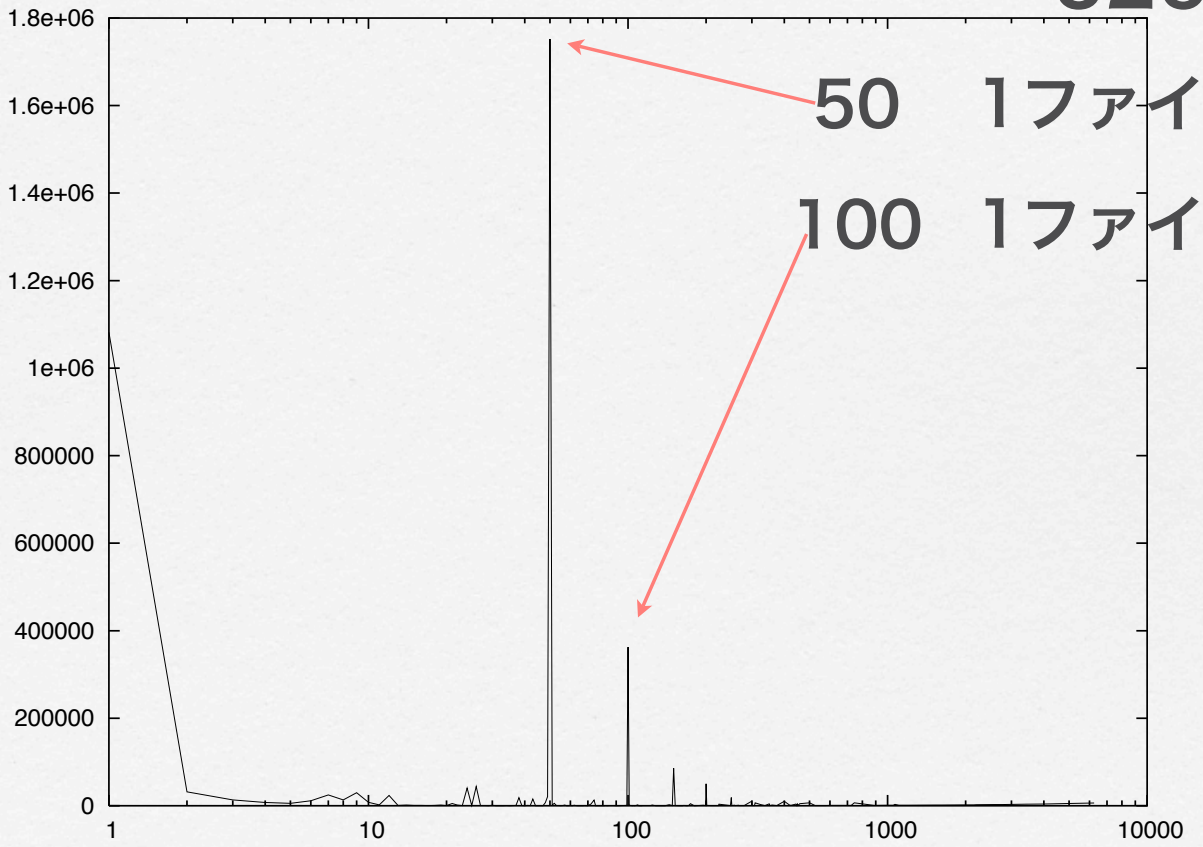
実験データ

- データ：HTMLファイル
- 前処理はしない
  - 空白文字もタグもそのまま用いる
- $F(f)$のグラフを作るところまで
  - `Reconstruct()`を実行しない

