A taxonomy of sublinear multiple keyword pattern matching algorithms

Citation for published version (APA):

Document status and date:
Published: 01/01/1995

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 11. Nov. 2019
A taxonomy of sublinear multiple keyword pattern matching algorithms

ISSN 0926-4515

All rights reserved
editors: prof.dr. J.C.M. Baeten
         prof.dr. M. Rem

Computing Science Report 95/13
Eindhoven, April 1995
A taxonomy of sublinear multiple keyword pattern matching algorithms

B.W. Watson & G. Zwaan
Faculty of Mathematics and Computing Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB
Eindhoven, The Netherlands
email: watson@win.tue.nl or wsinswan@win.tue.nl

April 21, 1995

Abstract

This paper presents a taxonomy of sublinear keyword pattern matching algorithms related to the Boyer-Moore algorithm [BM77] and the Commentz-Walter algorithm [CW79a, CW79b]. The taxonomy includes, amongst others, the multiple keyword generalization of the single keyword Boyer-Moore algorithm and an algorithm by Fan and Su [FS93, FS94]. The corresponding precomputation algorithms are presented as well. The taxonomy is based on the idea of ordering algorithms according to their essential problem and algorithm details, and deriving all algorithms from a common starting point by successively adding these details in a correctness preserving way. This way of presentation not only provides a complete correctness argument of each algorithm, but also makes very clear what algorithms have in common (the details of their nearest common ancestor) and where they differ (the details added after their nearest common ancestor). Introduction of the notion of safe shift distances proves to be essential in the derivation and classification of the algorithms. Moreover, the paper provides a common derivation for and a uniform presentation of the precomputation algorithms, not yet found in the literature.
## Contents

1 Introduction  
   1.1 Basic algorithm and derivation principles  
   1.2 The taxonomy  
   1.3 Overview  

2 The problem and some naive solutions  

3 Sublinear pattern matching algorithms  
   3.1 No lookahead at the unscanned part of the input string  
   3.2 Restriction to one symbol lookahead  
   3.3 Lookahead symbol is mismatching  
   3.4 The multiple keyword Boyer-Moore algorithm  
   3.5 The Commentz-Walter algorithm  
   3.6 Complete decoupling of recognized suffix and lookahead symbol  
   3.7 Discarding the lookahead symbol  
   3.8 One symbol lookahead at the unscanned part of the input string  

4 Precomputation  
   4.1 Precomputation of $\tau_P$  
   4.2 Precomputation of $d$-functions  
   4.3 Precomputation of $char_{cw}$, $char_{bm}$, and $char_a$  
   4.4 Precomputation of $msp$  

5 Conclusions  

6 Acknowledgements  

A Definitions and properties  

B Algorithm and problem details
1 Introduction

The keyword (or string) pattern matching problem can informally be described as the problem of finding all occurrences of keywords (strings) from a given set as substrings in a given (input) string. This problem is encountered in many areas and in several forms. In computing science, for instance, it plays a role in text search/analysis, lexical analysis, and data processing. In biology it is encountered in the analysis of, amongst others, DNA sequences. The problem can also be generalized to the matching of regular expressions, tree patterns, and graph patterns, none of which is treated here.

The keyword pattern matching problem has been extensively studied and a multitude of diverse solutions/algorithms exists. Single keyword algorithms are for instance described by Knuth, Morris, and Pratt [KMP77] and Boyer and Moore [BM77]; multiple keyword algorithms by Aho and Corasick [AC75], by Commentz-Walter [CW79a, CW79b], and by Fan and Su [FS93, FS94]. An overview of keyword pattern matching algorithms can be found in [Aho90].

Due to the diversity of the algorithms and their descriptions—that tend to be rather involved and verbal—it is hard to get a good overview and to make a sound comparison between algorithms. In order to fulfill these needs a taxonomy of keyword pattern matching algorithms was presented by Watson and Zwaan in [WZ92, WZ93]. Here, we focus our attention on a part of that taxonomy containing a family of multiple keyword pattern matching algorithms that have a matching time that may be sublinear in the length of the input string (taking the number of symbol comparisons as a measure of matching time). Amongst others, it comprises the multiple keyword generalization of the single keyword Boyer-Moore algorithm [BM77], the Commentz-Walter algorithm [CW79a, CW79b], and the algorithm by Fan and Su [FS93, FS94] (only after deriving this algorithm we found its description by Fan and Su). Both the Boyer-Moore and Commentz-Walter algorithms provided the inspiration for our derivations and classifying principle.

The main results of this paper are comprised in the taxonomy graph shown in figure 1 (a more detailed description of the graph is found further on in this section). This taxonomy graph can be viewed as an alternative table of contents to this paper. It was obtained in order to meet the following goals:

- the systematic and formal derivation of the algorithms from a common starting point through a series of refinements to either algorithm or problem
- to factor out common portions of (the derivations of) well-known algorithms in order to facilitate the understanding of these algorithms and their comparison
- the presentation of the algorithms in a common framework to permit an easier comprehension of and a better comparison between the algorithms

It is the first concise and systematic presentation of and comparison between the algorithms from the family considered here. Another taxonomy of (single keyword) pattern matching algorithms by Hume and Sunday [HS91] does not meet the goals set in this paper since there any derivations and proofs of algorithms are missing.

Moreover, we show that all functions that need to be precomputed for the pattern matching algorithms of this family can in a simple way be expressed in a small set of base functions. The definitions of the base functions can all be written in forms satisfying one general pattern. From this general pattern a precomputation algorithm scheme has been derived that can be instantiated for each base function to yield a precomputation algorithm for that function (sometimes, the resulting algorithm can be simplified and/or some trivial postprocessing has to be added to it). Hence, the precomputation algorithms can also be derived and presented in a uniform way. This includes formal derivations of the precomputation algorithms for the Boyer-Moore algorithm, the Commentz-Walter algorithm and the algorithm presented by Fan and Su not yet found in the literature.

The taxonomy given here differs from the corresponding part of the taxonomy in [WZ92, WZ93] in that we here present the correct multiple keyword generalization of the Boyer-Moore algorithm.
Figure 1: Taxonomy graph of keyword pattern matching algorithms from the Boyer-Moore family. Each vertex corresponds to an algorithm. Vertices are labelled with either an algorithm number or a subsection number preceded by § referring to an algorithm or a subsection in this paper. Some vertices are additionally labelled with literature references. Each edge corresponds to the addition of either a problem or algorithm detail (see appendix B for a complete list of details and their descriptions). The sequence of edge labels that are encountered when going from the top vertex to another vertex forms a characterization of the algorithm corresponding to that vertex. For instance, the algorithm of the vertex labelled §3.5 (the Commentz-Walter algorithm) is characterized by \((p, s, r, t, s, d, n, l, a, u, o, p, t, b, m, c, w, c, w)\). The dashed edge leading to [KMP77, AC75] indicates the path to the rest of the taxonomy presented in [WZ92, Z93] that contains the Knuth-Morris-Pratt and the Aho-Corasick algorithms and is omitted in this paper.
and a common ancestor of this algorithm and the Commentz-Walter algorithm. Because of this the
derivations and the structure of the taxonomy have changed. Moreover, we present a completely
new derivation of the precomputation algorithms.

An implementation as a C procedural library of almost all algorithms from [WZ92] (called
"Eindhoven Pattern Kit") and an analysis of their performance is given in [Wat94] (the imple­
mentation is in a somewhat rudimentary form being meant for the benchmarking only). Only the
Commentz-Walter algorithm and a common descendant of both the Boyer-Moore and Commentz­
Walter algorithm are implemented and discussed there. The performance results for these algo­
rithms conform with the qualitative predictions made here. An implementation of all algorithms
from this paper and from [WZ92] as a C++ class library is called "SPARE Parts: A C++
toolkit for String PAttern REcognition" and will be available at URL ftp://ftp.win.tue.nl
/pub/techreports/pi/pattm/spare/. This implementation is entirely based on the abstract algo­
rithms described in this paper—in fact it is a systematic translation of them (this in contrast
to for instance the implementations given in [HS91]).

1.1 Basic algorithm and derivation principles

The algorithms in this family traverse the input string in a direction that is opposite to the
direction in which keyword symbols are matched to symbols in the input string. In this paper we
choose to inspect suffixes of prefixes of the input string both in order of increasing length. This
algorithm will be the starting point of all further derivations. Choosing such a basic algorithm we
have the possibility to attain matching times that are sublinear in the length of the input string
(i.e. not all symbols of the input string are inspected). It is achieved by taking steps through
the input string of which the length is determined by a shift function based on the information
of the last matching attempt and possibly on additional information. The example in figure 2
illustrates this principle. Notice that not all symbols of the input string are scanned, although it
is possible that some symbols of the input string are scanned more than once. The most simple

\[ \ldots aaabcdabef \ldots \]
\[ \text{shift}\ 3 \]
\[ \ldots aaabcdabef \ldots \]
\[ \text{shift}\ 4 \]
\[ \ldots aaabcdabef \ldots \]

Figure 2: Example of larger shift distances

shift function is the function that always yields 1. However, one can imagine larger shifts being
possible. Ideally, such a shift would take us to the next occurrence of a match, but then calculating
the value of the shift function is equivalent to the pattern matching problem itself. Therefore,
we strive for shift functions that are easier to calculate and that do not exceed the ideal shift
(called safe shift functions) i.e. we aim at approximations of the ideal shift from below. The ideal
shift function is the minimum over a domain characterized by some predicate. We derive various
approximations from below by systematically weakening this predicate and derived predicates,
by applying rules for minimum and maximum over disjunctive or conjunctive domains, and by
enlarging domains. Considerations that play a role in these derivations are, for instance, whether
or not to look ahead at symbols of the unscanned part of the input string, what information to use
on the last scanned (non-matching) symbol, and the extent to which this information is coupled
with the information on the recognized suffix. Thus, we obtain several shift functions that meet
the aforementioned requirements leading to an equal number of algorithms amongst which are the
well-known Boyer-Moore and Commentz-Walter algorithms. The most simple weakening of the
predicate is the weakening to the predicate true yielding the shift function that is always equal
to 1. The reason to derive ever smaller shift functions is that a smaller shift functions usually takes
less precomputation time and less storage space (for instance, a one dimensional table instead of
a two dimensional table).

The techniques we use to systematically derive shift functions from the ideal shift function
enable us to clearly delineate the relations between the resulting shift functions and algorithms.
Furthermore, they may be used in the derivation of yet other members of this family. This truly
systematic approach, especially the predicate weakening technique, is not known from literature.
Among other things we derive that the Commentz-Walter algorithm [CW79a, CW79b] is not the
multiple keyword generalization of the Boyer-Moore algorithm [BM77]. In fact, we show that the
algorithms are incomparable (meaning that the shift distance in one algorithm is not always at
least the shift distance in the other) and that they have a faster common ancestor that combines
the properties of both. As it is, this common ancestor is derived first and subsequently the Boyer-
Moore algorithm and the Commentz-Walter algorithm are derived from it. Both algorithms also
have a common descendant (the algorithm that was incorrectly identified as the Boyer-Moore
algorithm in [WZ92, WZ93]). Furthermore, it is shown that the algorithm described by Fan
and Su [FS93, FS94] is an even faster ancestor of the common ancestor of the Boyer-Moore and
Commentz-Walter algorithms.

1.2 The taxonomy

The algorithms are derived from a common starting point by successively adding either algorithm
or problem details (see appendix B for a complete list of details and their descriptions). The only
problem detail considered here is the restriction to the one keyword case. Among the algorithm
details considered are restriction of nondeterminacy, introduction of the reverse trie, introduction
of a shift function, and choice of a particular shift function. Each addition of an algorithm
detail gives a new algorithm satisfying the same specification as the original algorithm; thus, the
correctness of the algorithms is preserved. The sequence of details introduced in the derivation
of an algorithm can be used to identify its similarities and its differences with other algorithms.
These ordered detail sequences are used to identify the algorithms and to give a taxonomy of
the algorithms in this family. The taxonomy is depicted as a graph in figure 1. Our method
of developing a taxonomy was inspired by the method described by Jonkers [Jon83]. There it is
applied to develop a taxonomy of garbage collection algorithms. The method is also applied to
attribute evaluation algorithms by Marcelis [Mar90]. Other examples of algorithm taxonomies are
found in [Br83, Dar78].

