УДК 004.627

О.В. Тарасов, Є.В. Онопко

Харківський національний економічний університет, Харків

БЛОЧНО-СТАТИСТИЧНИЙ МЕТОД СТИСНЕННЯ ІНФОРМАЦІЇ

Стиснення інформації є однією з тих проблем, яка нерозривно пов'язана з обробкою даних з використанням засобів обчислювальної техніки. Текст і звук, графіка і відео – для кожного з цих видів інформації існують свої найбільш відповідні методи стиснення. Метою процесу стиснення, як правило, є отримання компактнішого способу представлення початкових даних, який мінімізує об'єм займаної пам'яті за допомогою деякого їх перетворення.

Існує декілька різних підходів до проблеми стиснення інформації, які базуються або на складних математичних алгоритмах або засновані на властивостях інформаційного потоку і алгоритмічно досить прості. Проте усі способи стиснення можна розділити на дві категорії: оборотне і безповоротне стиснення. У першому випадку можливо повне і безпомилкове відновлення початкових даних, які були піддані стисненню. У другому – це неможливо. Проте з точки зору практичного застосування результат, отриманий в процесі стиснення, може бути цілком задовільний, наприклад, для графіки або звукових повідомлень.

Особливий інтерес викликає застосування методів стиснення у базах даних для зменшення об'єму текстових полів. По-перше, для таких даних неприпустимо використання алгоритмів безповоротного стиснення, а по-друге, застосування алгоритмів, закладених в широко відомих архіваторах, наприклад, RAR або ZIP, є неефективним з причини того, що стисненню піддаються текстові поля невеликого об'єму і в цьому випадку отриманий при стисненні результат, з урахуванням даних для декодування, може займати об'єм пам'яті більший, ніж початкові дані.

Одним з ефективних методів, який застосовується у базах даних для стиснення, є класичний алгоритм Хаффмана [1], що базується на знанні частоти розподілу символів в тексті і дає в результаті, набір кодів змінної довжини для усього вхідного алфавіту. В результаті цього, хоча окремі символи і рідко зустрічаються у тексті, з причини їх довгої кодової послідовності, середня довжина коду символу може бути досить великою.

Виникає питання. А чи не можна зменшити середню довжину коду, розбивши увесь вхідний алфавіт на дві або більше груп і провівши кодування по Хаффману для кожної з них окремо? Правда в цьому випадку у вхідному алфавіті з'являються нові символи, що характеризують перехід до тієї або іншої групи, але довжина коду у кожній окремій групі може бути зменшена. Був проведений експеримент, який для вхідного алфавіту з 80 символів визначив середню довжину коду символу для алгоритму Хаффмана, при розподілі частот символів, що підкоряються закону Ципфа [2].

Далі початковий алфавіт був розбитий на дві рівні по кількості символів групи. Сумарна імовірність символів в кожній групі, при цьому, склала: $P_1 = 0{,}862$ і $P_2 = 0{,}138$. Ймовірність переходів з групи в групу при цьому дорівнює $P_{12} = P_{21} = P_1 \times P_2 = 0{,}119$. Наступним кроком, символи в кожній групі були окремо піддані стисненню по методу Хаффмана. Результати експерименту наведені в табл.1.

Таблиця 1

Середня довжина коду символу

Для усього тексту	Для блоку №1 (P_1=0,862)	Для блоку №2 (P_2=0,138)
4,91	4,36	3,49
	Середня довжина=4,24	

Таким чином застосування блочно-статистичного методу стиснення, навіть при простому розбитті на дві групи, може значно збільшити міру стиснення в порівнянні з методом Хаффмана.

Список літератури

1. Huffman D.A. A method for Construction of Minimum-Redundancy Codes / D.A. Huffman // Proc.JRE. – 1952. – 40, № 9. – P.1098.

2. Zipf, G.K. Human Behavior and the Principle of least Eort / G.K. Zipf. – Addison-Wesley, New York, 1949.