Харківський національний університет радіоелектроніки

Зубань Юрій Олександрович

УДК 004.627

МОДЕЛІ І ЗАСОБИ СТИСКУ ДАНИХ В ІНФОРМАЦІЙНИХ СИСТЕМАХ

05.13.06 - автоматизовані системи управління
та прогресивні інформаційні технології

Автореферат дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків - 2004

Дисертацією є рукопис.

Робота виконана на кафедрі електроніки і комп'ютерної техніки Сумського державного університету Міністерства освіти і науки України.

Науковий керівник – доктор технічних наук, професор
Борисенко Олексій Андрійович,
Сумський державний університет,
завідувач кафедри електроніки і комп'ютерної техніки

Офіційні опоненти:

д-р техн. наук, професор Аліпов Микола Васильович, Харківський національний університет радіоелектроніки, професор кафедри електронних обчислювальних машин;

канд. техн. наук, доцент Губка Сергій Олексійович, Національний аерокосмічний університет ім. М.Є. Жуковського "ХАІ", заст. зав. кафедри інформаційних управляючих систем

Провідна установа –
Херсонський державний технічний університет, кафедра інформаційних технологій, Міністерство освіти і науки України, м. Херсон.

Захист відбудеться 03.11.2004 р. о 14 год. на засіданні спеціалізованої вченої ради Д 64.052.01 у Харківському національному університеті радіоелектроніки за адресою: 61166, м. Харків, пр. Леніна,14, тел. (057)-7021-451.

З дисертацією можна ознайомитися в бібліотеці Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, пр. Леніна, 14.

Автореферат розісланий 01.10.2004р.

Вчений секретар
спеціалізованої вченої ради В. М. Левикін

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Науково-технічний прогрес суспільства нерозривно пов'язаний з розвитком і впровадженням інформаційних технологій управління. Україна, будучи частиною світової суспільно-економічної системи, також реалізує програму загальної інформатизації. Будь-яка система управління, у тому числі автоматизована, не може працювати без інформації про стан керованого об'єкта і зовнішнього середовища, без передачі інформації про прийняті керуючі впливи. У технічних системах джерелом інформації про стан керованого процесу або об'єкта є, як правило, технічні засоби – вимірювальні прилади, датчики і перетворювачі. В АСУ – це, головним чином, системи нижчого рівня управління. Визначення оптимальних обсягів інформації, що надходить у різні органи управління, і розподілу потоків інформації в часі і просторі, – необхідна умова ефективного функціонування АСУ.

Розвиток вищезгаданих інформаційних мереж і систем привів до істотного зростання інформаційних потоків між територіально розміщеними джерелами й одержувачами повідомлень. Підвищення ефективності систем управління досягається за рахунок застосування відповідних економіко-математичних методів, а також використання технічних засобів збору, передачі, збереження й обробки інформації. Для підвищення ефективності використання комунікаційних та інформаційно-обчислювальних ресурсів зазначених систем важливу роль відіграють методи скорочення надмірності, що забезпечують стиск обсягу інформації, яка передається чи запам'ятовується. Це дозволяє істотно розвантажити канали зв'язку та системи обробки і збереження даних за рахунок виключення надлишкових повідомлень, що еквівалентно підвищенню пропускних здатностей інформаційних систем або збільшенню ємності запам'ятовувальних пристроїв.

Існує багато вагомих причин виділяти ресурси АСУ з розрахунку на стисле представлення, тому що більш швидка передача даних і скорочення простору для їх збереження дозволяють зберегти значні ресурси і поліпшити показники АСУ. Розроблення ефективних засобів стиску даних є частиною основних вимог до інформаційного забезпечення АСУ.

Розроблення таких способів представлення даних, при яких простір витрачається більш ощадливо, були започатковані в сорокових роках минулого століття. К. Шеннон розробив більшість базових понять, Д. Хаффманом був запропонований алгоритм одержання оптимального коду. У 1977 р. А. Лемпель і Я Зив запропонували ідею словникових методів стиску. Значний внесок у розвиток методів стиску зробили також Кричевський Р.Е., Рябко Б.Я., Амелькін В.А. та інші вчені.

Однак наявні розв’зання задачі стиску інформації для застосування в АСУ відрізняються, з одного боку підвищеною складністю алгоритмів і відповідно складністю апаратної реалізації, а з іншого – недостатнім ступенем стиску для ряду задач. Тому задача даної дисертаційної роботи, обумовлена необхідністю розроблення нових моделей, методів і алгоритмів стиску даних, є акту-

альною.

Зв'язок роботи з науковими програмами, планами, темами. Робота виконувалася згідно з планом науково-дослідних робіт Сумського державного університету в рамках держбюджетної теми: № 0197U016602 "Разработка алгоритмов синтеза и обработки двумерных изображений на основе метода срезов и локальных окон с использованием биномиальных систем счисления", над якою автор працював, обіймаючи посаду молодшого наукового співробітника.

Мета і задачі дослідження. Метою роботи є підвищення швидкості передачі даних у підсистемах АСУ і зменшення ємності пам'яті для їх збереження.

нен стиль структуры: Двухбайтовая
ета і задачі дослідження
Формулюють мету роботи і задачі, які необхідно вирішити для досягнення поставленої мети Не слід формулювати мету як “Дослідження ”, “Вивчення ”, тому що ці слова вказують на засіб досягнення мети, а не на саму мету

Для досягнення поставленої мети в роботі вирішуються такі задачі:

- розроблення математичних моделей джерел інформаційних масивів;
- розроблення методу стиску масивів кінцевих двійкових повідомлень;
- синтез алгоритмів кодування джерел структурної інформації та їх оцінка;
- синтез ефективних програм і пристроїв стиску на основі розроблених алгоритмів.

