The String Matching Algorithms Research Tool

Simone Faro¹, Thierry Lecroq², Stefano Borzı¹,
Simone Di Mauro¹, and Alessandro Maggio¹

¹ Università di Catania, Viale A.Doria n.6, 95125 Catania, Italy
² Université de Rouen, LITIS EA 4108, 76821 Mont-Saint-Aignan Cedex, France
faro@dmi.unict.it

Abstract. String matching is the problem of finding all occurrences of a given pattern in a given text. It is an extensively studied problem in computer science because of its direct application to several areas such as text, image and signal processing, speech analysis and recognition, data compression, information retrieval, computational biology and chemistry. Since 1970 more than 85 string matching algorithms have been proposed, and more than 50% of them in the last ten years. In this paper we present SMART, an efficient and flexible tool designed for developing, testing, comparing and evaluating string matching algorithms. It also provides the most comprehensive survey of online exact single string matching algorithms together with a set of corpora available for testing purposes.

Keywords: string matching, text processing, design and analysis of algorithms, testing framework, algorithms survey, experimental evaluation

1 Introduction

String matching is a very important subject in the wider domain of text processing. It consists in finding all occurrences of a given pattern in a given text. It is an extensively studied problem in computer science, mainly due to its direct applications in many areas related with information retrieval and information analysis. String matching algorithms are also basic components used in implementations of practical software existing under most operating systems. Moreover, they emphasize programming methods that serve as paradigms in other fields of computer science. Finally they also play an important role in theoretical computer science by providing challenging problems.

Applications require two kinds of solutions depending on which string, the pattern or the text, is given first. Algorithms based on the use of automata or combinatorial properties of strings are commonly implemented to preprocess the pattern and solve the first kind of problem. In this paper we are interested in this kind of problem, which is generally referred as online string matching. Recently Faro and Lecroq presented a comprehensive survey [17] of almost all online string matching algorithms appeared in literature up to 2010.

In 1991 Hume and Sunday presented an efficient framework [20] for testing string matching algorithms. It was developed in the C programming language and has been extensively used in the field during the last few decades. The authors compiled their framework using the stringsearch package[1] including the implementations of 37 string matching algorithms. Although their tool is very useful and simple to be used it presents some questionable points.

1 http://hackage.haskell.org/package/stringsearch-0.3.6.4
First of all, it was designed in order to maintain the preprocessing and the searching phase as separate functions. Although this allows an accurate measurement of the preprocessing time, it needs also the pattern to be copied to some local buffer, affecting the time measurement with an additional overhead which becomes negligible only if the length of the text is large enough. In addition all data related with the pattern are stored in common structure, thus guaranteeing a similar behavior of the algorithms. However it may slightly increase the access time to pattern information during the searching phase.

A useful feature of the tool from Hume and Sunday is that all algorithms are implemented in separate files. However their compilation is not independent from the whole framework. Thus when testing several algorithms there is always a risk that they meddle with each other.

We also noticed that the use of the `signed char` type does not always work properly, because indexing of arrays does not work with negative values. This could be avoided by forcing the cast to the `unsigned char` type all values which are used as indexes.

In this paper we introduce a new version of SMART (String Matching Algorithms Research Tool), an efficient and flexible framework designed for developing, testing, comparing and evaluating string matching algorithms. It includes most of the features which characterize the tool by Hume and Sunday, but includes a lot of other useful and interesting improvements. It allows the user to completely customize the testing environment, adding new algorithms and testing them for correctness, without the need of recompiling it. Moreover it includes the implementation of more than 120 algorithms, divided into more than 300 variants, and a large set of corpora, divided into categories, which can be used inside the testing environment. Finally it has been designed in order to provide a fair comparison between different solutions, thanks also to a large variety of experimental observations.

In May 2010 a preliminary version of SMART was released to the scientific community and referred in a technical report [13] were the authors discussed a comprehensive evaluation of almost all exact string matching algorithms known up to 2010. In the last six years it was used many times to perform experimental evaluation (see for instance [18, 2, 23, 15, 10]) and to test the performances of new algorithms.

One of the most interesting features of the new version of SMART is a practical and useful graphical interface which works over the standard framework and allows the use of all functionalities of the tool.

