АНОТАЦІЯ

Дослідження алгоритму посимвольного стиснення текстової інформації в адресних базах даних // Дипломна робота ОКР “Магістр” // Грицай Роман Вікторович // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп’ютерно–інформаційних систем і програмної інженерії, кафедра комп’ютерних наук, група СНм–51 // Тернопіль, 2014 // С. , рис. – , табл. – , кресл. – , додат. – , бібліогр. – .

Ключові слова: СТИСНЕННЯ ДАНИХ, МЕТОД, МОДЕЛЬ, ПІДХІД, АЛГОРИТМ ХАФФМАНА, БЛОЧНО-СТАТИСТИЧНИЙ АЛГОРИТМ, БАЗИ ДАНИХ.

В першому розділі проведено аналіз проблемати та поставлено завдання для наукового дослідження. Розглянутий алгоритм Хаффмана та інші найбільш популярні алгоритми стиснення даних такі як LZW, алгоритм Хаффмана, PPM, RLE та інші. Сформульована проблема економного зберігання великої кількості даних у реляційних базах даних. Наведена класифікація алгоритмів стиснення даних на алгоритми з втратами та без, наведені приклади використання кожного підходу.

В другому розділі описано теоретичні аспекти дослідження. Детально розглянуто алгоритм Хаффмана, розібраний приклад стиснення даних. Сформульована пропозиція для модифікації алгоритму Хаффмана для отримання більшої ефективності при стисненні даних. Побудовані блок-схеми класичного та модифікованого алгоритмів. Також розглянуті основні класи розробленого продукту, їх функції. Описані вхідні дані для проведення експериментів.

В третьому розділі обгрунтовано практичне застосування проведених досліджень. Описано дані, що використовувалися для проведення досліджень ефективності розробленої модифікації. Моделювання проведено за допомогою розробленого програмного комплексу, яке отримує на вхід текстовий документ,

проводить кодування по класичному алгоритму Хаффмана (будує кодове дерево та кодує кожен символ) та по його модифікації (розбиває таблицю символів на дві групи згідно їх частот, будує кодове дерево для кожної з них та кодує символи за допомогою них), підраховує розмір стиснених даних за оригінальним алгоритмом та модифікацією.

Описаний інтерфейс та порядок роботи програмного засобу, за допомогою якого були проведені ці експерименти. Була проведена оцінка адекватності отриманих результатів дослідження, також отримані результати були порівняні з класичним алгоритмом Хаффмана та зроблені висновки про перевагу модифікації.

Об'єкт дослідження – процес стиснення текстових даних за допомогою алгоритму Хаффмана та його модифікації.

Предмет дослідження – ефективність блочно-статистичного алгоритму у порівнянні з класичним алгоритмом Хаффмана при стисненні текстових даних.

Мета роботи: розробка модифікації алгоритму Хаффмана, яка є покращенням показників ефективності стиснення текстових даних в базах даних.

Основні результати: Розроблена модифікація алгоритму, створена програмна реалізація для проведення експериментів на ПК з використанням різних вхідних даних за допомогою мови програмування C#.

В результаті виконання роботи був детально розглянутий на прикладі алгоритм Хаффмана та інші найбільш популярні алгоритми стиснення даних такі як LZW, PPM, RLE. Була сформульована пропозиція для модифікації алгоритму Хаффмана для отримання більшої ефективності при стисненні даних. Проведені експерименти по дослідженню модифікації та зроблені висновки відносно його ефективності.

ABSTRACT

Research spelling compression algorithm textual information in address databases // Bachelor Thesis "Master" // Gritsay Roman. // Ternopil Ivan Pul'uy National Technical University, Department of Computer Information Systems and Software Engineering, Department of Computer Science, group SNm-51 // Ternopil, 2014 // P. , Fig - , Table - , Draw , Ref. - .

Information - is not just a scientific category, and commercial, which is the same principal factor of development, feedstock energy. Now for the recovery of stocks of raw materials and energy, mankind is in dire need of information. The information opens up new ways of more efficient and economical to obtain funds for further scientific and technological progress and development in all spheres of human activity. Any serious problem is unsolvable without processing large amounts of information and effective communication processes.

