UNIVERSITY OF JOENSUU

DEPARTMENT OF COMPUTER SCIENCE

Report Series A

Efficient parameterized string matching

Kimmo Fredriksson and Maxim Mozgovoy

A-2006-2

ACM F.2.2, H.3.3

ISBN 952-458-807-2

ISSN 0789-7316

Sublinear parameterized single and multiple string matching

Kimmo Fredriksson* and Maxim Mozgovoy
Department of Computer Science
University of Joensuu
{kfredrik,mmozgo}@cs.joensuu.fi

Abstract

We consider the following pattern matching problem. We have *pattern* $P[0 \ldots m-1]$ and *text* is $T[0 \ldots n-1]$, where the symbols of P and T are taken from two disjoint finite alphabets Σ of size σ and Λ of size λ. The pattern Pmatches the text substring $T[j \ldots j+m-1]$, iff for all $i \in \{0 \ldots m-1\}$ it holds that $M_j(P[i]) = T[j+i]$, where $M_j(\cdot)$ is one-to-one mapping on $\Sigma \cup \Lambda$ such that the mapping is identity on Σ, but on Λ can be different for each text position j. We give efficient algorithms that find all parameterized occurrences of P in T. The algorithms are based on generalizing Shift-Or and Backward DAWG Matching (BDM) algorithms. The latter can be used for searching r patterns simultaneously. The Shift-Or based algorithm runs in $O(n\lceil m/w \rceil)$ worst case time, while the average case for fixed alphabets and under some mild and realistic assumptions is $O(n \log_\sigma(m)/w)$, where w is the number of bits in computer word. The BDM based algorithm runs in $O(n \log_\sigma(rm)/m)$ average time. This is optimal within a constant factor. For general alphabets the times increase by a factor $O(\log(m))$.

Keywords: algorithms, parameterized string matching, bit-parallelism, suffix automaton

ACM Classification: F.2.2 [Analysis of algorithms and problem complexity]: Non-numerical algorithms and problems — *Pattern matching, Computations on discrete structures*; H.3.3 [Information storage and retrieval]: Information Search and Retrieval — *Search process*.

1 Introduction

In traditional string matching problem one is interested in finding the occurrences of a pattern P from a text T, where P and T are strings over some alphabet Σ. Many variations of this basic problem setting exist, such as searching multiple patterns simultaneously, and/or allowing some limited number of errors in the matches, and indexed searching, where T can be preprocessed to allow efficient queries of P. See e.g. [9, 11, 6] for an overview and references. Yet another variation is *parameterized matching* [4]. In this variant we have two disjoint alphabets, Σ for *fixed* symbols, and Λ for *parameter* symbols. In this setting we search *parameterized* occurrences of P, where the symbols from Σ must match exactly, while the symbols in Λ can be also renamed. This problem has important applications e.g. in software maintenance and plagiarism detection [4], where the symbols of the strings can be e.g. reserved words and identifier or parameter names of some (possibly tokenized) programming language source code. Hence one might be interested in finding code snippets that are the same up to some systematical variable renaming.

A myriad of algorithms have been developed for the classical problem, but only a few exist for parameterized matching. Linear exact on-line matching matching algorithms have been developed for single [3, 1] and multiple patterns [10]. Sublinear algorithm on average for single pattern was developed in [3]. Other algorithms exist for the off-line problem [4, 5]. In this paper we develop algorithms that run in sublinear time on average, are simple to implement and

*Supported by the Academy of Finland, grant 202281.

perform well in practice. Our algorithms are based on generalizing the well known Shift-Or [2] and Backward DAWG Matching algorithms [7]. Our algorithms generalize for the multipattern matching as well.

