Playing with patterns, searching for strings

Jaap van der Woude

87/13
This is a series of notes of the Computing Science Section of the Department of Mathematics and Computing Science of Eindhoven University of Technology.
Since many of these notes are preliminary versions or may be published elsewhere, they have a limited distribution only and are not for review.
Copies of these notes are available from the author or the editor.
§ 0

Pattern identification is a source of several instructive programming exercises. We shall present one such exercise that is especially interesting for problems dealing with periodicity. In particular it enables us to treat preprocessing and search in the Knuth-Morris-Pratt pattern search algorithm as a unity.

Some remarks on names and notations:
Let $\Sigma$ be a fixed alphabet. A word over $\Sigma$, i.e. an element of the free monoid $(\Sigma^*, \Lambda)$ generated by $\Sigma$, will be called a string. $\Sigma^* \setminus \{\Lambda\}$ will be denoted by $\Sigma^*$, the nonempty strings. (In the sequel, capitals refer to strings, lowercase letters to naturals (including 0) or functions, unless stated otherwise.)

Let $X$ be a string. In order to facilitate references to the length $|X|$ of $X$ and symbols occurring in $X$, we shall write $X(i : 0 \leq i < N)$. Then $|X| = N$ and $X(i)$ is the $i+1$st symbol of $X$.

If $n \leq N$ we shall write $X \downarrow n$ for $X(i : 0 \leq i < n)$, the prefix of $X$ with length $n$.

A string $X$ is called periodic if $X = P^m$ for some $P \in \Sigma^*$ and $m \geq 2$, where $P^m$ is the concatenation of $m$ copies of $P$. In that case, $P$ as well as $|P|$ is called a period of $X$. The period of $X$ is the smallest period (if any).

For strings $X$ and $Y$

$X \leq Y$ denotes "$X$ is a prefix of $Y$"

$X < Y$ means $X \leq Y \land X \neq Y$

So much for general remarks on names and notations.

First we shall state the basic problem (MPP) in its "historical" context, subsequently we shall give two applications

- Pattern-search (e.g. Knuth-Morris-Pratt)
- Periodicity-search (for all prefixes)
§ 1. The MPP problem

The problem we would like to consider evolved from the exercise below. We shall give some heuristics for this program evolution.

**Exercise ("carrécheck")**

A string $X$ is a *carré* if $X = P^2$ for some string $P \in \Sigma^+$. Derive a program to find every prefix of $X$ that is a carré.

The formal specification:

\[
\begin{aligned}
N & : \text{int} ; \; (N \geq 1) \\
X & (i : 0 \leq i < N) : \text{string} \\
c & (i : 1 \leq i \leq N) : \text{array of bool} \\
\text{CARRÉCHECK} & \{(A i : 1 \leq i \leq N : c(i) = \text{car}(X \downarrow i))\} \\
\end{aligned}
\]

where

\[
\begin{aligned}
\text{car}(X) & = (E P : P \in \Sigma^+ : X = PP).
\end{aligned}
\]

A standard feature in programming methodology is: weaken in the postcondition by replacement of a constant by a variable.

If we consider the term in definition (0) : $X = PP$ we might think of $P$ as being a constant. Symmetry tells us to replace $P$ by a variable twice. So $X = PP$ might be "generalized" to $X = PE \land X = FP$ for variables $E$ and $F$.

I.e. $P$ is a pre- and postfix of $X$.

So for strings $P$ and $X$ we define

\[
(1) \quad P \ pp X \ = \ (E E, F : E, F \in \Sigma^+ : X = PE \land x = FP)
\]

and note that $P \ pp X \land |X| = 2 \ast |P| \Rightarrow \text{car}(X)$.

The lack of reflexivity of $pp$, created by the domain in (1), seems unnatural, but it results from the definition of periodicity, i.e. the domain in definition (0).

Moreover, incorporating reflexivity of $pp$ leads to additional non-triviality analysis (in *mpp* below for example).

First we give a few properties of $pp$, the simple proofs are omitted. Let $H$, $P$ and $X$ be strings and $h, x \in \Sigma$, then

\[
\begin{aligned}
(2) & \quad H \ pp P \land P \ pp X \Rightarrow H \ pp X \\
(3) & \quad H \ pp X \land P \ pp X \land |H| < |P| \Rightarrow H \ pp P \\
(4) & \quad Hh \ pp Xx = Hh < Xx \land H \ pp X \land h = x
\end{aligned}
\]
Property 4 will be used for prefixes of a fixed string $X$:

(4') \[ X \downarrow k + 1 \quad \text{pp} \quad X \downarrow n + 1 = \ X \downarrow k \quad \text{pp} \quad X \downarrow n \land X(k) = X(n) \]

(4'') \[ \neg \ X \downarrow k + 1 \quad \text{pp} \quad X \downarrow n + 1 \iff \neg \ X \downarrow k \quad \text{pp} \quad X \downarrow n \lor X(k) \neq X(n) \]

By 2, the transitivity of pp, we feel invited to consider maximal pre- and postfixed. So define:

\[(5) \quad P \ \text{mpp} \ X = P \ \text{pp} \ X \land (A \ H : H \text{pp}X : H=P \lor H \text{pp}P) \]

Note that the following are equivalent

\[ a \ P \ \text{mpp} \ X \]
\[ b \ P \ \text{pp} \ X \land (A \ H : H \text{pp}X : |H| \leq |P|) \]
\[ c \ P \ \text{pp} \ X \land (A \ H : H \text{pp}X : H \leq P) \]

In section 6 we show that, given $P \ \text{mpp} \ X$, we have

\[ \text{car}(X) = \lfloor X \rfloor \mod (2 \ast (\lfloor X \rfloor - 1 \ast P)) = 0 \]

So indeed the generalization (weakening of carré) is fruitful. The carrécheck problem has evolved to the MPP problem:

Derive a program that calculates for every prefix of a given string its maximal pre- and postfix.

