

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра дискретной математики и алгоритмики

Аннотация к дипломной работе

**«Обобщенные алгоритмы для типичных задач на базе суффиксных
деревьев»**

Гутько Давид Николаевич

Научный руководитель – старший преподаватель кафедры дискретной
математики и алгоритмики ФПМИ Комаровский И. В.

Минск, 2024

Реферат

Дипломная работа, 39 страниц, 11 рисунков, 1 таблица, 9 источников

Ключевые слова: СУФФИКСНОЕ ДЕРЕВО, АЛГОРИТМ УККОНЕНА, ОПТИМИЗАЦИЯ, ОБОБЩЕННОЕ ПРОГРАММИРОВАНИЕ, C++20, ЗАДАЧА О ПОДСТРОКЕ, НАИБОЛЬШАЯ ОБЩАЯ ПОДСТРОКА.

Объектом исследования являются эффективные алгоритмы на строках, методы программирования C++.

Предметом исследования является эффективная реализация алгоритма Укконена и алгоритмы на базе суффиксных деревьев.

Целью работы является изучение алгоритмов на суффиксных деревьях, их программная реализация для решения прикладных задач, исследование различных оптимизаций алгоритма, обобщение алгоритма для различных видов строковых последовательностей.

В ходе работы были рассмотрены основные концепции суффиксных деревьев и их применение в некоторых базовых задачах. Был рассмотрен линейный алгоритм Укконена для построения суффиксного дерева. Были предложены оптимизации алгоритма, фокусирующиеся на структурах данных для хранения дуг в вершинах суффиксного дерева. Проведены эксперименты для различных длин последовательностей и мощностей алфавитов с выбором разных структур данных для хранения дуг. Шаблонный класс на языке программирования C++ был разработан для универсального применения этих оптимизаций в различных сценариях.

Полученный в результате работы программный продукт можно использовать для разработки высоконагруженных приложений в области обработки текстовых и других последовательностей данных.

Abstract

Diploma thesis, 39 pages, 11 figures, 1 table, 9 sources.

Keywords: SUFFIX TREE, UKKONEN'S ALGORITHM, OPTIMIZATION, GENERIC PROGRAMMING, C++20, SUBSTRING PROBLEM, LONGEST COMMON SUBSTRING.

The object of research is efficient string algorithms and C++ programming techniques.

The subject of study is an efficient implementation of the Ukkonen's algorithm and algorithms based on suffix trees.

The aim of this work is the study of algorithms on suffix trees, their software implementation for solving applied problems, the study of various optimizations of the algorithm, generalization of the algorithm for various types of string sequences.

In the course of the work, the basic concepts of suffix trees and their application to some basic problems were reviewed. The linear Ukkonen's algorithm for constructing a suffix tree was considered. Algorithm optimizations have been proposed that focus on data structures for storing arcs at the vertices of the suffix tree. Experiments were carried out for various sequence lengths and alphabets with a choice of different data structures for storing arcs. A template class in the C++ programming language was developed to universally apply these optimizations in various scenarios.

The resulting software can be used to develop high-load applications in the field of processing text and other data sequences.