Об'єкт дослідження – засоби стиску для інформаційних задач АСУ.

Предмет дослідження – моделі, методи і алгоритми стиску даних в інформаційних системах.

Методи дослідження базуються на положеннях теорії інформації, теорії кодування, теорії імовірностей, методологіях аналітичного та імітаційного моделювань.

Наукова новизна одержаних результатів. У ході вирішення поставлених задач автором особисто були отримані такі результати:

1. Вперше запропонована і досліджена математична модель відносної адресації для опису бінарного комбінаторного джерела інформації, що дозволяє описати рівноймовірні двійкові повідомлення імовірнісним джерелом з детермінованими параметрами. Це дає можливість використовувати для кодування відомі методи усунення статистичної надмірності.

2. Дістала подальшого розвитку математична модель бернуллiївського джерела інформаційних масивів кінцевих двійкових повідомлень. Модель дозволяє перейти до двох взаємозалежних джерел інформації, роздільне кодування яких дозволяє усунути структурну і ймовірнісну надмірності вхідних повідомлень. Крім того, роздільне кодування джерел інформації дає можливість використовувати модель для систем із захистом даних від несанкціонованого доступу.

3. Уперше розроблений і досліджений метод локальних зсувів для стиску інформаційних масивів, що дозволяє більш ефективно усувати надмірність у двійкових повідомленнях. Метод дозволяє виділити з повідомлень надмірність різного роду і застосувати для усунення кожного з них відповідний алгоритм.

Практичне значення одержаних результатів. Практична цінність роботи полягає в тому, що:

- розроблені і досліджені алгоритми кодування джерел інформації на основі методу локальних зсувів, що базуються на комбінаторному розкладанні масивів на класи еквівалентності;

- розроблено структури програм і пристроїв стиску на основі методу локальних зсувів. Апаратна реалізація засобів стиску дозволяє істотно підвищити швидкодію і надійність їх роботи;
- розроблено програми стиску графічних даних і двійкових послідовностей на основі методу локальних зсувів. Використання цих програм дозволило зробити експериментальну оцінку ефективності стиску даних запропонованим у роботі методом.

Результати дисертаційної роботи використані на Науково-виробничому колективному підприємстві "Преобразователь", м. Суми, а також у Сумському державному університеті в навчальному процесі з дисциплін "Методи і засоби стиску даних в інформаційних системах", "Системи передачі даних", "Системи відображення інформації" для дослідження методів обробки зображень, стиску даних.

Особистий внесок здобувача. Всі результати дисертаційної роботи одержані автором самостійно. У роботах, написаних у співавторстві, особисто автору належать: у [3] розроблені імітаційні моделі та проведений аналіз алгоритмів оптимального нерівномірного кодування у методі локальних зсувів; у [1] застосовано ОНК для методу локальних зсувів; у роботі [6] –математичний опис алгоритму стиску даних у методі локальних зсувів та синтез алгоритму кодування адрес символів з використанням комбінаторного розкладання на класи еквівалентності; у роботі [9] – модель відносної адресації та виведення формули для визначення імовірностей відносних адрес символів; в [7,10] – застосування методів стиску в прикладних задачах обробки зображень; у [8, 14] – аналіз умов застосування методів рівномірного і нерівномірного кодування; у [13] – розроблення алгоритму кодування.

Апробація результатів дисертації. Основні результати дисертації доповідалися та обговорювалися на конференціях:

- 4-му та 5-му Міжнародних форумах "Радіоелектроніка і молодь у XXI ст."(Харків, ХТУРЕ, 2000 р., 2001 р.);
- Міжнародній науково-технічній конференції молодих вчених "Оптоелектронні інформаційно-енергетичні технології" (Вінниця, ВДТУ, 2001 р.);
- Міжнародній науково-технічній конференції "Контроль і управління в складних системах" (Вінниця, ВДТУ, 2001 р.);
- Міжнародній науковій конференції "Современные методы кодирования в электронных системах" (Суми, СумДУ, 2002 р.);
- The European Material Conference. E-MRS 1999 Spring Meeting. – Strasbourg (France);
- Науково-технічній конференції викладачів, співробітників, аспірантів та студентів фізико-технічного факультету Сумського державного університету (Суми, 1999 – 2001).
- на наукових семінарах Сумського державного університету.

Публікації. Результати дисертації опубліковані в 14 наукових працях (6 наукових статтях у виданнях, перелік яких затверджений ВАКом України, та 8 тезах доповідей і праць конференцій).

Структура та обсяг дисертації. Робота складається із вступу, 4 розділів, загальних висновків, додатків. Загальний обсяг дисертації – 169 сторінок, у тому числі 130 сторінок основного тексту. Список використаних джерел містить 104 найменування. Кількість додатків – 6. Кількість рисунків і таблиць – відповідно 49 і 4.

ОСНОВНИЙ ЗМІСТ РОБОТИ

Вступ до дисертації містить обґрунтування актуальності теми, визначення об'єкта і предмета дослідження, формулювання мети і задач роботи, опис основних наукових результатів, їх новизни, достовірності та практичної цінності, а також відомості про публікації, впровадження, апробацію і структуру роботи.

У першому розділі на підставі аналізу літературних джерел, відомих теоретичних положень і технічних рішень обговорюються шляхи підвищення ефективності передачі даних в інформаційних системах, обґрунтовується вибір предмета дослідження і формулюються задачі дослідження.