The source file of SMART, together with a detailed description and documentation, can be found at [http://www.dmi.unict.it/~faro/smart/](http://www.dmi.unict.it/~faro/smart/). It is distributed using GitHub at [https://github.com/smart-tool](https://github.com/smart-tool).

The paper is organized as follows. In Section 2 we give a brief description of the SMART tool, including its main features and the principles of fair testing which are at the basis of the framework. We give also a list of all implemented algorithms and the corpora which are included in SMART. Then in Section 4 we give a brief survey of the last comprehensive experimental evaluation performed with SMART, and describe the directions of future works in Section 5.

---

2 The preliminary version of SMART included 85 different algorithms, divided in 130 variants.
3 The GitHub web page is accessible at [https://github.com](https://github.com)
2 The Smart Tool in Short

SMART is an open source software which provides a standard framework for researchers in string matching. It helps users to test, design, evaluate and understand existing solutions for the exact string matching problem. Moreover it provides the implementation of (almost) all string matching algorithms and a wide corpus of text buffers. The SMART source code can be downloaded at the web page [http://www.dmi.unict.it/~faro/smart/](http://www.dmi.unict.it/~faro/smart/) It is released under the GNU general public license.

SMART is written in the C language and can be compiled in any operating system with a standard gcc compiler. The tool uses shared memory for storing the text. Thus SMART requires the system to allow the allocation of shared memory. The default size of the text is 1MB, which is small enough to be supported by any system. However if one wants to use SMART for testing algorithms on larger texts, system settings for shared memory must be checked.

In the following sections we briefly describe SMART’s main features.

2.1 Implemented Algorithms

In the last 40 years tens of string matching algorithms (and even a larger number of variants thereof) have been proposed. SMART provides the implementation of more than 300 variants and more than 120 different algorithms. This number is on the rise thanks to the continuous contributions of the community.

All implemented algorithms can be divided into four classes (but further classifications are possible): characters comparison, deterministic automata, bit parallelism and packed string matching. Classical approaches to the problem make use of comparisons between characters or perform transitions on some kinds of deterministic automata. However in the last two decades a lot of work has been made in order to exploit the power of the word RAM model of computation to speed-up classical string matching algorithms. In this model, the computer operates on words of length \( \omega \), thus blocks of characters are read and processed at once. This means that usual arithmetic and logic operations on the words all take one unit of time. Most of the solutions which exploit the word RAM model are based on the bit-parallelism technique or on the packed string matching technique.

An almost comprehensive list of all algorithms implemented in the preliminary version of SMART (more than 85) can be found in [13,17]. In the present release of the software we have extended such list introducing some additional variants of previous solutions and the following new algorithms presented between 2010 and 2016. The comprehensive list of algorithms implemented in the new version of SMART (more than 120) can be found in [8].

- Three bit-parallel algorithms for exact searching of long patterns appeared in [17]. Two algorithms are modifications of the BNDM [22] algorithm and the third one is a filtration method which utilizes locations of \( q \)-grams in the pattern. Two algorithms apply a condensed representation of \( q \)-grams.
- Two generalizations of the Forward-SBNDM [12] algorithm presented in [23]. The first generalizes the algorithm by using \( q \)-grams while the second introduces also a \( q \)-gram lookahead approach.

---

4 [http://www.gnu.org/licenses/gpl.html](http://www.gnu.org/licenses/gpl.html)
• A solution based on a factorization of the pattern and on a particular encoding of the suffix automaton [16], which turns out to lead to longer shifts than that proposed by other known solutions which make use of suffix automata.
• A constant-space $O(n)$-time packed string matching algorithm [2] which runs in optimal $O(n/\alpha)$-time, where $\alpha = w/\log \sigma$, and even in real-time.
• Several variants of previous solutions obtained by using a general approach to string matching based on multiple sliding text-windows [14].
• Two algorithms, presented in [24], based on a combination of the Boyer-Moore [3] and Horspool [19] algorithms. It takes the maximum shift proposed by the two occurrence heuristics.
• Three improvements of the standard occurrence heuristics [4,5].
• An improvement of Quick-Search algorithm [25] which improves the shift performed by the occurrence heuristics by computing the shift to left performed by the reverse of the pattern at a given fixed distance from the current window.
• A combination of Skip-Search and the Hashq algorithms which computes buckets of positions for the fingerprint of each $q$-gram in the pattern. It was presented in [9].
• Improved versions of the Shift-Or and Shift-And algorithms [1] using a two way scan of the window and $q$-grams. They were presented in [6].