2003.12.12

産経新聞の記事ファイル50

526Kbyte



50　1ファイル1回

100　1ファイル2回

27

2003.12.12

Yahoo!検索結果

46：各ファイルごと

91：一致したカテゴリ数

913：各検索結果ごと

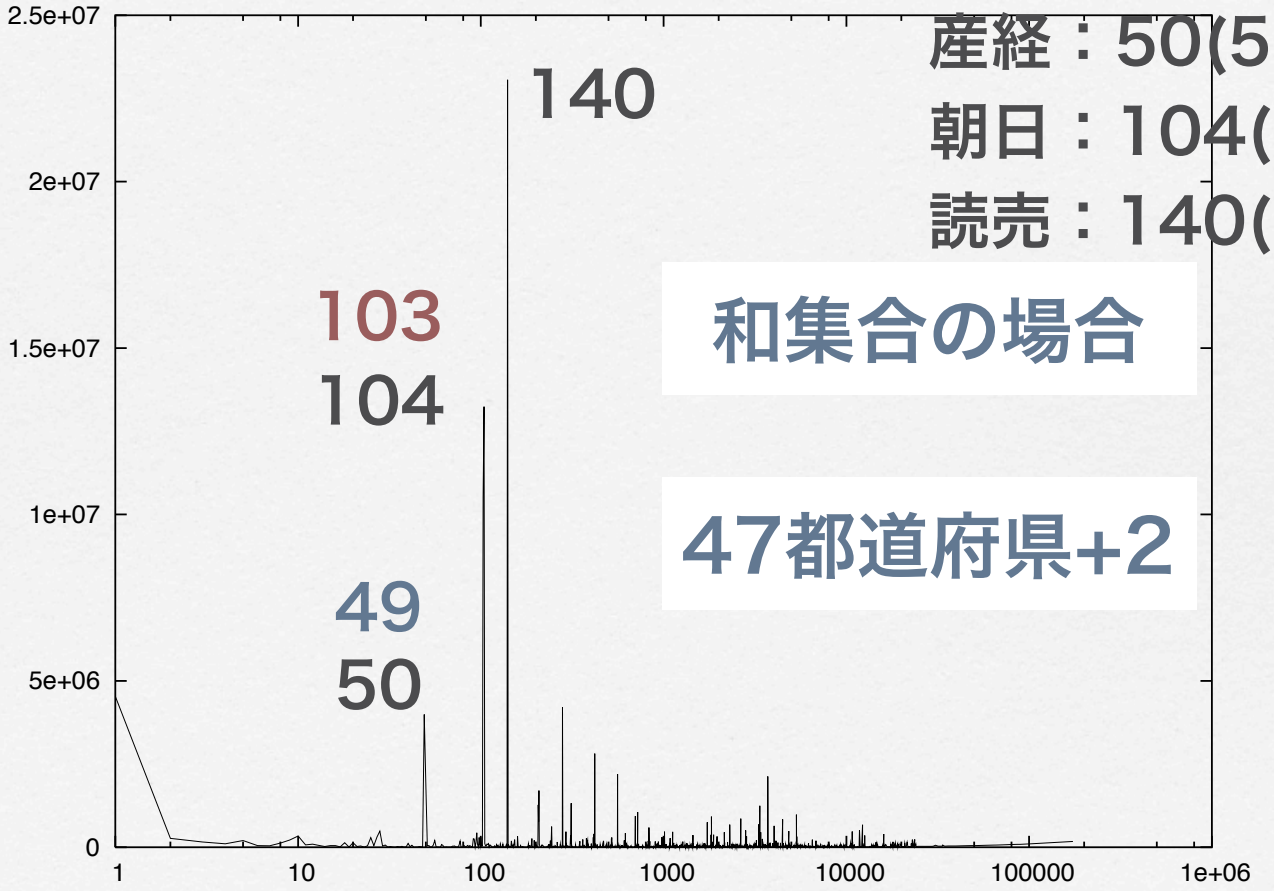
46ファイル
(1212Kbyte)
913件

複数の周期が異なる
パタンを発見

共通パタンと頻出
パタンの同時発見

3サイトのニュース記事

2003.12.12



産経：50(526Kb)

朝日：104(3412Kb)

読売：140(2324Kb)

和集合の場合

47都道府県+2

29

正規表現との関係

半構造化データに必須

- 正則パタン言語族は正規言語族の真の部分クラス⇒表現力に劣る
- 1回以上の繰り返し $(abx)^+$
  - 同一ファイル内で出現回数が増える
  - abは共通でかつ(ファイル内で)頻出
- 複数パタンの和 $(\alpha|\beta|\ldots)$