Проаналізовані основні математичні моделі, які використовуються для опису даних різного типу. Обґрунтовано вибір моделі Бернуллі, коли між символами в повідомленнях немає взаємозв'язків, як більш простої та зручної для застосування в інформаційних системах АСУ:

$$X = \{x_i, i = 1 \ldots |X|\},$$

$$\mathrm{P}(X) = \{p(x_i), i = 1 \ldots |X|\},$$

де X – алфавіт джерела інформації;

$\mathrm{P}(X)$ – множина імовірностей символів $p(x_i)$.

Перевагою даної моделі є простота і можливість переходу від неї до двох взаємозалежних джерел інформації, кодування яких може бути більш ефективне.

Проаналізовано існуючі методи стиску в АСУ, такі, як алгоритми оптимального нерівномірного кодування (ОНК) Хаффмана і Шеннона-Фано, арифметичне кодування, словникові та комбінаторні методи стиску. Будь-який метод стиску може бути описаний як процес кодування джерела інформації, який являє собою взаємно однозначне відображення $f : A \to V$, де $A = \{a_i\}$ – множина повідомлень, що стискаються, $V = \{v_i\}$ – результуюча множина двійкових слів $v_i = f(a_i)$. Причому стиск має місце, коли $|v_i| < |a_i|$.

Зроблено висновок, що з точки зору обробки різнопланової інформації, яка передається в АСУ, кращими є комбінаторні методи стиску, які відрізняються високою ефективністю і простотою та досить привабливі для передачі і збереження повідомлень, особливо на етапі, коли ще не виявлені

імовірнісні характеристики джерела інформації. Це має велике практичне значення для застосувань в АСУ, а необхідність розвитку цих методів стиску і відповідних моделей і алгоритмів визначає предмет і задачу даної роботи.

Другий розділ присвячений розробленню математичної моделі джерела масивів двійкових даних. У цій роботі масив даних, що стискається, розглядається як прямокутна матриця $\mathbf{A}=\{a_{ji}: j=1\ldots h, i=1\ldots n\}$ з n стовпцями і h рядками. Джерело повідомлень різного типу досить точно може бути математично описане за допомогою моделі Бернуллі.

Математична модель бернуллівського джерела двійкових послідовностей допускає представлення у вигляді двох взаємозалежних джерел: комбінаторного S і ймовірнісного B. Сутність такого перетворення полягає в тому, що відправна множина з 2^n двійкових повідомлень розбивається на $(n+1)$ класів еквівалентності, в яких утримуються повідомлення з однаковою кількістю одиниць. Представником класу еквівалентності є число k одиниць у двійкових кодових комбінаціях a_i.

Ентропія бернуллівського джерела інформації A дорівнює сумі ентропій імовірнісного B і комбінаторного S джерел:

$$H(A)=H(B)+H(S/B)$$

Із збільшенням довжини комбінацій n частина інформації, що вміщується в структурі повідомлення, зростає. При значній довжині повідомлень ентропія комбінаторного джерела S значно більша за ентропію імовірнісного джерела B.

З цього випливає, що ступінь стиску здебільшого залежить від оптимального кодування саме комбінаторного джерела S. При великій кількості можливих повідомлень кодування комбінаторного джерела відомими методами ускладнено. Це дає передумови для представлення двійкових комбінаторних послідовностей у вигляді, придатному для розроблення нових методів кодування.

Представлення повідомлень в адресному вигляді дозволяє досягти стиску. Адресна послідовність – це множина адрес позицій двійкової послідовності, у яких розміщені одиниці. Очевидно, що кількість адрес дорівнює кількості одиниць у послідовності і ступінь стиску збільшується з їх зменшенням. Запропоновано перехід до адресної форми представлення послідовностей. На відміну від існуючого абсолютного адресного способу представлення інформації про розміщення символів послідовності запропоновано новий – відносний (рис. 1).

Відносна форма адресації усуває природну надмірність абсолютних адрес Z_i^*, визначаючи лише відстань Z_i між сусідніми одиницями. При цьому за область, що адресується, використовується тільки та локальна частина слова, в якій можливе розташування даного символа замість усієї послідовності, як при абсолютній адресації.

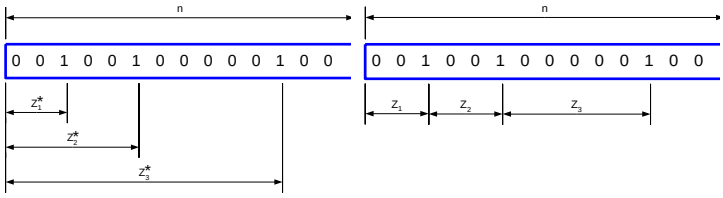

Рис. 1. Абсолютна і відносна адресація символів:

$$Z_i^* \in [1; n], \tag{1}$$

$$Z_3 \in (Z_1 + Z_2; n]. \tag{2}$$

Для визначення цієї області множина адресних послідовностей з параметрами n і k розбивається на підкласи еквівалентності, ознаками яких є значення першої адреси. Кількість підкласів еквівалентності та їх потужностей обмежена і визначається з параметрів n і k. Кожний з отриманих підкласів еквівалентності також може бути розбитий на підкласи. У результаті кожен підклас визначає можливі значення відповідних йому адрес. Це дає можливість аналізувати послідовність локально в межах підкласу.

Наслідком цього є те, що в кожної s-ї одиниці у слові a_i є діапазон вірогідних розміщень (ДВР) – множина позицій у послідовності, у яких можливе її розміщення. Ширина (потужність) ДВР N_s – це максимально можлива відстань між $(s-1)$ і s одиницею – кількість підкласів еквівалентності на даному етапі розбиття.