The formal specification

\[ \begin{align*}
\text{MP} \ &= \begin{cases}
\forall[N \ : \ \text{int} \ ; \ \{N \geq 1\}] \\
X(i : 0 \leq i < N) : \text{string} ;
\end{cases} \\
\text{MPP} \ &= \begin{cases}
\forall \forall \ (i : 1 \leq i \leq N) : \text{array of } [0..N) ;
\end{cases} \\
& \quad \text{MPP} \quad \{(A i : 1 \leq i \leq N : X \downarrow f(i) \text{mpp}X \downarrow i)\} \\
\end{align*} \]
§ 2. Solution to the MPP problem

Forced by the postcondition of MPP:

\[ R_0 \quad (A \; i : 1 \leq i \leq N : X \downarrow f(i) \text{mpp} X \downarrow i) \]

we choose the following invariants:

\[ P_0 \quad (A \; i : 1 \leq i \leq n : X \downarrow f(i) \text{mpp} X \downarrow i) \]

\[ P_1 \quad 1 \leq n \leq N. \]

**Approximation 0 (for MPP)**

\[ n := 1 ; f : (1) = 0 \quad \{ P_0 \land P_1 \} \]

\[ \text{do } n \neq N \rightarrow S_0 \{ X \downarrow k \text{ mpp} X \downarrow n +1 \} \]

\[ \quad ; f : (n +1) = k \quad \{ (P_0 \land P_1)_{n+1} \} \]

\[ \quad ; n := n +1 \quad \{ P_0 \land P_1 \} \]

\[ \text{od } \{ P_0 \land P_1 \land n = N, \text{ hence } R_0 \} \]

On cosmetical grounds, with 4' in mind, we consider a slightly different postcondition for \( S_0 \):

\[ R_1 \quad X \downarrow k +1 \text{ mpp} X \downarrow n +1 \]

By definition of \( \text{mpp} \) (version 5b) and by 4', \( R_1 \) equivales

\[ X \downarrow k \quad ppX \downarrow n \land X(k) = X(n) \land (A \; j : X \downarrow j \; pp \downarrow n +1 : j \leq k +1) \]

This leads us to a repetition for \( S_0 \) with guard \( X(k) \neq X(n) \) and invariants:

\[ Q_0 \quad (A \; j : X \downarrow j \; pp \downarrow n +1 : j \leq k +1) \]

\[ Q_1 \quad X \downarrow k \quad ppX \downarrow n \land k \geq 0 \]

**Approximation 1 (for \( S_0 \))**

\[ k := f(n) \quad \{ Q_0 \land Q_1 , \text{ see the note below} \} \]

\[ \quad \text{do } X(k) \neq X(n) \land k \neq 0 \]

\[ \rightarrow \text{ "decrease } k \text{ under invariance of } Q \text{ (and } P\)" \]

\[ \quad \text{od } \{ Q_0 \land Q_1 \land (X(k) = X(n) \lor k = 0) \} \]

The second conjunct in the guard is forced upon us by the wish to decrease \( k \), leaving \( k \geq 0 \) invariant.

**note**

\[ P_0 \]

\[ \Rightarrow \{ \text{instantiation at } n , \text{ def. mpp (5b)} \} \]

\[ X \downarrow f(n) \quad ppX \downarrow n \land (A \; j : X \downarrow j \; pp \downarrow n : j \leq f(n)) \]

\[ \Rightarrow \{ \text{by 4': } X \downarrow j +1 \; ppX \downarrow n +1 \Rightarrow X \downarrow j \; ppX \downarrow n \} \]

\[ X \downarrow f(n) \quad ppX \downarrow n \land (A \; j : X \downarrow j +1 \; ppX \downarrow n +1 : j \leq f(n)) \]

\[ \Rightarrow \{ \text{dummy change} \} \]

\[ X \downarrow f(n) \quad ppX \downarrow n \land (A \; j : X \downarrow j \; ppX \downarrow n +1 : j \leq f(n) +1) \]

\[ = \{ \text{def. } Q_0, Q_1 \} \]

\[ (Q_1 \land Q_0)_{f(n)} \]
Under assumption of $Q_0 \land Q_1 \land P_0 \land P_1$ and the guard, we study reduction of $k$. First with respect to $Q_0$:

