МЕТОД ЗА КОМПРЕСИРАНЕ НА ТЕКСТОВА ИНФОРМАЦИЯ ПРИ ПРЕДАВАНЕ И СЪХРАНЕНИЕ

РУМЕН ТЕНЕВ КОРТЕНСКИ

Технически университет – София

I. УВОД

Бурното развитие на науката и техниката води до натрупване на все по-големи количества информация. Необходимостта от бързото й предаване и съхранение в намалени обеми памет изисква търсенето на нови методи за компресиране.

В настоящата статия се предлага метод за компресиране без загуба на информация. Той съчетава предимствата на статистическото моделиране и компресирането с речници [1], [2], [3]. По същество компресията се осъществява със статичен речник, който е съставен на базата на статистическо моделиране за всеки език поотделно. Предварителната информация е езикът, на който е текстът. Речникът се състои от фиксиран брой групи от букви с най-висока вероятност за даден език.

II. СЪЩНОСТ НА МЕТОДА

Той се състои в разделяне на текстта на често срещани срички и групи от букви по предварително изготвена таблица. Тя е допълнение към международната таблица на ASCII-кодовете до 256 комбинации. Така новите комбинации са с дължина 8 бита. Особеност на метода е неговата езикова зависимост. В Таблица 1 е представено разширение на ASCII-таблицата с примерно избраните най-често срещани срички и групи от букви за английски език и десетичните им ASCII-кодове. След подробен компютърен анализ на големи текстове от различни области на познанието и консултации със специалисти-филолози са възможни някои промени в нея. При ползване на метода за текстове на други езици е необходимо друго допълнение на таблицата с най-често срещаните срички, изработено от специалисти по съответния език. За демонстрация на метода и оценка на ефективността от него той е приложен върху произволен текст с дължина една стандартна машинописна страница – 30 реда с 60 знака на ред. Поради ограничения обем на статията текстът не е приложен. Общият брой на символите е 1718 . След компресирането новите комбинации са 866 броя. В Таблица 1 срещу всяка група символи е показано колко пъти тя се среща в разглеждания текст. Всяка от тях има 8-битов код. Следователно новото количество информация е 866 x 8=6928 бита. За оценка на ефекта от компресирането е необходимо да се изчисли

Таблица 1

97. _a_ -12	132. ct -	163. id - 7	194._not_- 1	225. ta - 1
98. ad - 2	133. da - 1	164. ie -	195. nt - 4	226. tch - 1
99. ai -	134. de - 5	165._if_ - 3	196. n't - 7	227. te - 3
100. al - 1	135. di - 2	166. il - 1	197. od - 1	228. th -21
101._all_ - 2	136._do_ -2	167. im -	198._of_ - 2	229._the_-10
102. am -	137. do - 3	168._in_ - 2	199. ol - 1	230. ti - 2
103._an_ - 1	138.don't-	169. in - 7	200. om - 2	231.tial -
104. an - 1	139. ea - 2	170.ing_ - 11	201._on_ - 2	232.tion - 1
105._and_-18	140. ed -10	171._is_ - 1	202. on - 1	233._to_ -15
106. ap -	141. em - 4	172. is - 1	203. op -	234. to - 7
107._are_-	142. en - 5	173._it_ -17	204. or - 3	235. tr - 4
108._as_ - 1	143. er -10	174. kn - 3	205. ou - 6	236. tt - 1
109._at_ - 1	144. es - 2	175. la - 4	206. ow - 6	237.un - 1
110. at - 5	145.ev - 2	176. id - 2	207. pa - 5	238._up_ - 3
111. ba -	146.ex -	177. le - 2	208. pe - 1	239._us_ -
112. be -14	147. fa - 1	178. li - 7	209. ph -	240. us - 4
113. bi - 2	148. fe - 2	179. ll - 4	210. pl -	241. va -
114. ble -	149. fi -	180. lo -	211. pr -	142. ve -10
115. bo - 2	150. fo - 2	181. ly - 4	212. qu - 1	243. wa - 8
116. br - 1	151._for_- 3	182. ma -	213. ra -	244._was_-13
117. bu - 2	152. fu -	183_me_ - 3	214. re - 8	245._we_ -10
118._but_- 5	153. ga - 4	184. me - 4	215. ri - 6	246. we - 3
119. ca - 6	154. ge - 3	185. mi - 4	216. sa -11	247. wh - 6
120. ce - 1	155. ght - 7	186. mo - 3	217. se - 5	248. wi - 4
121. ch -	156. go - 4	187._my_ - 2	218. sh - 7	249. wn - 4
122. ci - 1	157. ha - 5	188. na -	219._she_- 1	250. wo - 5
=======	158._he_ - 4	189. nd - 5	220 si - 2	251._you_-2
128.cion -	159. he - 8	190. ne - 1	221. so - 4	252. ._ -12
129. ck - 1	160. hi - 3	191. ni - 1	222. ss - 5	253. ,_ -12
130. cl - 2	161. ho - 5	192._no_ - 2	223.st -3	254.._-_ - 2
131. co - 7	162._I_ - 7	193. no - 4	224.sw -	255. ?_ - 3

Таблица 2

Брой поддържани символи	Брой битове на стария код	Количество информация [bit]	Коефициент на компресия
32	5	8 590	0,19
64	6	10 308	0,33
97	6.6	11 339	0,39
128	7	12 026	0,42

количеството информация преди него. В Таблица 2 е представено количеството информация преди компресията в зависимост от броя на битовете, кодиращи един символ. Коефициентът на компресия $K_{comp.}$ се изчислява като единица минус отношението между новото и старото количество информация:

$$K_{comp.} = 1 - \frac{A}{B}$$, където

A – ново количество информация,

B – старо количество информация.