All algorithms are presented in a somewhat extended version of the guarded command language
of Dijkstra [Dij76] in order to avoid the peculiarities of any particular programming language. The
algorithms use string type variables in order to abstract from any implementation detail like, for
instance, indexing. Derivation of the algorithms is done a calculational way following Dijkstra's
style of program derivation.

1.3 Overview

In section 2 we give a formal definition of the pattern matching problem. From a trivial solution
to this problem we derive, by addition of a number of algorithm details, an algorithm that is the
starting point for the derivation of the algorithms in section 3. In section 3 we start by adding the
algorithm detail that states that shifts larger than one may be possible. Addition of this program
detail accounts for the possible sublinear matching time of all of the algorithms to be derived in
this section. Subsequently, we derive by systematic approximation from below of the maximal safe
shift distance the various algorithms of this family (in order of decreasing matching speed). The
The problem and some naive solutions

The keyword pattern matching problem is to find all occurrences of keywords from a set as sub­strings in an input string. Formally, given an alphabet \( V \) (a non-empty finite set of symbols), an input string \( S \in V^* \), and a finite non-empty pattern set \( P \subseteq V^* \), establish

\[ R : O = ( \cup l, v, r : lv = S : \{l\} \times (\{v\} \cap P) \times \{r\} ) \cdot \]

Throughout this paper we will adopt the convention that, unless stated otherwise, program variables and bound variables with names from the beginning of the Latin alphabet (i.e. \( a, b, c \)) will range over \( V \), while variables with names from the end of the Latin alphabet (i.e. \( l, q, r, u, v, w \)) will range over \( V^* \). A trivial (but unrealistic) solution to the problem is

Algorithm 2.1

\[
O := ( \cup l, v, r : lv = S : \{l\} \times (\{v\} \cap P) \times \{r\} )
\]

The sequence of details describing this algorithm is the empty sequence (sequences of details are introduced in subsection 1.2 and figure 1).

There are two basic directions in which to proceed while developing naive algorithms to solve this problem. Informally, a substring of \( S \) can be considered a "suffix of a prefix of \( S \)" or a "prefix of a suffix of \( S \)". Only the first possibility is considered here, since the second possibility only leads to algorithms that are the mirror images of algorithms obtained by following the first possibility (basically, it amounts to reversing all strings in the problem). Moreover, this is the way that the Boyer-Moore, Commentz-Walter, and Fan and Su algorithms treat substrings of input string \( S \).

Formally, we can consider "suffixes of prefixes of \( S \)" as follows:

\[
( \cup l, v, r : lv = S : \{l\} \times (\{v\} \cap P) \times \{r\} )
\]

=  
\[
\{ \text{introduce } u : u = lv \}
\]

=  
\[
( \cup l, v, r, u : ur = S \land lv = u : \{l\} \times (\{v\} \cap P) \times \{r\} )
\]

=  
\[
\{ l, v \text{ only occur in the latter range conjunct, so restrict their scope} \}
\]

=  
\[
( \cup u, r : ur = S : ( \cup l, v : lv = u : \{l\} \times (\{v\} \cap P) \times \{r\} ) )
\]

A simple non-deterministic algorithm is obtained by applying "examine prefixes of a given string in any order" (algorithm detail (p)) to input string \( S \). In the following algorithm we use a for-rof statement in order to express its non-determinism. Statement for \( x : P \rightarrow S \) rof amounts to executing statement list \( S \) once for each value of \( x \) that satisfies \( P \) initially, assuming only a finite number of such values exist. The order in which the values of \( x \) are chosen is arbitrary. It results in

Algorithm 2.2(p)

\[
O := \emptyset ;
\]

\[
\text{for } (u, r) : ur = S \rightarrow
O := O \cup ( \cup l, v : lv = u : \{l\} \times (\{v\} \cap P) \times \{r\} )
\]

\[
\text{rof} \{ R \}
\]
The update of \( O \) (with another quantifier) in the inner repetitions of algorithm (ps) can be computed with another non-deterministic repetition. This inner repetition would consider suffixes of \( u \). Thus by applying "examine suffixes of a given string in any order" (algorithm detail (s)) to string \( u \) we obtain algorithm

Algorithm 2.3(ps)

\[
O := \emptyset; \\
\text{for } (u, r) : ur = S \rightarrow \\
\quad \text{for } (l, v) : lv = u \rightarrow \\
\quad \quad O := O \cup \{l\} \times (\{v\} \cap P) \times \{r\} \\
\text{rof} \\
\text{rof}\{ R \}
\]

Algorithm (ps) consists of two nested non-deterministic repetitions. In each case, the repetition can be made deterministic by considering prefixes (or suffixes as the case is) in increasing (called detail (\( + \))) or decreasing (detail (\( - \))) order of length. This gives two binary choices. Since the Boyer-Moore and Commentz-Walter algorithms examine string \( S \) from left to right and the patterns in \( P \) from right to left we focus our attention on (the operators \( l, l, l \), and \( l \) are defined in definition A.1; relation \( \leq_p \) is defined in definition A.3):

Algorithm 2.4(p,s+)

\[
\begin{align*}
u, r := &\, e, S; \\
O := &\, \{e\} \times (\{e\} \cap P) \times \{S\}; \\
\{ \text{invariant: } O = (\cup \{x, y, z : xy = S \land xy \leq_p u \} : \{x\} \times (\{y\} \cap P) \times \{z\}) \}
\end{align*}
\]

\[
\text{do } r \neq e \rightarrow \\
\quad u, r := u(r[1], r[1]); l, v := u, e; O := O \cup \{u\} \times (\{e\} \cap P) \times \{r\}; \\
\quad \text{do } l \neq e \rightarrow \\
\quad \quad l, v := l[l], (l[l])v; \\
\quad \quad O := O \cup \{l\} \times (\{v\} \cap P) \times \{r\} \quad \text{od} \\
\text{od}\{ R \}
\]

This algorithm has running time \( O(|S|^2) \), assuming that intersection with \( P \) is a \( O(1) \) operation. We will now improve the running time of this algorithm. Consider the set of suffixes of keywords \( \text{suff}(P) \) (the functions \( \text{suff} \) and \( \text{pref} \) are defined in definition A.2). A string \( w \) is an element of \( \text{suff}(P) \) if and only if it can be extended on the left to a pattern in \( P \), i.e., \( \exists w' : w' \in V^* : w'w \in P \).

It follows that if \( w \notin \text{suff}(P) \) any extension of \( w \) on the left is not an element of \( \text{suff}(P) \) either. Consequently, the inner repetition in algorithm 2.4 can terminate as soon as \( (l[l])v \notin \text{suff}(P) \) holds, since then all suffixes of \( u \) that are equal to or longer than \( (l[l])v \) are not in \( \text{suff}(P) \) and hence not in \( P \). The inner repetition guard is therefore strengthened to

\[
l \neq e \text{ cand } (l[l])v \notin \text{suff}(P).
\]

Observe that \( v \in \text{suff}(P) \) is an invariant of the inner repetition. This invariant is initially established by the assignment \( v := \epsilon \) since \( P \neq \emptyset \) and thus \( \epsilon \in \text{suff}(P) \). Direct evaluation of \( (l[l])v \notin \text{suff}(P) \) is expensive. Therefore, it is done using the transition function \( \tau_P \) of the reverse trie [Fre60] corresponding to \( P \) where \( \tau_P : \text{suff}(P) \times V \rightarrow \text{suff}(P) \cup \{\bot\} \) is defined by

\[
\tau_P(w, a) = \begin{cases} 
aw & \text{if } aw \in \text{suff}(P) \\
\bot & \text{if } aw \notin \text{suff}(P)
\end{cases} \quad (w \in \text{suff}(P), a \in V).
\]

(algorithm detail (\( \tau \))). Since we usually refer to the trie corresponding to \( P \) we will write \( \tau \) instead of \( \tau_P \). Transition function \( \tau \) can be computed beforehand. Its precomputation is discussed in subsection 4.1. The guard becomes \( l \neq e \text{ cand } \tau(v, l[l]) \neq \bot \) yielding algorithm
3 Sublinear pattern matching algorithms

Algorithm 2.5($p_1$, $s_1$, $rt$)

\[ u, r := o, S; \quad O := \{e\} \times (\{e\} \cap P) \times \{S\}; \]
\[ \{ \text{invariant: } O = \{ (x, y, z : xyz = S \land xy \leq_p u : \{x\} \times (\{y\} \cap P) \times \{z\}) \} \}
\]
do $r \neq e$ \rightarrow
\[ u, r := u(r[1]), r[1]; \]
\[ l, v := u, e; \quad O := O \cup \{u\} \times (\{e\} \cap P) \times \{r\}; \]
do $l \neq e$ cand $\forall (u, l[1]) \neq \perp$ \rightarrow
\[ l, v := l[1], (l[1])v; \]
\[ O := O \cup \{l\} \times (\{v\} \cap P) \times \{r\} \]
\odo\{ u = lv \land v \in \text{suff}(P) \land (l = e \lor (l[1])v \not\in \text{suff}(P)) \}
\odo\{ R \}

This algorithm has $O(|S| \cdot (\max_p : p \in P : |p|))$ running time. It will serve as a starting point for the derivation of all algorithms in the following section.

3 Sublinear pattern matching algorithms

In this section we derive sublinear pattern matching algorithms starting with algorithm 2.5 by exploring the possibility of safely (without missing matches) making shifts of more than one symbol, i.e. replacing assignment $u, r := u(r[1]), r[1]$ by assignment $u, r := u(r[k]), r[k]$ for some $k$ satisfying

\[ 1 \leq k \leq (\min n : 1 \leq n \land \text{suff}(u(r[n]) \cap P \neq \emptyset : n) \]

(algorithm detail (ssd) (safe shift distance)). The upperbound is the distance to the next match, the maximal safe shift distance. A number $k$ satisfying this condition is called a safe shift distance.

Since computing the upperbound on $k$ is essentially the same as the problem we are trying to solve we aim at easier to compute approximations from below of the upperbound. These are derived by systematically weakening the predicate $\text{suff}(u(r[n])) \cap P \neq \emptyset$ in the range of the upperbound, weakening resulting predicates, applying rules for minimum and maximum over a disjunctive or conjunctive domain, and enlarging domains. Considerations that play a role in these derivations are, for instance, whether or not to look ahead at symbols of the unscanned part of the input string, whether to use or not to use information on the last scanned symbol (usually a non-matching symbol), and the extent to which this information is coupled with the information on the recognized suffix. Thus, several algorithms are obtained amongst which the generalized version of the Boyer-Moore algorithm [BM77], the Commentz-Walter algorithm [CW79a, CW79b], and the algorithm by Fan and Su [FS93, FS94]. We derive ever smaller shift functions since the smaller the shift function the less precomputation time and storage space for the functions constituting the shift function is usually needed. For instance, the algorithm by Fan and Su [FS93, FS94] is faster than the Commentz-Walter algorithm [CW79a, CW79b], but it needs a two dimensional table to store one of the functions constituting its shift function whereas the Commentz-Walter algorithm only needs one dimensional tables.