Let $j > 0$ then

$$X \downarrow j \ensuremath{pp} X \downarrow n + 1$$

$$\Rightarrow \{ Q_0 ; \text{by } 4'' \, , \, X(k) \neq X(n) \Rightarrow X \downarrow k + 1 \ensuremath{pp} X \downarrow n + 1 \}$$

$$X \downarrow j \ensuremath{pp} X \downarrow n + 1 \land j \leq k + 1 \land j \neq k + 1$$

$$\Rightarrow \{ j \geq 0 , 4'' \}$$

$$X \downarrow j - 1 \ensuremath{pp} X \downarrow n \land j - 1 < k$$

$$\Rightarrow \{ Q_1 , 3 \}$$

$$X \downarrow j - 1 \ensuremath{pp} X \downarrow k$$

$$\Rightarrow \{ P_0 , 0 \neq k < n \, , \, \text{def. } mpp \ 5a \}$$

$$j - 1 \leq f(k)$$

For $j = 0$, certainly we have $j \leq f(k)+1$. Hence $Q_0 \not\in 4)$ holds.

With respect to $Q_1$:

$$P_0 \land Q_1$$

$$\Rightarrow \{ k \neq 0 : P_0 \text{ instantiated at } k \}$$

$$X \downarrow f(k) \ensuremath{pp} X \downarrow k \land X \downarrow k \ensuremath{pp} X \downarrow n$$

$$\Rightarrow \{ 2 \}$$

$$X \downarrow f(k) \ensuremath{pp} X \downarrow n$$

$$= \{ \text{def. } Q_1 , f(k) \geq 0 \}$$

$$Q_1 \not\in 4(b)$$

This shows that "decrease $k$ under invariance ..." is established by $k := f(k)$. (Certainly $P_0 \land P_1$ is not affected.)

Because of the conjunct $k \neq 0$ in the guard, neither $R_1$ not the original postcondition of $S_0$ are met, but a mixture is:

for

$$Q_0 \land Q_1 \land X(k) = X(n) \Rightarrow R_1$$

and

$$Q_0 \land Q_1 \land X(k) \neq X(n) \land k = 0$$

$$\Rightarrow \{ 4'' \}$$

$$Q_0 \land Q_1 \land \neg X \downarrow k + 1 \ensuremath{pp} X \downarrow n + 1 \land k = 0$$

$$\Rightarrow \{ X \downarrow 0 \ensuremath{pp} X \downarrow n + 1 , \text{def. } mpp \}$$

$$X \downarrow k \ensuremath{mpp} X \downarrow n + 1 \land k = 0$$
This proves the following solution for MPP:

\[
\begin{align*}
  n &:= 1 ; f : (1) = 0 \{ P_0 \land P_1 \} \\
  \text{do } n \neq N \\
  \rightarrow [ k : \text{int} ; \\
  & \quad k := f(n) \{ Q_0 \land Q_1 \} \\
  & \quad \text{do } X(k) \neq X(n) \land k \neq 0 \\
  & \quad \rightarrow k := f(k) \\
  & \quad \text{od } \{ Q_0 \land Q_1 \land (X(k) = X(n) \lor k = 0) \} \\
  & \quad \text{if } X(k) = X(n) \rightarrow (X \downarrow k + 1 \text{ mpp } X \downarrow n + 1) k := k + 1 \\
  & \quad \text{fi } \{ X \downarrow k \text{ mpp } X \downarrow n + 1 \} \\
  & \quad ; f : (n + 1) = k \{ (P_0 \land P_1)^n + 1 \} \\
  \} \\
  \od \{ P_0 \land P_1 \land n = N, \text{ hence R_0} \}
\end{align*}
\]

For the complexity of the algorithm, consider \( k \) to exist outside the innerblock \((P_2: k = f(n))\).

A variant function that shows linearity is \( 2N - 2n + k \).
§ 3. Pattern search

Let \( P \in \Sigma^* \) be fixed, the pattern. Suppose we are interested in (all) occurrences of \( P \) in a string \( Z \).

Put \( Y = PZ \), then we are searching for numbers \( n \), such that

\[
  n \geq 2 \times |P| \land P \text{pp} Y \downarrow n
\]

In this setting the pattern might be represented by its length only.

Let \( X \) be a string, \( p \) a number \( 1 \leq p < |X| \).

As \( X \downarrow p \) is a postfix of \( X \downarrow n \) if \( p = n \lor X \downarrow p \text{pp} X \downarrow n \), we define (the occurrence of the pattern as postfix of \( X \downarrow n \)):

\[
  (6) \quad O(n) = p = n \lor X \downarrow p \text{pp} X \downarrow n
\]

Let \( f \) be as in the MPP problem. It seems reasonable to hope for suitable \( O \)-information in the \text{mpp}-knowledge recorded in \( f \). Indeed, for strings \( H, P \) and \( X \) we have

\[
  (7) \quad P \text{mp} P X \Rightarrow (H \text{pp} X = H = P \lor H \text{pp} P).
\]

Property 7, which is closely linked to 3, follows easily from 5a, 2.

