Інформаційна технологія для дослідження методів ущільнення даних

Лужецький В. А.[1], Михалевич О. В.[2]

[1]Проф., д.т.н., завідувач кафедри захисту інформації, Вінницький національний технічний університет
вул. Хмельницьке шосе, 95, м. Вінниця, Україна
[2]Старший інженер-програміст в Playtika, м. Київ, Україна, mikhal.alex@gmail.com

***Анотація* — Обгрунтовано необхідність створення інформаційної технології для дослідження методів ущільнення даних. Запропоновано узагальнену модель процесу ущільнення даних, що включає правила моделювання вихідної послідовності даних, їх подальше кодування та формування структури послідовності ущільнених даних. Для дослідження особливостей і характеристик процесу ущільнення даних запропоновано теоретико-множинну модель, що описує відповідну інформаційну технологію. Наведено структурно-функціональну модель інформаційної технології.**

Ключові слова: ущільнення даних,-інформаційна технологія, модель, процедура.

Information technology for the study of data compression methods

Luzhetsky V. A., Mykhalevych O. B.

[1] Prof., Head of Department of Information Protection, Vinnytsia National Technical University
Khmelnytske shose., 95, Vinnytsia, Ukraine,
[2]Senior Software Engineer at Playtika, Kiev, Ukraine, mikhal.alex@gmail.com

***Abstract* — The need for the creation of information technology to study the data compression methods is rejustified. The generalized model of the data compression process is proposed. This model includes the original data sequence modeling rules and their further formation sequence encoding the compressed data structure. To study the features and characteristics of the data compression process suggested set-theoretical model describing the appropriate information technology. Structural and functional model of information technology is given.**

Keywords: data compressors, information technology, model, procedure.

ВСТУП

Типовою є ситуація, коли розробники нових алгоритмів ущільнення даних вимушені створювати свої власні програмні засоби для дослідження цих алгоритмів. Крім того, порівняння різних алгоритмів бажано здійснювати на основі єдиної програмної реалізації базових складових алгоритмів. Тільки у цьому випадку може бути забезпечена коректність отриманих порівняльних оцінок.

Тому доцільно розробити інструментарій для дослідження відомих та нових методів ущільнення даних. Цей інструментарій реалізує інформаційну технологію для дослідження методів ущільнення даних (ITRC - Information technology for the research of data compression).

УЗАГАЛЬНЕНА МОДЕЛЬ ПРОЦЕСУ УЩІЛЬНЕННЯ ДАНИХ

Пропонується така узагальнена модель процесу ущільнення даних

$$\mathbf{DC} = \left\{\mathbf{P}, \mathbf{A}, R_M, R_C, \mathbf{P}_M, \mathbf{P}_C, \mathbf{P}^*, \mathrm{S}\right\},$$

де $\mathbf{P}$ - вихідна послідовність даних, що складається з символів алфавіту $\mathbf{A} = \{a_0, \mathrm{a}_1, \ldots \mathrm{a}_{n-1}\}$;

n - потужність алфавіту;

R_M - правило моделювання джерела даних;

R_C - правило кодування даних;

$\mathbf{P}_M$ - послідовність даних, що є результатом моделювання;

$\mathbf{P}_C$ - послідовність даних, що є результатом кодування;

P* - послідовність ущільнених даних;

S - правило формування структури послідовності **P***.

Згідно з цією моделлю вихідна послідовність символів $\mathbf{P}$ перетворюється в послідовність $\mathbf{P}_M$ з використанням правила моделювання R_M. Це описується відображенням $\mathbf{P} \xrightarrow{R_M} \mathbf{P}_M$.

Послідовність $\mathbf{P}_M$ кодується з використанням правил кодування R_C. Правилу кодування відповідає відображення $\mathbf{P}_M \xrightarrow{R_C} \mathbf{P}_C$.

