Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Generalized pattern matching and periodicity under substring consistent equivalence relations

Yoshiaki Matsuoka [a], Takahiro Aoki [b], Shunsuke Inenaga [a,*], Hideo Bannai [a], Masayuki Takeda [a]

[a] Department of Informatics, Kyushu University, Japan
[b] Department of Electrical Engineering and Computer Science, Kyushu University, Japan

ARTICLE INFO

Article history:
Received 1 October 2015
Received in revised form 2 February 2016
Accepted 12 February 2016
Available online xxxx

Keywords:
String matching algorithm
Generalized periodicity lemma
Order-preserving pattern matching
Parameterized pattern matching

ABSTRACT

Let $\approx$ be a *substring consistent equivalence relation* (*SCER*) on strings such that for any two strings x, y, $x \approx y$ implies that (1) $|x| = |y|$ and (2) $x[i..j] \approx y[i..j]$ for all $0 \leq i \leq j < |x|$. Examples of SCER are parameterized pattern matching and order-preserving pattern matching. We present a generalized and efficient algorithm for pattern matching with SCER $\approx$. Also, we show analogues of Fine and Wilf's periodicity lemma hold for SCER.

1. Introduction

An equivalence relation $\approx$ on strings is called a *substring consistent equivalence relation* (*SCER* in short), if for any two strings x, y, $\approx$ satisfies the following: $x \approx y$ implies that (1) $|x| = |y|$ and (2) $x[i..j] \approx y[i..j]$ for all $0 \leq i \leq j < |x|$. Many notions of non-standard string matching in applied fields such as bioinformatics are based on such equivalence relations; for instance, the isomorphism used for parameterized matching [3], DNA/RNA complements, palindrome pattern matching [8], and order-preserving matching [12,10].

The contribution of this paper is twofold: First, we propose a new generalized algorithm for pattern matching in terms of SCER $\approx$. Second, we report new results concerning periodicities of strings in terms of SCER $\approx$. Further details are described in the following subsections.

1.1. Algorithm for pattern matching with SCERs

The classical exact pattern matching problem is a task to find all occurrences of a query pattern P in a text string T, namely, to find all positions i of T such that $T[i..i + |P| - 1] = P$. This problem can be solved in $O(n + m)$ optimal time using e.g., well-known Morris and Pratt's algorithm [15], where m is the length of pattern P and n the length of text T.

In this paper, we consider the pattern matching problem with SCER $\approx$, namely, to find all beginning positions i of T such that $T[i..i + |P| - 1] \approx P$, which is a natural extension to the above exact pattern matching problem. Let x and y

* Corresponding author.
E-mail addresses: yoshiaki.matsuoka@inf.kyushu-u.ac.jp (Y. Matsuoka), inenaga@inf.kyushu-u.ac.jp (S. Inenaga), bannai@inf.kyushu-u.ac.jp (H. Bannai), takeda@inf.kyushu-u.ac.jp (M. Takeda).

http://dx.doi.org/10.1016/j.tcs.2016.02.017
0304-3975/

be any substrings of T and P of same length, respectively. Suppose that there is an algorithm which, after $\tau_{\approx}(n,m)$-time preprocessing on T and P, determines whether or not $x \approx y$ in $\xi_{\approx}(n,m)$ time provided that $x[0..|x|-2] \approx y[0..|y|-2]$ is already known. Then, we show that there exists a Morris–Pratt type algorithm which solves the pattern matching problem with $\approx$ in $O(\tau_{\approx}(n,m) + \xi_{\approx}(n,m)(n+m))$ time.

The motivation for our generalized pattern matching algorithm is to help further developments of pattern matching algorithms with new, unknown SCERs. Suppose that someday one will initiate a study on pattern matching w.r.t. a new equivalence relation R, which is yet unknown to us today. If it appears that R satisfies the condition of SCER, then one will not need to design an algorithm from scratch in order to solve the problem. Instead, one only needs to design an efficient method for $\tau_R(n,m)$ and $\xi_R(n,m)$ so that our generalized pattern matching algorithm can be directly applied.

1.1.1. Related work

For exact matching isomorphism $=$, it is clear that $\tau_{=}(n,m) = \xi_{=}(n,m) = O(1)$, and thus our result matches the optimal $O(n+m)$-time bound for $=$ (e.g., [15,11]).

Two strings x, y of same length are said to *order preserving match* (denoted $x \overset{\text{op}}{\approx} y$), if the ranks of the characters of x and y at each position are equal, namely, $x[i] \prec x[j] \Leftrightarrow y[i] \prec y[j]$ for all $0 \le i \le j \le |x|-1$, where $\prec$ denotes the lexicographical order of characters. If T and P are drawn from an integer alphabet of size $O(\text{poly}(n))$, then it is known that $\tau_{\overset{\text{op}}{\approx}}(n,m) = O(n)$ and $\xi_{\overset{\text{op}}{\approx}}(n,m) = O(1)$. Thus, our result matches the optimal $O(n+m)$-time bound [12,5] for order preserving matching. If T and P are drawn from a general ordered alphabet, then $\tau_{\overset{\text{op}}{\approx}}(n,m) = O(m \log m)$ and $\xi_{\overset{\text{op}}{\approx}}(n,m) = O(1)$. Thus, our result matches the $O(n + m\log m)$-time bounds [12,5,10], which are state-of-the-art solutions for order preserving matching in this setting.