It relates \( O(n) \) to \( f(n) \) as follows:

\[
  (6) \quad O(n) = p = n \lor X \downarrow p \text{pp} X \downarrow n
\]

\[
  = \{ X \downarrow f(n) \text{mp} P X \downarrow n \}, \text{7 with } H, P, X := X \downarrow p, X \downarrow f(n), X \downarrow n \}
\]

\[
  = (X \downarrow p = X \downarrow f(n) = p = f(n); 6)
\]

As \( f(n) < n \) (nonreflexivity of \( \text{pp} \)), \( O(n) \) depends only on \( f(n) \) and \( O(i : 1 \leq i < n) \). This settles pattern search as a simple extension of the MPP problem, by adding invariant

\[
  (A i : 1 \leq i \leq n : o(n) = O(n))
\]

and initialization \( o : (1) = (p = 1) \) \( (P_3) \)

extra statement : \( o : (n + 1) = (p = n + 1) \lor o(f(n + 1)) \) \( (P_3^* + 1) \)

immediately following the innerblock.
§ 4. Knuth - Morris - Pratt

The pattern search presented in the previous section has a serious drawback: storage linear in the length of the given string (concatenated with the pattern). Indeed, the algorithm needs $f(i : 1 \leq i \leq n)$ to calculate $f(n + 1)$ and $O(i : 1 \leq i \leq n)$ and $f(n + 1)$ to calculate $O(n + 1)$.

As, for fixed $p$, we are interested in $n$ such that $X \downarrow_p pp X \downarrow n$ (instead of $mpp$!) the information recorded in $f$ exceeds our needs: we might do with pre- and postfixes with lengths at most $p$. So we define: ($P$ and $p$ are not related!)

\begin{align}
(8) & \quad P \ p \pi X = P \ \text{pp} X \wedge \ 1 \ p \ 1 < p \\
(9) & \quad P \ m \pi X = P \ \pi X \wedge (A H : H \ p \pi X : H = P \lor H \ p \pi p)
\end{align}

The reader is urged to convince himself of the truth of the $\pi\pi$-versions of 2, 3, 5$^a$, $^b$, $^c$ (i.e. only $(m)pp$ replaced by $(\mu) \pi \pi$). Property 4, however, has a slightly different $\pi\pi$-version.

We shall only provide the $\pi\pi$-version of 4:

\begin{align}
(10') & \quad X \downarrow k \ p \pi X \downarrow n \wedge X(k) = X(n)
\end{align}

The $\mu\pi\pi$ problem is given by the postcondition

$\rho_0 \quad (A i : 1 \leq i \leq N : X \downarrow \phi(i) \mu\pi\pi X \downarrow i)$

For the solution of the $\mu\pi\pi$ problem we define invariants, (the obvious adaptations of the invariants for MPP)

$\pi_0 \quad (A i : 1 \leq i \leq n : X \downarrow \phi(i) \mu\pi\pi X \downarrow i)$

$\pi_1 \quad 1 \leq n \leq N$

$\psi_0 \quad (A j : X \downarrow j \ p \pi X \downarrow n + 1 : j \leq k + 1)$

$\psi_1 \quad X \downarrow k \ p \pi X \downarrow n \wedge k \geq 0$

Except from the obvious adaptations, $\mu\pi\pi$ differs from MPP only in the case analysis in the innerblock: the drawback of 10:

Certainly,

$\psi_0 \land \psi_1 \land X(k) = X(n) \land k < p - 1$

$\Rightarrow \{10', \ \text{def. } \mu\pi\pi\}$

$X \downarrow k + 1 \ \mu\pi\pi X \downarrow n + 1$

but

$\psi_0 \land \psi_1 \land X(k) = X(n) \land k = p - 1$

$\Rightarrow \{10', k + 1 = p\}$

$X \downarrow p \ \text{pp} X \downarrow n + 1$

$\Rightarrow \{\pi_0, \ \text{instantation at } p ; 2\}$

$X \downarrow \phi(p) \ \mu\pi\pi X \downarrow n + 1$
This shows that the alternative statement following the inner repetition should be changed (for the \( \mu \pi \pi \) problem) to

\[
\{ \psi_0 \land \psi_1 \land (X(k) = X(n) \lor k = 0) \}
\]

\( \land X(k) = X(n) \land k < \rho - 1 \rightarrow (X \downarrow k + 1 \mu \pi \pi X \downarrow n + 1) \} k := k + 1
\]

\( \land X(k) = X(n) \land k = \rho - 1 \rightarrow (X \downarrow \phi(p) \mu \pi \pi X \downarrow n + 1) \} k := \phi(p)
\]

\( \land X(k) \neq X(n) \rightarrow (k = 0 \land X \downarrow k \mu \pi \pi X \downarrow n + 1) \) skip

\( \land \}

This and the change from \( f \) to \( \phi \) make the solution of MPP to a solution of \( \mu \pi \pi \). The code will reappear in the Knuth-Morris-Pratt pattern search algorithm, so we shall leave it with this. Note that for calculation of \( \phi(n + 1) \) only \( \phi(n) \) and \( \phi(i : 1 \leq i \leq \rho) \) are needed : \( \phi(n) \) for initializing \( k \), \( \phi(i : 1 \leq i \leq \rho) \) in the inner repetition. I.e. \( \mu \pi \pi \) needs storage proportional to the "pattern length".