2 Preliminaries

We use the following notation. The *pattern* is $P[0 \ldots m-1]$ and the *text* is $T[0 \ldots n-1]$. The symbols of P and T are taken from two disjoint finite alphabets Σ of size σ and Λ of size λ. The pattern Pmatches the text substring $T[j \ldots j+m-1]$, iff for all $i \in \{0 \ldots m-1\}$ it holds that $M_j(P[i]) = T[j+i]$, where $M_j(\cdot)$ is one-to-one mapping on $\Sigma \cup \Lambda$. Moreover, the mapping must be identity on Σ, but on Λ can be different for each text position j. For example, assume that $\Sigma = \{\text{A,B}\}$, $\Lambda = \{\text{X,Y,Z}\}$ and $P = \text{AAZYZABXYZAX}$. Then P matches the text substring AAZYZABXYZAX with identity mapping, and AAXYXABZYXAZ with parameter mapping X $\mapsto$ Z, Y $\mapsto$ Y, and Z $\mapsto$ X. This mapping is simple with *prev* encoding [4]. For a string S, *prev*(S) maps all parameter symbols s in S to a non-negative integer p, where p is the number of symbols since the last occurrence of symbol s in S. The first occurrence of the parameter is encoded as 0. If s belongs to Σ, it is mapped to itself (s). For our example pattern, *prev*(P) = AA002AB055A4. This is the same as the encoding for the two example substrings, i.e. *prev*(AAZYZABXYZAX) = *prev*(AAXYXABZYXAZ). Hence the problem is reduced to exact string matching, where we match *prev*(P) against *prev*($T[j \ldots j+m-1]$) for all $j = 0 \ldots n-m$. The string *prev*(S) can be easily computed in linear time for constant size alphabets. The only remaining problem then is how to maintain *prev*($T[j \ldots j+m-1]$) (and any algorithmic parameters that depend on it) efficiently as j increases. The key is the following Lemma [4].

Lemma 1 . *Let $S' = prev(S)$. Then for $S'' = prev(S[j \ldots j+m-1])$ for all i such that $S[i] \in \Lambda$ it holds that $S''[i] = S'[i]$ iff $S'[i] < m$. Otherwise $S''[i] = 0$.*

We are now ready to present our algorithms. For simplicity we assume that Σ and Λ are finite constant size alphabets. For large alphabets all our time bounds hold if we multiply them by $O(\log(m))$. Moreover, without loss of generality, we assume the symbols to be atomic, i.e. they can be accessed in constant time.

3 Parameterized bit-parallel matching

In this section we present bit-parallel approach for parameterized matching, based in Shift-Or algorithm [2]. For the bit-parallel operations we adopt the following notation. A machine word has w bits, numbered from the least significant bit to the most significant bit. We use C–like notation for the bit-wise operations of words; & is bit-wise `and`, | is `or`, ^ is `xor`, $\sim$ negates all bits, << is shift to left, and >> shift to right, both with zero padding. For brevity, we make the assumption that $m \leq w$, unless explicitly stated otherwise. We first review the standard Shift-Or algorithm, and then show how it can be adapted to parameterized matching, and finally improve its average case running time.

3.1 Standard Shift-Or

The standard Shift-Or automaton is constructed as follows. The automaton has states $0, 1, \ldots, m$. The state 0 is the initial state, state m is the final (accepting) state, and for $i = 0, \ldots, m-1$ there is a transition from the state i to the state $i+1$ for character $P[i]$. In addition, there is a transition for every $c \in \Sigma$ from and to the initial state, which makes the automaton nondeterministic. The preprocessing algorithm builds a table B, having one bit-mask entry for each $c \in \Sigma$. For $0 \leq i \leq m-1$, the mask $B[c]$ has ith bit set to 0, iff $P[i] = c$. These correspond to the transitions of the implicit automaton. That is, if the bit i in $B[c]$ is 0, then there is a

Alg. 1 Shift-Or(T, n, P, m).

```
1   for i ← 0 to σ − 1 do B[i] ← ∼0 >> (w − m)
2   for i ← 0 to m − 1 do B[P[i]] ← B[P[i]] & ∼(1 << i)
3   D ← ∼0; mm ← 1 << (m − 1)
4   for i ← 0 to n − 1 do
5       D ← (D << 1) | B[T[i]]
6       if (D & mm) ≠ mm then report match
```