Figure 1 presents the temporal distribution of algorithms proposed in the last 26 years (1990-2016) and the percentage of algorithms belonging to each class, up to 2016. Observe that the number of proposed solutions have doubled in the last ten years, demonstrating the increasing interest in this issue.

The class of algorithms based on comparison of characters is the wider class and consists of almost 50% of all solutions. Also automata play a very important role in the design of efficient string matching algorithms and have been developed to design algorithms which have optimal sub-linear performance on average. Almost 20% of all algorithms in SMART are based on automata.

Bit-parallelism [1] takes advantage of the intrinsic parallelism of the bit operations of automata. It is interesting to observe also that almost 50% of solutions in the last ten years (and 31% all along) are based on bit-parallelism, and it seems that such number follows an increasing trend.
1. Brute-Force (BF)  no date
2. Deterministic-Finite-Automaton (DFA) 1970
4. Boyer-Moore (BM) 1992
5. Boyer-Moore-BM (BBM) 1992
6. B licens (BO) 1980
7. Galil-Seiferas (GS) 1981
8. Apostolico-Giancarlo (AG) 1986
10. Zhu-Takaoka (ZT) 1989
11. Shift-Or (SO) 1990
12. Shift-And (SA) 1990
13. Quick-Search (QS) 1990
15. Maximal-Shift (MS) 1991
16. Apostolico-Crochemore (AC) 1992
17. Two-Way (TW) 1992
18. Tuned-Boyer-Moore (TUNBM) 1992
19. Colussi (COL) 1992
20. Smith (SMITH) 1992
22. Raita (RAITA) 1992
23. S.M. on Ordered Alphabet (SMOA) 1992
24. Turbo-Boyer-Moore (TBM) 1992
25. Reverse-Factor (RF) 1993
26. Not-So-Naive (NSN) 1993
27. Reverse-Colussi (RCOL) 1994
28. Simon (SIM) 1994
29. Turbo-Reverse-Factor (TRF) 1994
30. Forward-DAWG-Matching (FDM) 1994
32. Skip-Search (SKIP) 1994
33. Alpha-Skip-Search (ASKIP) 1994
34. Knuth-Morris-Pratt Skip-Search (KMPS) 1994
35. Nondeterministic DDM (BNDM) 1994
36. Berry-Ravindran (BR) 1994
37. Backward-Oracle-Matching (BOM) 1994
38. Double Forward DAWG Matching (DFDM) 2000
39. BNDM for Long patterns (BNDML) 2000
40. Super Alphabet Simulation (SAS) 2002
41. Ahmed-Kaykobad-Chowdhury (ACK) 2003
42. Fast-Search (FS) 2004
43. Simplified BNDM (SBNDM) 2004
44. Two-way DNDM (TNDM) 2004
45. Long patterns BNDM (LBNDM) 2004
46. Shift Vector Matching (SVM) 2004
47. Forward-Fast-Search (FFS) 2004
48. Backward-Fast-Search (BFS) 2004
49. Tailed-Substring (TS) 2005
50. Sheik et al. (SSARS) 2005
51. Wide Window (WW) 2005
52. Linear DAWG Matching (LDM) 2005
53. BNDM with loop-unrolling (BNDM2) 2005
54. SBNDM with loop-unrolling (SBNDM2) 2005
55. BNDM with Horspool Shift (BNDMBM) 2005
56. Horspool with BNDM test (BMHBNDM) 2005
57. Forward DDM (FNDM) 2005
58. Bit parallel Wide Window (BWW) 2005
59. Average Optimal Shift-Or (AOSO) 2006
60. Fast Average Optimal Shift-Or (FAOSO) 2006
61. Thathoo et al. (TVSBS) 2006
62. Horspool using Probabilities (PBMH) 2006
63. Improved LDM (ILDM) 2006
64. Improved LDM 2 (ILDM2) 2006
65. Franck-Jennings-Smyth (FJS) 2007
66. 2-Block Boyer-Moore (2BLOCK) 2007
68. Horspool with q-grams (BSHq) 2008
69. Two Sliding Windows (TSW) 2008
70. Extended BOM (EBOM) 2008
71. Forward BOM (FBOM) 2008
72. Succinct BOM (SBDM) 2009
73. Forward BNDM (FBNDM) 2009
74. Forward Simplified BNDM (FSBNDM) 2009
75. Bit-Parallel Length Invariant (BLIM) 2009
76. Genomic Rapid Algo for S.M. (GRASPm) 2009
77. Simplified Extended BOM (SEBOM) 2009
78. Simplified Forward BOM (SFBOm) 2009
79. BNDM with q-grams (BNDMq) 2009
80. Simplified BNDM with q-grams (SBNDMq) 2009
81. FNDM with q-grams (UFNDMq) 2009
82. Small Alphabet Bit-Parallel (SABP) 2009
83. Packaged String Search (PSS) 2009
84. Streaming SIMD Extensions Filter (SEF) 2009
85. Bounded Boyer-Moore (BBM) 2009
86. Bounded Fast-Search (BFS) 2009
87. Bounded Forward-Fast-Search (FFFS) 2009
88. BNDM with Extended Shifts (BXS) 2009
89. BNDMq Long (BQL) 2009
90. q-Gram Filtering (QF) 2009
91. Bit-Parallel Wide-Window (BPWWW) 2010
92. Bit-Parallel Wide-Window (BPWW2) 2010
93. Factorized Shift-And (KSA) 2010
94. Factorized BNDM (KBNDM) 2010
95. Packed Belazzougui (PBB) 2010
96. Packed Belazzougui-Raffinot (PBR) 2010
97. FSBNDM with q-grams (FSBNDMq) 2010
98. Packed Crochemore-Perrin (SSECP) 2010
100. TVBSw with Multiple Windows (TVBSw) 2013
101. Max Shift Boyer-Moore (MSBM) 2013
102. Max Shift Horspool (MSH) 2013
103. Enhanced Two Sliding Windows (ETSW) 2013
104. Hashy using Multiple Hashing (MHASHq) 2013
105. Enhanced Berry-Ravindran (RSA) 2013
106. Backward SNK DAWG Matching (BSDM) 2013
107. Multiple Windows SBNDM (SBNDMw) 2013
108. Multiple Windows FSBNDM (FSBNDMw) 2013
109. Enhanced RS-A (ERSA) 2013
110. Improved Occurrence Heuristics (IOM) 2013
111. Worst Occurrence Heuristics (WOM) 2013
112. Jumping Occurrence Heuristics (JOM) 2013
113. Exact Packed String Matching (EPSM) 2014
114. Improved Two-Way Shift-And (TSA) 2014
115. Improved Two-Way Shift-Or (TSO) 2014
116. Two-Way Shift-And using q-grams (TSAs) 2014
117. Two-Way Shift-Or using q-grams (TSOq) 2014
118. Simple String Matching (SSM) 2014
119. Quantum Leap Quick-Search (QLQS) 2014
120. Enhanced ERS-A (EERSA) 2014
121. Four Sliding Windows (FSW) 2014
122. Skip-Search using q-grams (SKIPq) 2014
123. BNDMq with q-grams (BNDMqg) 2014
124. BNDMq with multiple windows (BNDMqgw)