Similar to § 3, we now transform \( \mu \pi \pi \) to a pattern search algorithm: Knuth-Morris-Pratt. With respect to occurrence of \( X \downarrow \rho \) as postfix of \( X \downarrow n + 1 \), note that

\( O(\rho + 1) \)

\( p = \rho + 1 \land X \downarrow \rho \pi \pi X \downarrow n + 1 \)

\( = (4'; \text{def. } \pi \pi) \)

\( p = \rho + 1 \lor (X \downarrow \rho - 1 \pi \pi X \downarrow n \land X(\rho - 1) = X(n)) \)

\( = (\text{def. of } \pi \pi : \psi_{\phi^{-1}}^{k} = \text{true; def. } \psi) \)

\( p = \rho + 1 \lor (\psi_{0} \land \psi_{1} \land X(k) = X(n))_{\rho - 1}^{k} \)

So calculation of \( O(\rho + 1) \) depends only on the postcondition of the inner repetition and occurrence is to be signalled in the second alternative (the occurrence at \( n + 1 = \rho \) is not relevant of course).

We are now ready for the Knuth-Morris-Pratt algorithm. As only \( \phi(i : 1 \leq i \leq \rho) \) and \( \phi(n) \) are needed to calculate \( \phi(n + 1) \), we have to distinguish between

preprocessing — "filling \( \phi \"

search — "signalling occurrences".

This separation of the two parts is inevitable, but an earlier separation is unnecessary, unelegant and confusing. In order to account for the reduced domain of \( \phi \) we modify \( \pi_0 \) to \( \pi_0 \), to "buffer" \( \phi(n) \) we add \( \pi_2 \)

\( \pi_0 \)  \( \{ \Theta i : 1 \leq i \leq \rho \min n : X \downarrow \phi(i) \pi \pi X \downarrow i \}
\)

\( \pi_2 \)  \( \downarrow k \mu \pi \pi X \downarrow n \)

and we take \( k \) outside the inner repetition.
The Knuth-Morris-Pratt pattern search algorithm we derived:

```plaintext
1{ k : int ;
   \phi(i : 1 \leq i \leq p) : array of [0 . . p - 1) ;
   n , k := 1 , 0 ; \phi(1) = 0 \{ \pi_0 \land \pi_1 \land \pi_2 \}
   ; do n \neq N
   \rightarrow (\pi_0 \land \pi_1 \land \pi_2 \land n \neq N , so \psi_0 \land \psi_1}
   do X(k) \neq X(n) \land k \neq 0
   \rightarrow k := \phi(k)
   od \{ \psi_0 \land \psi_1 \land (X(k) = X(n) v k = 0) \}
   if X(k) = X(n) \land k < p - 1 \rightarrow k := k + 1
   \{ X(k) = X(n) \land k = p - 1 \rightarrow k := \phi(p) ; "MATCH" \}
   \{ X(k) \neq X(n) \rightarrow \text{skip} \}
   fi \{ X \downarrow k \land \pi \pi X \downarrow n + 1 , so \pi_2 \}
   ; if n < p \rightarrow \phi : (n + 1) = k
   \{ n \geq p \rightarrow \text{skip} \}
   fi \{ (\pi_0 \land \pi_1 \land \pi_2)^*_n \}
   ; n := n + 1
   od \{ \pi_0 \land \pi_1 \land \pi_2 \land n = N , so \rho_0 \}
```

The interested reader might want to try a direct approach via $\mu \pi \pi \pi$. 
§ 5. Further remarks on pattern search

The second alternative statement in the algorithm above, distinguishing preprocessing and search, may also lead to a code with two (sequential) repetitions, one for each alternative. We chose for the form above to stress the uniformity: the difference between the parts is solely based upon coding, the genesis doesn't differentiate!

Several people noticed the strong resemblance of those parts, but in the literature we searched in vain for a presentation or derivation (at all) of the algorithm that did justice to that resemblance. ([IC 85] and [W 86] deserve some credit).

[Note that even in 1983 the preprocessing was said to be "complicated and difficult to understand" ([S 83] p. 242). As the two parts are almost identical such a statement is puzzling. Has it anything to do with the widespread chaotic algorithm presentation? (e.g. [KMP 77], [BM 77]).]

In our opinion, exploitation of pre- and postfixes simplified the "derivation" of the algorithm such that it becomes within reach of every freshmen course.

We conclude the discussion of pattern search with a remark on the Boyer-Moore fast pattern search ([BM 77]). As this algorithm is slightly beyond the scope of this paper, we shall only hint at its relation with the MPP problem.

Consider $X \in \Sigma^*$ and pattern $X \downarrow p$. In the Knuth-Morris-Pratt pattern search we decided to build up pre- and postfixes bit by bit, but we could have been greedier:

To that end consider the (linear-search-like) invariant

$PBM \quad (A i : X \downarrow p \, \text{pp} \, X \downarrow i : i > n)$

As $X \downarrow p \, \text{pp} \, X \downarrow n + 1 \Rightarrow \{4\} \; X(p-1) = X(n)$ we first check $X(n)$ as a candidate for the end of a pattern occurrence.

Let $s = (\text{MAX} j : 0 \leq j \leq p \land X(n) = X(j) : j) \; \text{max} -1$

Then $PBM_{s+p-s}$ holds.

