УДК 681.39

Володимир Лужецький
Валентина Каплун

КЛАСИФІКАЦІЯ МЕТОДІВ УЩІЛЬНЕННЯ, ЩО БАЗУЮТЬСЯ НА ОБЧИСЛЕННІ ВІДХИЛЕНЬ

Загальновідомі класифікації не дають змоги повною мірою враховувати можливості різних моделей джерел інформації. Основні відомі методи ущільнення використовують або статистичні характеристики даних, що ущільнюються, або словниковий принцип. Крім того, сучасні мікропроцесори ефективніше здійснюють арифметичні операції над числами. Тому останнім часом здійснюються пошуки підходів щодо ущільнення даних, які базуються на їх представленні як цілих чисел.

В даній статті розглядається числова модель джерела інформації. Тобто дані, які підлягають ущільненню, незалежно від їх фактичного змісту, розглядаються як цілі числа. Пропонується класифікація методів ущільнення даних, основою яких є обчислення відхилень.

Існує багато вагомих причин виділяти ресурси ПЕОМ з розрахунку на стисле представлення, оскільки швидше передавання даних і зменшення простору для їх зберігання дають змогу заощадити значні засоби і часто покращити показники ПЕОМ. Будь-який спосіб, підхід і алгоритм, що реалізує ущільнення даних, призначений для зменшення обсягу вихідного потоку інформації в бітах за допомогою її оборотного або необоротного перетворення. З цієї точки зору всі способи ущільнення можна розділити на дві категорії: *ущільнення без втрат* (оборотне) і *ущільнення з втратами* (необоротне) [1]. Під ущільненням з втратами розуміють таке перетворення вхідного потоку даних, при якому вихідний потік, оснований на певному форматі інформації, достатньо схожим за зовнішніми характеристиками на вхідний потік об'єкта, проте відрізняється від нього обсягом. Ущільнення без втрат завжди приводить до зниження об'єму вихідного потоку інформації без втрати інформаційної структури. Більш того, з вихідного потоку за допомогою декомпресуючого алгоритму можна одержати вхідний, і лише після процесу розпаковування дані придатні для обробки відповідно до їх внутрішнього формату.

Крім того, методи ущільнення можуть бути статистичними або трансформуючими, а також можуть обробляти дані потоками або блоками. З цієї точки зору існують три види стратегії ущільнення [2]. Перша з них – *перетворення потоку*, коли кодування даних, які надходять, здійснюється за допомогою вже відпрацьованих раніше даних, тобто за таблицею. При цьому ніякі імовірності

не обчислюють. У результаті перетворення вхідного потоку може бути сформовано декілька вихідних, навіть якщо сумарний обсяг потоків збільшується, їх структура покращується, і наступне ущільнення можна здійснити простіше, швидше і краще.

Друга стратегія – *перетворення блоків*, коли вхідні дані розбиваються на блоки, які потім трансформуються цілком. Як і у попередньому випадку, в результаті можуть сформуватися декілька блоків, і не дивлячись на те, що сумарна довжина блоків може не зменшитися, їх структура значно покращиться для наступного ущільнення.

Третя стратегія – *статистична*. В ній розглядають *адаптивну* (потокову) і *блокову*. При використанні статистичної адаптивної стратегії обчислюються імовірності для даних, що надходять, на основі статистики за попередньо обробленими даними і здійснюється кодування з використанням цих обчислених імовірностей. У блоковій стратегії окремо кодуються і додаються до ущільнених блоків їх статистики.

Авторами розглядається числова модель джерела інформації. Тобто дані, які підлягають ущільненню, незалежно від їх фактичного змісту, розглядаються як цілі числа. При цьому весь потік вхідної інформації представляється як послідовність *n*-розрядних двійкових чисел.

