Δομές Δεδομένων και Αλγόριθμοι

Χρήστος Γκόγκος

ΤΕΙ Ηπείρου

Χειμερινό Εξάμηνο 2014-2015
Παρουσίαση 20. Huffman codes

Κωδικοποίηση σταθερού μήκους

- Αν χρησιμοποιηθεί κωδικοποίηση σταθερού μήκους δηλαδή όλα τα σύμβολα έχουν κωδικοποίηση με το ίδιο μήκος m και το πλήθος των συμβόλων που πρέπει να κωδικοποιηθούν είναι n τότε ισχύει ότι $m = \lceil \log_2 n \rceil$.
- Παραδείγματα κωδικοποιήσεων σταθερού μήκους είναι η κωδικοποίηση ASCII (8 bits) και η κωδικοποίηση UNICODE (16 bits).
- Η κωδικοποίηση Huffman **δεν** είναι κωδικοποίηση σταθερού μήκους.

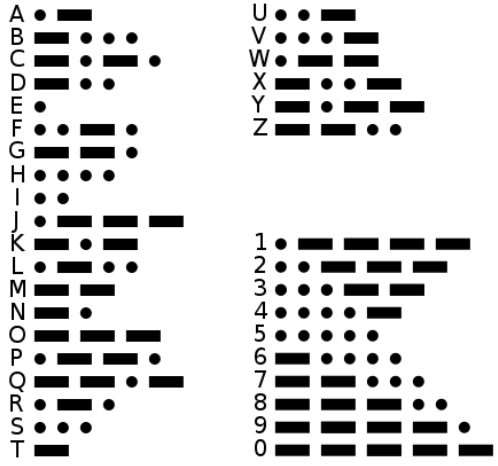
Απόσπασμα κωδικοποίησης ASCII

ASCII Code: Character to Binary

0	0011 0000	O	0100 1111	m	0110 1101
1	0011 0001	P	0101 0000	n	0110 1110
2	0011 0010	Q	0101 0001	o	0110 1111
3	0011 0011	R	0101 0010	p	0111 0000
4	0011 0100	S	0101 0011	q	0111 0001
5	0011 0101	T	0101 0100	r	0111 0010
6	0011 0110	U	0101 0101	s	0111 0011
7	0011 0111	V	0101 0110	t	0111 0100
8	0011 1000	W	0101 0111	u	0111 0101
9	0011 1001	X	0101 1000	v	0111 0110
A	0100 0001	Y	0101 1001	w	0111 0111
B	0100 0010	Z	0101 1010	x	0111 1000
C	0100 0011	a	0110 0001	y	0111 1001
D	0100 0100	b	0110 0010	z	0111 1010
E	0100 0101	c	0110 0011	.	0010 1110
F	0100 0110	d	0110 0100	,	0010 0111
G	0100 0111	e	0110 0101	:	0011 1010
H	0100 1000	f	0110 0110	;	0011 1011
I	0100 1001	g	0110 0111	?	0011 1111
J	0100 1010	h	0110 1000	!	0010 0001
K	0100 1011	I	0110 1001	'	0010 1100
L	0100 1100	j	0110 1010	"	0010 0010
M	0100 1101	k	0110 1011	(	0010 1000
N	0100 1110	l	0110 1100	)	0010 1001
				space	0010 0000

http://www3.amherst.edu/~jcook15/binarycode.html

Κωδικοποίηση μεταβλητού μήκους - ο κώδικας Morse



http://en.wikipedia.org/wiki/Morse_code

- Ο κώδικας Morse προτάθηκε από τον Samuel Morse στα μέσα του 19ου αιώνα και χρησιμοποιήθηκε στον τηλέγραφο.
- Κάθε χαρακτήρας αναπαρίσταται από μια μοναδική ακολουθία από τελείες και παύλες.
- Η διάρκεια κάθε παύλας είναι τριπλάσια της διάρκειας κάθε τελείας
- Κάθε τελεία ή παύλα ακολουθείται από μια παύση ίση με τη διάρκεια της τελείας.
- Το συνηθέστερο γράμμα της Αγγλικής αλφαβήτου, το E, έχει και τη συντομότερη κωδικοποίηση, μια τελεία.
- Γράμματα όπως το Y που δεν είναι συχνά σε τυπικά κείμενα στην αγγλική γλώσσα έχουν κωδικοποιήσεις με μεγαλύτερο μήκος.

