



УКРАЇНА

(19) UA (11) 83013 (13) C2
(51) МПК (2006)
G06F 5/00
G06F 17/21

МІНІСТЕРСТВО ОСВІТИ
І НАУКИ УКРАЇНИ

ДЕРЖАВНИЙ ДЕПАРТАМЕНТ
ІНТЕЛЕКТУАЛЬНОЇ
ВЛАСНОСТІ

ОПИС ДО ПАТЕНТУ НА ВІНАХІД

(54) СПОСІБ ГРАМАТИКО-СТАТИСТИЧНОГО ДИНАМІЧНОГО СТИСКАННЯ ТЕКСТОВИХ ПОВІДОМЛЕНЬ

1

(21) а200505435
(22) 07.06.2005
(46) 10.06.2008, Бюл.№ 11, 2008 р.
(72) ЧЕРНЕГА ВІКТОР СТЕПАНОВИЧ, UA
(73) СЕВАСТОПОЛЬСЬКИЙ НАЦІОНАЛЬНИЙ
ТЕХНІЧНИЙ УНІВЕРСИТЕТ, UA
(56) UA 64907 A, 15.03.2004
JP 6324841, 25.11.1994
JP 1060072, 27.03.1986
Чернега В.С. Сжатие информации в компьютер-
ных сетях. - Севастополь: СевГТУ, 1997. - С. 214,
102-106
(57) Спосіб граматико-статистичного динамічного
стискання текстових повідомлень, що включає:
граматичний розбір повідомлення, що кодується,
рекурентну процедуру об'єднання одиночного вхід-
ного сигналу-символу або групи сигналів-
символів з наступним сигналом-символом вхідного
потoku, зіставлення кода утвореного рядка з кода-
ми з рядків, які проіндексовані і знаходяться в та-
блицях кодування кодера і декодера, та в яких роз-
міщена початкова інформація, при відсутності його
в таблиці кодування вносять код утвореного непо-
рівняного рядка в таблицю кодування з черговим
індексом, відділення кода символу, що призвів до

2

непорівняного кода рядка, і видачу на вихід коде-
ра індексу зіставленого рядка максимальної дов-
жини, а процедура утворення і зіставлення рядків
починається спочатку з останнього вхідного сим-
волу текстового повідомлення, що призвів до
утворення непорівняного рядка, який **відрізня-**
ється тим, що як початкову інформацію в таблиці
кодування розміщують з відповідними індексами
афіксальні морфеми, які найчастіше зустрічаються
в текстових повідомленнях, а також, якщо код до-
датково утвореного непорівняного рядка є почат-
ком кодової комбінації однієї з розміщених мор-
фем, то додатково виконують об'єднання кода
непорівняного рядка з кодом наступного символу
вхідного потоку, порівнюють після кожного приєд-
нання чергового сигналу-символу сформований
рядок з усіма можливими варіантами афіксальних
морфем, одночасно перевіряють, чи не скінчилася
морфема або чи не відбувся незбіг сформованого
рядка з морфемою, при збігу кодів рядка і морфе-
ми на вихід кодера виводиться індекс співставле-
ної морфеми, а у випадку незбігу рядка ні з однією
з морфем вносять код утвореного непорівняного
рядка в таблицю кодування з черговим індексом.

Спосіб відноситься до області стискання тек-
стових повідомлень і призначений для використан-
ня в комп'ютерних системах і мережах, WEB-
вузлах, архіваторах.

Існує спосіб динамічного стискання текстових
повідомлень LZ78 [Чернега В.С. Стискання інфор-
мації в комп'ютерних мережах /В.С.Чернега -
Севастополь: СевГТУ, 1997. - 214с.], відповідно з
яким складається таблиця (словник) підрядків си-
мволів, які зустрічалися в попередній частині по-
відомлення, що кодується. Виділені із вхідного
повідомлення підрядки кодуються дуплетом (i, C),
де i - індекс, що відповідає найдовшому підрядку в
словнику, що збігається з виділеним підрядком, а
C - код символу вхідного потоку, що йде за збіж-
ним підрядком. Цей дуплет потім стає новим запи-
сом у словнику, якому приписується черговий ін-
декс таблиці підрядків. Недоліком способу є

невисока швидкодія і відсутність стискання на ета-
пі заповнення словника.