Нехай початкова послідовність чисел Q складається з K елементів q_i:

$$Q = \{q_1, q_2, ..., q_K\}.$$

Така послідовність має певні характеристики, які можуть бути враховані для її ущільнення. Серед характеристик послідовності можна виділити такі:

– мінімальний і максимальний елементи послідовності:

$$q_{\min} = \min\{q_i \mid i = \overline{1,K}\}; \quad q_{\max} = \max\{q_i \mid i = \overline{1,K}\};$$

– діапазон значень елементів вхідної послідовності Q, :

$$D = q_{\max} - q_{\min},$$

– центр діапазону, який обчислюється за формулою

$$C = q_{\min} + \frac{D}{2};$$

– середнє значення елементів діапазону, що обчислюється як середнє арифметичне значення вхідних чисел:

$$m = \frac{\sum_{i=1}^{K} q_i}{K};$$

– накопичене середнє значення:

$$M_1 = 0; \; M_{l+1} = \frac{\sum_{i=1}^{l} q_i}{l}, \; l = 1,2,...,K-1.$$

Весь діапазон D значень елементів вхідної послідовності може бути розбитий на N піддіапазонів D_j $(j = \overline{1,N})$, величина кожного з яких визначається так:

$$h = \frac{D}{N} = \frac{q_{\max} - q_{\min}}{N},$$

а елементи вхідної послідовності належать одному з інтервалів:

$$\left(q_{\min} + (j-1) \cdot h; q_{\min} + j \cdot h\right).$$

Тоді вхідна послідовність Q розіб'ється на N підпослідовностей:

$$Q = \{Q_1, Q_2, ..., Q_N\},\ \text{причому}\ Q_i \cap Q_j = \varnothing, (i \neq j; i = \overline{1,N}; j = \overline{1,N}).$$

При цьому кожна з підпослідовностей буде мати свої власні характеристики:

– мінімальне і максимальне значення елементів l-го підділапазону:

$$q^l_{\min} = \min\{q_i \subset Q_l \mid i = 1 \div K\};$$

$$q^l_{\max} = \max\{q_i \subset Q_l \mid i = 1 \div K\};$$

– середнє значення елементів підділапазону:

$$m_l = \frac{\sum_{i=1}^{K} q_i}{k_i}, (q_i \subset Q_l),$$

– центр підділапазону:

$$C_l = q^l_{\min} + \frac{h}{2}, \quad l = \overline{1,N}.$$

Вхідну послідовність чисел можна розбити на групи по k елементів у кожній:

$$G_j = \{q_{i_1}, q_{i_1+1}, ..., q_{i_k}; i_1 = (j-1)\cdot k; i_k = j \cdot k\},\ j=1, \ldots, K/k,$$

$$G_i \cap G_j = \varnothing, (i \neq j; i = \overline{1, K/k}; j = \overline{1, K/k}).$$

У кожній з отриманих груп елементів можна визначити такі характеристики:

– мінімальне і максимальне значення елементів у групі:

$g^j_{\min} = \min\{g_i \subset G_j \mid i = \overline{1,k}\};\ g^j_{\max} = \max\{g_i \subset G_j \mid i = \overline{1,k}\};\ j=1,\ldots,K/k;$

– середнє значення елементів групи:

$$m_{G_j} = \frac{\sum_{i=1}^{k} g_i}{k}, (g_i \subset G_j);\ j=1,\ldots,K/k.$$

Вищенаведені характеристики пропонується покласти в основу методів ущільнення, що базуються на відхиленні елементів послідовності Q від цих характеристик. При цьому ущільнення досягається за рахунок зберігання обчислених відхилень, які, за певних умов, можуть бути меншими, ніж самі елементи початкової послідовності. Після перетворення вхідного потоку може бути сформовано декілька вихідних потоків зі структурою, яка буде кращою за структуру вхідного потоку, і наступне ущільнення бути досить ефективним.

Оскільки існує певна множина таких методів, то у роботі пропонується класифікація методів ущільнення, що якраз і базуються на обчисленні і зберіганні цих відхилень.