Alg. 2 Encode(P, m).

```
1   for i ← 0 to m − 1 do if P[i] ∈ Λ then prv[P[i]] ← −1
2   for i ← 0 to m − 1 do
3       if P[i] ∈ Λ then
4           if prv[P[i]] = −1 then P'[i] ← σ else P'[i] ← i − prv[P[i]] + σ
5           prv[P[i]] ← i
6       else
7           P'[i] ← P[i]
8   return P'
```

transition from the state i to the state $i + 1$ with character c. The bit-vector D encodes the states of the automaton. The ith bit of the state vector is set to 0, iff the state i is active, i.e. the pattern prefix $P[0 \dots i]$ matches the current text position. Initially each bit is set to 1. For each text symbol c the vector is updated by $D \leftarrow (D << 1) \mid B[c]$. This simulates all the possible transitions of the nondeterministic automaton in a single step. If after the update the mth bit of d is zero, then there is an occurrence of P. Alg. 1 gives the code. If $m \leq w$, then the algorithm runs in time $O(n)$.

3.2 Parameterized Shift-Or

In order to generalize Shift-Or for parameterized matching, we must take care of three things: (i) Pmust be encoded with *prev*; (ii) *prev*$(T[j \dots j + m - 1])$ must be maintained in $O(1)$time per text position; (iii) the table Bmust be built so that all parameterized pattern prefixes can be searched in parallel. The items (i) and (ii) are trivial, while (iii) is a bit more tricky. To compute *prev*(P) we just maintain an array *prv*$[c]$ that for each symbol $c \in \Lambda$ stores the position of its last occurrence. Then *prev*(P) can be computed in $O(m)$ time by a linear scan over P. To simplify indexing in the array B, we assume that $\Sigma = \{0 \dots \sigma - 1\}$, and map the *prev* encoded parameter offsets into the range $\{\sigma \dots \sigma + m - 1\}$. Alg. 2 gives the pseudo code. The text is encoded in the same way, but the encoding is embedded into the search code, see Alg. 3. The only difference is that we apply Lemma 1 to reset offsets that are greater than $m - 1$ (i.e. offsets that are for parameters that are outside of the current text window) to zero. Otherwise the search algorithms is exactly the same as for normal Shift-Or.

The tricky part is the preprocessing phase. We denote the *prev* encoded pattern as P'. At first P' is preprocessed just as P in the normal Shift-Or algorithm. This includes the parameter offsets, which are handled as any other symbol. However, this is not enough. We illustrate the problem by an example. Let $P =$ XAXAX and $T =$ ZZAZAZAZ. In encoded forms these are $P' =$ 0A2A2 and $T' =$ 01A2A2A2. Clearly P has two (overlapping) parameterized matches in T. However, P' does not match in T' at all.

The problem is that as the algorithm searches all the m prefixes of the pattern in parallel, then some non-zero encoded offset p (of some text symbol) should be interpreted as zero in some cases. These prefixes have lengths from 1 to m, and in able to successfully apply Lemma 1 we should be able to apply it in parallel to all m substrings. In other words, any non-zero parameter offset p must be treated as zero for all pattern prefixes whose length h is less than p, since by Lemma 1 the parameter with offset p is dropped out of the window of length h.

This problem can be solved as follows. The bit-vector $B[\sigma + i]$ is the match vector for offset i. If the j bit of this vector is zero, it means by definition that $P'[j] = i$. If any of the i least significant bits of $B[\sigma]$ are zero, we clear the corresponding bits of $B[\sigma + i]$ as well. More

Alg. 3 P-Shift-Or(T, n, P, m).

```
1   P' ← Encode(P, m)
2   for i ← 0 to σ + m − 1 do B[i] ← ∼0 >> (w − m)
3   for i ← 0 to λ − 1 do prv[σ + i] ← −∞
4   for i ← 0 to m − 1 do B[P'[i]] ← B[P'[i]] & ∼(1 << i)
5   for i ← 1 to m − 1 do B[σ + i] ← B[σ + i] & (B[σ] | (∼0 << i))
6   D ← ∼0; mm ← 1 << (m − 1)
7   for i ← 0 to n − 1 do
8       c ← T[i]
9       if c ∈ Λ then
10          c ← i − prv[T[i]] + σ
11          if c > σ + m − 1 then c ← σ
12          prv[T[i]] ← i
13      D ← (D << 1) | B[c]
14      if (D & mm) ≠ mm then report match
```

precisely, we set

$$B[\sigma + i] \leftarrow B[\sigma + i] \ \&\ (B[\sigma] \mid (\sim 0 << i)).$$

This means that the offset i is treated as offset i for prefixes whose length is greater than i, and as zero for the shorter prefixes, satifying the condition of Lemma 1.