In the derivations we use part of the postcondition of the inner repetition in algorithm 2.5 ($u = lv \land v \in \text{suff}(P)$). Adding $l, v := e, e$ to the initial assignments in algorithm 2.5 turns $u = lv \land v \in \text{suff}(P)$ into an invariant of the outer repetition. Due to the dependence of the upperbound on $l, v$, and $r$ we will aim at shift functions $k$ that depend on $l, v$, and $r$ and write $k(l, v, r)$. Hence, we arrive at the following algorithm scheme for all algorithms to be derived in this section:
Algorithm 3.1\((p, S, RT, SSD)\)

\[
\begin{align*}
u, r &:= \epsilon, S; \quad O := \{\epsilon\} \times (\{\epsilon\} \cap P) \times \{S\}; \quad l, v := \epsilon, \epsilon; \\
\{ \text{invariant: } O = (U \times y, z) : x y = S \land x y \leq u \times (\{y\} \cap P) \times \{z\}) \\
\land u = lv \land v \in \text{suff}(P) \land (l \in \epsilon \text{ cor } (l|1)v \notin \text{suff}(P)) \}
\end{align*}
\]

\[
do r \neq \epsilon \rightarrow \\
u, r := u(r|k(l,v,r)), r|k(l,v,r); \\
l, v := u, \epsilon; \quad O := O \cup \{u\} \times (\{\epsilon\} \cap P) \times \{r\}; \\
do l \neq \epsilon \text{ cand } \tau(v, l|1) \neq \bot \rightarrow \\
l, v := l|1, (l|1)v; \\
O := O \cup \{l\} \times (\{v\} \cap P) \times \{r\}
\od
\od\{ R \}
\]

Particular algorithms are obtained by substituting their shift functions for \(k(l,v,r)\). Such a substitution may not always yield an algorithm that exactly corresponds with its original description in the literature; sometimes an additional transformation of the resulting algorithm is needed (for instance, a phase shift of the repetition; see [WZ92] for a phase shifted version of the algorithm scheme). Conjunct \(l = \epsilon \text{ cor } (l|1)v \notin \text{suff}(P)\) is added to the invariant in order to stress that provided \(l\) is nonempty symbol \(l|1\) is non-matching.

### 3.1 No lookahead at the unscanned part of the input string

In this subsection we derive an approximation from below of the upperbound on \(k\) that does not depend on \(r\) and that will be a starting point of most of our further derivations. In terms of algorithms this means that we refrain from looking ahead at the symbols of \(r\), the yet unscanned part of the input string (algorithm detail \((NLAt)\) \((\text{No Lookahead at Unscanned part of the input string})\)). This is in accordance with most of the algorithms we are aiming at. One symbol lookahead at the unscanned part of the input string is discussed in subsection 3.8. We derive

\[
\begin{align*}
(M) &\quad \big(\min n : 1 \leq n \land \text{suff}(u(r|n)) \cap P \neq \emptyset : n\big) \\
= &\quad \big\{ \text{domain split, } r|n = r \text{ if } n \geq |r| \big\} \\
= &\quad \big(\min n : 1 \leq n \leq |r| \land \text{suff}(u(r|n)) \cap P \neq \emptyset : n\big) \\
\geq &\quad \big(\min n : 1 \leq n \leq |r| \land \text{suff}(uV^n) \cap P \neq \emptyset : n\big) \\
= &\quad \big(\min n : 1 \leq n \leq |r| \land \text{suff}(uV^n) \cap P \neq \emptyset : n\big) \\
= &\quad \big(\min n : 1 \leq n \land \text{suff}(uV^n) \cap P \neq \emptyset : n\big) \\
= &\quad \big(\min n : 1 \leq n \land \text{suff}(uV^n) \cap P \neq \emptyset : n\big)
\end{align*}
\]

Since the last formula is to be the starting point of our further derivations we will from here on aim at shift functions \(k\) being dependent only on \(u\), i.e. on \(l\) and \(v\) (remember \(u = lv \land v \in \text{suff}(P)\)). We will write \(k(l,v)\) instead of \(k(l,v,r)\).

### 3.2 Restriction to one symbol lookahead

In all derivations in this subsection and the following subsections we assume

\[
u = lv \land v \in \text{suff}(P).
\]
Restriction to one symbol lookahead (l|1, the last symbol of u scanned in the inner loop) leads to the algorithm by Fan and Su [FS93, FS94]. It is obtained by weakening the predicate in the approximation of the upperbound in subsection 3.1 in the following way:

\[
suff(uV^n) \cap P \neq \emptyset = \{ u = lv \}
\]

\[
suff(luV^n) \cap P \neq \emptyset 
\Rightarrow \{ l = (l|1)(l|1), l|1 \in V^*, \text{monotonicity of } suff \text{ and } \}
\]

\[
suff(V^*(l|1)uV^n) \cap P \neq \emptyset = \{ \text{property A.4} \}
\]

\[
V^*(l|1)uV^n \cap V^*P \neq \emptyset
\]

\[
= \{ l = \varepsilon: \text{property A.5.i; } l \neq \varepsilon: \text{property A.5.ii} \}
\]

\[
V^*(l|1)uV^n \cap P \neq \emptyset \lor uV^n \cap V^*P \neq \emptyset
\]

Notice that we have obtained a weaker predicate solely by discarding any information on l|1. The only information on l that is still taken into account is l|1 being either empty or consisting of one symbol. In the latter case we say to have one symbol lookahead. Observe that the symbol is the last symbol of u scanned in the inner loop and that it is a non-matching symbol. After substituting the weaker predicate we obtain shift distance \( k_{opt}(l, u) \) where \( k_{opt} \in V^* \times suff(P) \to N \) is defined for \( x \in V^* \) and \( y \in suff(P) \) by

\[
k_{opt}(x, y) = (\text{MIN } n : n \geq 1 \land (V^*(x|1)yV^n \cap P \neq \emptyset \lor yV^n \cap V^*P \neq \emptyset)) : n).
\]

Function \( k_{opt} \) can be expressed as follows

\[
k_{opt}(x, y) = \begin{cases} 
d_{opt}(x|1, y) \min d_{sp}(y) & x \neq \varepsilon 
d_{i}(y) \min d_{sp}(y) & x = \varepsilon 
\end{cases}
\]

where \( d_{opt} \in V \times suff(P) \to N \) is defined by

\[
d_{opt}(a, y) = (\text{MIN } n : n \geq 1 \land V^*aV^n \cap P \neq \emptyset : n) \quad (a \in V, y \in suff(P)),
\]

\( d_{sp} \in suff(P) \to N \) is defined by

\[
d_{sp}(y) = (\text{MIN } n : n \geq 1 \land yV^n \cap V^*P \neq \emptyset : n) \quad (y \in suff(P)),
\]

(function \( d_2 \) in [CW79a, CW79b]), and \( d_{i} \in suff(P) \to N \) is defined by

\[
d_{i}(y) = (\text{MIN } n : n \geq 1 \land V^*yV^n \cap P \neq \emptyset : n) \quad (y \in suff(P)),
\]

(function \( d_1 \) in [CW79a, CW79b]). Functions \( d_{opt} \) and \( d_{i} \) account for occurrences of \( ay \) and \( y \), respectively, within some keyword (i.e. infix of some keyword), whereas function \( d_{sp} \) accounts for occurrences of suffixes of \( y \) as proper prefixes of some keyword. Precomputation of \( d_{opt}, d_{sp}, \) and \( d_{i} \) is discussed in subsection 4.2.

Calculating the shift distance in this way is referred to as algorithm detail (OPT) and results in algorithm \( \{P, S, \delta, RT, SSD, NLAU, OPT\} \). We arrived at this algorithm not knowing it had already been described by Fan and Su in [FS93, FS94]. From their informal description it undoubtedly follows that they describe the same algorithm though their formal treatment of the algorithm and, especially, the precomputation is rather involved. Finally, notice that to store function \( d_{opt} \) one needs a two dimensional table, whereas functions \( d_{i} \) and \( d_{sp} \) only need one dimensional tables. In the following subsections we derive shift functions smaller than \( k_{opt} \) that are expressed solely in functions needing one dimensional tables for storage.
3.3 Lookahead symbol is mismatching

We derive an approximation from below of $d_{opt}$ that yields an algorithm that is the common ancestor of the multiple keyword generalization of the Boyer-Moore algorithm [BM77] and the Commentz-Walter algorithm [CW79a, CW79b]. Essentially, the resulting shift function is not based on the identity of the lookahead symbol $l_1$ but only uses the fact that the lookahead symbol is mismatching, as is done in the Boyer-Moore shift function. In this way one might say that the recognized suffix and the (mismatching) lookahead symbol have to some extent been decoupled.

We start by weakening the predicate from $d_{opt}$. Assume $l \neq \varepsilon$ and $(l_1)v \not\in \text{suff}(P)$. We derive

$$V^*(l_1)v^n \cap P \neq \emptyset$$

$$\Rightarrow \{ v \in V^{|v|}, \text{monotonicity of } \cap \}$$

$$V^*(l_1)V^{|v|+n} \cap P \neq \emptyset \wedge V^*(l_1)v^n \cap P \neq \emptyset$$

$$\Rightarrow \{ (l_1)v \not\in \text{suff}(P), \text{so } l_1 \in \{ a \mid a \in V \wedge av \not\in \text{suff}(P) \} \}, \text{definition } MS$$

$$V^*(l_1)V^{|v|+n} \cap P \neq \emptyset \wedge V^*(V \setminus MS(v))v^n \cap P \neq \emptyset$$

where $MS \in \text{suff}(P) \to V$ is defined by

$$MS(y) = \{ a \mid a \in V \wedge ay \in \text{suff}(P) \} \quad (y \in \text{suff}(P)).$$

The first conjunct will lead to a shift component based on the identity of the lookahead symbol that is identical to a component of the Commentz-Walter shift function. The second conjunct will lead to a shift component—based on the recognized suffix and the fact that the lookahead symbol is mismatching—that is identical to a component of the Boyer-Moore shift function. Replacing the range predicate of $d_{opt}$ by the last predicate of the preceding derivation we proceed

$$(\text{MIN } n : n \geq 1 \wedge V^*(l_1)V^{|v|+n} \cap P \neq \emptyset \wedge V^*(V \setminus MS(v))v^n \cap P \neq \emptyset : n)$$

$$\geq \{ \text{MIN with conjunctive range} \}$$

$$(\text{MIN } n : n \geq 1 \wedge V^*(l_1)V^{|v|+n} \cap P \neq \emptyset : n)$$

$$\max (\text{MIN } n : n \geq 1 \wedge V^*(V \setminus MS(v))v^n \cap P \neq \emptyset : n)$$

$$= \{ \text{change of bound variable: } m = |v| + n; \text{definition } d_{ui} \text{ (after derivation)} \}$$

$$(\text{MIN } m : m \geq |v| + 1 \wedge V^*(l_1)V^m \cap P \neq \emptyset : m - |v|) \max d_{ui}(v)$$

$$\geq \{ \text{enlarging domain} \}$$

$$(\text{MIN } m : m \geq 1 \wedge V^*(l_1)V^m \cap P \neq \emptyset : m - |v|) \max d_{ui}(v)$$

$$= \{ l \neq \varepsilon, \text{definition of char}_{cw} \text{ (after derivation)} \}$$

$$\text{char}_{cw}(l_1, |v|) \max d_{ui}(v)$$

where $d_{ui} \in \text{suff}(P) \to \mathbb{N}$ is defined by

$$d_{ui}(y) = (\text{MIN } n : n \geq 1 \wedge V^*(V \setminus MS(v))v^n \cap P \neq \emptyset : n) \quad (y \in \text{suff}(P))$$

and $\text{char}_{cw} \in V \times \mathbb{N} \to \mathbb{N}$ is defined by

$$\text{char}_{cw}(a, z) = (\text{MIN } n : n \geq 1 \wedge V^*a^n \cap P \neq \emptyset : n - z) \quad (a \in V, z \in \mathbb{N}).$$

This results in shift distance $k_{mcw}(l,v)$ where $k_{mcw} \in V^* \times \text{suff}(P) \to \mathbb{N}$ is defined by

$$k_{mcw}(x, y) = \{ (\text{char}_{cw}(x[1], |y|) \max d_{ei}(y)) \min d_{sp}(y) \mid (x \in V^+, y \in \text{suff}(P)) \}$$

$$\min d_{ei}(y) \quad (x = \varepsilon, y \in \text{suff}(P)).$$

Precomputation of $\text{char}_{cw}$ is discussed in subsection 4.3 and of $d_{ui}$, $d_{sp}$, and $d_i$ in subsection 4.2.