Further increase in the volume of information in a larger extent increase the efficiency and relevance of information provision. According to UNESCO, about half of the employed population of most developed countries directly involved in the production and dissemination of information. In some countries, up to half of the national product due to the information activities of society. Computerization has evolved into an important resource, has been a factor of production activities at all levels.

Information resources - intellectual product of the most skilled and creatively active part of the workforce. In the last quarter of the twentieth century information resources reached record volumes so that were introduced the concept of "information explosion", "information revolution." Confirmation of this is to increase the flow of information from the beginning of this century, more than 30 times. Thus the necessary inventions and the use of innovative methods and tools of perception, transmission, processing, storage and dissemination of information that can handle large amounts of information in real time.

Currently it is implemented - based computers an information industry that defined the transition to paperless communication technologies based videophones, facsimile transmission of documents, emails, teleconferences, local and wide area data networks, satellite communications, databases and databases, information retrieval systems, computer workstations.

Due to the growing avalanche of information flows in various spheres of human activity, the question arises, how and by what means can be represented in the computer so varied and numerous information and use it successfully. The most perfect and progressive forms of information and knowledge in computer is a database and knowledge base. Their main task - to provide users with the necessary information, ie the possibility to answer information requests from users to the database and knowledge base in order to obtain the desired information.

By the mid-sixties of the twentieth century in the world used files, with all their shortcomings, to store information. In these "databases" information is often collapsed because of the inability of a large number of simultaneous users, insufficient search.

Since the mid-sixties to 1980 began the use of non-relational databases. Developers and users understand that only use files are very costly to produce, and began to look for ways to solve problems. This was originally developed hierarchical data model, then appeared the network data model, and finally - relational. But all these models have a common problem that lies in the efficient storage and transmission of data.

Because of the large volume of information has become important question of their economic storage and transmission. Interest in the problem of data compression in the database was originally driven by the desire to reduce the amount of physical databases. Price subsystem IO was the bulk of the cost of equipment. Therefore, the proper integration of database techniques lossless data achieved significant savings.

The first section analyzes problematy and tasked to research. The algorithm of Huffman and other most popular data compression algorithms such as LZW, Huffman algorithm, PPM, RLE and others. Formulated the problem of economical storage of

large amounts of data in relational databases. The classification algorithms for data compression algorithms are lossy and without, are examples of each approach.

The second section describes the theoretical aspects of the study. Considered in detail Huffman algorithm, parsed data compression example. Suggestions for modifications Huffman algorithm for more efficient data compression. Built flowchart classic and modified algorithms. Also the basic classes of the developed product, their function. We describe the input data for the experiments.

In the third section reasonably practical applications of the research. We describe the data used for research effective policy modification. The modeling was done using the developed software system that reads a text document, conducts classical coding algorithm Huffman (building code tree and encodes each character) and on its modifications (symbol table splits into two groups according to their frequency, building code tree for each with them, and encodes characters using them), calculates the size of compressed data to the original algorithm and modification.

Described interface and operations of the software, with which these experiments were conducted. It was evaluated the adequacy of the obtained research results, and the results were compared with the classical algorithm and Huffman conclusions about the superiority of modification.

The object of the study - the process of compression of text data using Huffman algorithm and its modifications.

The subject of the study - the efficiency of block-statistical algorithm in comparison with the classical Huffman in compression of text data.

The purpose of the modified Huffman algorithm development is to improve the performance of compression of text data in databases, which will reduce the cost of storage and data transmission.

Developed the modification of the algorithm, created software implementation of the algorithm to run experiments on PC using variety of input data with programming language C #.

As a result of the work we have discussed the Huffman algorithm on example and other most popular compression algorithms such as LZW, PPM, RLE. Proposal for algorithm modification was formulated to get greater efficiency in data

compression. Experiments were held to study algorithm modification and conclusions about its efficiency were made.

Keywords: DATA COMPRESSION METHODS, MODELS, APPROACH, ALGORITHM HUFFMAN BLOCK-STATISTICAL ALGORITHMS, DATABASES.