In other words: the first candidate $m$ to satisfy $X \downarrow p \, \text{pp} \, X \downarrow m$ is $m = n + p - s$.

If $s < p - 1$ we can "leap further", if $s = p - 1$ we check $X(n - 1)$, etcetera.

This requires knowledge of the occurrences of values and periodicities in the pattern.

The reader is challenged to give a "derivation" of the Boyer-Moore fast pattern search based on this early deviation of the MPP problem.
§ 6. Periodicity search

In section 1 we "generalized" the carrécheck exercise to the MPP problem, and we promised to show that a solution for carrécheck is found as soon as MPP is solved.

We shall keep our promise in the following way:
- we give a variant of carrécheck
- we proclaim an enrichment of MPP that solves the (variant) exercise
- we perform some string-mathematics to prove that the exercise is solved by that enrichment of MPP.

For fixed \( m \geq 2 \) consider the following postcondition

\[
R = (\forall i : 1 \leq i \leq N : c(i) = (\exists P : P \in \Sigma^* : X \downarrow i = P^m)) \\
\wedge (\forall i : 1 \leq i \leq N : \text{per}(i) = i \min(\text{MIN}p : p \text{ period of } X \downarrow i : p))
\]

In case \( m = 2 \), the first conjunct of \( R \) is just the postcondition of carrécheck.

The second conjunct of \( R \) means:
\( \text{per}(i) \) is the period of \( X \downarrow i \) if \( X \downarrow i \) is periodic, otherwise \( \text{per}(i) = i \).

Obviously we should extend invariant \( P \) for the MPP problem with a conjunct \( P_4 \) to get an invariant for the new problem.

\[
P_4 \quad R^N_4
\]

Initialization of \( P_4 \):
\( ; c : (1) = \text{false} ; \text{per} : (1) = 1 \).

The outer repetition should contain an establishment of \( P_{an+1}^n \).

So, following "\( f : (n + 1) = k \ (P_{an+1}^n) \)" and before "\( n := n + 1 \)", we proclaim the statement list:

\[
\begin{align*}
; c : (n + 1) &= (n + 1) \mod (m \ast (n + 1 - f (n + 1))) = 0 \\
; \text{if } (n + 1) \mod (n + 1 - f (n + 1)) &= 0 \implies \text{per} : (n + 1) = n + 1 - f (n + 1) \\
; \text{if } (n + 1) \mod (n + 1 - f (n + 1)) &= 0 \implies \text{per} : (n + 1) = n + 1 \\
\end{align*}
\]

\( \text{fi } (P_{an+1}^n , \text{ see corollary 5 to follow }) . \)

Indeed a minor adaptation, but it takes a proof!

The string-mathematics to follow is quite elementary, and has nothing to do with programming and - methodology. So we adopt a more conventional mathematical style, but (for the convenience of non-mathematicians) we still take small steps in the proofs.

The basic idea is to squeeze periodicity information out of \( \text{pp} \) or \( \text{mpp} \) knowledge.

Let \( D, Y \in \Sigma^* \) with \( D \ \text{pp} \ Y \). Then there are \( E, F \in \Sigma^* \) such that \( Y = DE \land Y = FD \).

Lemmas 1 and 2 tell us about (almost) periodicity of \( Y \). (They are well-known, e.g. see [L 79] Ch. 11.5). Much more can be said if \( D \ \text{mpp} \ Y \), some of which is done in 3.4.5.
Lemma 1 Let $D \in \Sigma^*$ and $E, F \in \Sigma^*$ such that $DE = FD$.
Then there are $L \in \Sigma^*$, $K \in \Sigma^*$ and $n \geq 0$ with
1. $D = F^* L$ and $L < F$ (hence $DE = FD = F^{*+1} L$)
2. $E = KL$ and $F = LK$.

Proof Certainly, there are $L \in \Sigma^*$ and $n \geq 0$ such that $D = F^* L$ and $1L < 1F1$.
Then $F^* LE = DE = FD = F^{*+1} L$, so $LE = FL$.
As $1L < 1F1 = 1E1$, there are $K, K^1 \in \Sigma^*$ with $LK = F$ and $E = K^1 L$.
Hence, $L(K^1 L) = LE = FL = (LK)L$ and it follows that $K^1 = K$.

Lemma 2 Let $D, F \in \Sigma^*$ with $DF = FD$. Then there is a $P \in \Sigma^*$ such that
$D, F \in \{P^m \mid m \geq 0\}$. If $D, F \in \Sigma^*$ it follows that $DF$ is periodic with period at
most $\gcd(1D1, 1F1)$.

Proof By induction to the length of $DF$.
If $D = \Lambda$ or $F = \Lambda$ the existence of $P$ is obvious.
Let $D, F \in \Sigma^*$. By 1, there are $K, L, n$ such that $D = F^* L$, $F = KL$ and $F = LK$.
Hence $KL = LK$.
As $D \neq \Lambda$, $1KL < 1F1 < 1DF1$, so by induction there is a $P \in \Sigma^*$ (as $K \in \Sigma^*$ even $P \in \Sigma^*$) such that $K, L \in \{P^m \mid m \geq 0\}$.
Consequently, $D, F \in \{P^m \mid m \geq 0\}$ which proves the first part.
If $D, F \in \Sigma^*$ then $DF \in \{P^{m+2} \mid m \geq 0\}$ while $P \in \Sigma^*$.
Note that $1P1$ divides $1D1$ and $1F1$.