Table 1. The list of those string matching algorithms implemented in SMART which were published 1970-2016. Each algorithm is associated to its acronym used in SMART.
In packed string matching, multiple characters are packed into one larger word, so that the characters can be compared in bulk rather than individually. In this context, if the characters of a string are drawn from an alphabet of size $\sigma$, then $\lfloor w/ \log \sigma \rfloor$ different characters fit in a single word, using $\lfloor \log \sigma \rfloor$ bits per characters. Although algorithms in this class appeared in the last four years they turn out to be among the fastest solutions \cite{21,10}, reaching in some cases the optimal $O(n \log \sigma/w)$ time complexity \cite{2}.

2.2 Algorithm Testing and Evaluation

The main command provided by the tool, \texttt{smart} indeed, is used for running experimental tests. The experimental settings could be almost completely customized. The easiest way to use \texttt{SMART} is to run a single search for a custom pattern and a custom text. To this purpose one should use the \texttt{-simple} parameter followed by the pattern and the text. Otherwise it is possible to select the corpus which will be used to compute the experimental results by typing the parameter \texttt{-text} followed by the name of the selected corpus (for ex. \texttt{smart -text genome}). It is also possible to select more than one corpus by typing the name of the corpora, separated by a dash symbol (for ex. \texttt{smart -text genome-protein}). You can also type the parameter \texttt{-text all} in order to run experimental tests for all corpora, in which case the corpora will be processed one after another.