За прототип узятий спосіб динамічного коду-
вання LZW [Welch T.A. A technique for high-
performanse data compression //IEEE Computer.-
1984. -Vol.17. -N6.- P.8-19], заснований на словни-
ковому методі кодування [Російськомовний опис
способу викладений у кн. Чернега В.С. Стискання
інформації в комп'ютерних мережах. -
Севастополь: СевДТУ, 1997 на стор.102-106]. Да-
ний спосіб включає процедуру граматичного роз-
бору повідомлення, що кодується, рекурентну
процедуру об'єднання одиночного вхідного симво-
лу чи групи символів з наступним символом вхід-
ного потоку, зіставлення заново утвореного рядка
з рядками, що знаходяться в таблиці кодування,
внесення заново утвореного непорівняного рядка
в таблицю кодування з черговим індексом, що є
кодом цього рядка, відділення символу, що призвів

(13) C2

(11) 83013

(19) UA

до непорівнянного рядка і видачі на вихід кодової комбінації зіставленого рядка максимальної довжини, при цьому символ, що призвів до непорівнянності утвореного ним рядка, береться в якості чергового одиночного вхідного символу кодера, а процедура утворення рядків і пошук порівнянного рядка максимальної довжини повторюється спочатку.

Недоліком приведеного способу є розширення вихідної послідовності в порівнянні з вхідною на етапі заповнення таблиці кодування, що призводить до зниження коефіцієнта стиснення текстового повідомлення в цілому.

В основу винаходу поставлена задача підвищення ступеня стиснення текстових повідомлень на початковому етапі процедури кодування. Поставлена задача розв'язується наступним чином. У таблиці кодування (словнику) кодера і декодера в процесі початкової ініціалізації розміщуються афіксальні морфеми, що найбільш часто зустрічаються в текстових повідомленнях [Росінська О.А. Граматика української мови для учнів, абітурієнтів і студентів. - Донецьк: ТОВ ВКФ "БАО", 2004. - 288с.], а в процесі рекурентного об'єднання вхідних символів у рядок та зіставлення з рядками словника при відсутності його в словнику додатково здійснюється об'єднання не зіставленого рядка з наступним символом вхідного потоку і зіставлення цього рядка з афіксальними морфемами словника, причому якщо додатково утворений рядок є початком однієї з морфем, то до нього приєднується наступний символ вхідного потоку і процедура зіставлення повторюється знову, а у випадку збігу додатково утвореного рядка з однією з морфем, що

знаходиться в словнику, на вихід кодера видається індекс морфеми, а процедура утворення і зіставлення рядків починається заново з чергового вхідного символу текстового повідомлення.

Проілюструємо процедуру стиснення текстових повідомлень на прикладі стискування вхідної послідовності "примирення пройшло", що здійснюється по способу LZW і запропонованому способу. Нехай таблиці кодування з використанням способу LZW після початкової ініціалізації містять тільки одиночні ASCII-символи.

Передбачається також, що при початковій ініціалізації при стиснанні по запропонованому способу в таблиці кодування (словник) на кодуючій і декодуючій сторонах, окрім одиночних ASCII-символів, занесені афіксальні морфеми (префікси, суфікси, закінчення), що найбільш часто зустрічаються в текстових повідомленнях на українській (чи іншій) мові, а також інші граматичні компоненти мови, наприклад, сполучники, частки. Припустимо, що в таблиці кодування, у результаті початкової ініціалізації, було заповнено 350 рядків (з 0-го рядка по 255-й займали одиночні символи, інші рядки зайняті морфемами і синтаксичними компонентами). Зокрема, нехай суфікс "ення_" має індекс 280, "ення_" - 281, "ло_" - 300, "лось_" - 301, префікс "пре" - 320, "при" - 321, "про" - 322 і т.д. Тут символ "_" означає пропуск. Лапки в приведені прикладі використані тільки для зручності виділення в даному описі морфем при читанні і не входять у повідомлення, що кодується. Процедура стиснення двома способами ілюструється наступною таблицею.

Таблиця

Порівняння способів стиснення текстових повідомлень

Крок	Вхідне повідомлення	Спосіб LZW			Новий спосіб		
		Вихід кодера LZW	Рядок таблиці	Індекс	Вихід кодера	Рядок таблиці	Індекс
1	п						
2	р	п	пр	256		пр	351
3	и	р	ри	257	321(при)		
4	м	и	им	258			
5	и	м	ми	259	м	ми	351
6	р	и	ир	260	и	ир	352
7	е	р	ре	261	р	ре	353
8	н	е	ен	262			
9	н	н	нн	263			
10	я	и	ня	264			
11	_	е	я_	265	280(ення)		
12	п	_					
13	р						
14	о	256	про	266	322(про)		
15	й	о	ой	267			
16	ш	й	йш	268	й	йш	354
17	л	ш	шл	269	ш	шл	355
18	о	л	ло	270	300(ло_)		
19	_	о	о_	271			
20		_					

При надходженні на вхід кодера першого символу "п" кодера по обом способам об'являють його поточним рядком і шукають у словниках співпадаючий рядок. В зв'язку з тим, що всі одиночні символи присутні в словнику, кодери поєднують цей символ з наступним символом із вхідного потоку і процедура пошуку порівнянного рядка в словнику повторюється. У результаті утвориться рядок "пр".