Формування структури послідовності **P*** з вказівкою додаткової інформації, необхідної для відновлення послідовності **P**, здійснюється на підставі правила S. Таким чином процес ущільнення даних описується композицією відображень

$$DC = R_M \circ R_C \circ S .$$

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ТА ЇЇ СТРУКТУРНО-ФУНКЦІОНАЛЬНА МОДЕЛЬ

Для дослідження особливостей і характеристик процесу ущільнення даних, що відповідає певному алгоритму, пропонується інформаційна технологія, яка описується такою моделлю

$$\mathbf{ITRC}=\{\mathbf{D}, \mathbf{M}, \mathbf{C}, \mathbf{F}, \mathbf{CH}\}$$

де **D** – множина ущільнюваних даних;

M={M_1, M_2, … } – множина алгоритмів моделювання джерела даних;

C={C_1, C_2, … } – множина алгоритмів кодування даних;

F={F_1, F_2, … } – множина функцій конструювання методу ущільнення;

CH – множина характеристик методу ущільнення даних.

Множину алгоритмів моделювання джерела даних складають такі алгоритми: M_1 – перетворення Барроуза-Уїлера або BW-перетворення (BWT) [1], [2]; M_2 – алгоритм "купи книжок" (Move To Front - MTF) [3]; M_3 – алгоритм формування числової моделі [4]; M_4 - алгоритми контекстного моделювання [5-8] та ін..

Множина алгоритмів кодування складається з алгоритмів, що забезпечують формування кодів Хаффмана [9], Еліаса [12], Голомба [13], Фібоначчі [2], [14], [15], а також алгоритм арифметичного кодування [10], [11].

Функція конструювання F_i методу ущільнення описує послідовність реалізації процедур моделювання та кодування.

Як характеристики методу ущільнення даних використовуються: коефіцієнт ущільнення даних; час ущільнення даних; обсяг оперативної пам’яті, що використовується для зберігання проміжних даних та програмного коду, що реалізує метод ущільнення.

Описаній теоретико-множинній моделі інформаційної технології відповідає структурно-функціональної модель, що зображена на рис. 1.

Джерело даних має дві складові. Перша – це набір файлів *Canterbury Corpus* [16], Що використовується для універсальних архіваторів. Друга складова – це програмний засіб, що генерує дані з заданим законом розподілу.

Процедура введення даних передбачає вибір конкретних даних, що підлягають ущільненню та передачу їх для подальших перетворень.

Процедури моделювання і кодування передбачають реалізації перетворень за алгоритмами, що вибираються з відповідних баз за допомогою процедури конструювання методу ущільнення.

Визначення характеристик моделі джерела даних і коефіцієнта ущільнення забезпечуються відповідними процедурами.

Процедура візуалізації результатів досліджень забезпечує їх подання у формах, зручних для сприйняття людиною.

ВИСНОВКИ

Запропонована інформаційна технологія для дослідження методів ущільнення даних надасть можливість досліджувати відомі та нові методи ущільнення із забезпеченням коректності отриманих порівняльних характеристик.

REFERENCES

[1] Burrows M. A Block-sorting Lossless Data Compression Algorithm [Text] / M. Burrows, D. J. Wheeler // SRC Research Report 124. - Palo Alto: Digital Systems Research Center, 1994. – 18 p.

[2] Bastys, R. Fibonacci Coding Within the Burrows-Wheeler Compression Scheme [Text] / R. Bastys // Electronics and Electrical Engineering. – Kaunas: Technologija, 2010. – № 1(97). – P. 28–32.

[3] Рябко Б. Я. Сжатие данных с помощью стопки книг [Текст] / Б. Я. Рябко // Проблемы передачи информации. - 1980. - Т. 16, Вып. 2. - С. 16-21.

[4] Лужецький В. А. Розробка та дослідження методів адаптивного ущільнення даних на основі лінійної форми фібоначчі [Текст] / В. А. Лужецький, Л. А. Савицька // Східно-європейський журнал передових технологій. – 2015. – №1/9 (73). – С. 16-22.

[5] Cleary J. G. Data Compression Using Adaptive Coding and Partial String Matching [Text] / J. G. Cleary, I. H. Witten // IEEE Trans. Commun. – 1984. – V. 32, №4. – P. 396–402.