Η κωδικοποίηση Huffman

Η κωδικοποίηση Huffman προτάθηκε από τον David Huffman το 1952 και στηρίζεται στην ιδέα της αντιστοίχισης μικρότερων σε μήκος κωδικοποιήσεων σε συχνά εμφανιζόμενα σύμβολα και μεγαλύτερων σε μήκος κωδικοποιήσεων σε λιγότερο συχνά εμφανιζόμενα σύμβολα. Συνεπώς πρόκειται για μια κωδικοποίηση μεταβλητού μήκους.

Τυπικοί λόγοι συμπίεσης που επιτυγχάνονται με τη κωδικοποίηση Huffman είναι από 20% έως 80%.

Το πρόβλημα της κωδικοποίησης μεταβλητού μήκους

Πως μπορούμε να γνωρίζουμε πόσα bits του κωδικοποιημένου μηνύματος αντιστοιχούν στο πρώτο σύμβολο, πόσα στο δεύτερο σύμβολο κ.ο.κ.;

Για παράδειγμα η ακόλουθη κωδικοποίηση

- Α →0
- Β →1
- Γ →01
- Δ →10
- Ε →11

δε μπορεί να χρησιμοποιηθεί για να αντιστοιχηθεί μονοσήμαντα μια ακολουθία δυαδικών ψηφίων όπως η 011011 στα κατάλληλα σύμβολα καθώς μπορεί να αποκωδικοποιηθεί ως ΓΔΕ (01-10-11), ΑΕΓΒ (0-11-01-1), ΑΒΒΑΕ (0-1-1-0-11), ...

Κωδικοποίηση μεταβλητού μήκους ελεύθερη από προθέματα - prefix free

Κανένας κωδικός δε θα πρέπει να αποτελεί πρόθεμα άλλου κωδικού. Αυτό επιτυγχάνεται με αναπαράσταση των συμβόλων δημιουργώντας σταδιακά ένα δυαδικό δένδρο στο οποίο οι αριστερές ακμές αναπαρίστανται με μηδέν και οι δεξιές ακμές αναπαρίστανται με ένα (ή και αντίστροφα). Η δημιουργία του δένδρου γίνεται με τον αλγόριθμο του Huffman.

Σε μια prefix free κωδικοποίηση, μια ακολουθία δυαδικών ψηφίων εξετάζεται από αριστερά προς τα δεξιά μέχρι να βρεθεί η πρώτη υποακολουθία δυαδικών ψηφίων που αποτελεί κωδικοποίηση για κάποιο σύμβολο. Στη συνέχεια αντικαθίσταται η υποακολουθία με το σύμβολο και η διαδικασία επαναλαμβάνεται μέχρι το τέλος της ακολουθίας δυαδικών ψηφίων.

Αλγόριθμος του Huffman για την κωδικοποίηση ενός αλφαβήτου με n σύμβολα

1. Αρχικοποίηση n κόμβων. Κάθε κόμβος έχει ως βάρος τη συχνότητα εμφάνισης του αντίστοιχου συμβόλου.
2. Εύρεση των 2 κόμβων με τα μικρότερα βάρη και δημιουργία νέου κόμβου με βάρος το άθροισμα των βαρών των 2 κόμβων και αριστερό και δεξιό παιδί το καθένα από τους 2 κόμβους. Η διαδικασία επαναλαμβάνεται μέχρι όλοι οι κόμβοι να ενσωματωθούν σε ένα δυαδικό δένδρο.