У загальному випадку обчислення відхилень елементів послідовності може бути здійснено або від розглянутих вище характеристик (статистик), або від елементів, які вибрані за певними правилами і не враховують цих статистик. Такі правила, наприклад, можуть визначати значення сусідніх елементів вхідної послідовності, значення певним чином введених констант чи дискретні значення функцій апроксимації. Таким чином, усі методи ущільнення, що базуються на обчисленні відхилень, поділяються на такі, що враховують статистики, і такі, що їх не враховують.

Окрему групу методів, що враховують числові статистичні характеристики, утворюють методи, які базуються на розбитті значень елементів послідовності Q на *піддіапазони*.

Діапазон D чисел вхідної послідовності розбивається на N піддіапазонів, в результаті чого вхідна послідовність розпадається на N підпослідовностей. В ущільненій послідовності $Q_{ущ.}$ кожний елемент вхідної послідовності Q представляється набором значення відхилення від певної характеристики піддіапазону і деякою додатковою інформацією, яка необхідна для однозначного відновлення первинного повідомлення. Характеристиками піддіапазону, як відзначалося вище, можуть бути середні значення елементів піддіапазонів, визначені за формулами (2), і значення центрів піддіапазонів, визначені за формулами (3).

У першому випадку ущільнена послідовність $Q_{ущ.}$ повинна, крім самих відхилень, містити ще й номери піддіапазонів c_i ($i = \overline{1, K}$), яким вони належать, і самі середні значення m_j ($j = \overline{1, N}$) піддіапазонів, за допомогою яких обраховувались відповідні відхилення (рис.1).

Поле відхилень	c_1	...	c_K	m_1	...	m_N

Рис. 1. **Структура ущільненої інформації у випадку використання відхилення від середнього значення у піддіапазоні**

При використанні центрів піддіапазонів ущільнена послідовність $Q_{ущ.}$ повинна містити лише номери піддіапазонів c_i ($i = \overline{1, K}$), яким вони належать, а самі значення центрів піддіапазонів залежать тільки від розрядності чисел n і можуть бути обчислені заздалегідь (рис.2).

Поле відхилень	c_1	...	c_K

Рис. 2. **Структура ущільненої інформації у випадку використання відхилення від центрів піддіапазонів**

Отже, серед методів, що враховують числові статистичні характеристики елементів піддіапазонів даних вхідної послідовності, можна виділити такі, що базуються на обчисленні відхилень від центрів піддіапазонів, і такі, що основані на обчисленні відхилень від середніх значень елементів піддіапазонів.

Одним із недоліків таких методів є необхідність зберігати номери піддіапазонів, яким належать числа з вхідної послідовності. Його можна усунути, якщо визначати відхилення від характеристик *групи сусідніх елементів* послідовності Q. При цьому вхідна послідовність Q розбивається на групи G_j по k елементів у кожній.

Як відзначалося вище, такими характеристиками можуть бути максимальне або мінімальне значення елементів відповідної групи, визначене за формулами (4). Тоді в ущільненій послідовності $Q_{ущ.}$ кожний елемент послідовності Q представляється як відхилення від вибраної характеристики, тобто ущільнена послідовність складатиметься тільки зі значень відхилень.

У випадку використання середнього значення чисел у групі, обчисленого за формулами (5), в ущільненій послідовності $Q_{ущ.}$ кожний елемент послідовності Q представляється відхиленням від середнього значення m_{G_j} ($j = \overline{1, K/k}$) і самим значенням цього середнього значення для подальшого однозначного відновлення вхідної послідовності (рис. 3).

Поле відхилень	m_{G_1}	...	$m_{G_{K/k}}$

Рис. 3. **Структура ущільненої інформації у випадку використання відхилення від середніх значень у групах**

Таким чином, серед методів, що враховують числові статистичні характеристики елементів у групах, на які розбивається вхідна послідовність чисел, можна виділити такі, що базуються на обчисленні відхилень від мінімальних або максимальних елементів у групах, і такі, що основані на обчисленні відхилень від середніх значень елементів у групах.

Отже, при застосуванні майже всіх попередніх методів в ущільненій послідовності доводиться зберігати інформацію або про належність до певного піддіапазону, або саму характеристику піддіапазону чи групи, або і те, й інше разом.