For each input file, \texttt{SMART} generates sets of $r$ patterns of fixed length, randomly extracted from the text, where the length of the patterns ranges over the set of values $\{2^k \mid 1 \leq k \leq 12\}$, so that running times can be easily reported in a log-scale plot. The value $r$ is set to 500 by default, but it is allowed to use the parameter \texttt{-pset} in order to modify the size of the set of patterns generated by the tool (for ex. you can type \texttt{smart -text genome -pset 100}). You can also use the parameter \texttt{-short} in order to perform experimental tests on short patterns, whose length ranges over the set of values $\{2 + \ell \mid 0 \leq \ell \leq 30\}$. If necessary, it is also allowed to restrict the pattern’s length to a given range by using the parameter \texttt{-plen} and indicating an upper bound and a lower bound for such lengths (for ex. you can type the command \texttt{smart -text genome -plen 8 64}).

Many algorithms are very slow under particular conditions. In order to avoid excessive running times during the experimental evaluation it is possible to set a \textit{time bound} which cannot be exceeded by any single run. This can be done by using the parameter \texttt{-tb}, followed by a value expressed in milliseconds. By default such bound is set to 300 ms.

The \texttt{SMART} tool has been developed in order to follow the principles of fair testing in string matching. In particular the experimental testing is based on the following features:

\textit{Algorithm verification}

The tool verifies that all tested algorithms work properly. This verification is done by counting the number of matches returned by the procedure and testing whether the search stops properly at the end of the text. Since all searched patterns are always randomly extracted from the text, it is guaranteed that the number of occurrences is always equal or greater than 1. It is not uncommon that some algorithms do not work for specific input parameters. For instance a $q$-gram based algorithm does not work for patterns shorter than $q$ characters. The framework also provides a control
mechanism able to distinguish a malfunctioning from a not working instance, in which case an error message is returned by the tool.

Clean time measuring
The SMART tool has been designed in order to rule out from time measurements all disturbing events. To this purpose the reading of data (text and patterns) belongs to an outer part of the test setting, so that times spent to reading is excluded from time measurements. Moreover printing of matches is not performed during time measurement, since printing produces also an additional overhead, which is partly unsynchronized. Finally the framework has been developed in order that the time measurement itself does not disturb the work of algorithms.

Fair comparison
Since in SMART the measuring is focused on performance, the tested algorithms have been implemented in a uniform way, using the same standard for processing string characters, compare them, performing automata transitions, computing matches and for more other common tasks. Then SMART uses the -O3 level of optimization, which is the highest optimization level. All algorithms are available online and can be analyzed and improved by the community. During time measurement all tested algorithms share the same input data in order to allow a fair comparison. Moreover the SMART tool has been designed in order to ensure that no residual information in cache, during multiple executions of the same algorithm, may be used in the next attempt affecting the time measurement.

Algorithm preprocessing
It has never been clear in the literature if preprocessing should be included in the measurements. The work done in preprocessing is only a proportion of the whole task, and it may depend on the pattern length and on the alphabet size. According to Horspool [19] in string matching the timings do not include the work of initializing tables, however for an online algorithm, preprocessing time is spent to compute useful informations which are then used for speeding up the search process. Moreover, it is also true that in most cases the longer is the preprocessing time the faster is the searching. This line of reasoning finds its borderline case in an offline algorithms, where a preprocessing of the text leads to an extremely fast searching phase. For this reason the SMART tool has been developed in order to include the preprocessing time in the performance measurement.

However the tool can be set up in order to separate time measurements in searching and preprocessing times. This can be done by using the parameter -pre. By default SMART produces only a single time measurement which includes preprocessing and searching time.

Stability measurement
It is also useful to find out how accurately repeatable the results are. To this purpose using only average running times easily hides important details. It is common, in fact, that for some algorithms running times move away from the mean value. This is what in general we associate with the stability of a given algorithm: the smaller is the standard deviation of the running times the superior is its stability. Thus the SMART tool can be set up in order to compute also standard deviation values of running times (use the parameter -std). In addition the tool also allows the computation of
the best and worst running times obtained during the experimental evaluation (use the parameter \texttt{-dif}).