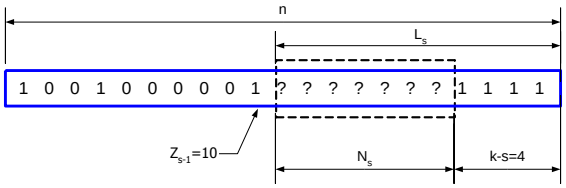

Рис.2. Діапазон вірогідних розміщень (ДВР)

Ширина ДВР визначається виразом

$$N_s = n - Z_{s-1} - (k - s), \tag{3}$$

де Z_{s-1} – абсолютна адреса попередньої одиниці.

При числі одиниць у послідовності більше 1 відбуваються локальні зсуви – визначення і перерахування параметрів ДВР для кожної наступної одиниці і подальше кодування значення адреси.

Доведено твердження, що ймовірності значень відносних адрес визначаються виразом

$$p(m, s, L_s) = C_{L_s - m}^{k-s} \Big/ C_{L_s}^{k-s+1}, \tag{4}$$

де m – адреса позиції символа в діапазоні його можливих розміщень; s – номер одиничного символа; k – кількість одиниць у послідовності; L_s – число позицій послідовності, в яких можуть розміщуватися одиниці з s-ї по k-ту.

Вираз **Ошибка! Источник ссылки не найден.** є відношенням потужності множини адресних послідовностей зі значенням m s-ї адреси до потужності множини можливих адресних послідовностей з $k-s+1$ одиниць, адреси яких $Z_i \in [1; L_s]$. Характер залежності ймовірності відносних адрес від m, s та L_s наведено на графіку (**Ошибка! Источник ссылки не найден.**).

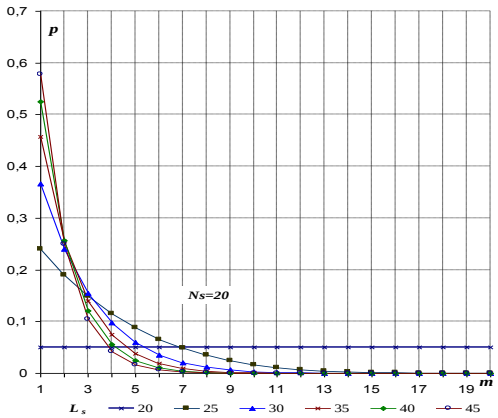

Рис. 3. Імовірності відносних адрес при $N_s = 20$ і різних значеннях m і L_s

Для апаратної реалізації алгоритму формулу можна подати в рекурентному вигляді, що значно спрощує процес обчислення, особливо при великих n. Доведено, що

$$p(1, s, L_s) = \frac{(k-s+1)}{L_s}, \tag{5}$$

$$\frac{p(m, s, L_s)}{p(m-1, s, L_s)} = \frac{(L_s - m - k + s + 1)}{(L_s - m + 1)}. \tag{6}$$

З **Ошибка! Источник ссылки не найден.** і **Ошибка! Источник ссылки не найден.** випливає, що розподіл імовірностей відносних адрес не завжди є рівномірним, отже, відносна адреса може містити надмірність, яку можна усунути методами ОНК.

Відомо також, що ефективність стиску можна помітно підвищити, якщо перед кодуванням перетворити повідомлення, перетворивши його з послідовності символів із сильними статистичними зв'язками в послідовність символів зі слабкими статистичними зв'язками, тобто виконати декореляцію.

Для визначення ступеня кореляції використаний параметр ρ, який дорівнює частці символів, що збіглися, у двох сусідніх рядках. Якщо позначити поточний рядок як $\mathbf{a}=\{a_i : i=1\ldots n\}$, а попередній як $\mathbf{b}=\{b_i : i=1\ldots n\}$, то

$$D_i^j = k'P^m, \qquad (7)$$

де $\mathbf{d}$ – рядок, отриманий у результаті порівняння двох рядків $\mathbf{a}$ і $\mathbf{b}$ за допомогою операції логічної рівнозначності послідовно до всіх їх елементів $A_{эф}=A_{Ra}+1{,}3A_{Th}+0{,}086A_K$. Для двовимірного інформаційного масиву параметр ρ визначається як середнє значення для всіх рядків.

Для усунення надмірності, викликаної збігом елементів сусідніх рядків, запропоновано алгоритм, що базується на операції логічної нерівнозначності. Результатом є рядок $\mathbf{c}=\{a_i \oplus b_i : i=1\ldots n\}$ (

Рис. 4. Декореляція рядків ($\rho=\frac{11}{14}$)

Рис.5. Вхідне і вихідне зображення при декореляції рядків ($\rho=99\%$)

У результаті декореляції рядків кількість k одиниць у них зменшується до 0 при $\rho=1$ або збільшується до n при $\rho=0$. У загальному випадку число одиниць після декореляції дорівнює

$$k=n(1-\rho). \qquad (8)$$

Як випливає з **Ошибка! Источник ссылки не найден.** і **Ошибка! Источник ссылки не найден.**, при високому ступені кореляції даних число одиниць після декореляції істотно зменшується, що підвищує ступінь стиску. Аналогічний підхід можна застосувати не тільки до рядків, але і до стовпців інформаційного масиву.

Отримані вище моделі дозволили розробити метод локальних зсувів для стиску двійкових масивів даних. Суть методу полягає в застосуванні декореляції для усунення взаємозв'язків між елементами інформаційного масиву (може бути пропущене за відсутності кореляції); побудові комбінаторної та ймовірнісної моделей за параметрами n і k; оптимальному кодуванні ймовірнісного і комбінаторного джерел інформації; кодуванні двійкових комбінаторних послідовностей як послідовностей відносних адрес менш ймовірних символів.

Вхідний масив, що складається з h рядків a_i кінцевої довжини n, стискається шляхом кодування кожного з рядків (**Ошибка! Источник ссылки не найден.**). Кожен рядок a_i подається у вигляді адрес 1 або 0. Якщо $k>n-k$, то зазначають адреси нулів, у іншому випадку – адрес одиниць. Для усунення структурної надмірності відносних адрес використовуються методи ОНК.