По LZW-способу: у зв'язку з відсутністю в словнику рядка, що співпадає з "пр", він заноситься у вхідний словник з черговим індексом 256. Потім виконується відділення останнього символу ("р"), що призвів до непорівнянного рядка, на вихід видається код символу "п", а відділений символ "р" використовується для утворення наступного рядка "ри", який заноситься в словник з індексом 257. Аналогічним способом відбувається кодування всього повідомлення. З таблиці видно, що на 14-м кроці кодування кодер виявляє в словнику порівнянний рядок, що складається з двох символів "пр" і видає на вихід двійкову кодову комбінацію індексу цього рядка 256. При цьому він формує трьохсимвольний рядок "про" і заносить його в словник з черговим індексом 266.

При стисканні по запропонованому способу: після формування рядка "пр" і занесення його в словник з черговим індексом 351 відділення останнього символу (у нашому прикладі "р"), що призвів до утворення непорівнянного рядка, не виконується, а здійснюється аналіз на предмет збігу сформованого рядка ("пр") з початком однієї з афіксальних морфем, що занесена в таблицю на етапі початкової ініціалізації. У випадку відсутності в словнику морфеми з таким початком, подальша процедура виконується по способу LZW. Якщо ж сформований рядок збігається з початком однієї з морфем (як у нашому прикладі), то з вхідного потоку береться наступний символ і процедура зіставлення утвореного рядка з початком морфеми словника повторюється. У випадку збігу вхідного рядка з однією з морфем (у нашому прикладі "при") на вихід кодера видається її код (у нашому прикладі 321), а описана вище рекурентна процедура утворення і зіставлення рядків продовжується з наступного вхідного символу (у нашому прикладі "м") доти, поки не буде виявлений символ кінця повідомлення.

Коефіцієнт стискання вхідного повідомлення визначається по формулі

$$K_{\text{ст}} = N_p N_{\text{бр}} / (N_{\text{см}} N_{\text{бсм}}),$$

де N_p - кількість закодованих рядків; $N_{\text{бр}}$ - кількість біт, що затрачуються на кодування рядка символів (довжина індексу в бітах); $N_{\text{см}}$ - число

закодованих символів повідомлення; $N_{\text{бсм}}$ - кількість біт, затрачуваних на кодування одиночних символів у первинному ASCII-коді.

Для нашого приклада число закодованих символів вхідного повідомлення з врахуванням двох пропусків $N_{\text{см}} = 19$, а кількість біт на символ первинного коду $N_{\text{бсм}} = 8$. Як видно з таблиці, кількість рядків, закодованих по способу LZW, дорівнює 18, а кількість рядків, закодованих по запропонованому способі - 9. Якщо таблиця кодування містить 4096 рядків (найбільше часто використовуваний випадок), то число біт, затрачуваних на кодування рядка $N_{\text{бр}} = 12$.

Таким чином, коефіцієнт стискання по способу LZW текстового повідомлення довжиною 19 символів дорівнює

$$K_{\text{ст LZW}} = 18 \times 12 / (19 \times 8) = 1,42.$$

Тобто, на початковій стадії кодування має місце розширення повідомлення.

Коефіцієнт стискання $K_{\text{ст зс}}$ того ж повідомлення по запропонованому способі дорівнює

$$K_{\text{ст зс}} = 9 \times 12 / (19 \times 8) = 0,71.$$

Таким чином, підвищення ефективності стискання повідомлення $\xi_{\text{ст}}$, запропонованим способом складає

$$\xi_{\text{ст}} = K_{\text{ст LZW}} / K_{\text{ст зс}} = 1,42 / 0,72 = 2.$$

Очевидно, що підвищення ступеня стискування в запропонованому способі досягається за рахунок трохи більших витрат пам'яті на збереження морфемних рядків у словнику. Однак при сучасних досягненнях мікроелектроніки, у тому числі, в області створення елементів пам'яті, декілька кілобайт пам'яті не грають істотної ролі.

Варто помітити, що при досить довгому вхідному повідомленні в таблиці кодування по способу LZW можуть бути поступово сформовані афіксальні морфеми, що зустрічаються кілька разів у вхідному повідомленні, що кодується, після чого коефіцієнт стискання на цій стадії кодування буде наближатися до коефіцієнта, що досягається в запропонованому способі. Однак у середньому коефіцієнт стискання запропонованого способу буде завжди вище за рахунок виключення явища розширення вихідної послідовності на початковій стадії стискання.

Таким чином, запропонований спосіб є найбільш ефективним при передачі коротких текстових запитів, збереженні і видачі довідок, а також різних текстових підказок, що широко використовуються в сучасних комп'ютерних і Інтернет-додатках.