2.3 Output Formats

The \textsc{Smart} tool associates to any experimental test a unique alphanumeric code on 13 characters, beginning with the prefix \texttt{EXP} and followed by a string of 10 numbers computed from the timestamp. At the end of the execution of an experimental test \textsc{Smart} stores experimental data in the directory \texttt{results/EXPCODE}, where \texttt{EXPCODE} is the unique code associated with the experimental test. Files containing experimental data are named with the name of the corpus which has been selected. The system can store experimental data in different formats: simple text, \LaTeX, xml, html and php format.

Files in xml format report data in a structured way suitable to be processed or included in other documents, while html files present data in a tabular format. An additional \texttt{index.html} file is generated which contains the list of all html pages containing the experimental results computed during the test. Figure 2 shows a portion of the HTML output produced by \textsc{Smart}.

Finally, \LaTeX files can be generated to make easy the inclusion of tables containing the experimental results in \LaTeX source files.

2.4 Text Corpora

\textsc{Smart} comes with a set of corpora which can be selected in for running experimental results. The corpora are stored in a directory named \texttt{“data”}, and each corpus consists of a set of texts stored in a sub-directory with the same name of the corpus. Each sub-directory contains an index file (\texttt{index.txt}) containing the names and a description of all the files contained in the corpus. It is possible to select the corpus which will be used to compute the experimental results by typing the parameter \texttt{-text} followed by the name of the selected corpus.

The \textsc{Smart} tool allows to set an upper bound dimension of the text size used during the experimental results. By default this upper bound dimension is set to 1MB. This means that (at most) the first 1MB of the selected corpus will be used for testing. The default upper bound dimension can be changed by using the parameter \texttt{-tsize}, followed by an integer value which indicate the dimension, in Mbytes, which will be used (for ex. \texttt{smart -text genome -tsize 5}).

Then all files listed in the index will be loaded in a text buffer, one by one, until the upper bound is reached. If the upper bound is larger than the whole size of the corpus the list of files is processed again in order to fill the whole buffer. In details, \textsc{Smart} provides the following set of 15 corpora.

(i) \textbf{englishTexts}, a set of english texts (6.1 MB) over an alphabet of 94 characters.
   It includes two text of size 3.9 MB and 2.4 MB, respectively.
(ii) \textbf{italianTexts}, a set of italian texts (5 MB) over an alphabet of 120 characters.
   It includes seven texts whose size ranges from 281 KB to 1.5 MB.
(iii) \textbf{frenchTexts}, a set of french texts (6.6 MB) over an alphabet of 119 characters.
   It includes seven texts whose size ranges from 631 KB to 1.2 MB.

\footnote{Note that during the experimental evaluation the text buffer is stored in shared memory, thus if you set the upper bound to a value \( K \) MB it is necessary to ascertain your system allows the allocation of at least \( K \) MB of shared memory}
Figure 2. A portion of the HTML output produced by SMART. Experimental data are reported in tabular and in graphical form. Here we observe average running times, worst running times and best running times of Horspool and Boyer-Moore compared on a Rand2 text buffer.

(iv) chineseTexts, a set of Chinese texts (5.7 MB) over an alphabet of 160 characters. It includes five texts whose size ranges from 745 KB to 2.3 MB.

(v) genome, a set of DNA sequences (4.4 MB) over an alphabet of 4 characters. It includes Complete genome of the E. Coli bacterium (4.4 MB).

(vi) protein, a set of protein sequences (3.1 MB) over an alphabet of 20 characters. It includes a protein sequence from the Human genome (3.1 MB).

(vii) midimusic, a set of midi sequences (2.7 MB) over an alphabet of 117 characters. It includes 206 midi files on Johann Sebastian Bach work (1685-1750) whose size ranges from 4 KB to 205 KB.

(viii) randσ, random texts (5 MB) over an alphabet of size σ with a uniform distribution, where σ ranges over the values \{2, 4, 8, 16, 32, 64, 128, 256\}.
Files (i) and (v) are from the Large Canterbury Corpus, files (ii), (iii) and (iv) are from the Gutenberg project, while file (vi) is from the Protein Corpus. Finally, files in (vii) are from the Johann Sebastian Bach Midi Page.

In addition the SMART tool provides a simple way for adding new corpora to the default set. This can be done by simply introducing a new sub-directory named with the name of the corpus, and containing the set of selected files together with an index file listing their names.