Рис. 6. Формат вхідних і стиснутих даних

Для підвищення швидкодії методу локальних зсувів можна аналізувати значення кількості одиниць у послідовностях і при значеннях, що близькі і дорівнюють $k = \frac{n}{2}$, не проводити кодування комбінаторного джерела. Ширина діапазону значень k, при яких кодування не виконується $\left[\frac{n}{2} - \mu; \frac{n}{2} + \mu\right]$, визначається значенням μ і задається користувачем.

У третьому розділі проведено аналіз алгоритмів синтезу ОНК на основі розроблених програмних моделей. Результат аналізу виявив, що для застосування в методі локальних зсувів ефективніше використовувати алгоритм побудови кодового дерева „згори-вниз", який застосовано і в методі Шеннона-Фано. На основі його модифікації і враховуючи характерну для методу локальних зсувів функціональну залежність для імовірностей відносних адрес, розроблено алгоритм кодування джерела відносних адрес.

Один з етапів алгоритму – розподіл множини відносних адрес символа на дві підмножини з близькими сумарними імовірностями. Для зручності замість множини дробових чисел – імовірностей, обчислених за формулою **Ошибка! Источник ссылки не найден.**, використано множини цілих чисел – чисельники з виразу **Ошибка! Источник ссылки не найден.**:

$$\Lambda = \left\{\lambda_i : \lambda_i = C_{q-i}^{v-1}, i = 1 \ldots (q - v + 1)\right\}, \tag{9}$$

де $q = L_s$, $v = k - s + 1$.

Параметри q і v визначають елементи множини Λ і залежать від параметрів k, s, L_s одиниці, що кодується. Оскільки знаменники у всіх імовірностей однакові, то значення елементів множини Λ являють собою імовірності відповідних їм адрес.

Параметр φ у процесі розподілу визначає відношення сумарної імовірності елементів другої підмножини до сумарної ймовірності елементів усієї множини:

$$\varphi(\Lambda_{q,v}, x) = \frac{C_{q-x}^{v}}{C_q^v}, \tag{10}$$

$$\begin{cases} \varphi(\Lambda, 1) = \left(1 - \frac{v}{q}\right), \\ \varphi(\Lambda, x) = \varphi(\Lambda, x-1)\left(1 - \frac{v}{q - x + 1}\right). \end{cases} \tag{11}$$

Параметр γ визначає відношення сумарної імовірності елементів першої підмножини до сумарної імовірності елементів множини: $\gamma(\Lambda_{q,v},x)=1-\varphi(\Lambda_{q,v},x)$.

Розподіл множини відбувається шляхом послідовного перебору елементів і рекурентного обчислення параметра φ на кожному кроці. Процес розподілу закінчується при знаходженні значення φ, найбільш близького до $\frac{1}{2}$.

Після кожного етапу розподілу вихідна множина Λ обмежується ліворуч або праворуч значеннями крайніх елементів t_L або t_R. При цьому множина розглянутих елементів позначається відповідно Θ^L, Θ^R або Θ^{LR}. Для загального випадку

$$\varphi(\Theta^{LR},x) \quad =1-\frac{\gamma(\Lambda_{q-t_L,v},x)}{1-\frac{\varphi(\Lambda_{q,v},t_R)}{\varphi(\Lambda_{q,v},t_L)}}, \tag{12}$$

$$\varphi(\Theta^L,t_R)=\frac{\varphi(\Lambda_{q,v},t_R)}{\varphi(\Lambda_{q,v},t_L)}. \tag{13}$$

Синтезовано алгоритм кодування на основі комбінаторного розкладання на класи еквівалентності:

1. Обчислюється φ для Λ з параметрами (q,v), $t_L=0$, $t_R=q-v+1$. При цьому $\Theta^L=\Lambda$, $\varphi(\Theta^L,t_R)=0$, $\gamma(\Theta^L,t_R)=1$. Знаходимо таке значення x', при якому $\left|\frac{1}{2}-\varphi(\Lambda,x')\right|$ досягає мінімального значення.
2. Якщо елемент у першій підмножині $z\le x'+t_L$, то у вихідний код записується 0. Змінюються параметри множини $t_R=x'+t_L$, $\varphi(\Theta^L,t_R)=\varphi(\Theta^L,x')$.
3. Якщо елемент у другій підмножині $z>x'+t_L$, то у вихідний код записується 1. Змінюються параметри множини: значення t_L збільшується на x', $\Theta^L=\Lambda$ з параметрами $(q-t_L,v)$, якщо $\varphi(\Theta^L,t_R)\neq 0$, то $\varphi(\Theta^L,t_R)$ зменшується в $\varphi(\Theta^L,x')$ разів.
4. Якщо множина складається з двох елементів $t_R-t_L=2$ і кодується останній $z=t_R$, то у вихідний потік записується 1, інакше – 0. Процес кодування закінчений.
5. Якщо множина складається з одного елемента $t_R-t_L=1$, то у вихідний потік нічого не записується. Процес кодування закінчений.
6. Обчислюється φ. Знаходиться таке значення x', при якому $\left|\frac{1}{2}-\varphi(\Theta^L,x')\right|$ досягає мінімального значення.
7. Перехід до п.2.

Розроблений алгоритм не потребує обчислення імовірностей значень, які кодуються, що значно спрощує алгоритм стиску і відповідно підвищує його швидкодію.

У роботі розроблено алгоритм визначення параметрів моделі ймовірнісного джерела B, який є адаптивним і дозволяє побудувати адекватну модель для його опису. Застосування класичних методів статистичного кодування даного джерела у відповідності до отриманої моделі дозволяє підвищити ефективність методу локальних зсувів, особливо при малій довжині послідовностей n.