Lemma 3 Let $Y = FD$ and $D \text{ mpp } Y$. Then $F$ is not periodic.

Proof By definition of (mpp), $F \neq \Lambda$. So, by 1, there are $L, n$ with $D = F^* L$ and $L < F$.
Suppose $F$ is periodic, say $F = Q^m$ for some $Q \in \Sigma^*$, $m \geq 2$.
Then $QF = FQ$, hence $QL < QF = FQ$. As also $L < F < QF$, it follows that $FQ$ has both $L$ and
$QL$ as prefix. Since $1L1 < 1QL1$ we have $L < QL$ and, equivalently, $Q^m Q^{m-1} L < Q^m Q^m L$.
As $Y = FD = Q^m Q^m L$, it follows that $Q^m Q^{m-1} L \text{ pp } Y$. However, since
$m \geq 2$, $1D1 = 1Q^m L1 < 1Q^m Q^{m-1} L1$ which contradicts $D \text{ mpp } Y$. This falsifies periodicity of $F$. \[ \square \]
Lemma 4 Let $Y = FD$, $D \mathbin{mpp} Y$. Let $P \mathbin{pp} Y$ and $|P| \geq |F|$, then there is a $k \geq 0$ such that $D = F^k P$. (i.e. all pre-postfixes of $Y$ with length $|F|$ are known).

Proof As $D \mathbin{mpp} Y$, $F \neq \Lambda$, so there are $n \geq 0$ and $L < F$ with $D = F^n L$.
Because $|L| < |F| \leq |P|$ and $P$ is a postfix of $D$, $L$ is a postfix of $P$. Hence there are a $k : 0 \leq k \leq n$ and $H < F$ with $D = F^k H P$. Let $H G = F$ then $P = GF^{n-k-1} L$.
As $|H| + |G| = |F| \leq |P| = |G| + |F^{n-k-1} L|$, $|H| \leq |F^{n-k-1} L|$.
Since $H < F^{n-k}$ and $F^{n-k-1} L < F^{n-k}$ it follows that $H \leq F^{n-k-1} L$, so $GH \leq GF^{n-k-1} L = P$.
On the other hand $HG = F \leq P$, so $GH = HG$.
As $H < F$, and, by 2, $F$ is not periodic it follows from 2 that $H = \Lambda$, which shows $D = F^k P$.

Corollary 5 Let $D \mathbin{mpp} Y$, say $Y = FD$. Let $m \geq 2$, then

$$(EC : Y = C^m) \text{ iff } |Y| \mod (m \ast |F|) = 0.$$  

In particular, $Y$ is a carré iff $|Y| \mod 2 |F| = 0$, and $Y$ is periodic iff $D \neq \Lambda$ and $|Y| \mod |F| = 0$.

Proof By 1, $Y = F^{n+1} L$ and $L < F$, so the if-part is obvious.
Let $Y = C^m$; note that since $m \geq 2$, $C^{m-1} \mathbin{pp} Y$. As $|D| \geq |C^{m-1}|$, $|F| \leq |C| \leq |C^{m-1}|$, so by 4, $D = F^k C^{m-1}$.
Hence $F^{k+1} C^{m-1} = Y = C^{m-1}$ and $C = F^{k+1}$, so $Y = F^m * (k+1)$.
The remark on $Y$ being a carré is an instantiation for $m = 2$.
Finally, as $Y \neq F$, $|Y| \mod |F| = 0 = (EC : m \geq 2 : |Y| \mod (m \ast |F|) = 0)$.

Note that if $Y$ is periodic, $|Y| - |D|$ is the period.
Acknowledgements

The carré problem and related other problems were "en vogue" in the environment of the Eindhoven University of Technology. So, inevitably, I was contaminated too.

I would like to thank all colleagues that contributed to the genesis of this paper by comments and interest.

References

[BM77] Boyer, R.S. and Moore J.S.,
A fast string searching algorithm,

[C85] Chengdian, C.,
A derivation of the Knuth-Morris-Pratt pattern matching program,
EUT-report 85-Wsk-02,
Eindhoven University of Technology (1985).

[KMP77] Knuth, D.E., Morris, J.H. and Pratt, V.R.,
Fast pattern matching in strings,

[L79] Lallement, G.,
Semigroups and combinatorial applications,

[S83] Sedgewick, R.,
Algorithms,
Addison - Wesley, (1983).