Η κωδικοποίηση κάθε συμβόλου προκύπτει διαβάζοντας τα δυαδικά ψηφία που βρίσκονται πάνω στις ακμές διανύοντας το δένδρο από την κορυφή προς το αντίστοιχο σύμβολο που αποτελεί φύλλο του δένδρου.

Παράδειγμα κωδικοποίησης Huffman (1/2)

Έστω ένα αλφάβητο με τα 5 σύμβολα Α, Β, Γ, Δ, Ε με συχνότητες εμφάνισης του κάθε συμβόλου 0.35, 0.1, 0.2, 0.2, 0.15 αντίστοιχα.

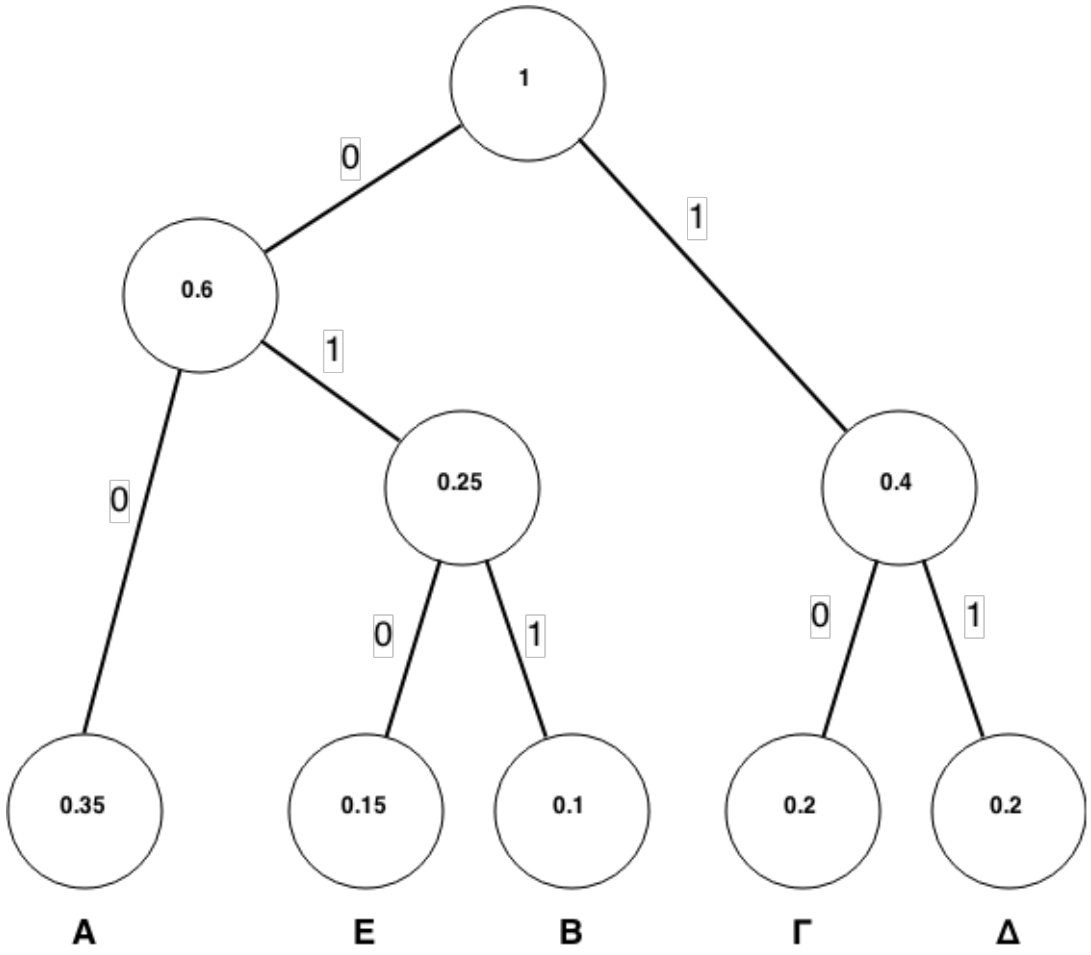
Κωδικοποίηση σταθερού μήκους

Σε περίπτωση που επιλέγει κωδικοποίηση σταθερού μήκους απαιτούνται τουλάχιστον 3 bits για κάθε σύμβολο καθώς το πλήθος των συμβόλων είναι 5. Μια πιθανή κωδικοποίηση μπορεί να είναι η ακόλουθη: Α →000, Β →001, Γ →010, Δ →011, Ε →100.