2.5 Adding New Algorithms

The SMART tool is not only a framework for testing all known string matching algorithms. It provides also an easy and fast way for assisting researchers to develop and test new efficient algorithms. It is possible to add new string matching algorithms to SMART, testing them for correctness and compare their efficiency against the previous solutions.

The following few requirements must be guaranteed: a new algorithm must be implemented in the C programming language and must include the header file “include/main.h”. The main method must be defined as

\[
\text{int search(unsigned char *x, int m, unsigned char *y, int n)}
\]

where \(x\) maintains the pattern, \(y\) maintains the text, while \(m\) and \(n\) are their lengths, respectively. The method must return the number of occurrences of the pattern in the text. In addition, if the algorithm does not run under particular conditions (for instance when the length of the pattern is less than a given value), it is required the algorithm to return the value \(-1\).

Since preprocessing time is computed separately from searching time, it is required to arrange the code concerning the preprocessing phase between the macros \texttt{BEGIN_PREPROCESSING} and \texttt{END_PREPROCESSING}, while the code concerning the searching phase must be arranged between the macros \texttt{BEGIN_SEARCHING} and \texttt{END_SEARCHING}. Figure presents the C code of the Horspool algorithm in a format suitable for inclusion in SMART.

Before compiling the C file, copy the header file \texttt{main.h} (which is stored in the folder \texttt{source/algos/include}) in the same directory. Then put the compiled binary file in the directory “\texttt{source/bin}”. Before running a new experimental setting you can test the correctness of your algorithm by executing the command “./test algoname”, where \texttt{algoname} is the name of the binary file of your algorithm. Then it will be possible to include the new algorithm in SMART by typing the command “./select -add algoname”.

3 A Graphical User Interface

The new version of SMART comes with a useful Graphical User Interface (SMARTGUI) which could be used for running experimental results. It is implemented in C++ using the Qt WebKit, one of the major engine to render webpages and execute JavaScript code. Figure shows a screenshot of the SMARTGUI where reporting the average

---

6 http://www.data-compression.info/Corpora/CanterburyCorpus/
7 http://www.gutenberg.org
8 http://data-compression.info/Corpora/ProteinCorpus/
9 http://www.bachcentral.com
#include "include/define.h"
#include "include/main.h"

int search(unsigned char *P, int m, unsigned char *T, int n) {
  int i, s, count, hbc[SIGMA];

  BEGIN_PREPROCESSING
  for(i=0;i<SIGMA;i++) hbc[i] = m;
  for(i=0;i<m-1;i++) hbc[P[i]] = m-i-1;
  END_PREPROCESSING

  BEGIN_SEARCHING
  s = 0;
  count = 0;
  while(s <= n-m) {
    i = 0;
    while( i<m && P[i]==T[s+i] ) i++;
    if( i==m ) count++;
    s += hbc[T[s+m-1]];
  }
  END_SEARCHING
  return count;
}

Figure 3. The C code of the Horspool algorithm for string matching.

running times of the Horspool and Boyer-Moore algorithms when compared on a genome sequence.

The central part of SMARTGUI is dedicated to report graphs of running times. The text output is reported next to the graphs, giving a familiar feedback to users which execute SMART using the terminal. When several experimental evaluations are executed, SMARTGUI organizes the graphs using tabs.

SMARTGUI has been developed to make easier the use of SMART. It allows to view in real-time all the results of the experimental evaluations and to compare the single algorithms of the experiments through the tabs.

In addition SMARTGUI makes easy the customization of the tests through input text parameters and checkbox to choose the output format (txt, \LaTeX, pdf, etc.) and the text buffers. It also allows to select algorithms or add new algorithms and run tests on it. During any test SMARTGUI shows the status through a progress bar. When a single experimental evaluation ends SMARTGUI shows more statistics through a web view opening the html format results.

Executable binaries of SMARTGUI are available at the SMART web page for Windows, Linux and Mac. It can be downloaded and installed in any folder of your computer, even different from your main SMART folder. At the first execution of SMARTGUI it is necessary to link the main SMART folder using the setup procedure available at the menu button setup smart gui.

4 Experimental Evaluation

In this section we discuss experimental results which could be performed using SMART. The details of the experimental evaluation can be analyzed at the SMART web page (http://www.dmi.unict.it/~faro/smart/).