Alg. 3 gives the complete code. The algorithm clearly runs in $O(n\lceil m/w \rceil)$ worst case time. For long patterns one can search just a length w prefix of the pattern, and verify with the whole pattern whenever the prefix matches, giving $O(n)$ average time. However, note that a long variable name (string) is just one symbol (token) in typical applications, hence w bits is usually plenty. Finally, note that for unbounded alphabets we cannot use arrays for *prv* and B. We can use balanced trees instead, but then the time bounds must be multiplied by $O(\log(m))$.

3.3 Skipping text symbols

Standard Shift-Or can be improved to run in optimal $O(n \log_\sigma(m)/m)$ average time [8]. The algorithm takes a parameter q, and from the original pattern generates a set $\mathcal{P}$ of q new patterns $\mathcal{P} = \{P^0, \ldots, P^{q-1}\}$, each of length $m' = \lfloor m/q \rfloor$, where $P^j[i] = P[j + iq]$ for $i = 0 \ldots \lfloor m/q \rfloor - 1$. In other words, the algorithm generates q different alignments of the original pattern P, each alignment containing only every qth character. The total length of the patterns in $\mathcal{P}$ is $q\lfloor m/q \rfloor \leq m$. For example, if $P =$ ABCDEF and $q = 3$, then $P^0 =$ AD, $P^1 =$ BE and $P^2 =$ CF. Assume now that P occurs at $T[i..i + m - 1]$. From the definition of P^j it directly follows that $P^j[h] = T[i + j + hq]$, where $j = i \mod q$ and $h = 0 \ldots m' - 1$. This means that we can use the set $\mathcal{P}$ as a filter for the pattern P, and that the filter needs only to scan every qth character of T. All the patterns must be searched simultaneously. Whenever an occurrence of P^j is found in the text, we must verify if P also occurs, with the corresponding alignment.

This method clearly works for parameterized matching as well. We generate the set of patterns $\mathcal{P}$, and also *prev*-encode them. For example for $P =$ AAZYZABXYZAX and $q = 3$ we process the pattern set $prev(\{$AYBZ,AZXA,ZAYX$\}) = \{$A0B0,A00A,0A00$\}$. In the search phase the text is also encoded on-line, encoding only every qth symbol, but assuming that they are consecutive. In other words, every parameter offset is effectively divided by q to agree with the encoding of the patterns. Finally, the verification phase checks if $prev(P) = prev(T[v \ldots v + m - 1]$, where v is the starting position of a potential match.

The search of the pattern set can be done using the parameterized Shift-Or algorithm. This is possible by concatenating and packing the set of patterns into a single machine word [8, 2]. Another alternative is to use the parameterized version [10] of Aho-Corasic algorithm. Both lead to the same average case running time, but the latter does not require that $m \leq w$, as it is not based on bit-parallelism. We denote the Shift-Or based algorithm as PFSO. Alg. 4 shows the pseudo code for the filtering phase, and Alg. 5 the verification code.

The filtering time is $O(n/q)$. The filter searches the exact matches of q patterns, each of length $\lfloor m/q \rfloor$. We are not able to analyze the exact effect of the parameter alphabet to the

Alg. 4 Average-Optimal-P-Shift-Or(T, n, P, m, q).

```
1   P' ← Encode(P, m)
2   h ← 0; mm ← 0
3   for j ← 0 to q − 1 do
4       for i ← 0 to ⌊m/q⌋ − 1 do
5           P[j][i] ← P[iq + j]
6       h ← h + ⌊m/q⌋
7       mm ← mm | (1 << (h − 1))
8       P[j] ← Encode( P[j], ⌊m/q⌋)
9   for i ← 0 to σ + ⌊m/q⌋q − 1 do B[i] ← ∼0 >> (w − ⌊m/q⌋q)
10  for j ← 0 to q − 1 do
11      for i ← 0 to ⌊m/q⌋ − 1 do
12          B[P[j][i]] ← B[P[j][i]] & ∼(1 << (j⌊m/q⌋ + i))
13  for j ← 0 to q − 1 do
14      msk ← ∼0
15      for i ← 1 to ⌊m/q⌋ − 1 do
16          msk ← msk ^ (1 << (j⌊m/q⌋ + i − 1))
17          B[σ + i] ← B[σ + i] & (B[σ] | msk)
18  for i ← 0 to λ do prv[σ + i] ← −∞
19  D ← ∼0; i ← 0
20  while i < n do
21      c ← T[i]
22      if c ∈ Λ then
23          c ← i/q − prv[T[i]] + σ
24          if c > σ + ⌊m/q⌋ − 1 then c ← σ
25          prv[T[i]] ← i/q
26      D ← ((D & ∼mm) << 1) | B[c]
27      if (D & mm) ≠ mm then Verify(T, i, n, P', m, q, D, mm)
28      i ← i + q
```

