БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра многопроцессорных систем и сетей

Аннотация к дипломной работе

«Обработка больших массивов данных»

Козлова Виталия Олеговна

Научный руководитель - профессор, доктор техн. наук Буза М.К.

2016

Реферат

Дипломная работа, 46 с., 30 рис., 10 источников.

Ключевые слова: БОЛЬШИЕ ДАННЫЕ, АЛГОРИТМЫ СЖАТИЯ, КОЭФФИЦИЕНТ СЖАТИЯ, ВРЕМЯ ВЫПОЛНЕНИЯ, АЛГОРИТМЫ СТЕММИНГА.

Объект исследования – текстовая информация и алгоритмы сжатия применимые к данному типу информации.

Цель работы – разработка программного приложения для сжатия текстовой информации и анализ полученных результатов.

За время работы реализованы следующие задачи: изучены методы обработки больших массивов данных; изучены основные алгоритмы сжатия текстовой информации без потерь; проведены исследования по результатам работы алгоритмов на основе четырех характеристик: время выполнения алгоритма, коэффициент сжатия информации, объем сжимаемых данных и язык представления текстовых данных; проведен анализ и сравнение алгоритмов в различных условиях; поставлены опыты с последовательным использованием нескольких алгоритмов для одного блока данных и выявлены лучшие комбинации; изучены основные алгоритмы стемминга, позволяющие сжимать текстовую информацию с потерями, при сохранении семантики; проведен анализ алгоритмов стемминга; выявлены лучшие алгоритмы по времени выполнения и по коэффициенту сжатия среди алгоритмов стемминга; даны рекомендации по применению алгоритмов.

Работа имеет большое практическое значение в сфере информационных технологий: выбор алгоритмов для построение архиваторов, улучшения существующих решений для сжатия данных, построение поисковых систем и компоновка данных для последующего анализа.

Результаты данной работы могут быть использованы при выборе наиболее подходящего алгоритма для сжатия текстовой информации в зависимости от языка представления информации, а также при задачах оптимизации алгоритмов сжатия или необходимости использования нескольких алгоритмов сжатия последовательно.

Результаты тестирования алгоритмов стемминга могут быть применены при проектировании систем поиска.

Abstract

Diploma thesis, 46 pages, 30 figures, 10 sources.

Key words: BIG DATA, COMPRESSION ALGORITHMS, COMPRESSION RATIO, RUNNING TIME, STEMMING ALGORITHMS.

Object of research - text information and compression algorithms are applicable to this type of information.

The purpose of work - development of a software application for compressing text and analysis of the results.

During the work completed the following task: explored methods of processing large amounts of data; explored the basic lossless compression algorithms; conducted research on the results of the algorithms based on four characteristics: running time, aspect ratio information, the amount of compressed data, and the language of the text data; analyzed and compared algorithms under different conditions; experimented with consistent use of multiple algorithms for a data block and found the best combinations; explored basic Stemming algorithms allowing compress text information with losses, while maintaining the semantics; identified the best algorithms for run-time and compression ratio among Stemming algorithms; provides recommendations on the use of algorithms.

The work is of great practical importance in the sphere of information technologies: the choice of algorithms for the construction of archives, improve existing solutions for data compression, the construction of the search engines and linking of data for further analysis.

The results of this work can be used in selecting the most suitable algorithm for compressing textual information depending on the language of presentation information, as well as to resolve optimization problems or the need to use several compression algorithms sequentially. Stemming algorithms test results can be applied in the design of search system.