Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт (национальный
исследовательский университет)»
Физтех-школа радиотехники и компьютерных технологий
Кафедра теоретической и прикладной информатики

Направление подготовки: 03.03.01 Прикладные математика и физика
Направленность (профиль) подготовки: Инфокоммуникационные системы и технологии

СЖАТЫЕ СТРУКТУРЫ ДЫННЫХ В ЗАДАЧЕ ПОЛНОТЕКСТОВОГО ПОИСКА

(бакалаврская диссертация)

Студент:
Соколов Вадим Андреевич

(подпись студента)

Научный руководитель:
Неганов Алексей Михайлович

(подпись научного руководителя)

Москва 2021

Аннотация

Цели и задачи работы.

Данная работа посвящена исследованию структур данных, использующих сжатые индексы (succinct index) для хранения текстовой информации.

Целью данной работы является проверка эффективности различных методов сжатия данных. Исследуется применимость сжатых индексов на практике при работе с данными определенного типа. Производится сравнение как с традиционными решениями (suffix array), так и с более современными (radix tree), использующими подход для индексирования, отличный от исследуемых структур данных.

Полученные результаты.

Удалось получить сравнительные характеристики работы традиционных структур данных. Были измерены и проанализированы:

1. объем потребляемой памяти;

2. время, требуемое для вставки / поиска подстроки;

На языке Go реализована сжатая структура данных succinct suffix array. Приведены результаты потребления памяти для хранения сжатого индекса и скорости поиска подстроки в тесте.

Содержание

1 Введение **4**

2 Постановка задачи **5**

3 Измерения **6**

3.1 Экспериментальная платформа 6

3.2 Suffix array 6

3.3 Radix tree 6

4 Оценка результата **6**

5 Выводы **6**

6 Заключение **6**

1 Введение

Работа с текстовыми данными находит применение в широком спектре задач современной компьютерной индустрии. Существует ряд проблем, связанных с поиском информации в поисковых сервисах. Рост количества информации в Интернете приводит к дополнительным издержкам при хранении и поиске данных. В связи с этим существует необходимость исследования различных способов уменьшения потребляемой памяти без существенных затрат на поиск данных.

Одним из возможных решений такого рода задач является применение сжатых структур данных (succinct data structures). В зависимости от степени сжатия информации структуры данных различаются на имплицитные, сжатые и компактные. Сжатые структуры используют близкое к теоретически минимальному количеству информации для хранения данных. Кроме того, в отличие от архивов и других сжатых представлений, остается возможность эффективно выполнять операции поиска. Предположим, что для хранения некоторого количества данных требуется Z бит. Сжатые структуры данных занимают $Z + o(Z)$ бит. Например структура данных, занимающая $Z + \ln(Z)$ бит памяти, является сжатой.

Данные не всегда сжимаемы. Кроме того, не любые данные целесообразно сжимать с точки зрения эффективности их использования в несжатом виде. В этой работе предлагается рассмотреть сжатие индекса суффиксного массива, построенного для различных тектовых данных. При этом сам текст остается в несжатом виде.

Для того чтобы представить данные в сжатом виде, необходимо подготовить их в специальном промежуточном формате. В этой работе используется алгоритм Элиас-Фано, позволяющий сжимать возрастающие последовательности неотрицательных целых чисел. Исследование направлено на изучение потребления памяти для сжатого представления суффиксного массива. Реализованы функции поиска подстроки, и произведен анализ их эффективности.

2 Постановка задачи

...

3 Измерения

3.1 Экспериментальная платформа

3.2 Suffix array

3.3 Radix tree

4 Оценка результата

5 Выводы

6 Заключение