Notice that we have

$$k_{opt}(x, y) \geq k_{mcw}(x, y) \quad (x \in V^+, y \in \text{suff}(P)).$$
Approximation from below of $k_{opt}$ by $k_{bmew}$ is referred to as algorithm detail (BMCW). We chose this name to reflect that essential ideas from both the Boyer-Moore and Commentz-Walter algorithms are introduced. In the next two subsections these algorithms are derived from the algorithm presented in this subsection and characterized by detail sequence ($P_+,S_+,RT,SSD,NLAU,OPT,BMCW$).

3.4 The multiple keyword Boyer-Moore algorithm

We proceed by deriving the multiple keyword generalization of the Boyer-Moore algorithm [BM77] from the algorithm in subsection 3.3. It only differs from the algorithm there in the way the lookahead symbol is taken into account. Assuming $l \neq \varepsilon$ we derive

$$char_{ew}(l|v|v) = \begin{cases} \text{definition } char_{ew} \\ (\text{MIN } n : n \geq 1 \land V^*(l|v|v \cap P \neq \emptyset : n - |v|) \\ (\text{MIN } n : n \geq 1 \land V^*(l|v|v \cap V^*P \neq \emptyset : n - |v|) \\ \{ P \neq \emptyset, V^*(l1)|v|v \cap V^*p \neq \emptyset \text{ for all } p \in P, \text{ nonempty domain} \} \\ (\text{MIN } n : n \geq 1 \land V^*(l|v|v \cap V^*P \neq \emptyset : n - |v|) \\ \{ l \neq \varepsilon, \text{ definition of } char_{bm} \text{ (after derivation)} \} \\ char_{bm}(l|1|) - |v| \end{cases}$$

where $char_{bm} \in V \rightarrow \mathbb{N}$ is defined by

$$char_{bm}(a) = (\text{MIN } n : n \geq 1 \land V^*aV^n \cap V^*P \neq \emptyset : n) \quad (a \in V).$$

It results in shift distance $k_{bm}(l,v)$ where $k_{bm} \in V^* \times \text{suff}(P) \rightarrow \mathbb{N}$ is defined by

$$k_{bm}(x,y) = \begin{cases} \{ (char_{bm}(x|1| - |y|) \max d_{ev}(y)) \min d_{sp}(y) \\ d_{v}(y) \min d_{p}(y) \} \\ (x \in V^+, y \in \text{suff}(P)) \\ (x = \varepsilon, y \in \text{suff}(P)). \end{cases}$$

Precomputation of $char_{bm}$ is discussed in subsection 4.3 and of $d_{ev}, d_{sp}$, and $d_v$ is discussed in subsection 4.2. Approximating $k_{bmew}$ from below by $k_{bm}$ is referred to as algorithm detail (BM). It results in the multiple keyword generalization of the regular Boyer-Moore algorithm [BM77]. The algorithm is characterized by detail sequence ($P_+,S_+,RT,SSD,NLAU,OPT,BMCW,BM$). The regular Boyer-Moore algorithm can be obtained by restricting $P$ to one keyword (problem detail (OKW) (one keyword)). Notice that we have

$$k_{bmew}(x,y) \geq k_{bm}(x,y) \quad (x \in V^*, y \in \text{suff}(P)).$$

Inequality can only occur if the lookahead symbol does not occur in any keyword except as the last symbol.

The formula for the Boyer-Moore shift function given here differs from the ones given in [BM77] and [Aho90]. We will show that all formulas are equivalent. First we define

$$m_p = \begin{cases} 1 \\ (\text{MIN } p : p \in P \cap |p|) \end{cases} \quad \text{if } \varepsilon \in P \quad \text{if } \varepsilon \notin P.$$ 

We now have for all $a \in V$

$$char_{bm}(a) \leq m_p.$$ 

Since for all $y \in \text{suff}(P)$ we have $(\forall n : 1 \leq n < m_p - |y| : yV^n \cap V^*P = \emptyset)$ it follows that

$$m_p - |y| \leq d_{sp}(y) \quad (y \in \text{suff}(P)).$$
Finally, we derive for \( x \in V^+ \) and \( y \in \text{suff}(P) \)

\[
\kappa_{bm}(x,y) = \begin{cases} \text{definition } \kappa_{bm} \\ ((\text{char}_{bm}(x,|y|) \max d_{vi}(y)) \min d_{sp}(y) \end{cases}
\]

\[
= \begin{cases} \{ \text{definition } d_{vi} \} \\ ((d_{sp}(y) \geq m_P - |y| \geq \text{char}_{bm}(x,|y|)) \max d_{vi}(y)) \min ((\text{char}_{bm}(x,|y|) \max d_{sp}(y)) \end{cases}
\]

\[
= \begin{cases} \{ \text{distributivity} \} \\ (\text{char}_{bm}(x,|y|)) \max (d_{vi}(y) \min d_{sp}(y)). \end{cases}
\]

The last formula in the preceding derivation coincides with the ones in [BM77] and [Ah09].

### 3.5 The Commentz-Walter algorithm

Instead of approximating \( \text{char}_{cw} \) in \( \kappa_{bm} \) from below by \( \text{char}_{bm} \) we now approximate \( d_{vi} \) in \( \kappa_{bm} \) from below by \( d_{vi} \). This results in the Commentz-Walter algorithm [CW79a, CW79b]. We derive

\[
d_{vi}(v) = \begin{cases} \text{definition } d_{vi} \} \\ (\text{MIN } n : n \geq 1 \land V^*(V \setminus MS(v))V^n \cap P \neq \emptyset : n) \end{cases}
\]

\[
\geq \begin{cases} \text{MIN } n : n \geq 1 \land V^*V^n \cap P \neq \emptyset \Rightarrow V^*V^n \cap P \neq \emptyset \} \\ \text{MIN } n : n \geq 1 \land V^*V^n \cap P \neq \emptyset : n) \end{cases}
\]

\[
= \begin{cases} \text{definition } d_{i} \} \\ d_{i}(v). \end{cases}
\]

This results in shift distance \( \kappa_{cw}(l,v) \) where \( \kappa_{cw} \in V^* \times \text{suff}(P) \rightarrow \mathbb{N} \) is defined by

\[
\kappa_{cw}(x,y) = \begin{cases} \{ \text{char}_{cw}(x,|y|) \max d_{vi}(y)) \min d_{sp}(y) \} \\ d_{vi}(y) \min d_{sp}(y) \end{cases} \begin{cases} (x \in V^+ \land y \in \text{suff}(P)) \\ (x = y, y \in \text{suff}(P)). \end{cases}
\]

Precomputation of \( \text{char}_{cw} \) is discussed in subsection 4.3 and of \( d_{i} \) and \( d_{sp} \) in subsection 4.2. Approximating \( \kappa_{bm} \) from below by \( \kappa_{cw} \) is referred to as algorithm detail (cw). It results in the Commentz-Walter algorithm [CW79a, CW79b] that is characterized by detail sequence \( (P, S, R, T, S, S, N, L, A, U, O, P, T, B, M, C, W, C, W) \). Notice that we have

\[
\kappa_{bm}(x,y) \geq \kappa_{cw}(x,y) \quad (x \in V^+, y \in \text{suff}(P)).
\]

Such a comparison can not be made between \( \kappa_{bm} \) and \( \kappa_{cw} \) as the following example shows.

**Example 3.1** Let \( V \in \{a,b,c,d\} \), \( P = \{cababa\} \), and \( x \in V^+ \). Shift functions \( \kappa_{bm} \) and \( \kappa_{cw} \) are incomparable since

\[
\kappa_{opt}(xa,a) = +\infty \min 6 = 6
\]

\[
\kappa_{bm}(xa,a) = (+\infty \max 4) \min 6 = 6
\]

\[
\kappa_{bm}(xa,a) = ((6 - 1) \max 4) \min 6 = 5
\]

\[
\kappa_{cw}(xa,a) = (+\infty \max 2) \min 6 = 6
\]

and

\[
\kappa_{opt}(xa,a) = +\infty \min 6 = 6
\]

\[
\kappa_{bm}(xa,a) = ((2 - 1) \max 4) \min 6 = 4
\]

\[
\kappa_{bm}(xa,a) = ((2 - 1) \max 4) \min 6 = 4
\]

\[
\kappa_{cw}(xa,a) = ((2 - 1) \max 2) \min 6 = 2.
\]

It also follows that in some cases \( \kappa_{bm} \) is smaller than \( \kappa_{opt} \) and that in some cases \( \kappa_{bm} \) and \( \kappa_{cw} \) are smaller than \( \kappa_{bm} \). □
It is not possible that both $k_{bmcw}(x, y) > k_{bm}(x, y)$ and $k_{bmcw}(x, y) > k_{cw}(x, y)$ hold for some $x \in V^+$ and $y \in \text{suff}(P)$ since the first inequality implies $\text{char}_{cw}(x, |y|) = +\infty$ and this in its turn implies $k_{bmcw}(x, y) = d_{sp}(y) = k_{cw}(x, y)$.

### 3.6 Complete decoupling of recognized suffix and lookahead symbol

The derivations in the previous subsections effect an ever stronger decoupling of the recognized suffix $v$ and the lookahead symbol $l|1$ in the subsequent shift functions. By approximating $d_v$ in $k_{bm}$ from below by $d$, or $\text{char}_{cw}$ in $k_{cw}$ by $\text{char}_{bm}$ (or both in $k_{bmcw}$) we obtain a complete decoupling. It results in shift distance $k_{dal}(l, v)$ where $k_{dal} \in V^* \times \text{suff}(P) \rightarrow N$ is defined by

$$k_{dal}(x, y) = \begin{cases} \max \{ \text{char}_{bm}(x|1) - |y|, d(y) \} \min d_{sp}(y) & (x \in V^+, y \in \text{suff}(P)) \\ d(y) \min d_{sp}(y) & (x = \epsilon, y \in \text{suff}(P)) \end{cases}$$

Precomputation of $\text{char}_{bm}$ is discussed in subsection 4.3 and of $d$ and $d_{sp}$ in subsection 4.2. The algorithm can be characterized by detail sequences $(P+S+, RT, SSD, NLAU, OPT, BMCW, BM, CW)$ and $(P+S+, RT, SSD, NLAU, OPT, BMCW, CW, BM)$.

### 3.7 Discarding the lookahead symbol

We weaken the predicate in the range of $k_{opt}$ by weakening its first disjunct to $V^* \cup V^n \cap P \neq \emptyset$ due to $V^* |l|1 \subseteq V^*$ and the monotonicity of $\cap$. This weakening step is referred to as discarding the lookahead symbol $l|1$. The shift distance corresponding to this weakening is $k_{nla}(v)$ where $k_{nla} \in \text{suff}(P) \rightarrow N$ is defined by

$$k_{nla}(y) = d_i(y) \min d_{sp}(y) \quad (y \in \text{suff}(P)).$$

Notice that this shift function can also be viewed as an approximation from below of $k_{dal}$. Precomputation of $d_i$ and $d_{sp}$ is discussed in subsection 4.2. Approximating $k_{opt}$ from below by $k_{nla}$ is referred to as algorithm detail (NLA) (No LookAhead at mismatching symbol) and results in algorithm $(P+S+, RT, SSD, NLA, OPT, BMCW, BM)$.