У четвертому розділі проведено аналіз ефективності стиску даних на основі розроблених моделей джерел інформаційних масивів і алгоритмів їхнього кодування методом локальних зсувів. Аналіз показав, що метод дозволяє виконувати стиск як типізованих, так і нетипізованих даних заданого класу краще, ніж відомі на сьогоднішній день методи.

В експериментах при стиску графічних даних ступінь кореляції рядків змінювався від $\rho = 1\%$ до $\rho = 99\%$. Експеримент показав, що серед алгоритмів стиску, які використовуються у форматах JBIG, TIFF CCITT Group 3, TIFF CCITT Group 3-2D, TIFF CCITT Group 4, TIFF LZW, TIFF RLE, коефіцієнт стиску методом локальних зсувів є найкращим [4]. Найбільш близьким щодо ефективності стиску є алгоритм JBIG, що на сьогоднішній день визнаний міжнародним індустріальним стандартом для стиску подібних зображень. Для наочності на **Ошибка! Источник ссылки не найден.** показано значення різниці коефіцієнтів стиску JBIG і методу локальних зсувів у залежності від ступеня кореляції рядків.

Рис. 7. Різниці коефіцієнтів стиску JBIG і методу локальних зсувів
у залежності від ступеня кореляції рядків

Для аналізу результатів з погляду ефективності стиску, отриманих при стисканні бінарних файлів, використовувався параметр Δ, що визначає якість коду і розраховується за формулою

$$\Delta_V = \frac{l_{cp}(V) - H(A)}{H(A)}, \qquad (14)$$

де $H(A)$ — ентропія первинного алфавіту; $l_{cp}(V)$ – середня довжина коду.

Для порівняння були використані найбільш поширені й ефективні алгоритми стиску, застосовувані у форматах архівів RAR і ZIP. Для їх одержання використовувалася програма WINRAR 3.00. У всіх випадках вибирався найкращий алгоритм стиску.

При стиску методом локальних зсувів середнє значення параметра Δ дорівнює 1,54%, що приблизно в 4 і 6 разів краще, ніж в інших порівнюваних методах відповідно (**Ошибка! Источник ссылки не найден.**).

Рис.8. Порівняння параметрів Δ при різному значенні імовірності 1

Синтезована структурна і функціональна схеми системи стиску можуть бути використані для апаратної реалізації розроблених алгоритмів. Розглянуті структурні блоки можна побудувати з використанням ПЛІС. Надійність і швидкодія таких пристроїв істотно вищі, ніж реалізованої програмної моделі. Це дає можливість припускати, що вони можуть бути використані в інформаційних системах, які працюють у реальному режимі часу.

У додатку наведені програмні інтерфейси імітаційних моделей, фрагменти нетипізованих даних, результати їх стиску, фрагменти графічних даних з різним ступенем кореляції рядків і результати їх стиску.

ВИСНОВКИ

У дисертаційній роботі наведено теоретичне обґрунтування та нове розв'язання наукової задачі, що полягає у створенні сукупності моделей та алгоритмів стиску даних, орієнтованих на застосування в інформаційних системах. У роботі отримані такі результати:

1. Обґрунтовано доцільність застосування комбінаторних методів стиску в АСУ, особливо для задач, коли невідомі ймовірнісні характеристики джерела інформації. Показано, що стиск даних комбінаторними методами дає теоретично кращий результат, ніж статистичними методами, що оперують ймовірнісними властивостями джерела інформації.

2. Уперше запропонована математична модель відносної адресації символів у двійкових комбінаторних послідовностях, що дозволяє їх адекватно описувати у вигляді, зручному для оптимального нерівномірного кодування. Дана модель дозволяє перейти від рівноймовірної комбінаторної моделі джерела двійкових послідовностей до ймовірнісного джерела відносних адрес символів з детермінованим розподілом імовірностей. Розроблена модель дала подальший розвиток розкладанню бернуллієвського джерела двійкових повідомлень і стала основою для методу локальних зсувів для стиску інформаційних масивів.

3. Розроблено математичну модель і алгоритм декореляції для рядків і стовпців двовимірних інформаційних масивів. Застосування декореляції є одним із методів попередньої обробки даних, що дозволяють усунути або істотно послабити взаємозв'язок між елементами масиву, який стискається. Вона дозволяє істотно підвищити ефективність стиску даних на основі розроблених моделей.

4. Уперше запропонований і розроблений метод локальних зсувів для стиску масивів двійкових даних. Метод дозволяє виділити з даних надмірність статистичного і структурного роду і застосувати для її усунення відповідні методи кодування. Метод використовує бернуллієвську модель для опису вихідних повідомлень. Для стиску використовуються розкладання бернуллієвського джерела інформації на два взаємозалежних і їх роздільне кодування. Основний ефект стиску досягається при

кодуванні комбінаторного джерела інформації, для опису якого використовується розроблена модель відносної адресації.

5. Розроблено алгоритм кодування ймовірнісного джерела відносних адрес для усунення надмірності адресних послідовностей, що не потребує обчислення імовірностей. Це значно спрощує алгоритм стиску і відповідно підвищує його швидкодію. В алгоритмі використовується комбінаторне розкладання адрес на класи еквівалентності.

6. Розроблено програми стиску графічних даних і двійкових послідовностей на основі методу локальних зсувів. Використання цих програм дозволило зробити експериментальну оцінку ефективності стиску запропонованим у роботі методом. Розроблений пакет програм застосовується в навчальних цілях у Сумському державному університеті з дисциплін “Методи і засоби стиску даних в інформаційних системах”, “Системи передачі даних” для дослідження методів обробки зображень, стиску графічних і нетипізованих даних.

