

DES BORNES EXACTES DE LA COMPLEXITE DES ALGORITHMES SEQUENTIELS DE RECHERCHE D'UN MOTIF

Christophe Hancart

Résumé

Dans cet article, nous étudions la complexité exacte du problème du calcul de toutes les occurrences d'un motif dans un texte par des algorithmes qui, comme celui de Knuth, Morris et Pratt, n'ont accès au texte qu'au travers une fenêtre glissante de longueur 1, en supposant que l'alphabet est inconnu des algorithmes et contient au moins deux lettres.

Nous donnons une famille d'algorithmes linéaires en temps et en espace qui réalisent le calcul en effectuant au plus $\lfloor (2 - \frac{1}{m})n \rfloor$ comparaisons entre caractères du motif et caractères du texte, et en comparant chaque caractère du texte au plus $\min\{1 + \lfloor \log_2 m \rfloor, \text{card}(A)\}$ fois, où m est la longueur du motif, n celle du texte et A l'alphabet. Nous montrons que ces bornes sont les bornes exactes du problème. Les algorithmes optimaux résultent d'une implémentation d'automates finis.

Abstract

String matching is the problem of finding all the occurrences of a string, called the pattern, within another, called the text. We give sharp bounds when only one head moving from the left to the right of the text is used, for the general case where the alphabet is unknown to the algorithms : $\lfloor (2 - \frac{1}{m})n \rfloor$ comparisons between pattern characters and text characters, and $\min\{1 + \lfloor \log_2 m \rfloor, \text{card}(A)\}$ comparisons on each text character, where m is pattern length, n is text length, and A is the alphabet. These results improve the previous results given by Knuth, Morris and Pratt, and, more recently, by Simon.

1 Introduction

Le problème de la *recherche d'un motif* consiste à localiser toutes les occurrences d'un mot x , appelé le *motif*, de longueur non nulle m dans un mot y , appelé le *texte*, de longueur n , où les deux mots sont considérés sur un alphabet donné A .

Received by the editors April 1993, revised February 1994.

Communicated by M. Boffa.

AMS Mathematics Subject Classification : 68Q20, 68Q25, 68R15, 68U15.

Keywords : string matching, string searching, text editing, computational complexity, worst case behavior, finite automata.