Alg. 5 Verify(T, i, n, P, m, q, D, mm).

```
1   D ← (D & mm) ^ mm
2   while D ≠ 0 do
3       s ← ⌊log2(D)⌋
4       c ← −(⌊m/q⌋ − 1) q − ⌊s/⌊m/q⌋⌋
5       if P = Encode(T[i + c . . . i + c + m − 1], m) then report match
6       D ← D & ∼(1 << s)
```

probability that two randomly picked symbols match. However, if we assume that a constant fraction ε of the pattern positions are randomly selected to have a randomly selected symbol from Σ, then the probability that P^j occurs in a given text position is $O((1/\sigma)^{\lfloor \varepsilon m/q \rfloor})$. A brute force verification cost is in the worst case $O(m)$ (but only $O(1)$ on average). To keep the total time at most $O(n/q)$ on average, we select q so that $n/q = mn/\sigma^{\varepsilon m/q}$, i.e. $q = O(m/\log_\sigma(m))$. The total average time is therefore $O(n \log_\sigma(m)/m)$. This is optimal [12] within a constant factor.

4 Parameterized backward trie matching

We now present an algorithm based on Backward DAWG Matching (BDM) [7]. BDM is optimal on average, i.e. it runs in $O(n \log_\sigma(m)/m)$ average time. We call our parameterized version of BDM as Parameterized Backward Trie Matching, PBTM for short. In the preprocessing phase PBTM builds a trie for the encoded suffixes of the reversed pattern. Trie is a rooted tree, where each edge is labeled by a symbol. The edges of the path from the root node to some leaf node then spells out the string of symbols stored into that leaf. The pattern in reverse is denoted by P^r. The set of its suffixes is $\{P^r[i \ldots m-1] \mid 0 \leq i < m\}$ (note that this corresponds to the prefixes of the original pattern). Each suffix is then encoded with *prev*, and the encoded strings are inserted into a trie. For example, if $P = \text{AZBZXBXY}$, then the set of stored strings is $\{00b20b2a, 0b20b2a, b00b2a, 00b2a, 0b2a, b0a, 0a, a\}$, The trie allows efficient searching of any pattern substring that occurs in P^r. A brute force algorithm for this takes $O(m^2)$ time, but can be improved to $O(m)$ by using efficient suffix tree construction algorithms for parameterized

Alg. 6 PBTM(T, n, P, m).

```
1   root ← EncSTrie(P^r)
2   for i ← 0 to λ − 1 do prv[σ + i] ← −∞
3   i ← 0
4   while i < n − m do
5       j ← m; shift ← m; u ← root
6       while u ≠ null do
7           c ← T[i + j − 1]
8           if c ∈ Λ then
9               c ← m − j − prv[T[i + j − 1]] + σ
10              if c > σ + m − 1 then c ← σ
11              prv[T[i + j − 1]] ← m − j
12          j ← j − 1
13          u ← child(u, c)
14          if u ≠ null AND issuffix(u) then
15              if j > 0 then shift ← j else report match
16      for k ← i + j to i + m − 1 do if T[k] ∈ Λ then prv[T[k]] ← −∞
17      i ← i + shift
```

strings [5]. An alternative to the trie is suffix array, i.e. the trie can be replaced with sorted array of *prev* encoded suffixes of the reverse pattern. Following an edge in the trie can then be simulated by a binary search in the array. We call the resulting algorithm PBAM. The benefit is that the array based method is easy to implement space efficiently since only one pointer is needed for each suffix.