### 3.8 One symbol lookahead at the unscanned part of the input string

In this subsection we consider looking ahead at the first symbol of the unscanned part $r$ of the input string. The first symbol of $r$ will be taken into account independently of the other available information. In this way we obtain stronger variants of all of the shift functions derived thus far. Assuming $r \neq \epsilon$ we derive

$$\begin{align*}
(MIN \ n : 1 \leq n & \land \text{suff}(u(r|n)) \cap P \neq \emptyset : n) \\
= & \{ \text{domain split, } 1 \leq n \leq |r|: r|n = (r|1)((r|1)|(n-1)), |r| < n: r|n = r \} \\
(MIN \ n : 1 \leq n & \leq |r| \land \text{suff}(u(r|1)|(r|1)|(n-1))) \cap P \neq \emptyset : n) \\
\min (MIN \ n : |r| < n \land \text{suff}(ur) \cap P \neq \emptyset : n) \\
\geq & \{ 1 \leq n \leq |r|: (r|1)|(n-1) \in V^{n-1}, \text{monotonicity of suff and cap} \} \\
(MIN \ n : 1 \leq n & \leq |r| \land \text{suff}(u(r|1)V^{n-1}) \cap P \neq \emptyset : n) \\
\min (MIN \ n : |r| < n \land \text{suff}(ur) \cap P \neq \emptyset : n) \\
= & \{ r \neq \epsilon, r \in (r|1)V^{r|1-1} \land \text{suff}(ur) \cap P \neq \emptyset \Rightarrow \text{suff}(u(r|1)V^{r|1-1}) \cap P \neq \emptyset \} \\
(MIN \ n : 1 \leq n & \leq |r| \land \text{suff}(u(r|1)V^{n-1}) \cap P \neq \emptyset : n) \\
\geq & \{ u \in V^*, \text{monotonicity of suff and } \cap \} \\
(MIN \ n : 1 \leq n & \leq |r| \land \text{suff}(V^* (r|1)V^{n-1}) \cap P \neq \emptyset : n) \\
\geq & \{ \text{enlarging domain, changing bound variable: } m = n - 1 \} 
\end{align*}$$
Precomputation

In this section we derive algorithms for the precomputation of the functions used in the pattern matching algorithms in sections 2 and 3. The algorithms are correct due to their formal derivation. This cannot always be said about the algorithms found in the literature, mostly due to the absence of any formal derivation (see for instance the single keyword Boyer-Moore precomputation algorithms given in [BM77], [KMP77], and [Ryt80], where each article shows the preceding article to give an incorrect precomputation algorithm). Moreover, we give the first formal derivation of the precomputation algorithms for the Boyer-Moore family of algorithms. They can, amongst others, be specialized to a correct precomputation algorithm for the single keyword Boyer-Moore algorithm. In fact, we show that the definition of all $d$-functions introduced in section 3 can be rewritten into a form in accordance with one general pattern. Subsequently, a general precomputation algorithm scheme for this general pattern is derived that can be instantiated for every $d$-function.

### 4.1 Precomputation of $\tau_P$

The transition function $\tau_P \in \text{suff}(P) \times V \rightarrow (\text{suff}(P) \cup \{\bot\})$ of the reverse trie corresponding to $P$ is defined by

$$\tau_P(u,a) = \begin{cases} au & \text{if } au \in \text{suff}(P) \\ \bot & \text{if } au \notin \text{suff}(P) \end{cases} \quad (u \in \text{suff}(P), a \in V).$$

Since $\text{suff}$ is idempotent and the definition of $\tau_P$ only depends on $\text{suff}(P)$, we have $\tau_P = \tau_{\text{suff}(P)}$. Set $P$ being nonempty we have $\text{suff}(P) = \{\varepsilon\} \cup \text{suff}(P)$ and $\tau_{\text{suff}(P)} = \tau_{\varepsilon \cup \text{suff}(P)}$. These
observations lead to the following algorithm (cf. [AC75], section 3, algorithm 2) to compute $\tau_P$ in which variable $\tau$ is used to calculate and store $\tau_P$ thereby viewing $\tau$ as a set of ordered pairs (the usual notion of a function) and abbreviate statements like $\tau := \tau + \{(x,a), y\}$ to $\tau(x,a) := y$:

$$
\text{tau} := \emptyset; \\
\text{for } a : a \in V \rightarrow \tau(a, a) := \bot \text { rof}; \\
\text{P}_d, P_r := \emptyset, P; \\
\text{do } P_r \neq \emptyset \rightarrow \\
\text{p} : p \in P_r; \\
\text{u, v} := p, \epsilon; \\
\text{if } \tau(v, u[1]) = \bot \rightarrow \\
\text{tau}(v, u[1]) := (u[1])v; \\
\text{for } a : a \in V \rightarrow \tau((u[1])v, a) := \bot \text { rof} \\
\text{od}; \\
\text{P}_d, P_r := P_d + \{p\}, P_r - \{p\} \\
\text{od} \{ \text{tau} = \tau_P \}
$$

In the algorithms we use $+$ for the union of disjoint sets and $-$ for the difference of a set and a subset of it. Notice that the algorithm does a depth first traversal of the reverse trie. Also notice that variable $P_d$ is only needed to formulate an invariant for $\tau$, so it may safely be removed from the algorithm. Furthermore, the states of the reverse trie are represented by strings. In practice, one can resort to a more suitable representation, for instance a representation by natural numbers. We will not elaborate this here.

### 4.2 Precomputation of d-functions

In this subsection we show that all $d$-functions introduced in section 3 can be written according to a general pattern. For this general pattern a general precomputation algorithm is derived. In order to obtain a precomputation algorithm for a particular $d$-function one only has to instantiate the general precomputation algorithm and possibly simplify the resulting algorithm.

The general pattern we strive for is a function $d : V \times \text{suff}(P) \rightarrow \mathbb{N}$ defined for $a \in V$ and $y \in \text{suff}(P)$ by

$$
d(a, y) = \left( \text{MIN} t : t \in \text{suff}(P) \setminus \{\epsilon\} \wedge Q(a,t) \wedge R(a,y,t) \wedge y <_P t : |t| - |y| \right)
$$

where $Q$ is a predicate on $V \times V^*$ and $R$ a predicate on $V \times V^* \times V^*$. Why both $Q$ and $R$ are introduced will become clear when we derive an algorithm scheme for the computation of $d$. We will now show that all $d$-functions introduced in section 3 can be expressed in this pattern. In the following derivations let $a \in V$ and $y \in \text{suff}(P)$.

- We derive

$$
d_{\text{opt}}(a, y) = \left( \text{MIN} n : n \geq 1 \wedge V^*ay V^n \cap P \neq \emptyset : n \right)
$$

\{ property A.4 \}
\[(\text{MIN } n : n \geq 1 \land ayV^n \cap \text{suff}(P) \neq \emptyset : n)\]
\[= \quad \{ \text{change of bound variable: } n = |s| \} \]
\[(\text{MIN } s : s \in V^+ \land ay \in \text{suff}(P) : |s|)\]
\[= \quad \{ \text{change of bound variable: } t = ys \} \]
\[(\text{MIN } t : t \in \text{suff}(P) \setminus \{\varepsilon\} \land at \in \text{suff}(P) \land y <_p t : |t| - |y|)\]

Hence, we have expressed \(d_{sp}\) according to the general pattern with \(Q(a, t) = at \in \text{suff}(P)\) and \(R(a, y, t) = \text{true}\). Notice that \(at \in \text{suff}(P) \equiv \tau(t, a) \neq 1\).

- Assuming \(y = \varepsilon\) we derive
  \[d_{sp}(\varepsilon)\]
  \[= \quad \{ \text{definition } d_{sp} \} \]
  \[(\text{MIN } n : n \geq 1 \land V^n \cap V^*P \neq \emptyset : n)\]
  \[= \quad \{ \text{calculus} \} \]
  \[\left(\text{MIN } t : t \in P : (\text{MIN } n : n \geq 1 \land V^n \cap V^*t \neq \emptyset : n)\right)\]
  \[= \quad \{ \text{domain split: } P = P \setminus \{\varepsilon\} \cup (P \cap \{\varepsilon\}) \} \]
  \[(\text{MIN } t : t \in P \setminus \{\varepsilon\} : |t|) \min (\text{MIN } t : t \in P \cap \{\varepsilon\} : 1)\]

Assuming \(y \neq \varepsilon\) we derive
  \[d_{sp}(y)\]
  \[= \quad \{ \text{definition } d_{sp} \} \]
  \[(\text{MIN } n : n \geq 1 \land yV^n \cap V^*P \neq \emptyset : n)\]
  \[= \quad \{ \text{property A.4} \} \]
  \[(\text{MIN } n : n \geq 1 \land \text{suff}(yV^n) \cap P \neq \emptyset : n)\]
  \[= \quad \{ y \neq \varepsilon, \text{ hence } \text{suff}(yV^n) = yV^n \cup \text{suff}((y|1)V^n), \text{ set calculus} \} \]
  \[(\text{MIN } n : n \geq 1 \land yV^n \cap P \neq \emptyset : n)\]
  \[\min (\text{MIN } n : n \geq 1 \land \text{suff}((y|1)V^n) \cap P \neq \emptyset : n)\]
  \[= \quad \{ \text{change of bound variable: } n = |s|, \text{ property A.4} \} \]
  \[(\text{MIN } s : s \in V^+ \land ys \in P : |s|) \min (\text{MIN } n : n \geq 1 \land (y|1)V^n \cap V^*P \neq \emptyset : n)\]
  \[= \quad \{ \text{change of bound variable: } t = ys, y \in \text{suff}(P), y|1 \in \text{suff}(P), \text{ definition } d_{sp} \} \]
  \[(\text{MIN } t : t \in P \land y <_p t : |t| - |y|) \min d_{sp}(y|1)\]
  \[= \quad \{ \text{rewriting in order to obtain general pattern} \} \]
  \[(\text{MIN } t : t \in \text{suff}(P) \setminus \{\varepsilon\} \land t \in P \land y <_p t : |t| - |y|) \min d_{sp}(y|1)\]

Although the derived definition of \(d_{sp}\) is recursive and \(d_{sp}\) does not have an argument \(a \in V\) one can still discern the general pattern with \(Q(a, t) = t \in P\) and \(R(a, y, t) = \text{true}\). Precomputation of \(d_{sp}\) can be done according to the general precomputation algorithm without an iteration over \(V\), followed by a breadth first traversal of the reverse trie.
We derive
\[ d_i(y) \]
\[
\begin{align*}
&= \{ \text{definition } d_i \} \\
&= (\MIN n : n \geq 1 \land V^*yV^n \cap P \neq \emptyset : n) \\
&= \{ \text{property A.4} \} \\
&= (\MIN n : n \geq 1 \land yV^n \cap \suff(P) \neq \emptyset : n) \\
&= \{ \text{change of bound variable: } n = |s| \} \\
&= (\MIN s : s \in V^+ \land ys \in \suff(P) : |s|) \\
&= \{ \text{change of bound variable: } t = ys \} \\
&= (\MIN t : t \in \suff(P) \setminus \{\epsilon\} \land y <_P t : |t| - |y|)
\end{align*}
\]
Although \( d_i \) does not have an argument \( a \in V \) its definition still matches the general pattern with \( Q(a, t) = \text{true} \) and \( R(a, y, t) = \text{true} \).

We derive
\[ d_{e^i}(y) \]
\[
\begin{align*}
&= \{ \text{definition } d_{e^i} \} \\
&= (\MIN n : n \geq 1 \land V^*(V \setminus MS(y))yV^n \cap P \neq \emptyset : n) \\
&= \{ \text{property A.4} \} \\
&= (\MIN n : n \geq 1 \land (V \setminus MS(y))yV^n \cap \suff(P) \neq \emptyset : n) \\
&= \{ \text{change of bound variable: } n = |s| \} \\
&= (\MIN s : s \in V^+ \land (V \setminus MS(y))ys \cap \suff(P) \neq \emptyset : |s|) \\
&= \{ \text{change of bound variable: } t = ys \} \\
&= (\MIN t : t \in \suff(P) \setminus \{\epsilon\} \land (V \setminus MS(y))t \cap \suff(P) \neq \emptyset \land y <_P t : |t| - |y|) \\
&= \{ \text{definition } MS \} \\
&= (\MIN t : t \in \suff(P) \setminus \{\epsilon\} \land MS(t) \cap (V \setminus MS(y)) \neq \emptyset \land y <_P t : |t| - |y|)
\end{align*}
\]
Apart from the fact that \( d_{e^i} \) does not have an argument \( a \in V \) its definition matches the general pattern with \( Q(a, t) = \text{true} \) and \( R(a, y, t) = MS(t) \land (V \setminus MS(y)) \neq \emptyset \).

