Matching Patterns with Variables under Edit Distance

Paweł Gawrychowski Florin Manea **Stefan Siemer**

University of Wrocław, Göttingen University

gawry@cs.uni.wroc.pl, florin.manea@cs.uni-goettingen.de, stefan.siemer@cs.uni-goettingen.de

RP 2022

Introduction

Pattern Matching

Pattern Matching Problem

Find occurrences of a pattern α in a word w.

Example: $\alpha = \mathtt{bab}$, $w = \mathtt{aaaababbb}$

aaaababbb

bab

Pattern Matching with Variables

Exact Matching Problem for P: $\texttt{Match}_P$

Input: A pattern $\alpha \in P$, with $|\alpha| = m$, a word w, with $|w| = n$.

Question: Is there a substitution h with $h(\alpha) = w$?

Example: $\alpha = \mathtt{x}_1\mathtt{x}_1\mathtt{babx}_2\mathtt{x}_2$, $w = \mathtt{aaaababbb}$

$\mathtt{x}_1 \to \mathtt{aa}$ aaaababbb

$\mathtt{x}_2 \to \mathtt{b}$ aaaababbb

Example: $\alpha = \mathtt{x}_1\mathtt{babx}_2$, $w = \mathtt{aaaababbb}$

Edit Distance

Edit Distance

String metric

For words $u, w \in \Sigma^\star$, the *edit distance* between u and w is defined as the minimal number $d_{\mathrm{ED}}(u, w)$ of letter insertions, letter deletions, and letter to letter substitutions which one has to apply to u to obtain w.

Example: $u = \mathtt{abbab}, \quad w = \mathtt{baaba}, \quad \Delta = 3$

$$\mathtt{abbab} \xrightarrow[\textit{deletion}]{} \mathtt{bbab} \xrightarrow[\textit{substitution}]{} \mathtt{baab} \xrightarrow[\textit{insertion}]{} \mathtt{baaba}$$

Pattern Matching with Variables and Edit Distance

Approximate Matching Decision Problem for P: $\texttt{MisMatch}_P$

Input: A pattern $\alpha \in P$, with $|\alpha| = m$, a word w, with $|w| = n$, an integer $\Delta \leq m$.

Question: Is $d_{\texttt{ED}}(\alpha, w) \leq \Delta$?

Example: $\alpha = \texttt{x}_1\texttt{x}_1\texttt{bab}\texttt{x}_2\texttt{x}_2$, $w = \texttt{aaababbb}$, $\Delta = 1$

$\texttt{x}_1 \to \texttt{aa}$

$\texttt{x}_2 \to \texttt{b}$

aaaababbb

aaababbb

For Hamming Distance results see paper at MFCS 2021.

Pattern Matching with Variables and Edit Distance

Approximate Matching Minimisation Problem for P: $\texttt{MinMisMatch}_P$

Input: A pattern $\alpha \in P$, with $|\alpha| = m$, a word w, with $|w| = n$.

Question: Compute $d_{\texttt{ED}}(\alpha, w)$.

Example: $\alpha = \texttt{x}_1\texttt{x}_1\texttt{babx}_2\texttt{x}_2$, $w = \texttt{aaababbb}$

$\texttt{x}_1 \to \texttt{aa}$ $\texttt{aaaa}\texttt{bab}\texttt{bb}$

$\texttt{x}_2 \to \texttt{b}$ $\texttt{aaababbb}$

Regular

Regular Pattern Definition

Definition

$$\alpha \in \texttt{Reg} \text{ if } \alpha = w_0 \prod_{i=1}^{M} (x_i w_i), \text{ with } w_i \in \Sigma^\star.$$

Example: $\alpha = \texttt{abx}_1\texttt{abx}_2\texttt{x}_3\texttt{baab}$.

Context

- $\texttt{MisMatch}_{\texttt{Reg}}$ with distance $\Delta = 0$ in $\mathcal{O}(n)$.

- $\texttt{MisMatch}_{\texttt{Reg}}$ (HD) in $\mathcal{O}(n\Delta)$.

- $x_1 w_1 x_2$ known as Landau and Vishkin Algo in $\mathcal{O}(n\Delta)$.

- We extend the idea of Landau and Vishkin to achieve $\mathcal{O}(n\Delta)$.

Algorithm outline

- Interpret regular variables as free insertions on that position.

- First dynamic programming in $\mathcal{O}(nm)$.

- For increasing distance calculate maximal matching expansion on a diagonal in the DP matrix (Landau Vishkin idea).
- Consider surrounding diagonals and furthest previously reached variable + LCP extensions.
- Compute Δ distance increment steps for n diagonals, hence $\mathcal{O}(n\Delta)$.

