MINISTERUL EDUCAŢIEI ŞI CERCETĂRII AL REPUBLICII MOLDOVA
Universitatea Tehnică a Moldovei
Facultatea Calculatoare, Informatică şi Microelectronică
Departamentul Ingineria Software şi Automatică

Admis la susţinere
Şef departament:
FIODOROV Ion dr., conf.univ.

„___" _____ 2025

ANALIZA ALGORITMILOR DE COMPRESIE FĂRĂ PIERDERI DE DATE BAZAȚI PE DICȚIONAR

Teza de master

Masterand: _____ **Macovei Nichita, TI-231M**

Coordonator: _____ **Marusic Galina, dr., conf. univ.**

Consultant: _____ **Cojocaru Svetlana, asist.univ.**

Chişinău, 2025

REZUMAT

Teza de master propusă cercetează domeniul algoritimilor de compresie fără pierderi de date, accentuându-se pe algoritmi bazați pe dicționar. Lucrarea este compusă din lista de abrevieri, introducerea, trei capitole, concluzii, bibliografia din 21 de titluri și 8 anexe.

Capitolul "ANALIZA DOMENIULUI PRIVIND ALGORITMI DE COMPRESIE FĂRĂ PIERDERI DE DATE BAZAȚI PE DICȚIONAR" descrie algoritmi care vor fi cercetate – LZ77, LZSS, LZ78, LZW, LZMW și LZAP. Sunt descrise în detalii modul de lucru a metodelor de compresie și decompresie și structurile de date utilizate; sunt evidențiate avantajele și dezavantajele fiecărui algoritm. Capitolul oferă exemple de utilizare practică a algoritmilor, precum aplicații și formate de fișiere. Scopul capitolului este de a oferi o bază teoretică solidă pentru implementarea algoritmilor.

Capitolul "METODE PRIVIND IMPLEMENTAREA ALGORITMILOR DIN DOMENIUL DE STUDIU" detaliază procesul de implementare a algoritmilor în limbajul de programare C++, utilizând o arhitectură modulară bazată pe clase. Limbajul C++ a fost ales pentru eficiența sa și pentru capacitățile extinse de manipulare a fișierelor și datelor. În cadrul aplicației dezvoltate fiecare algoritm este definit printr-o clasă specifică, iar testele sunt realizate printr-o interfață text simplă. Implementarea se concentrează pe crearea unui instrument flexibil, capabil să comprime și să decomprimate fișiere utilizând diferiți algoritmi, măsurând în același timp performanța acestora.

Capitolul "ANALIZA ALGORITMILOR CU APLICAȚIA ELABORATĂ" analizează performanța algoritmilor cu ajutorul aplicației create prin testarea lor pe colecții diferite de fișiere diverse, cum ar fi texte, imagini și fișiere binare. Sunt evaluate rata de compresie și viteza procesului, iar rezultatele sunt comparate pentru a evidenția situațiile în care fiecare algoritm excelează.

ABSTRACT

The proposed master's thesis researches the field of lossless compression algorithms, focusing on dictionary-based algorithms. The work is composed of the list of abbreviations, the introduction, three chapters, conclusions, the bibliography of 21 titles and 8 appendices.

The chapter «DOMAIN ANALYSIS OF DICTIONARY BASED LOSSLESS COMPRESSION ALGORITHMS» describes algorithms that will be studied - LZ77, LZSS, LZ78, LZW, LZMW and LZAP. The working process of the compression and decompression methods and data structures used are described in detail. The advantages and disadvantages of the algorithm are highlighted. The chapter provides examples of practical use of the algorithms, as well as applications and the file formats. The aim of the chapter is to provide a solid theoretical basis for implementing algorithms.

The chapter «METHODS REGARDING THE IMPLEMENTATION OF ALGORITHMS IN THE FIELD OF STUDY» details the process of implementing algorithms in the C++ programming language, using a modular architecture based on classes. The C++ language was chosen for its efficiency and extensive file and data manipulation capabilities. Within the developed application, each algorithm is defined by a specific class, and the tests are performed through a simple text interface. The implementation focuses on creating a flexible tool capable of compressing and decompressing files using different algorithms while measuring their performance.