Having expressed all \( d \)-functions from section 3 in the general pattern we proceed by giving a rather straightforward and nondeterministic algorithm to compute \( d \) which will serve as a starting point for further algorithm derivations (notice that program variable \( \text{dee} \) is used to compute and finally store function \( d \)).
Precomputation

for \( y : y \in \text{suff}(P) \rightarrow \)
\[ \begin{align*}
  & \text{for } a : a \in V \rightarrow \text{dec}(a, y) := +\inf \text{ rof} \\
  & \text{rof};
\end{align*} \]
\[ \text{for } t : t \in \text{suff}(P) \setminus \{\varepsilon\} \rightarrow \]
\[ \begin{align*}
  & \text{for } a : a \in V \rightarrow \\
  & \quad \text{if } Q(a, t) \rightarrow \text{for } y : y \in \text{suff}(P) \land y <_P t \rightarrow \\
  & \quad \quad \text{if } R(a, y, t) \rightarrow \text{dec}(a, y) := \text{dec}(a, y) \min(|t| - |y|) \\
  & \quad \quad \text{rof} \\
  & \quad \quad \text{fi} \\
  & \quad \text{rof} \\
  & \text{rof}\{ \text{dec} = d \}.
\end{align*} \]

For \( d \) and \( d_{sp} \) the loop iterating over \( V \) can be omitted. For \( d_{sp} \) the following additional breadth first traversal of the reverse trie (\( \text{suff}(P) \)) is needed to complete the computation:

\[ \begin{align*}
  & n := 1; \\
  & \text{do } \text{suff}(P) \cap V^n \neq \emptyset \rightarrow \\
  & \quad \text{for } t : t \in \text{suff}(P) \cap V^n \rightarrow \text{deesp}(t) := \text{deesp}(t) \min \text{deesp}(t) + 1 \text{ rof} \\
  & \quad \text{od}\{ \text{deesp} = d_{sp} \}. \\
\end{align*} \]

First, we concentrate on making the innermost repetition deterministic. Define \( sp \in \text{suff}(P) \rightarrow \mathcal{P}(\text{suff}(P)) \) by

\[ sp(t) = \{ y \mid y \in \text{suff}(P) \land y <_P t \} \quad (t \in \text{suff}(P)), \]

the set of all suffixes of keywords that are a proper prefix of \( t \). Notice that for all \( t \in \text{suff}(P) \setminus \{\varepsilon\} \) \( sp(t) \) is finite, nonempty, and linearly ordered with respect to \( \leq_P \). Therefore, we can define \( msp \in \text{suff}(P) \setminus \{\varepsilon\} \rightarrow \text{suff}(P) \) by

\[ msp(t) = \{ \text{MAX}_{<_P} y : y \in \text{suff}(P) \land y <_P t : y \} \quad (t \in \text{suff}(P) \setminus \{\varepsilon\}) \]

being the maximal element of \( sp(t) \). In literature \([\text{AC75}, \text{BM77}, \text{CW79a}, \text{KMP77}]\) function \( msp \) is known as the failure function corresponding to the reverse trie. For \( t \in \text{suff}(P) \setminus \{\varepsilon\} \) we derive a recursive definition of \( sp(t) \) in terms of function \( msp \):

\[ \begin{align*}
  & sp(t) \\
  & \quad = \{ \text{definition } sp \} \\
  & \quad \quad \{ y \mid y \in \text{suff}(P) \land y <_P t \} \\
  & \quad = \{ t \in \text{suff}(P) \setminus \{\varepsilon\}, \text{property A.6} \} \\
  & \quad \quad \{ y \mid y \in \text{suff}(P) \land (y = msp(t) \lor y <_P msp(t)) \} \\
  & \quad = \{ \text{definition } sp \text{ and } msp \} \\
  & \quad \quad \{msp(t)\} \cup sp(msp(t))
\end{align*} \]

Provided \( msp \) is already computed (precomputation of \( msp \) is discussed in subsection 4.4) the innermost repetition traversing the set \( sp(t) \) can be replaced by the following deterministic repetition (variable \( \bar{v} \) is a ghost variable needed to express the invariant):
4 Precomputation

\[ v := t; \{ v := \emptyset; \} \]
\{ invariant: \( v \in sp(t) \cup \{ t \} \land sp(t) = \emptyset \cup sp(v) \land \emptyset \cap sp(v) = \emptyset \} \]
do \( v \neq \varepsilon \rightarrow \{ sp(v) = (msp(v)) \cup sp(msp(v)) \} \]
\[ v := msp(v); \{ v := v + \{ v \}; \} \]
if \( R(a, v, t) \rightarrow dee(a, v) := dee(a, v) \min(|t| - |v|) \)
| \( \neg R(a, v, t) \rightarrow \text{skip} \)
fi
od{ \( v = \varepsilon, \) so \( sp(v) = \emptyset \) and \( sp(t) = \emptyset \) }

The invariant expresses that we have a bipartition of \( sp(t) \) in \( v \) (elements of \( sp(t) \) that have already contributed to the computation of \( d \)) and \( sp(v) \) (the other elements of \( sp(t) \)).

In case \( R(a, y, t) = \text{true} \) for all \( a, y, \) and \( t \) the inner repetition can be made more efficient. Notice that this can be done for all presented \( d \)-functions except \( d_v \). In the following assume that \( R(a, y, t) = \text{true} \) for all \( a, y, \) and \( t \). Suppose that for some \( v \in sp(t) \) in the above repetition we have \( |t| - |v| \geq dee(a, v) \). From the structure of the algorithms we infer that \( dee(a, v) = |t_0| - |v| \) for some \( t_0 \in \text{suff}(P) \setminus \{ \varepsilon \} \) with \( |t_0| \leq |t| \) that has already contributed to the computation of \( d \). Therefore, for all \( s \in sp(v) \) we have \( dee(a, s) \leq |t_0| - |s| \) leading to

\[ dee(a, s) \leq \{ \} \]
\[ |t_0| - |s| \]
\[ = \{ dee(a, v) = |t_0| - |v| \} \]
\[ dee(a, v) + |v| - |s| \]
\[ \leq \{ dee(a, v) \leq |t| - |v| \} \]
\[ |t| - |s| \]

Hence, the contribution of \( t \) will not change the already computed value of \( dee(a, s) \) for \( s \in sp(v) \) and the inner repetition can be terminated. This yields the following repetition using an additional boolean variable \( \text{contributes} \) (notice that ghost variable \( \overline{v} \) is omitted):

\[ v := t; \text{contributes} := \text{true}; \]
do \( v \neq \varepsilon \land \text{contributes} \rightarrow \]
\[ v := msp(v); \]
if \( |t| - |v| < dee(a, v) \rightarrow dee(a, v) := |t| - |v| \)
| \( |t| - |v| \geq dee(a, v) \rightarrow \text{contributes} := \text{false} \)
fi
od.

Variable \( \text{contributes} \) can be removed resulting in the following repetition

\[ v := t; \]
do \( v \neq \varepsilon \rightarrow \]
\[ v := msp(v); \]
if \( |t| - |v| < dee(a, v) \rightarrow dee(a, v) := |t| - |v| \)
| \( |t| - |v| \geq dee(a, v) \rightarrow v := \varepsilon \)
fi
od.

In order to further exploit this phenomenon the elements of \( \text{suff}(P) \setminus \{ \varepsilon \} \) are dealt with in order of increasing length, i.e. the outermost repetition of the general precomputation algorithm does a breadth first traversal of the reverse trie. This results in the following algorithm:
for \( y : y \in \text{suff}(P) \rightarrow \)
for \( a : a \in V \rightarrow \text{dee}(a, y) := +\infty \)
rof;

\( n := 1; \)
do \( \text{suff}(P) \cap V^n \neq \emptyset \rightarrow \)
for \( t : t \in \text{suff}(P) \cap V^n \rightarrow \)
for \( a : a \in V \rightarrow \)
if \( Q(a, t) \rightarrow v := t; \)
do \( v \neq e \rightarrow \)
\( v := \text{msp}(v); \)
if \( |t| - |v| < \text{dee}(a, v) \rightarrow \text{dee}(a, v) := |t| - |v| \)
\( \mid |t| - |v| \geq \text{dee}(a, v) \rightarrow v := e \)
\fi
\od
\fi
rof;
\( n := n + 1 \)
od.

In this optimized breadth first precomputation algorithm for each node \( v \) in the reverse trie and each symbol from \( V \) the step from \( v \) to \( \text{msp}(v) \) is done at most two times. Therefore, the precomputation time is \( O(|\text{suff}(P)| \cdot |V|) \). If the traversal of \( V \) can be omitted (as is the case for the precomputation of \( d_i \) and \( d_{sp} \)) it is \( O(|\text{suff}(P)|) \).

The breadth first precomputation algorithm for \( d_i \) can be simplified further by observing that since \( Q(a, t) = \text{true} \) the steps taken from \( t \) are always preceded by the steps taken from \( \text{msp}(t) \) (provided \( \text{msp}(t) \neq e \)) since \( |\text{msp}(t)| < |t| \). So, only the contribution of \( t \) to \( d_i(\text{msp}(t)) \) has to be considered. This results in the following precomputation algorithm for \( d_i \) (program variable \( \text{dee}(i) \) is used to compute and finally store \( d_i \)):

for \( y : y \in \text{suff}(P) \rightarrow \text{dee}(i)(y) := +\infty \)
rof;

\( n := 1; \)
do \( \text{suff}(P) \cap V^n \neq \emptyset \rightarrow \)
for \( t : t \in \text{suff}(P) \cap V^n \rightarrow \)
\( \text{dee}(i)(\text{msp}(t)) := \text{dee}(i)(\text{msp}(t)) \min(|t| - |\text{msp}(t)|) \)
rof;
\( n := n + 1 \)
od.

Notice that the breadth first traversal in this algorithm may be replaced by an arbitrary traversal of the reverse trie.

### 4.3 Precomputation of \( \text{char}_{cw} \), \( \text{char}_{bm} \), and \( \text{char}_{la} \)

Function \( \text{char}_{cw} \) can be expressed as

\[
\text{char}_{cw}(a, z) = \begin{cases} 
+\infty & \text{if } \overline{\text{char}}_{cw}(a) = +\infty \\
\overline{\text{char}}_{cw}(a) - z & \text{if } \overline{\text{char}}_{cw}(a) \neq +\infty 
\end{cases} \quad (a \in V, z \in \mathbb{N})
\]

where function \( \overline{\text{char}}_{cw} \in V \to \mathbb{N} \) is defined by

\[
\overline{\text{char}}_{cw}(a) = \left( \text{MIN} n : n \geq 1 \land V^a V^n \cap P \neq \emptyset : n \right) \quad (a \in V).
\]

From the definition \( \overline{\text{char}}_{cw} \) it immediately follows that that its computation can be interwoven with the precomputation of the reverse trie \( \tau \).
Next, we derive for $a \in V$

$$\text{char}_{bm}(a) = \{ \text{definition char}_{bm} \}$$

$$(\text{MIN } n : n \geq 1 \land V^a V_n \cap V^P \neq \emptyset : n)$$

$$= \{ \text{property A.5.iii} \}$$

$$(\text{MIN } n : n \geq 1 \land (V^a V_n \cap P \neq \emptyset \lor aV_n \cap V^P \neq \emptyset) : n)$$

$$= \{ \text{disjunctive range, definition char}_{cw} \}$$

$$\overline{\text{char}_{cw}}(a) \min (\text{MIN } n : n \geq 1 \land V^a V_n \cap V^P \neq \emptyset : n)$$

$$= \{ \text{definition m}_{P} (\text{subsection 3.4}) \}$$

$$\overline{\text{char}_{cw}}(a) \min m_{P}.$$ 

Finally, we derive for $a \in V$

$$\text{char}_{la}(a) = \{ \text{definition char}_{la} \}$$

$$(\text{MIN } n : 0 \leq n \land V^a V_n \cap V^P \neq \emptyset : n)$$

$$= \{ \text{domain split: } n \geq 1 \lor n = 0, \text{ definition char}_{bm} \}$$

$$\text{char}_{bm}(a) \min (\text{MIN } n : n = 0 \land V^a a \cap V^P \neq \emptyset : 0)$$

$$= \{ \text{property A.5.ii} \}$$

$$\text{char}_{bm}(a) \min (\text{MIN } n : n = 0 \land (a \in \text{suff}(P) \lor \epsilon \in P) : 0)$$

We conclude that char_{cw}, char_{bm}, and char_{la} can be computed from char_{cw}.

