Finding patterns in strings

Author(s):
Shi, Fei

Publication Date:
1997

Permanent Link:

https://doi.org/10.3929/ethz-a-001735097 →

DISS. ETH No. 12014

Finding Patterns in Strings

A dissertation submitted to the
Swiss Federal Institute of Technology
(ETH)
Zurich

for the Degree of
Doctor of Technical Sciences

presented by
Fei SHI
Master of Science, Peking University

born December 1, 1959
citizen of the
People 's Republic of China

accepted on the recommendation of

Prof. Dr. Peter Widmayer, examiner
Prof. Dr. Esko Ukkonen, co-examiner

1997

Abstract

String pattern matching is one of the central and most widely studied problems in theoretical computer science. Solutions to the problem play an important role in many areas of science and information processing. The simplest form of the problem is to locate all occurrences of a string (called the pattern) as a substring in another string (called the text). Another variant of the problem is the approximate string matching problem: given a string P (the pattern), a string T (the text), a positive number k, and a distance metric d, find all substrings x of T such that $d(P, x) \leq k$, that is, the distance between the pattern P and the substring x is at most k.

The topic of this thesis is algorithms and data structures for the exact and approximate string matching problems and their variants. Specifically, we study the following problems in this thesis:

- the longest common substring problem,
- the (exact) multiple string matching problem,
- the approximate string matching problem,
- the approximate dictionary matching problem,
- the approximate multiple string matching problem, and
- the two-dimensional pattern matching problem.

Our algorithms for the longest common substring problem, the approximate string matching problem, the approximate multiple string matching problem, and for the two-dimensional pattern matching problem, have obvious advantages over the previously known solutions to the same problems. We will also present a data structure that efficiently supports the approximate multiple string matching. Our generalized suffix array is useful for solving many string matching problems. To the best of our knowledge, our work on the approximate dictionary matching problem was the first attempt at attacking this problem.

Kurzfassung

Textuelles Pattern Matching ist eines der zentralen und ausgiebig bearbeiteten Probleme der theoretischen Informatik. Die einfachste Variante dieses Problems ist die Aufgabe, alle Vorkommen eines Textstückes (Pattern genannt) in einem anderen Textstück (Text genannt) zu bestimmen. Eine andere Frage ist die des approximativen Pattern Matchings: Gegeben sei das Pattern P und der Text T, $k > 0$ und ein Abstandsmass d. Man finde nun alle Teilstücke x von P, für die $d(P, x) \leq k$ gilt, d. h. der Abstand zwischen dem Pattern P und dem Teilstück x höchstens k ist.

Diese Arbeit dreht sich um Algorithmen und Datenstrukturen für das exakte und das approximative Pattern Matching Problem und ihre Varianten. Genauer untersuchen wir hier die folgenden Probleme:

- die Suche nach dem längsten gemeinsamen Textteilstück,
- das (exakte) Pattern Matching in mehreren Texten,
- das approximative Pattern Matching,
- das approximative Matching für mehrere Pattern,
- das approximative Pattern Matching in mehreren Texten und
- zweidimensionales Pattern Matching.

Jeder unserer Algorithmen für die Suche nach dem längsten gemeinsamen Textteilstück, das approximative Pattern Matching, das approximative Pattern Matching in mehreren Texten und zweidimensionales Pattern Matching hat offensichtliche Vorteile gegenüber den bisher bekannten Resultaten für dieselben Probleme. Ausserdem geben wir eine Datenstruktur an, die approximative Pattern Matching in mehreren Texten auf effiziente Weise unterstützt. Unsere verallgemeinerten Suffixarrays sind für die Lösung vieler Pattern Matching Probleme nützlich. Soweit uns bekannt, ist der Teil der Arbeit über das approximative Matching für mehrere Pattern der erste Lösungsversuch dieses Problems.