Otimização de Aho-Corasick para contar palavras em um texto

Emerson Leonardo Lucena [1], **Rohit Gheyi** [2]

RESUMO

O algoritmo de Aho-Corasick é usado para reconhecer todas as ocorrências de um conjunto de palavras em um texto. Entretanto, em casos onde a entrada contenha uma grande quantidade de casamentos com os padrões do dicionário, o tempo de execução do procedimento irá aumentar. Nós iremos alterar o algoritmo Aho-Corasick para que não dependa do número de ocorrências das palavras do dicionário. Nós iremos descrever a alteração, provar sua corretude, e realizar comparações com o algoritmo original. O novo algoritmo poderá ser importante na otimização de aplicações em áreas diversas, como processamento de linguagem natural, análise de sequências de DNA, e pesquisa em sistemas de arquivos.

Palavras-chave: palavras-chave, ocorrências, strings, string matching, pattern matching, Aho-Corasick.

[1] Aluno do curso de Ciência da Computação, Unidade Acadêmica de Sistemas e Computação, UFCG, Campina Grande, PB, e-mail: emerson.lucena@ccc.ufcg.edu.br

[2] Doutor, Professor Associado, Unidade Acadêmica de Sistemas e Computação, UFCG, Campina Grande, PB, e-mail: rohit@dsc.ufcg.edu.br

Otimização de Aho-Corasick para contar palavras em um texto

ABSTRACT

The Aho-Corasick algorithm is used to recognize all occurrences of a set of strings in a text. However, when the input contains a large amount of matches with the dictionary patterns, the procedure runtime will increase. We will modify the Aho-Corasick algorithm so that it will not depend on the number of occurrences of the dictionary words. We will describe the modification, prove its correctness, and perform comparisons with the original algorithm. The new algorithm can be useful for optimizing applications in many areas, such as natural language processing, DNA sequence analysis, and file system search.

Keywords: keywords, occurrences, strings, string matching, pattern matching, Aho-Corasick.