### 4.4 Precomputation of m_{sp}

We conclude this section with the derivation of an algorithm computing m_{sp}. We start by deriving

- for $a \in \text{suff}(P) \cap V$

$$m_{sp}(a) = \{ \text{definition m}_{sp} \}$$

$$(\text{MAX}_{\leq_{P}} y : y \in \text{suff}(P) \land y <_{P} a \vdash y)$$

$$= \{ y <_{P} a \equiv y = \epsilon, \epsilon \in \text{suff}(P) \}$$

$$\epsilon$$

- and for $ax \in \text{suff}(P)$, $a \in V$, and $x \in \text{suff}(P) \setminus \{ \epsilon \}$

$$m_{sp}(ax) = \{ \text{definition m}_{sp} \}$$

$$(\text{MAX}_{\leq_{P}} y : y \in \text{suff}(P) \land y <_{P} ax \vdash y)$$

$$= \{ \text{domain split, } \epsilon <_{P} ax, \epsilon \in \text{suff}(P) \}$$
5 Conclusions

In this paper we derived and presented a taxonomy of sublinear keyword pattern matching algorithms closely related to the Boyer-Moore algorithm [BM77] and the Commentz-Walter algorithm [CW79a, CW79b]. It includes, amongst others, the multiple keyword generalization of the single keyword Boyer-Moore algorithm and the algorithm presented by Fan and Su [FS93, FS94]. We presented the algorithms within a common framework permitting an easier comprehension of and a better comparison between the algorithms. This was achieved by the systematic and formal derivation of the algorithms from a common starting point and by factoring out of common portions of the derivations. The derivations were done through series of refinements to either

\[ (\text{MAX}_{\leq_p} y : y \in \text{suff}(P) \land y \leq_p ax \land y \neq \varepsilon : y) \max_{\leq_p} \varepsilon \]

\[ = \{ \text{change of bound variable: } y = ay' \} \]

\[ (\text{MAX}_{\leq_p} y' : ay' \in \text{suff}(P) \land y' <_p x : ay') \max_{\leq_p} \varepsilon \]

\[ = \{ ay' \in \text{suff}(P) \Rightarrow y' \in \text{suff}(P), x \in \text{suff}(P) \setminus \{\varepsilon\}, \text{definition } sp \} \]

\[ (\text{MAX}_{\leq_p} y' : ay' \in \text{suff}(P) \land y' \in sp(x) : ay') \max_{\leq_p} \varepsilon \]

From this it follows that \( msp(ax) \) can be computed by a linear search in downward order over \( sp(x) \) (remember that \( sp(x) \) is linearly ordered with respect to \( \leq_p \)) starting with \( msp(x) \). Provided the computation of \( msp \) is done using a breadth first traversal of the reverse trie (\( \text{suff}(P) \)) the value of \( msp \) is already computed for all elements of \( sp(x) \cup \{x\} \) and can therefore be used to implement the linear search over \( sp(x) \). This results in the following algorithm (variable \( emsp \) is used to compute and finally store \( msp \)):

for \( a : a \in \text{suff}(P) \cap V \rightarrow emsp(a) := \varepsilon \) if; \n
\( n := 1; \)

\( \{ \text{invariant: } 1 \leq n \leq \left( \text{MAX}_{p} p : p \in P : |p| \right) + 1 \land \left( \forall y : y \in \text{suff}(P) \setminus \{\varepsilon\} \land |y| \leq n : emsp(y) = msp(y) \} \} \)

\( \text{do } \text{suff}(P) \cap V^n \neq \emptyset \rightarrow \)

\( \text{for } t : t \in \text{suff}(P) \cap V^n \rightarrow \)

\( \text{for } a : a \in V \rightarrow \)

\( \text{if } at \in \text{suff}(P) \rightarrow v := emsp(t); \)

\( \{ \text{linear search } \}

\( \text{do } av \notin \text{suff}(P) \land v \neq \varepsilon \rightarrow v := emsp(v) \) od;

\( \text{if } av \in \text{suff}(P) \rightarrow emsp(at) := av \)

\( \mid av \notin \text{suff}(P) \land v = \varepsilon \rightarrow emsp(at) := \varepsilon \)

\( \text{fi}

\( \mid at \notin \text{suff}(P) \rightarrow \text{skip} \)

\( \text{fi}

\( \text{rof; } \)

\( n := n + 1 \)

\( \text{rof.} \)

od.

This breadth first algorithm computing \( msp \) can be combined with the breadth first algorithm computing the \( d \)-functions. The precomputation time now is \( O(|\text{suff}(P)|^2 \cdot |V|) \). This can be reduced to \( O(|\text{suff}(P)| \cdot |V|) \) at the expense of \( O(|\text{suff}(P)| \cdot |V|) \) additional storage space by also computing and storing the transition function \( \gamma_r \) of the reverse trie where \( \gamma_r \in V \times \text{suff}(P) \rightarrow \text{suff}(P) \) is defined by

\[ \gamma_r(a, x) = (\text{MAX}_{\leq_p} y : y \in \text{suff}(P) \land y \leq_p ax : y) \quad (a \in V, x \in \text{suff}(P)). \]

The details of this approach can be found in [WZ92].

5 Conclusions
algorithm or problem. A refinement to the algorithm/problem is referred to as the introduction of an algorithm/problem detail. The sequence of details that are subsequently introduced in a derivation characterizes the algorithm obtained by that derivation. Detail sequences can therefore be used to classify the algorithms in the taxonomy. Algorithms can now be compared by looking at their detail sequences. The taxonomy graph in figure 1 constitutes a concise presentation and classification of the pattern matching algorithms discussed, vertices representing algorithms and edges representing the addition of an algorithm or problem detail. It can be viewed as an alternative table of contents to this paper. Our results show how fruitful the applied method of developing a taxonomy is (it was inspired by the method described by Jonkers [Jon83]).

Introduction of the notion of safe shift distances proved to be essential for the derivation of the various algorithms. All algorithms are characterized by a—systematically derived and more or less easy to compute—approximation from below of the maximal safe shift distance, computation of the latter being equivalent to the keyword pattern matching problem itself. The systematic derivation provided a means to compare the algorithms and their matching speeds, and to get a better understanding of the algorithms and their interrelations. Perhaps this better understanding will help further the use of the algorithms from this family. Our derivations show the Commentz-Walter algorithm not to be the multiple keyword generalization of the Boyer-Moore algorithm (as was the original intention of Commentz-Walter) and that such a generalization can indeed be obtained. Of the algorithms presented the algorithm by Fan and Su [FS93, FS94] is the fastest (at the expense of additional precomputation time and additional storage requirements), followed by the common ancestor of the Boyer-Moore and Commentz-Walter algorithms, and then by both the multiple keyword generalization of the Boyer-Moore algorithm [BM77] and the Commentz-Walter algorithm [CW79a, CW79b]. The latter two are incomparable in matching speed. It is clear that we have not derived and presented all algorithms of the Boyer-Moore family. Our derivation method, however, clearly indicates how yet other members of this family of algorithms may be derived.

Apart from giving a taxonomy of pattern matching algorithms from the Boyer-Moore family we presented the first formally derived and therefore correct precomputation algorithms (this can not always be said about the algorithms found in the literature, mostly due to the absence of any formal derivation; see for instance the many solutions for the Boyer-Moore precomputation that have been published, corrected and republished). In fact, we showed that most of the precomputation algorithms can be obtained as instantiations of a general precomputation algorithm scheme derived for a general function pattern in which most components of the various shift functions can be expressed. Thus, we provided a common framework for the precomputation algorithms as well.

6 Acknowledgements

We thank Kees Hemerik and Frans Kruseman Aretz for their careful reading of earlier drafts of this paper and their constructive criticisms. We also thank Anne Kaldewaij and Lex Bijlsma for reading the final version.

A Definitions and properties

This section provides a series of definitions and properties which are used throughout this paper. In the following let V be an alphabet.

For a string $w \in V^*$, $w^R$ denotes the reversal of $w$. For any language $L \subseteq V^*$ we define $L^R = \{w^R \mid w \in L\}$ (the reversal of language L).
Definition A.1 The infix operators $\|, |, |, \| : V^* \times N \to V^*$ are defined by

\[
\begin{align*}
\|0 &= \varepsilon & (v \in V^*) \\
|k+1 &= \varepsilon & (k \geq 0) \\
(aw)|k+1 &= a(w|k) & (k \geq 0, a \in V, w \in V^*) \\
\|0 &= v & (v \in V^*) \\
|k+1 &= \varepsilon & (k \geq 0) \\
(aw)|k+1 &= w|k & (k \geq 0, a \in W, w \in V^*)
\end{align*}
\]

Define $|$ as $v|k = (v_R|k)^R$ and as $v|k = (v^R|k)^R$. The operators $\|, |, |, \|$ are called "left take," "left drop," "right take," and "right drop" respectively.

Definition A.2 Functions $\text{pref} : P(V^*) \to P(V^*)$ and $\text{suff} : P(V^*) \to P(V^*)$ are defined by

\[
\begin{align*}
\text{pref}(L) &= \{ w \mid w \in V^* \land (\exists x : x \in V^* : wx \in L) \} \\
\text{suff}(L) &= (\text{pref}(LR))^R = \{ w \mid w \in V^* \land (\exists x : x \in V^* : wx \in L) \}.
\end{align*}
\]

For $w \in V^*$ we will write $\text{pref}(w)$ (suff$(w)$) instead of $\text{pref}([w])$ (suff$(\{w\})$).

Definition A.3 The relations $\leq_p$ and $\leq_s$ over $V^* \times V^*$ are defined by $u \leq_p v \equiv u \in \text{pref}(v)$ and $u \leq_s v \equiv u \in \text{suff}(v)$.

The following two properties are used in the derivation of the Commentz-Walter precomputation algorithm.

Property A.4 Let $A, B \subseteq V^*$. Then $\text{pref}(A) \cap B \neq \emptyset \equiv A \cap BV^* \neq \emptyset$ and $\text{suff}(A) \cap B \neq \emptyset \equiv A \cap V^B \neq \emptyset$.

Property A.5 Let $A, B \subseteq V^*$ and $a \in V$. Then

\[
\begin{align*}
i. V^*A \cap V^*B \neq \emptyset & \equiv V^*A \cap B \neq \emptyset \lor A \cap V^*B \neq \emptyset \\
ii. V^*aA \cap V^*B \neq \emptyset & \equiv V^*aA \cap B \neq \emptyset \lor A \cap V^*B \neq \emptyset \\
iii. V^*A \cap V^*B \neq \emptyset & \equiv V^*A \cap B \neq \emptyset \lor A \cap V^B \neq \emptyset
\end{align*}
\]

Property A.6 For $x, y \in \text{suff}(P)$ and $y \neq \varepsilon$ we have

\[
x <_p y \equiv x \leq_p \text{msp}(y).
\]

Proof

Let $x, y \in \text{suff}(P)$ and $y \neq \varepsilon$. We derive

\[
x <_p y \\
= \{ \text{definition of } <_p \text{ and } \text{pref}, x \in \text{suff}(P) \} \\
x \in \text{pref}(y) \setminus \{y\} \cap \text{suff}(P) \\
\Rightarrow \{ \text{pref}(y) \setminus \{y\} \cap \text{suff}(P) \text{ is nonempty } (y \neq \varepsilon), \text{finite and linearly ordered w.r.t. } \leq_p \} \\
x \leq_p \text{MAX}_{<_p} w : w \in \text{pref}(y) \setminus \{y\} \cap \text{suff}(P) : w \\
= \{ y \neq \varepsilon, \text{definition of msp} \} \\
x \leq_p \text{msp}(y) \\
\Rightarrow \{ y \neq \varepsilon, \text{msp}(y) <_p y \text{ (by definition of msp), transitivity of } <_p \} \\
x <_p y
\]

\[\square\]
B Algorithm and problem details

In this appendix we list the algorithm and problem details introduced in this paper with a short description.

P Examine prefixes of a given string in any order.