Тъй като ASCII-таблицата е допълнена след 97-мия символ, то най-вярната оценка ще бъде относно стара таблица, поддържаща 97 символа. Тъй като $2^{6,6} = 97$, условно приемаме, че броят на битовете на стария код е 6,6 . Оттук за разглеждания примерен текст получаваме $K_{comp.}=0,39$. Както става ясно, този коефициент е хипотетичен. Реалният коефициент на компресия ще се получи в зависимост от дължината на стария код, наличието на допълнителна информация (за цвят, шрифт и др.) и вида на предавания текст.

III. КОДИРАНЕ И ДЕКОДИРАНЕ

Ясно е, че при прилагане на новия метод е необходимо кодиране и декодиране на информацията съгласно новата ASCII-таблица (Таблица 1). Предварително кодерът и декодерът получават информация на какъв език ще бъде текстът. Те активират съответна ASCII-таблица. В зависимост от скоростта на предаване по канала за връзка кодирането и декодирането могат да бъдат по време на предаването или по друго време. В зависимост от икономическата рентабилност те могат да се осъществяват по хардуерен или софтуерен път. Методът не изисква сложни изчисления както Хъфмановото [4] и статистическото компресиране, нито големи обеми памет както компресирането с речници. Елементарното

кодиране и декодиране води до използване на малко компютърни ресурси – прости процесори, малко време и памет.

IV. ЗАКЛЮЧЕНИЕ

С предложения метод текстовата информация се компресира с около 1/3 . Прилагането му ще доведе до намаляване на времето за предаване и до намаляване на обема памет за съхранение. Той използва малко компютърни ресурси. Това го прави изключително икономичен и непретенциозен. Компресирането се осъществява само върху ASCII-кодовете на символите. Следователно методът е неефективен при предаване на графична информация. Ако съществуват битове за цвят на символите и друга допълнителна информация, те се запазват, с което ефектът от компресирането относително намалява. Методът е приложим там, където се цели бързо предаване или съхранение в малък обем памет, а шрифтът и цветът са без значение. Такива места са телексните връзки, военни предавания за съкратено време, връзки с космически апарати, системи с ограничен обем памет и др.

След оценка на ефекта от метода в реални условия при продължително използване може да се експериментира кодирането по таблици с 512 или 1024 нови комбинации.

V. ЛИТЕРАТУРА

1. Ziff J., A. Lempel. A universal algorithm for sequential data compression. IEEE Transaction on Information Theory, vol. IT - 23, pp. 337 - 343, May 1977.
2. Ziff J., A. Lempel. Compression of individual sequences via variable-rate coding. IEEE Transaction on Information Theory, vol. IT 24, pp. 530 - 536, September 1978.
3. Митев Д., С. Минчев. Съвременно компресиране на данни, II част. С., InfoPress, 1994.
4. Митев Д., С. Минчев. Съвременно компресиране на данни, I част. С., InfoPress, 1994.

METHOD FOR TEXT INFORMATION COMPRESSING DURING TRANSFER AND STORING

ROUMEN TENEV KORTENSKY
Tecnical University, Sofia - 1156, BULGARIA

RESUME

The rapid development of science and technology leads to the accumulation of great quantities of information. The nessesity of quick transfer and storage in smaller volumes of memory requires search for new methods of compressing.

The study proposes a method for compression without infornation losses. It combines the advantages of the statistic modeling and the compression using dictionary [1], [2], [3]. In fact the compression is accomplished by static dictionary which has been compiled separately for the different languages. Preleminary information is the language of the particular text. The dictionary consist of fixed combinations of letters with the highest probbability to be encountered in the particular language.

The main point of the method is dividing the text into often encountered syllables and combination of letters according to an earlier chosen table. It is an addition to the international ASCII-code table, completing it to 256 combinations. This allows the text to be compressed by about 1/3. The new combinations are 8-bit long. This method does not require complicated calculations in contrast to Huffman's [4] and statistical compression, neither big memory volume as the compression with dictionary does. Elementary coding and decoding leads to the usage of small computer resources - simple processors, less time and memory. All this makes it very simple and economical.

After a long term assessment of the method's efficiency in real conditions is carried out coding according to tables with 512 or 1024 new combinations can be experimented.

REFERENCES

1. Ziff J., A. Lempel. A universal algorithm for sequential data compression. IEEE Transaction on Information Theory, vol. IT - 23, pp. 337 - 343, May 1977.
2. Ziff J., A. Lempel. Compression of individual sequences via variable-rate coding. IEEE Transaction on Information Theory, vol. IT 24, pp. 530 - 536, September 1978.
3. Mitev, D, Mintchev, S. Modern data compression. Vol. II, Sofia, Infopress. 1994
4. Mitev, D, Mintchev, S. Modern data compression. Vol. I, Sofia, Infopress. 1994