The chapter «ANALYSIS OF ALGORITHMS WITH ELABORATE APPLICATION» analyzes the performance of algorithms with the help of the application created by testing them on different collections of various files such as texts, images and binary files. Compression rate and process speed are evaluated, and results are compared to highlight situations where each algorithm excels.

CUPRINS

LISTA DE ABREVIERI

LZ – Lempel-Ziv

PNG – acronim PNG is Not GIF

LZSS – Lempel-Ziv-Storer-Szymansky

ASCII – American Standart Code for Information Interchange

ARJ – Archiver by Robert Jung

RAR – Roshal Archiver

LZW – Lempel-Ziw-Welch

GIF – Graphics Interchange Format

LZMW – Lempel-Ziv-Miller-Wegman

LZAP – Lempel-Ziv All Prefixes

OSI – Organizația Internațională de Standartizare

RAM – Random Access Memory; memoria cu acces aleator

MB – megabyte; megaoctet

INTRODUCERE

În epoca contemporană aproape toate aspectele vieții zi-de-zi a unui om involvă, în mod sau altul, interacțiunea cu informația – prognoza meteo, știri, YouTube, navigare prin GPS, lucru cu documente digitale, sisteme de plăți, etc. Din acest motiv perioada curentă a istoriei omenirii se numește "era informațională" sau "era digitală".

Informația este date care au fost procesate, categorizate și stocate. Cât procesarea, atât și stocarea informației este efectuată de dispositive electronice – calculatoare personale, smartphone-uri, tablete, etc. În adiție, aceste dispositive se ocupă și cu transmiterea informației între ele.

Cu evoluția tehnologiilor creștea și necesitatea societății în informație. Dispozitivele de calcul au început să devină mai râspândite – din departamente militare la alte ministere, la instituții științifice, la instituții de învățământ și companii private, și, până la urmă, la utilizatorii personali. Împreună cu creșterea mulțimii de domenii în care s-a început utilizarea calculatoarelor a apărut și necesitatea de dispozitive mai performante, ceea ce a rezultat în evoluție rapidă a dispozitivelor de calcul – în anii 1970-1980 în microprocesoare nou-dezvoltați numărul de tranzistori se dubla aproximativ fiecare doi ani.

Creșterea puterii de calcul a dus la creșterea volumului de date procesate. La aceasta etapă stocarea și transmiterea volumurilor mari de informație a devenit problematic din punct de vedere costurilor și timpului. Pentru rezolvarea acestor probleme a fost inventată compresia datelor.

Compresia este transformarea informației din forma ei inițială în forma care ocupă pe disc un volum mai mic în comparație cu informație originală. Compresia datelor se efectuază de un algoritm care calculează în ce mod informația inițială este codată la compresie.

Pe măsura evoluției tehnico-științifice au fost inventați diferiți algoritmi care sunt utilizați în diferite situații pentru diferite tipuri de date. Acești algoritmi diferă în mai multe caracteristici, însă criteriu care este considerat principal este tipul de compresie – cu sau fără pierderi de date în procesul comprimării.

Lucrarea dată explorează domeniul algoritmilor de compresie a datelor fără pierderi. Prima etapă a cercetării este selecția un număr limitat de algoritmi a căror metoda de comprimare a datelor se bazează pe utilizarea dicționarelor. Astfel de algoritmi parcurg date de intrare, le separă în subșiruri care sunt înscrise într-o structură de date – dicționar, care se folosește în proces de comprimare și decomprimare.

A doua etapă – implementarea algoritmilor selectate în cod. La aceasta etapă va fi realizat un program care va conține aceste algoritmi și va putea să comprime și să comprime fișiere și director.