Two strings x, y of same length are said to *parameterized match* if there is a renaming bijection over the alphabet which transforms x into y. We write $x \overset{\text{pr}}{\approx} y$ iff x and y parameterized match. Amir et al. [1] showed $\tau_{\overset{\text{pr}}{\approx}}(n,m) = O(n + m\log\pi)$ and $\xi_{\overset{\text{pr}}{\approx}}(n,m) = O(\log\pi)$, where $\pi = \min\{|\Sigma|, m\}$. Then, our result matches the $O((n+m)\log\pi)$ bound [1]. It has been shown that this bound is optimal in the comparison model [1].

1.2. SCER version of Fine and Wilf's periodicity lemma

Periodicity of strings has been a central topic of stringology and combinatorics of words for decades, not only as mathematical interests but as powerful algorithmic tools on strings [6,13].

A positive integer p is said to be a *period* of a string w iff $w[i] = w[i+p]$ for all possible positions i in w. Probably the most well-known, and the most utilized, result on periodicities in strings is Fine and Wilf's *periodicity lemma* [7]: If two periods p, q of a string w of length n satisfies $p + q - \gcd(p,q) \le n$, then $\gcd(p,q)$ is also a period of w. Castelli et al. [4] extended Fine and Wilf's periodicity lemma for three periods, and then Justin [9] extended it to more than three periods.

We introduce three different definitions of periods of strings w.r.t. SCER $\approx$:

- The first kind, called *block-based* $\approx$*-period*, is based on the factorization of the string into blocks of length p each, possibly with the last block shorter than p.
- The second kind, called *sliding-window-based* $\approx$*-period*, is based on the sliding window of length p over the string.
- The third kind, called *border-based* $\approx$*-period*, is based on the border of length $n-p$ of the string, where a border is a suffix of the string which is also a prefix.

Although all these definitions turn out to be equivalent w.r.t. the exact equality $=$, it is not the case with SCER $\approx$ in general (see our example in Section 4 for order-preserving isomorphism $\overset{\text{op}}{\approx}$). For any set $\mathcal{P}$ of positive integers, let $b_{\text{opt}}(\mathcal{P})$ be the bound of Fine and Wilf's periodicity lemma for exact matching isomorphism $=$. Namely, if $\mathcal{P} = \{p, q\}$, then $b_{\text{opt}}(\mathcal{P}) = p + q - \gcd(p,q)$. $b_{\text{opt}}(\mathcal{P})$ for $|\mathcal{P}| \ge 3$ can be found in [9]. Then, we prove the following:

(1) For any SCER $\approx$, if a string w has a set $\mathcal{P}$ of block-based $\approx$-periods satisfying $b_{\text{opt}}(\mathcal{P}) \le n$, then $\gcd(\mathcal{P})$ is also a block-based $\approx$-period of w (Theorem 2).
(2) For any SCER $\approx$, if a string w has a set $\mathcal{P}$ of sliding-window-based $\approx$-periods satisfying $b_{\text{opt}}(\mathcal{P}) + \gcd(\mathcal{P}) - 1 \le n$, then $\gcd(\mathcal{P})$ is also a sliding-window-based $\approx$-period of w (Theorem 4).
(3) For order-preserving matching isomorphism $\overset{\text{op}}{\approx}$, if a string w has a set $\mathcal{P}$ of border-based $\overset{\text{op}}{\approx}$-periods satisfying $\gcd(\mathcal{P}) = 1$ and $b_{\text{opt}}(\mathcal{P}) + 1 \le n$, then $\gcd(\mathcal{P})$ is also a border-based $\overset{\text{op}}{\approx}$-period of w (Theorem 6). We also show that if $\gcd(\mathcal{P}) \ge 2$ in the above setting, then there exists an infinitely long string which has all border-based $\overset{\text{op}}{\approx}$-periods in $\mathcal{P}$ but does not have $\gcd(\mathcal{P})$ as its border-based $\overset{\text{op}}{\approx}$-period (Theorem 7).

We also show that all the above results (1)–(3) are optimal for order-preserving isomorphism $\overset{\text{op}}{\approx}$ (Theorems 3, 5, and 6).

Please cite this article in press as: Y. Matsuoka et al., Generalized pattern matching and periodicity under substring consistent equivalence relations, Theoret. Comput. Sci. (2016), http://dx.doi.org/10.1016/j.tcs.2016.02.017