S Examine suffixes of a given string in any order.

+ Examine the strings from a given set in order of increasing length (this program detail can only be applied after, for instance, program details P and S).

RT Usage of the transition function of the reverse trie corresponding to the set of keywords to check whether a string, that is a suffix of some keyword, preceded by a symbol is again a suffix of some keyword.

SSD Allow any shift distance at least one that is safe, i.e. that does not cause the omission of any matches.

NLAU No lookahead at the symbols of the unscanned part of the input string when computing a safe shift distance.

OLAU One symbol lookahead at the unscanned part of the input string when computing a safe shift distance.

OPT When computing a safe shift distance use the recognized suffix and only the immediately preceding (mismatching) symbol, strictly coupled.

NLA When computing a safe shift distance do not look at the symbols preceding the recognized suffix.

BMCW When computing a safe shift distance on the one hand use the recognized suffix and the fact that the symbol preceding it is mismatching, and on the other hand, but strictly independent, the identity of that symbol.

BM Lessen the contribution of the symbol preceding the recognized suffix to the shift distance in case it does not occur in any keyword.

CW When computing a safe shift distance do not use the fact that the symbol preceding the recognized suffix is mismatching (use the recognized suffix and the symbol preceding it independently).

OKW The set of keywords contains only one keyword (in contrast to the preceding program details this is a problem detail).

References


Computing Science Reports

In this series appeared:

93/01 R. van Geldrop
93/02 T. Verhoeff
93/03 T. Verhoeff
93/04 B.H.L. Aants
J.H.M. Korst
P.J. Zwietering
93/05 J.C.M. Baeten
C. Verhoef
93/06 J.P. Veltkamp
93/07 P.D. Moerland
93/08 J. Verhoosel
93/09 K.M. van Hee
93/10 K.M. van Hee
93/11 K.M. van Hee
93/12 K.M. van Hee
93/13 K.M. van Hee
93/14 J.C.M. Baeten
J.A. Bergstra
93/15 J.C.M. Baeten
J.A. Bergstra
R.N. Bol
93/16 H. Schepers
J. Hooman
93/17 D. Alstein
P. van der Stok
93/18 C. Verhoef
93/19 G.J. Houben
93/20 F.S. de Boer
93/21 M. Codish
D. Danz
G. Fillé
M. Bruynooghe
93/22 E. Poll
93/23 E. de Kogel
93/24 E. Poll and Paula Severi
93/25 H. Schepers and R. Gerth
93/26 W.M.P. van der Aalst
93/27 T. Klok and D. Kratsch
93/28 F. Kamareddine and R. Nederpelt
93/29 R. Post and P. De Bra
93/30 J. Dezgun
T. Klok
D. Kratsch
H. Müller

Department of Mathematics and Computing Science
Eindhoven University of Technology

Deriving the Aho-Corasick algorithms: a case study into the synergy of programming methods, p. 36.

A continuous version of the Prisoner’s Dilemma, p. 17

Quicksort for linked lists, p. 8.

Deterministic and randomized local search, p. 78.

A congruence theorem for structured operational semantics with predicates, p. 18.

On the unavoidability of metastable behaviour, p. 29

Exercises in Multiprogramming, p. 97

A Formal Deterministic Scheduling Model for Hard Real-Time Executions in DEDOS, p. 32.


Systems Engineering: a Formal Approach Part II: Frameworks, p. 44.


A Trace-Based Compositional Proof Theory for Fault Tolerant Distributed Systems, p. 27.

Hard Real-Time Reliable Multicast in the DEDOS system, p. 19.

A congruence theorem for structured operational semantics with predicates and negative premises, p. 22.

The Design of an Online Help Facility for ExSpect, p. 21.


Freeness Analysis for Logic Programs - And Correctness, p. 24

A Typechecker for Bijective Pure Type Systems, p. 28.

Relational Algebra and Equational Proofs, p. 23.

Pure Type Systems with Definitions, p. 38.


Multi-dimensional Petri nets, p. 25.

Finding all minimal separators of a graph, p. 11.

A Semantics for a fine λ-calculus with de Bruijn indices, p. 49.

GOLD, a Graph Oriented Language for Databases, p. 42.

On Vertex Ranking for Permutation and Other Graphs, p. 11.

A general conservative extension theorem in process algebra, p. 17.

Temporal operators viewed as predicate transformers, p. 11.

Job Shop Scheduling by Constraint Satisfaction, p. 22.

A taxonomy of finite automata construction algorithms, p. 87.

A taxonomy of finite automata minimization algorithms, p. 23.

A precise clock synchronization protocol, p. 17.

A precise clock synchronization protocol, p. 17.

Design and Analysis of Dynamic Leader Election Protocols in Broadcast Networks, p. 73.

A precise clock synchronization protocol, p. 17.

An object-oriented paradigm, p. 28.

A precise clock synchronization protocol, p. 17.


Graph Isomorphism Models for Non Interleaving Process Algebra, p. 18.


Time and the Order of Abstract Events in Distributed Computations, p. 29.

94/10 T. Verhoef
94/11 J. Peleska
C. Huizing
C. Petersohn
94/12 T. Kloks
D. Kratsch
H. Müller
94/13 R. Seljée
94/14 W. Peremans
94/15 R.J.M. Vaessens
E.H.L. Aarts
J.R. Lenstra
94/16 R.C. Backhouse
H. Doornbos
94/17 S. Mauw
M.A. Reniers
94/18 F. Kamareddine
R. Nederpelt
94/19 B.W. Watson
94/20 R. Bloo
F. Kamareddine
R. Nederpelt
94/21 B.W. Watson
94/22 B.W. Watson
94/23 S. Mauw and M.A. Reniers
94/24 D. Dams
O. Grumberg
R. Gerth
94/25 T. Kloks
94/26 R.R. Hoogerwoord
94/27 S. Mauw and H. Mulder
94/28 C.W.A.M. van Overveld
M. Verhoefen
94/29 J. Hooman
94/30 J.C.M. Baeten
J.A. Bergstra
Gh. Ştefănescu
94/31 B.W. Watson
R.E. Watson
94/32 JJ. Vereijken
94/33 T. Laan
94/34 R. Bloo
F. Kamareddine
R. Nederpelt
94/35 J.C.M. Baeten
S. Mauw
94/36 F. Kamareddine
R. Nederpelt
94/37 T. Basten
R. Bol
M. Voorhoeve
94/38 A. Bijlenga
C.S. Scholten

The testing Paradigm Applied to Network Structure. p. 31.
A New Method for Integrity Constraint checking in Deductive Databases. p. 34.
Ups and Downs of Type Theory. p. 9.
Job Shop Scheduling by Local Search. p. 21.
Mathematical Induction Made Calculational, p. 36.
An Algebraic Semantics of Basic Message Sequence Charts, p. 9.
Refining Reduction in the Lambda Calculus. p. 15.
The performance of single-keyword and multiple-keyword pattern matching algorithms. p. 46.
Beyond β-Reduction in Church's λ→, p. 22.
The design and implementation of the FIRE engine: A C++ toolkit for Finite automata and regular Expressions.
An algebraic semantics of Message Sequence Charts, p. 43.
Abstract Interpretation of Reactive Systems: Abstractions Preserving VCTL*, ECTL* and CTL*, p. 28.
Kωβ-free and Wω-free graphs, p. 10.
On the foundations of functional programming: a programmer's point of view, p. 54.
Correctness of Real Time Systems by Construction, p. 22.
Process Algebra with Feedback, p. 22.
A Boyer-Moore type algorithm for regular expression pattern matching. p. 22.
A formalization of the Ramified Type Theory, p.40.
The Barendregt Cube with Definitions and Generalised Reduction, p. 37.
Delayed choice: an operator for joining Message Sequence Charts, p. 15.
Canonical typing and Π-conversion in the Barendregt Cube, p. 19.
Simulating and Analyzing Railway Interlockings in ExSpect, p. 30.
Point-free substitution, p. 10.
<table>
<thead>
<tr>
<th>No.</th>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>94/39</td>
<td>A. Blokhuis T. Kloks</td>
<td>On the equivalence covering number of splitgraphs, p. 4.</td>
</tr>
<tr>
<td>94/40</td>
<td>D. Alstein</td>
<td>Distributed Consensus and Hard Real-Time Systems, p. 34.</td>
</tr>
<tr>
<td>94/41</td>
<td>T. Kloks D. Kratsch</td>
<td>Computing a perfect edge without vertex elimination ordering of a chordal bipartite graph, p. 6.</td>
</tr>
<tr>
<td>94/43</td>
<td>R.C. Backhouse M. Bijsterveld</td>
<td>Category Theory as Coherently Constructive Lattice Theory: An Illustration, p. 35.</td>
</tr>
<tr>
<td>94/45</td>
<td>G.J. Hoopen</td>
<td>Tutorial voor de ExSpect-bibliotheek voor &quot;Administratieve Logistiek&quot;, p. 43.</td>
</tr>
<tr>
<td>94/46</td>
<td>R. Bloo F. Kamareddine R. Nederpelt</td>
<td>The L-cube with classes of terms modulo conversion, p. 16.</td>
</tr>
<tr>
<td>94/47</td>
<td>R. Bloo F. Kamareddine R. Nederpelt</td>
<td>On (\Pi)-conversion in Type Theory, p. 12.</td>
</tr>
<tr>
<td>94/48</td>
<td>Mathematics of Program Construction Group</td>
<td>Fixed-Point Calculus, p. 11.</td>
</tr>
<tr>
<td>94/50</td>
<td>H. Geuvers</td>
<td>A short and flexible proof of Strong Normalization for the Calculus of Constructions, p. 27.</td>
</tr>
<tr>
<td>94/51</td>
<td>T. Kloks D. Kratsch H. Müller</td>
<td>Listing simplicial vertices and recognizing diamond-free graphs, p. 4.</td>
</tr>
<tr>
<td>94/52</td>
<td>W. Penczek R. Kuiper</td>
<td>Traces and Logic, p. 81</td>
</tr>
<tr>
<td>95/01</td>
<td>J.J. Lukkien</td>
<td>The Construction of a small Communication Library, p. 16.</td>
</tr>
<tr>
<td>95/02</td>
<td>M. Bezem R. Bol J.F. Groote</td>
<td>Formalizing Process Algebraic Verifications in the Calculus of Constructions, p. 49.</td>
</tr>
<tr>
<td>95/03</td>
<td>J.C.M. Baeten C. Verhoef</td>
<td>Concrete process algebra, p. 134.</td>
</tr>
<tr>
<td>95/04</td>
<td>J. Hidders</td>
<td>An Isotopic Invariant for Planar Drawings of Connected Planar Graphs, p. 9.</td>
</tr>
<tr>
<td>95/05</td>
<td>P. Severi</td>
<td>A Type Inference Algorithm for Pure Type Systems, p. 20.</td>
</tr>
<tr>
<td>95/07</td>
<td>G.A.M. de Bruyn G.S. van Rooijen</td>
<td>Drawing Execution Graphs by Parsing, p. 10.</td>
</tr>
<tr>
<td>95/08</td>
<td>R. Bloo</td>
<td>Preservation of Strong Normalization for Explicit Substitution, p. 12.</td>
</tr>
<tr>
<td>95/09</td>
<td>J.C.M. Baeten J.A. Bergstra</td>
<td>Discrete Time Process Algebra, p. 20</td>
</tr>
<tr>
<td>95/10</td>
<td>R.C. Backhouse R. Verhoeven O. Weber</td>
<td>Mathpad: A System for On-Line Preparation of Mathematical Documents, p. 15</td>
</tr>
<tr>
<td>Year</td>
<td>Author(s)</td>
<td>Title</td>
</tr>
<tr>
<td>------</td>
<td>----------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>95/11</td>
<td>R. Seljee</td>
<td>Deductive Database Systems and integrity constraint checking, p. 36.</td>
</tr>
<tr>
<td>95/12</td>
<td>S. Mauw and M. Reniers</td>
<td>Empty Interworkings and Refinement Semantics of Interworkings Revised, p. 19.</td>
</tr>
</tbody>
</table>