7. Синтезовано структурну і функціональну схеми системи стиску, що можуть бути використані для апаратної реалізації розроблених алгоритмів. Розглянуті структурні і функціональні блоки можна побудувати з використанням ПЛІС. Надійність і швидкодія таких пристроїв будуть істотно вищі, ніж реалізованої програмної моделі. Це дає можливість використання в інформаційних системах реального часу.

Алгоритми стиску даних на основі методу локальних зсувів застосовані Науково-виробничим колективним підприємством "Преобразователь" (м. Суми) у проекті автоматизованої системи управління, контролю і обліку електроенергії. Розроблені програми і алгоритми використовуються в навчальному процесі Сумського державного університету.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1.Борисенко А.А., Зубань Ю.А. Сжатие информации методом локальных сдвигов // Вісник Сумського державного університету. - 2000. - №16. - С.70-72.

2.Зубань Ю.А. Анализ электронограмм с помощью метода конструируемых локальных окон // Вісник Сумського державного університету. - 2000. - №17. - С.75-78.

3.Борисенко А.А., Зубань Ю.А. Оптимальное неравномерное кодирование в методе локальных сдвигов // Вісник Сумського державного університету. - 2002. - №1(34). - С.68-71.

4.Зубань Ю.А. Метод локальных сдвигов в задачах сжатия графической информации // Вісник Сумського державного університету. - 2002. - №12(45). - С.174-177.

5.Зубань Ю.А. О повышении эффективности сжатия данных без потерь информации // Сборник научных трудов "АСУ и приборы автоматики". - Харьков: ХНУРЕ, 2003. - №123. - С.53-57.

6.Борисенко А.А., Зубань Ю.А. Метод сжатия на основе комбинационного разложения передаваемых сообщений на классы эквивалентности // Вісник Сумського державного університету. - 2003. - №11(57). - С.88–99.

7.Зубань Ю.А., Протасова Т.А., Бражник И.Е. К задаче обработки изображений // Сборник научных трудов. - Харьков: ХГТУРЭ, 2000.- Ч.1.- С.173-174.

8.Кулик И.А., Зубань Ю.А. Универсальный метод оптимального кодирования на основе метода локальных сдвигов // Сборник научных трудов. - Харьков: ХГТУРЭ, 2001.- Ч.2.- С.132-133.

9.Кулик И.А., Зубань Ю.А. Повышения скорости передачи данных на основе сжатия информации методом локальных сдвигов // Матеріали VI Міжнародної конференції "Контроль і управління в складних системах" (КУСС-2001).– Винница, 2001.– С.147–150.

10. Borisenko A.A., Zuban Y.A. Application the method of local windows at physical researches // The European Material Conference. E-MRS 1999 Spring Meeting. – Strasbourg (France).

11. Зубань Ю.А. О возможности сжатия телевизионных изображений // Научно-техническая конференция преподавателей, сотрудников, аспирантов и студентов физико-технического факультета. - Сумы: Изд-во СумГУ, 2000. - С.22.

12. Зубань Ю.А. Сжатие изображений методом локальных сдвигов // Збірник тез доповідей Міжнародної науково-технічної конференції молодих вчених "Оптоелектронні інформаційно-енергетичні технології".– Вінниця: ВДТУ, 2001.- С.27.

13. Зубань Ю.А., Козачек А.В. Повышение быстродействия алгоритма построения кода Хаффмена // Научно-техническая конференция преподавателей, сотрудников, аспирантов и студентов физико-технического факультета. - Сумы: Изд-во СумГУ, 2001. - С.32.

14. Зубань Ю.А., Падалко А.В., Ярута Р.Н. Равномерное и неравномерное кодирование в методе локальных сдвигов // Международная научная конференция “Современные методы кодирования в электронных системах”.– Сумы: Изд-во СумГУ, 2002.– С.65.

АНОТАЦІЯ

Зубань Юрій Олександрович. Моделі і алгоритми стиску даних в інформаційних системах. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 - автоматизовані системи управління та прогресивні інформаційні технології. - Харківський національний університет радіоелектроніки, Харків, 2004.

Дисертація присвячена питанням розроблення засобів стиску даних в інформаційних системах. Запропоновані математичні моделі стали основою методу локальних зсувів для стиску інформаційних масивів. Розроблений метод локальних зсувів для стиску масивів двійкових даних дозволяє виділити надмірність статистичного і структурного роду і застосувати для її усунення відповідні методи кодування. Розроблені алгоритми кодування значно спрощують алгоритм стиску і відповідно підвищують його швидкодію. Синтезована структурна і функціональна схеми системи стиску можуть бути використані для апаратної реалізації розроблених алгоритмів. Це дає можливість їх використання в інформаційних системах реального часу.

Ключові слова: комбінаторне джерело, адресна послідовність, кодування, стиск, розкладання, класи еквівалентності.

АННОТАЦИЯ

Зубань Юрий Александрович. Модели и алгоритмы сжатия данных в информационных системах. – Рукопись.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.06 - автоматизированные системы управления и прогрессивные информационные технологии. - Харьковский национальный университет радиоэлектроники, Харьков, 2004.

Диссертация посвящена разработке новых моделей, методов и алгоритмов сжатия данных в информационных системах.