Κωδικοποίηση Huffman

Σε περίπτωση που επιλεγεί η κωδικοποίηση Huffman μια κωδικοποίηση μπορεί να είναι η ακόλουθη: Α →00, Β →011, Γ →10, Δ →11, Ε →010. Το αναμενόμενο μήκος ανά χαρακτήρα ισούται με τα άθροισμα των γινομένων που προκύπτουν πολλαπλασιάζοντας για κάθε χαρακτήρα το μήκος σε bits της κωδικοποίησης του επί τη συχνότητα εμφάνισης του αντίστοιχου χαρακτήρα. Άρα το αναμενόμενο μέσο μήκος ανά χαρακτήρα του παραδείγματος είναι 2*0.35 + 3*0.1 + 2*0.2 + 2*0.2 +3*0.15 = 2.25. Ο δε λόγος συμπίεσης στο συγκεκριμένο παράδειγμα είναι $\frac{3-2.25}{3} = 0.25$.

Παράδειγμα κωδικοποίησης Huffman (2/2)



Καθώς δεν υπάρχει κάποιο μονοπάτι από την ρίζα προς ένα φύλλο που να συνεχίζει σε άλλο φύλλο του δένδρου δεν μπορεί μια κωδικοποίηση να αποτελεί προοίμιο (prefix) μιας άλλης κωδικοποίησης.

Δυναμική κωδικοποίηση Huffman

- Τυπικά, στην κωδικοποίηση Huffman πραγματοποιείται προ-επεξεργασία των δεδομένων στην οποία υπολογίζεται η συχνότητα εμφάνισης του κάθε συμβόλου. Στη συνέχεια κατασκευάζεται το δένδρο Huffman. Ωστόσο, μαζί με το κωδικοποιημένο μήνυμα πρέπει να περιλαμβάνεται και ένα λεξικό που θα περιέχει την κωδικοποίηση κάθε συμβόλου που έχει προκύψει έτσι όπως αποτυπώνεται στο δένδρο Huffman.
- Προκειμένου να μην αποθηκεύεται η κωδικοποίηση του δένδρου στο κείμενο κωδικοποίησης η κωδικοποίηση μπορεί να ενημερώνεται κάθε φορά που προστίθεται ένα νέο σύμβολο. Αλγόριθμοι που υλοποιούν αυτή την ιδέα είναι ο αλγόριθμος Vitter και ο αλγόριθμος FGK. Οι τεχνικές αυτές είναι γνωστές ως dynamic ή adaptive κωδικοποιήσεις Huffman και βρίσκουν εφαρμογή σε περιπτώσεις που δεν είναι γνωστή εκ των προτέρων η κατανομή των συμβόλων που πρέπει να κωδικοποιηθούν.

Εφαρμογές της κωδικοποίησης Huffman

- Η κωδικοποίηση Huffman επιτρέπει τη μη απωλεστική (lossless) συμπίεση δεδομένων.
- Χρησιμοποιείται σε διάφορα μορφότυπα συμπίεσης gzip, pkzip, bzip2 κ.α.
- Χρησιμοποιείται σε διάφορα μορφότυπα εικόνων όπως το jpeg και το png.
- Χρησιμοποιείται στη συμπίεση ήχου mp3.
- Η κωδικοποίηση Huffman βρίσκει εφαρμογή σε προβλήματα που μπορούν να μοντελοποιηθούν ως δένδρα απόφασης (decision trees).