[6] Solving the Problems of Context Modeling by Charles Bloom [Електронний ресурс]. Режим доступу: http://www.cbloom.com/papers/ppmz.pdf

[7] Moffat A. Implementing the PPM Data Compression Scheme [Text] / A. Moffat // IEEE Transactions on Communications. - 1990. – V. 38, № 11. – P. 1917-1921.

[8] Bell T. Modeling for Text Compression [Text] / T. Bell, I. Witten, J. Cleary // ACM Computing Surveys. - 1989. - V. 21, №4. - P. 557-591.

[9] Huffman, D. A. A Method for the Construction of Minimum-Redundancy Codes [Text] / D. A. Huffman // Proceedings of the Institute of Electrical and Radio Engineers. – 1952. – V. 40, №9. – P. 1098–1101.

[10] Witten, I. Arithmetic Coding for Data Compression [Text] / I. Witten, R. Neal, J. Cleary // Communications of the ACM. – 1987. – V. 30, № 6. – P. 520–540.

[11] Langdon, G. G. An introduction to arithmetic coding [Text] / G. G. Langdon // IBM J. Res. Dev. – 1984. – V. 28, №2. – P. 135-149.

[12] Elias, P. Universal codeword sets and representations of the integers [Text] / P. Elias // IEEE Trans. Inf. Theory. – 1975. – V. IT–21, №2. – P. 194–203.

.

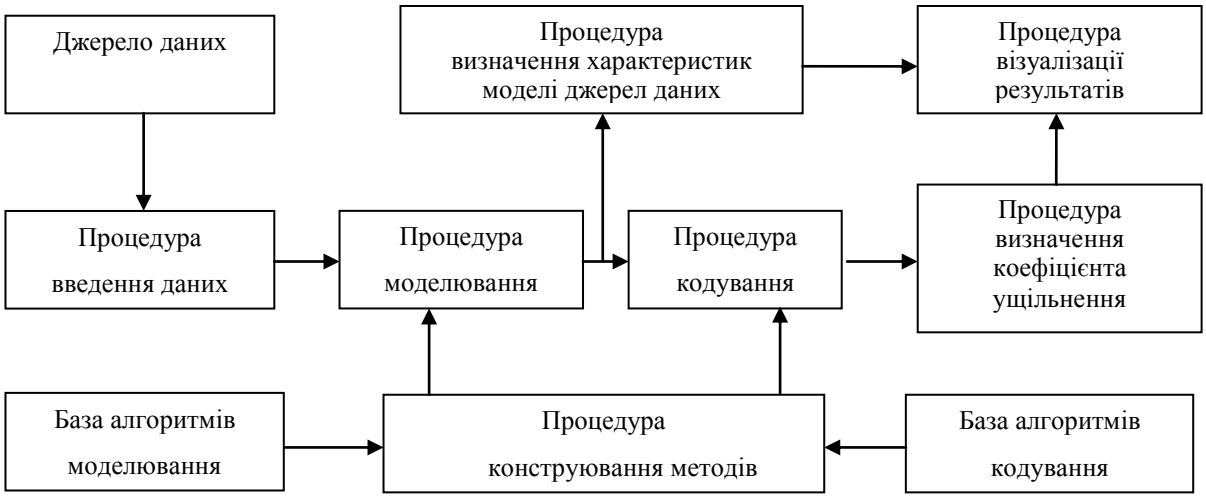

Рисунок 1 Структурно-функціональна модель інформаційної технології

[13] Golomb, S. W. Run-length encodings [Text] / S. W. Golomb // IEEE Transactions on Information Theory. – 1966. – V. 12, № 3. – P. 399–401.

[14] Стахов А. П. Коды золотой пропорции [Текст] / А. П. Стахов. - М.: Радио и связь, 1984. - 150 с.

[15] Лужецький В. А. Високонадійні математичні Фібоначчі-процесори: Монографія [Текст] / В. А. Лужецький. - Вінниця: Універсум-Вінниця, 2000. - 248 с.

[16] The canterbury corpus: [Електронний ресурс]. – Режим доступу: http://corpus.canterbury.ac.nz/descriptions/