Ultima etapă – analiza algoritmilor implementați. Aceasta va fi realizată prin comprimarea și decomprimarea unui set de fișiere de control. Etapa va conține interpretarea rezultatelor și tragerea concluziilor.

BIBLIOGRAFIA

[1] P. Deutsch, 'DEFLATE Compressed Data Format Specification version 1.3', RFC Editor, RFC1951, May 1996. doi: 10.17487/rfc1951.

[2] K. Sayood, *Introduction to data compression*, 3rd ed. in Morgan Kaufmann series in multimedia information and systems. Amsterdam Boston: Elsevier, 2006.

[3] D. Salomon, *Data compression: the complete reference*, 3. ed. New York Berlin Heidelberg: Springer, 2004.

[4] J. Ziv and A. Lempel, 'A universal algorithm for sequential data compression', *IEEE Trans. Inform. Theory*, vol. 23, no. 3, pp. 337–343, May 1977, doi: 10.1109/TIT.1977.1055714.

[5] J. A. Storer and T. G. Szymanski, 'Data compression via textual substitution', *J. ACM*, vol. 29, no. 4, pp. 928–951, Oct. 1982, doi: 10.1145/322344.322346.

[6] 'Allegro 5 on GitHub - lzss.c'. [Online]. Available: https://github.com/liballeg/allegro5/blob/4.4/src/lzss.c

[7] 'kext_tools on GitHub - compression.c'. [Online]. Available: https://github.com/opensource-apple/kext_tools/blob/bc71a85/compression.c

[8] 'LZ-78 - Data Compression'. [Online]. Available: https://www.stringology.org/DataCompression/lz78/index_en.html

[9] S. M. Choudhary, A. S. Patel, and S. J. Parmar, 'Study of LZ77 and LZ78 Data Compression Techniques', vol. 4, no. 3, 2015.

[10] 'LZW - Wikipedia'. [Online]. Available: https://en.wikipedia.org/wiki/Lempel%E2%80%93Ziv%E2%80%93Welch

[11] V. S. Miller and M. N. Wegman, 'Variations on a theme by Ziv and Lempel (data compression)', in *IEEE International Conference on Communications, - Spanning the Universe.*, Philadelphia, PA, USA: IEEE, 1988, pp. 390–394. doi: 10.1109/ICC.1988.13597.

[12] J. A. Storer, *Data Compression: Methods and theory*. in Principles of computer science series, no. 13[a]. Rockville,Md: Computer Science Press, 1988.

[13] B. Stroustrup, *The C++ programming language*, 3rd ed. Reading, Mass: Addison-Wesley, 1997.

[14] *ISO/IEC 14882:2024*. [Online]. Available: https://www.iso.org/standard/83626.html

[15] 'Compressor on GitHub'. [Online]. Available: https://github.com/Kotbenek/Compressor/tree/main/src

[16] 'lzss on GitHub - lzss.cpp'. [Online]. Available: https://github.com/tiagotmartinez/lzss/blob/main/cpp/lzss.cpp

[17] 'Fowler-Noll-Vo hash function - Wikipedia'. [Online]. Available: https://en.wikipedia.org/wiki/Fowler%E2%80%93Noll%E2%80%93Vo_hash_function

[18] T. Bell, I. H. Witten, and J. G. Cleary, 'Modeling for text compression', *ACM Comput. Surv.*, vol. 21, no. 4, pp. 557–591, Dec. 1989, doi: 10.1145/76894.76896.

[19] 'The University of Auckland - Calgary Corpus'. [Online]. Available: https://corpus.canterbury.ac.nz/descriptions/#calgary

[20] R. Arnold and T. Bell, 'A corpus for the evaluation of lossless compression algorithms', in *Proceedings DCC '97. Data Compression Conference*, Snowbird, UT, USA: IEEE Comput. Soc. Press, 1997, pp. 201–210. doi: 10.1109/DCC.1997.582019.

[21] 'The University of Auckland - Large Corpus'. [Online]. Available: - https://corpus.canterbury.ac.nz/descriptions/#large