Для усунення цього недоліку можна використати метод, у якому відхилення будуть обчислюватись не від певної характеристики піддіапазону або групи, яким належить елемент, а від накопиченого середнього значення M_l ($l = \overline{0, K-1}$), яке розраховується за формулою (1). У цьому випадку кожний елемент вхідної послідовності Q в ущільненій послідовності $Q_{ущ.}$ буде представлятися тільки відхиленням від накопиченого середнього значення елементів послідовності.

Як було сказано вище, обчислення відхилень може здійснюватись не обов'язково від числових статистичних характеристик вхідної послідовності Q.

Як числа, відхилення від яких обчислюють і зберігають, можуть використовуватись *сусідні елементи* вхідної послідовності, які підлягають ущільненню. У цьому випадку кожному елементу вхідної послідовності Q в ущільненій послідовності $Q_{ущ.}$ буде відповідати значення його відхилення від попереднього елемента (перший елемент при цьому залишиться без змін). Серед недоліків такого підходу можна виділити те, що отримане відхилення може виявитись також досить великим, якщо сусідні числа знаходяться на великій відстані одне від одного у діапазоні.

Для усунення такого недоліку можна використати вибраний за певними правилами набір констант. У такому разі обчислюється відхилення від тієї константи, відстань до якої є найменшою. При використанні такого підходу числа

вхідної послідовності Q в ущільненій послідовності $Q_{ущ.}$ представляються не тільки самими відхиленнями, а й константами C_j ($j = \overline{1, N}$), які були використані для їх обчислення. При цьому, якщо як константи вибрати числа, сформовані і упорядковані заздалегідь, ще до здійснення процесу ущільнення, і які не залежать від самих значень вхідних чисел, а залежать тільки від розрядності n чисел вхідної послідовності, то зберігати самі константи немає необхідності, а можна зберігати лише їх номери. Структура ущільненої послідовності у цьому випадку показана на рис. 4.

Поле відхилень	C_1	...	C_N

Рис. 4. **Структура ущільненої інформації у випадку використання відхилення від констант**

Ще одну групу методів, основаних на обчисленні відхилень без урахування числових характеристик, складають методи, що обраховують відхилення з вхідної послідовності від значень деякої апроксимуючої функції $y=f(x)$ у відповідних точках. У такому випадку кожний елемент вхідної послідовності Q в ущільненій послідовності $Q_{ущ.}$ буде представлятися відповідним відхиленням від значень апроксимуючої функції. Але крім самих відхилень в ущільненій послідовності повинна бути присутня інформація з характеристиками самої апроксимуючої функції (коефіцієнти, степені поліномів тощо) (рис. 5).

Поле відхилень	*Характеристики апроксимуючої функції*

Рис. 5. **Структура ущільненої інформації у випадку використання відхилення від апроксимуючої функції**

З огляду на введені поняття пропонується класифікація методів ущільнення, що базуються на обчисленні відхилень, яка наведена на рис. 6.

Кожний із запропонованих методів забезпечує ущільнення тільки у разі певних властивостей вхідної послідовності цілих чисел. Тому вибір конкретного методу повинен здійснюватись, виходячи з цих умов. Однак може бути використана додаткова процедура перетворення первинної послідовності до послідовності, яка матиме властивості, що задовольнять конкретному методу ущільнення.

Рис. 6. **Класифікація методів ущільнення, що базуються на обчисленні відхилень**

ЛІТЕРАТУРА

1. *Балашов К.Ю.* Сжатие информации: анализ методов и подходов. – Минск: Ин-т техн. кибернетики НАН Беларуси, 2000. – Вып.6. – 42 с.

2. *Ватолин Д., Ратушняк А., Смирнов М., Юкин В.* Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео. – М.: ДИАЛОГ-МИФИ, 2003. – 384 с.

3. *Фомин А.А.* Основы сжатия инфорации. – Санкт-Петербург, 1998.

Надійшла 9 жовтня 2007 р.