[W86] Wiltink, G.,
Knuth-Morris-Pratt, private communication,
(1986).
In this series appeared:

<table>
<thead>
<tr>
<th>No.</th>
<th>Author(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>85/01</td>
<td>R.H. Mak</td>
<td>The formal specification and derivation of CMOS-circuits</td>
</tr>
<tr>
<td>85/02</td>
<td>W.M.C.J. van Overveld</td>
<td>On arithmetic operations with M-out-of-N-codes</td>
</tr>
<tr>
<td>85/03</td>
<td>W.J.M. Lemmens</td>
<td>Use of a computer for evaluation of flow films</td>
</tr>
<tr>
<td>85/04</td>
<td>T. Verhoef</td>
<td>Delay insensitive directed trace structures satisfy the foam rubber wrapper postulate</td>
</tr>
<tr>
<td></td>
<td>H.M.J.L. Schols</td>
<td></td>
</tr>
<tr>
<td>86/01</td>
<td>R. Koymans</td>
<td>Specifying message passing and real-time systems</td>
</tr>
<tr>
<td>86/02</td>
<td>G.A. Bussing</td>
<td>ELISA, A language for formal specifications of information systems</td>
</tr>
<tr>
<td></td>
<td>K.M. van Hee</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M. Voorhoeve</td>
<td></td>
</tr>
<tr>
<td>86/03</td>
<td>Rob Hoogerwoord</td>
<td>Some reflections on the implementation of trace structures</td>
</tr>
<tr>
<td>86/04</td>
<td>G.J. Houben</td>
<td>The partition of an information system in several parallel systems</td>
</tr>
<tr>
<td></td>
<td>J. Paredaens</td>
<td></td>
</tr>
<tr>
<td></td>
<td>K.M. van Hee</td>
<td></td>
</tr>
<tr>
<td>86/05</td>
<td>Jan L.G. Dietz</td>
<td>A framework for the conceptual modeling of discrete dynamic systems</td>
</tr>
<tr>
<td></td>
<td>Kees M. van Hee</td>
<td></td>
</tr>
<tr>
<td>86/06</td>
<td>Tom Verhoeoff</td>
<td>Nondeterminism and divergence created by concealment in CSP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>86/07</td>
<td>R. Gerth</td>
<td>On proving communication closedness of distributed layers</td>
</tr>
<tr>
<td></td>
<td>L. Shira</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Author(s)</td>
<td>Title</td>
</tr>
<tr>
<td>------</td>
<td>-------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>86/09</td>
<td>C. Huizing, R. Gerth, W.P. de Roever</td>
<td>Full abstraction of a real-time denotational semantics for an OCCAM-like language</td>
</tr>
<tr>
<td>86/10</td>
<td>J. Hooman</td>
<td>A compositional proof theory for real-time distributed message passing</td>
</tr>
<tr>
<td>86/11</td>
<td>W.P. de Roever</td>
<td>Questions to Robin Milner - A responders commentary (IFIP86)</td>
</tr>
<tr>
<td>86/12</td>
<td>A. Boucher, R. Gerth</td>
<td>A timed failures model for extended communicating processes</td>
</tr>
<tr>
<td>86/14</td>
<td>R. Koymans</td>
<td>Specifying passing systems requires extending temporal logic</td>
</tr>
<tr>
<td>87/01</td>
<td>R. Gerth</td>
<td>On the existence of a sound and complete axiomatizations of the monitor concept</td>
</tr>
<tr>
<td>87/02</td>
<td>Simon J. Klaver, Chris F.M. Verbeeme</td>
<td>Federative Databases</td>
</tr>
<tr>
<td>87/03</td>
<td>G.J. Houben, J. Paredaens</td>
<td>A formal approach to distributed information systems</td>
</tr>
<tr>
<td>87/04</td>
<td>T. Verhoeff</td>
<td>Delay-insensitive codes - An overview</td>
</tr>
<tr>
<td>87/05</td>
<td>R. Kuiper</td>
<td>Enforcing non-determinism via linear time temporal logic specification</td>
</tr>
<tr>
<td>Year</td>
<td>Author(s)</td>
<td>Title</td>
</tr>
<tr>
<td>------</td>
<td>---------------------------------</td>
<td>-------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>87/06</td>
<td>R. Koymans</td>
<td>Temporal logica specificatie van message passing en real-time systemen (in Dutch)</td>
</tr>
<tr>
<td>87/07</td>
<td>R. Koymans</td>
<td>Specifying message passing and real-time systems with real-time temporal logic</td>
</tr>
<tr>
<td>87/08</td>
<td>H.M.J.L. Schols</td>
<td>The maximum number of states after projection</td>
</tr>
<tr>
<td>87/09</td>
<td>J. Kalisvaart</td>
<td>Language extensions to study structures for raster graphics</td>
</tr>
<tr>
<td></td>
<td>L.R.A. Kessener</td>
<td></td>
</tr>
<tr>
<td></td>
<td>W.J.M. Lemmens</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M.L.P van Lierop</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F.J. Peters</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H.M.M. van de Wetering</td>
<td></td>
</tr>
<tr>
<td>87/10</td>
<td>T. Verhoeff</td>
<td>Three families of maximally nondeterministic automata</td>
</tr>
<tr>
<td>87/11</td>
<td>P. Lemmens</td>
<td>Eldorado ins and outs. Specifications of a data base management toolkit according to the functional model</td>
</tr>
<tr>
<td>87/12</td>
<td>K.M. van Hee</td>
<td>OR and AI approaches to decision support systems</td>
</tr>
<tr>
<td></td>
<td>A. Lapinski</td>
<td></td>
</tr>
<tr>
<td>87/13</td>
<td>J. van der Woude</td>
<td>Playing with patterns, searching for strings</td>
</tr>
</tbody>
</table>