We now show how this can be used for efficient search. Assume that we are scanning the text window $T[i \ldots i + m - 1]$ *backwards*. The invariant is that all occurrences that start before the position i are already reported. The text window is *prev*-encoded (backwards as well) as we go, and the read substring of this window is matched against the trie. This is continued as long as the substring can be extended without a mismatch, or we reach the beginning of the window. If the whole window can be matched against the trie, then the pattern occurs in that window. If the pattern does not match, some of the occurrences may still overlap with the current window. However, in this case one of the suffixes stored into the trie must match, since the reverse suffixes are also the prefixes of the original pattern. The algorithm remembers the longest such suffix, that is not the whole pattern, found from the window. The window is then shifted so that its starting position will become aligned with the last symbol of that suffix. This is the position of the next possible pattern occurrence. If the length of that longest suffix was ℓ, the next window to be searched is $T[i + m - \ell \ldots i + m - 1 + m - \ell]$. This process is repeated until the whole text is scanned.

Some care must be taken to be able to do the encoding of the text window in $O(1)$ time per read symbol. To achieve constant time per symbol we must use an auxiliary array *prv* (as before) to store the position of the last occurrence for each symbol. We cannot afford to initialize the whole array for each window, so before shifting the window we rescan the symbols just read in the current window, and reinitialize the array only for those symbols. This ensures $O(1)$ total time for each symbol read. Alg. 6 gives the code.

The average case running time of this algorithm depends on how many symbols x are examined in each window. Again, if we make the simplifying assumption that a constant fraction of the pattern positions are randomly selected to have a randomly selected symbol from Σ, then the original analysis of BDM holds for PBTM as well, and the average case running time is $O(n \log_\sigma(m)/m)$. For general alphabets and for the PBAM version the time must be multiplied by $O(\log(m))$. Finally, this algorithm can be easily modified to search r patterns simultaneously. Basically, if all the patterns are of the same length, this generalization requires just storing all the suffixes of all the patterns into the same trie. This results in $O(n \log_\sigma(rm)/m)$ average time. With modest additional complexity patterns of different lengths can be handled as well in the same way as with regular BDM [6].

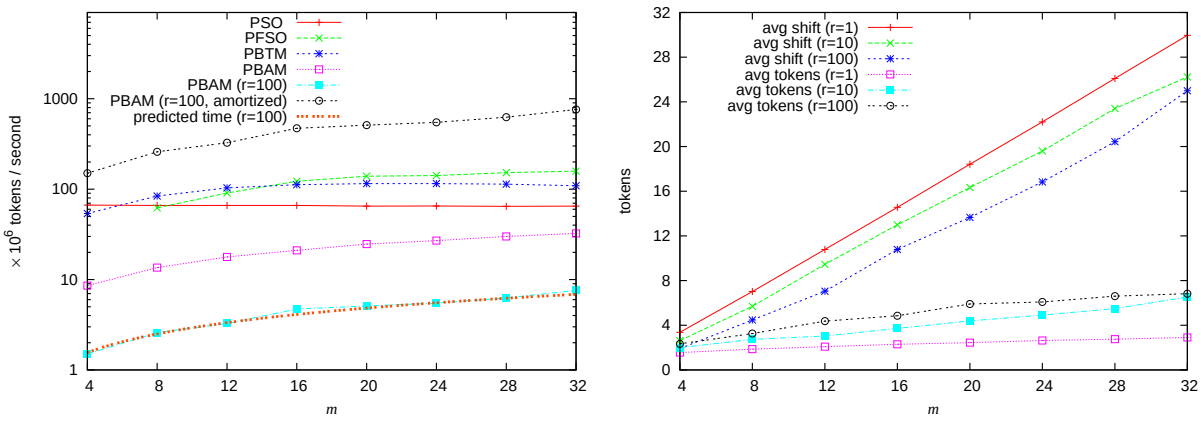

Figure 1: Left: the search speed in 10^6 tokens / second. Right: the average shift and average number of tokens inspected in each window of length m.