Rectangular lower bound

Hardness follows directly from rectangular lower bound of edit distance (Backurs, Indyk). $\alpha = xuy$, where u is a string of terminals and x and y are variables.

Theorem

$\texttt{MisMatch}_{\texttt{Reg}}$ *can not be solved in time* $\mathcal{O}(|w|^h \Delta^g)$ *(or* $\mathcal{O}(|w|^h |\alpha|^g)$*) where* $h + g = 2 - \epsilon$ *with* $\epsilon > 0$*, unless the Orthogonal Vectors Conjecture fails*

Repeated Variables

1Var Definition

Definition

$\alpha \in \mathtt{1Var}$ if there exists exactly one variable $\mathtt{x}_1$ with several occurences.

Example: $\alpha = \mathtt{abx}_1\mathtt{x}_1\mathtt{abx}_1\mathtt{x}_1\mathtt{x}_1\mathtt{baabx}_1 \in 1\mathit{Var}$.

Algorithm outline

Theorem

$\texttt{MisMatch}_{\texttt{1Var}}$ and $\texttt{MinMisMatch}_{\texttt{1Var}}$ can be solved in $O(n^{3|\alpha|_{x_1}})$ time, where x_1 is the single variable occurring in α.

- Brute force possible intervals for all occurrences of $\texttt{x}_1$.
- Edit Distance Median-String problem for selected intervals (Sankoff 75).

Reduction

Theorem

$\texttt{MisMatch}_{\texttt{1Var}}$ *is* $W[1]$*-hard w.r.t. the number of occurrences of the single variable* x *of the input pattern* α*.*

Median String: $\texttt{MS}$

Input: k strings $w_1, \ldots, w_k \in \sigma^*$ and an integer Δ.

Question: Does there exist a string s such that $\sum_{i=1}^{k} d_{\texttt{ED}}(w_i, s) \leq \Delta$?
(The string s for which $\sum_{i=1}^{k} d_{\texttt{ED}}(w_i, s)$ is minimum is called the median string of the strings $\{w_1, \ldots, w_k\}$.)

PM instance: w encodes the k strings seperated by long borders of two fresh symbols. α encodes the same long borders with x_1 in between.

Overview

Result Overview

Class	$\texttt{Match}(w, \alpha)$	$\texttt{MisMatch}(w, \alpha, \Delta)$ for $d_{\texttt{HAM}}(\cdot,\cdot)$	$\texttt{MisMatch}(w, \alpha, \Delta)$ for $d_{\texttt{ED}}(\cdot,\cdot)$
Reg	$O(n)$ [folklore]	$O(n\Delta)$, matching cond. lower bound	$O(n\Delta)$, matching cond. lower bound
1Var ($\texttt{var}(\alpha) = \{x\}$)	$O(n)$ [folklore]	$O(n)$	$O(n^{3\lvert\alpha\rvert_x})$ W[1]-hard w.r.t. $\lvert\alpha\rvert_x$
NonCross	$O(nm \log n)$ [1]	$O(n^3 p)$	NP-hard
1RepVar k=# x-blocks	$O(n^2)$ [1]	$O(n^{k+2}m)$ W[1]-hard w.r.t. k	NP-hard for $k \geq 1$
kLOC	$O(mkn^{2k+1})$ [2] W[1]-hard w.r.t. k	$O(n^{2k+2}m)$ W[1]-hard w.r.t. k	NP-hard for $k \geq 1$
kSCD	$O(m^2 n^{2k})$ [1] W[1]-hard w.r.t. k	NP-hard for $k \geq 2$	NP-hard for $k \geq 1$
kRepVar	$O(n^{2k})$ [1] W[1]-hard w.r.t. k	NP-hard for $k \geq 1$	NP-hard for $k \geq 1$

[1] Henning Fernau, Florin Manea, Robert Mercas, and Markus L. Schmid. Pattern matching with variables...

[2] Joel D. Day, Pamela Fleischmann, Florin Manea, and Dirk Nowotka. Local patterns

[3] Daniel Reidenbach and Markus L. Schmid. Patterns with bounded treewidth.

Outlook

Extensions & Outlook

- Matching lower and upper bounds for non-cross (HD).
- Enumeration algorithms
- Bounding the number of variables.
- Other metrics (Damerau-Levenshtein).
- Combining exact and approximate matching.
- Restrictions on Variables (e.g. RegEx membership).
- Analysing use cases (e.g. database theory, learning theory).

Extensions & Outlook

- Matching lower and upper bounds for non-cross (HD).
- Enumeration algorithms
- Bounding the number of variables.
- Other metrics (Damerau-Levenshtein).
- Combining exact and approximate matching.
- Restrictions on Variables (e.g. RegEx membership).
- Analysing use cases (e.g. database theory, learning theory).

Thank you.