A recent survey [13] already presents an extensive experimental evaluation of almost all string matching algorithms used for searching in different texts. The authors used SMART for computing the experimental results, indeed.
Figure 4. A screenshot of the Graphical User Interface of Smart. Here we observe average running times of Horspool and Boyer-Moore algorithms when compared on a genome sequence.

Other more recent experimental evaluations using the Smart tool appeared in [16,14,2,11] where new efficient algorithms were presented.

Thus in this section we do not give another extensive experimental evaluation of string matching algorithms, but we focus our attention on new general experimental observations which Smart allows to do. In particular Smart allows to analyze string matching algorithms from three different points of view: their efficiency, their stability and their flexibility.

Efficiency Measurement
The efficiency of an algorithm is evaluated in Smart by computing the mean of running times over a large set of attempts.

Most string matching algorithms are characterized by a performance plot with a decreasing trend. Thus on average the performances increases when the length of the pattern increases. Almost all efficient algorithms, in fact, are based on a sliding window approach whose shift is at most long as the length of the pattern. This behavior is evident, for instance, in Figure 2 and in Figure 4 where we show the experimental results of the well known Horspool [19] and Boyer-Moore [3] algorithms, which are comparison based algorithms based on the occurrence heuristic. Also automata show such decreasing trend, however almost all of them show a decrease in performances in the case of longer patterns. This is due to the size explosion of the underlying automaton and of the correspondent preprocessing time.

Algorithms based on q-grams are quite efficient on average. In general their performance increases when the length of the pattern increases or when the value of q increases. However, on the other hand in the case of short patterns their performances drastically decreases when the value of q increases.
Stability Measurement

In SMART the stability of an algorithm is computed as the standard deviation of running times observed during the evaluation. Such value shows how much variation exists from the average, i.e. the mean of the running times. A low standard deviation indicates that the running times tend to be very close to the mean, underlying a high stability of the algorithm. On the other hand a high standard deviation indicates that the running times are spread out over a large range of values, thus indicating a low stability. It turns out from our observations that almost all algorithms have a low stability for short patterns while their stability increases when the length of the pattern increases. Such behavior becomes more evident for larger alphabets.

Sometimes an opposite behavior can be observed when searching on texts over a small alphabet like DNA sequences. This is the case, for instance, of some comparison based algorithms based on the occurrence heuristic whose stability decreases when the length of the pattern gets shorter. This trend can be explained by the reduced combination of strings which could be obtained when the pattern is short and the alphabet is small.

Flexibility Measurement

Flexibility is used as an attribute of various types of systems. In the field of string matching, it refers to algorithms that can adapt when changes in the input data occur. Thus a string matching algorithm can be considered flexible when, for instance, it maintains good performances for both short and long patterns, or in the case of both small and large alphabets.

As already observed most string matching algorithms obtain good performances only in the case of long patterns sacrificing their performance for short ones. This is a common behavior, for instance, for all algorithm which make use of a sliding window approach. Such approach allows the pattern to slide along the text by performing subsequent shifts. Each shift can be at most as long as the length of the pattern. It turns out that statistically the shift increases when the length of the pattern increases, or when the size of the alphabet increases. Although bit-parallel algorithms are designed to be extremely efficient in the case of long patterns, also this class of algorithms suffers of a lack in flexibility.

Only packed string matching algorithms turn out to have good performances for short patterns. This is the case of the SSECP algorithm [2] whose performances degrade when the length of the pattern increases. It is also the case of the EPSM algorithm [11], whose flexibility is obtained only by combining different algorithms, depending on the length of the pattern.

5 Conclusions, Future Works and Acknowledgements

We presented SMART (http://www.dmi.unict.it/~faro/smart) a flexible testing and evaluation tool for single exact string matching algorithms. It contains the implementation of almost all string matching algorithms appeared since 1970 up to 2016. The tool helps researchers in the filed in various way and we encourage them to contribute to the project by providing their own code for testing. Many improvements are possible to enhance SMART, including its adjustment in order to work with 128 bit processors.

We wish to thank Prof. Jorma Tarhio, Dr. M. Oğuzhan Külekci and Dr. Arseny Kapoulkine for helpful comments and suggestions in improving the SMART frame-
work. We wish also to thanks the authors of several string matching algorithms for having provided their codes for inclusion in SMART.

References