5 Experimental results

We have implemented the algorithms in C++, and compiled them with Borland C++Builder 6. We performed the experiments on the AMD Sempron 2600+ (1.88 GHz) machine with 768 MB RAM, running Windows XP. A tokenized string of concatenated Java source files (taken from various open source projects, such as jPOS, smppapi, and TM4J) was used as a text to be searched. The tokenization procedure (based on JavaCC[1] parser) converted an input file into a sequence of two-byte codes, representing single characters, reserved Java words and distinct identifiers. The initial string had a size of 5.48MB, and after encoding it consisted of 1259799 tokens, including 51 reserved Java words and 10213 unique identifiers. A set of 100 patterns for each length reported was randomly extracted from the input text. We report the average number of tokens searched per second for each algorithm.

Fig. 1 summarizes the results. PSO denotes the basic parameterized shift-or algorithm, PFSO the fast parameterized shift-or, PBTM the parameterized backward trie matching algorithm, and PBAM the suffix array version of PBTM. For short patterns plain PSO and PBTM give the best results. For longer patterns PFSO wins in case of optimal q selection. For $m \in \{8, 12, 16, 20, 24, 28, 32\}$ we used $q = \{2, 3, 4, 4, 4, 5, 6\}$, respectively. For short patterns PBTM is faster than PFSO. For long patterns PBTM suffers from the large alphabet size. In our implementation we used arrays to implement the trie nodes and for long patterns the trie requires a lot of inititalization time and memory, not fitting into the CPU cache. PBAM does not have this flaw, but the binary search step needed for each accessed text symbol makes it comparatively slow.

We also experimented with the multipattern version of PBAM, searching $r = 100$ patterns simultaneously. The plot shows that while the raw speed is reduced, the amortized speed per pattern is clearly better than for any of the single pattern matching algorithms. The time also coincides nicely with the the theoretical curve $O(n \log_\sigma(rm) \log_2(rm)/m)$, supporting our analysis. This is also clear given the right plot, showing the average number of tokens inspected in each text window, and the average shift for $r = 1, 10, 100$. These behave like in random texts supporting our assumptions in the analysis.

6 Conclusions

We have shown how two well-known algorithms, namely Shift-Or and BDM, can be generalized for parameterized matching. The algorithms are easy to implement, and work well in practice.

[1] `https://javacc.dev.java.net/`

We have concentrated on obtaining fast average case times. However, the worst case times can be improved to $O(n)$ using known results [7, 8] and standard tricks.

References

[1] A. Amir, M. Farach, and S. Muthukrishnan. Alphabet dependence in parameterized matching. *Inf. Process. Lett.*, 49(3):111–115, 1994.

[2] R. A. Baeza-Yates and G. H. Gonnet. A new approach to text searching. *Commun. ACM*, 35(10):74–82, 1992.

[3] B. S. Baker. Parameterized pattern matching by Boyer-Moore-type algorithms. In *Proceedings of ACM-SODA'95*, pages 541–550, 1995.

[4] B. S. Baker. Parameterized duplication in strings: algorithms and an application to software maintenance. *SIAM J. Comput.*, 26(5):1343–1362, 1997.

[5] R. Cole and R. Hariharan. Faster suffix tree construction with missing suffix links. In *Proceedings of ACM-STOC'00*, pages 407–415, Portland, Oregon, 2000.

[6] M. Crochemore and W. Rytter. *Text algorithms*. Oxford University Press, 1994.

[7] M. Crochemore *et al.* Speeding up two string matching algorithms. *Algorithmica*, 12(4/5):247–267, 1994.

[8] K. Fredriksson and Sz. Grabowski. Practical and optimal string matching. In *Proceedings of SPIRE'2005*, LNCS 3772, pages 374–385. Springer–Verlag, 2005.

[9] D. Gusfield. *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge University Press, Cambridge, 1997.

[10] R. M. Idury and A. A. Schäffer. Multiple matching of parameterized patterns. *Theor. Comput. Sci.*, 154(2):203–224, 1996.

[11] G. Navarro and M. Raffinot. *Flexible Pattern Matching in Strings*. Cambridge University Press, 2002.

[12] A. C. Yao. The complexity of pattern matching for a random string. *SIAM J. Comput.*, 8(3):368–387, 1979.