2003.12.12

九大サイト内

九大トップページから深さ3まで

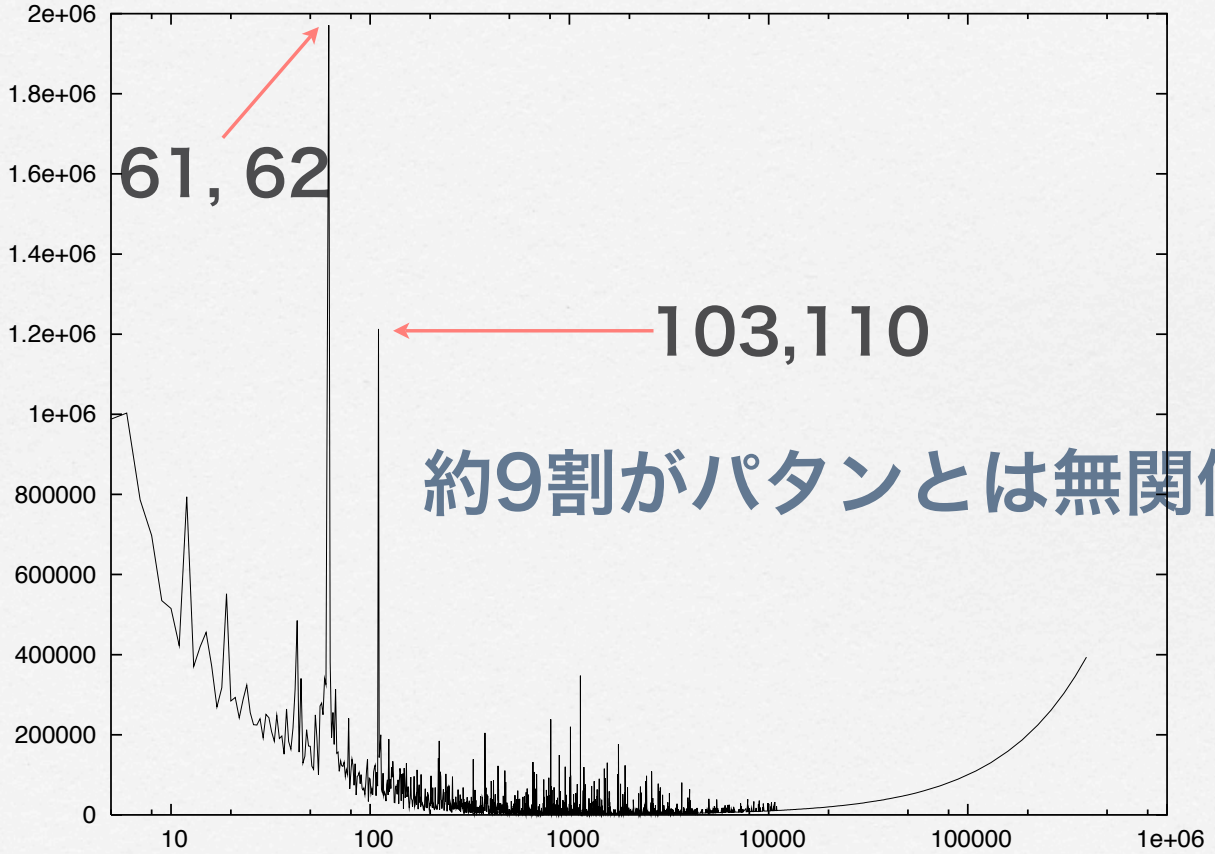
リンクをたどり598(5584Kb)ファイル

http://www.kyushu-u.ac.jp/

31

2003.12.12

九大サイト内(Cont.)



約9割がパタンとは無関係

九大サイト内(Cont.)

33

まとめ

34

まとめ

- テンプレート発見問題の定式化
  - 正例とベキ分布
- アルゴリズムの提案と計算量
  - 頻度 f を持つ部分文字列による共通部分列の表現
  - $F(f)$ による部分文字列増幅法
  - 計算量　$O(nm|t|^2 \log |t|)$

まとめ(Cont.)

- 情報抽出の実験
  - HTMLファイルを利用
- 検証された結果
  - ベキ分布の検証
  - $F(f)$による共通部分列の抽出
  - 繰り返しや和の演算にも対応可能
  - ノイズ耐性の高さ

今後の課題

- 近似文字列への拡張
  - ○○%の違いを許容
- ゲノム、異常値検出等への応用
- 再現率と適合率による評価
- 誤り率の平均的解析よる見積もり

37