Результаты анализа современного состояния и тенденций развития информационных систем показали актуальность проблемы сжатия данных. Это позволяет значительно разгрузить каналы связи, системы обработки и хранения данных за счет исключения избыточных сообщений, что эквивалентно повышению пропускных способностей информационных систем или увеличению емкости запоминающих устройств. В работе обоснована целесообразность применения комбинаторных методов сжатия в АСУ, особенно для задач, когда неизвестны вероятностные характеристики источника информации. Показано, что сжатие данных комбинаторными методами дает тео-

нен стиль структуры: Двухбайтовы
ві інші - обсягом до 0,5 сторінки машинописного тексту (до 1200 друкованих знаків)

нен стиль структуры: Двухбайтовы
нотація англійською або російською мовою повинна бути розгорнутою, обсягом 2 сторінки машинописного тексту (до п'яти тисяч друкованих знаків

ретически лучший результат, чем статистическими методами, оперирующими вероятностными свойствами источника информации.

Предложен метод локальных сдвигов для сжатия массивов двоичных данных. Метод позволяет выделить из сжимаемых данных избыточность статистического и структурного рода и применить для ее устранения соответствующие методы кодирования. Метод использует бернуллиевскую модель для описания исходных сообщений. Чтобы повысить адекватность описания данных моделью, возможно применение декорреляции как этапа предварительной обработки. Для сжатия используется разложение бернуллиевского источника информации на два взаимосвязанных, раздельное кодирование которых позволяет устранить структурную и вероятностную избыточность исходных сообщений. Кроме того, раздельное кодирование источников информации дает возможность использовать модель для систем с защитой данных от несанкционированного доступа. Основной эффект сжатия достигается при кодировании комбинаторного источника информации, для описания которого используется разработанная модель относительной адресации, позволяющая адекватно описывать последовательности в виде, удобном для их оптимального кодирования. Данная модель позволяет перейти от равновероятной комбинаторной модели источника двоичных последовательностей к вероятностному источнику относительных адресов символов с детерминированным распределением вероятностей генерируемых значений. Это дает возможность эффективно применять алгоритмы ОНК для кодирования комбинаторного источника без ограничения на мощность его алфавита. Применение декорреляции является одним из методов предварительной обработки данных, позволяющих устранить или существенно ослабить взаимосвязи между элементами сжимаемого массива. Она позволяет существенно повысить эффективность сжатия данных на основе разработанных моделей.

Разработанный алгоритм кодирования вероятностного источника относительных адресов не требует вычисления вероятностей значений кодируемых значений. Это значительно упрощает алгоритм сжатия и соответственно повышает его быстродействие. Алгоритм разработан на основе модификации известных методов ОНК и с учетом характерной для метода локальных сдвигов функциональной зависимости для вероятностей относительных адресов. В алгоритме используется комбинаторное разложение кодируемых адресов на классы эквивалентности. Генерируемый код является неравномерным, префиксным и оптимальным с точки зрения информационной нагрузки на каждый символ.

Разработаны программы сжатия графических данных и двоичных последовательностей на основе метода локальных сдвигов. Использование этих программ позволило произвести экспериментальную оценку эффективности сжатия предложенным в работе методом. Разработанный пакет программ применяется в учебных целях в Сумском государственном университете по дисциплинам

“Методы и средства сжатия данных в информационных системах”, “Системы передачи данных” для исследования методов обработки изображений, сжатия графических и нетипизированных данных.

Проведенный анализ эффективности сжатия данных на основе разработанных моделей источников информационных массивов и алгоритмов их кодирования методом локальных сдвигов показал, что метод позволяет производить сжатие как типизированных, так и нетипизированных данных лучше, чем известные на сегодняшний день методы. Применение разработанных методов кодирования в сочетании с алгоритмом декорреляции позволило превзойти по степени сжатия специализированные алгоритмы, разработанные для сжатия графических данных.

Синтезированы структурная и функциональная схемы системы сжатия, которые могут быть использованы для аппаратной реализации разработанных алгоритмов. Рассмотренные структурные и функциональные блоки можно построить с использованием ПЛИС. Кроме того, сами структуры кодирующих устройств обладают достаточной простотой и наглядностью. Надежность и быстродействие таких устройств будут существенно выше, чем реализованной программной модели. Это дает возможность предполагать, что они могут быть использованы в информационных системах, работающих в реальном режиме времени.

Результаты диссертационной работы в виде метода локальных сдвигов для сжатия данных с программной реализацией использованы в Научно-производственном коллективном предприятии “Преобразователь” (г. Сумы) в информационных каналах автоматизированной системы учета электроэнергии для промышленных предприятий. Благодаря использованию метода локальных сдвигов обеспечено хранение большего объема оперативных данных без изменения объема запоминающих устройств.

Ключевые слова: комбинаторный источник, адресная последовательность, кодирование, сжатие, разложение, классы эквивалентности.

ABSTRACT

Zuban Yury Alexandrovich. Models and algorithms for data compression for information systems. – Manuscript.

Thesis for a candidate degree of technical sciences by specialty 05.13.06 – automation control systems and progressive information technologies. – Kharkov national university of radioelectronics, Kharkov, 2004.

The dissertation is devoted to questions of development of means of compression given in information systems. The offered mathematical models have formed the basis for a method of local shifts for compression of information files. The developed method of local shifts for compression of files of the binary data allows to allocate redundancy of a statistical and structural sort and to apply to its elimination the appropriate methods of coding. The developed algorithms of coding considerably simplify algorithm of compression and, accordingly, raises its speed. The synthesized structural and functional circuits of system

of compression can be used for hardware realization of the developed algorithms. It enables their uses in information systems of real time.

Key words: combinatorial source, address sequence, coding, compression, decomposition, classes of equivalence.

Підп. до друку 15.09.2004 р. Формат 60×90/16.
Обл.-вид. Арк. 0,9. Наклад 100 прим.
Ум. друк. арк. 1,1. Замовл. № 445

Друкарня СумДУ. 40007, м. Суми, вул. Римського-Корсакова, 2