Τμήμα Πληροφορικής Α.Π.Θ. 2010-2011

Ανάλυση κατακερματισμού με ανοικτή διεύθυνση

Δομές Δεδομένων *Απόστολος Ν. Παπαδόπουλος*

1 Εισαγωγή

Στο τμήμα αυτό θα μελετήσουμε τη δομή δεδομένων του κατακερματισμού με ανοικτή διεύθυνση (open addressing), και για την ακρίβεια θα δώσουμε κλειστούς τύπους για το μέσο αριθμό των ελέγχων (probes) που αναμένονται να πραγματοποιηθούν για ανεπιτυχή και επιτυχή αναζήτηση. Σημειώνεται ότι μία ανεπιτυχής αναζήτηση έχει ως αποτέλεσμα να μη βρεθεί το στοιχείο που ψάχνουμε, σε αντίθεση με την επιτυχή αναζήτηση όπου το στοιχείο που αναζητούμε είναι ένα από τα στοιχεία του πίνακα. Η συζήτησή μας βασίζεται στην ανάλυση που παρουσιάζεται στο βιβλίο [1].

2 Ανάλυση

Έστω πίνακας κατακερματισμού μεγέθους m. Αν ο πίνακας έχει αποθηκευμένα n στοιχεία, τότε ορίζουμε ως $\alpha = n/m$ τον παράγοντα φόρτωσης (load factor) του πίνακα. Είναι προφανές ότι η τιμή του α βρίσκεται μεταξύ του 0 (άδειος πίνακας) και του 1 (γεμάτος πίνακας). Στη συνέχεια θα αποδείξουμε μερικά πολύ χρήσιμα θεωρήματα που μας επιτρέπουν να αξιολογήσουμε την απόδοση του κατακερματισμού ανοικτής διεύθυνσης. Θα βρούμε κλειστούς τύπους για: i) το μέσο πλήθος probes για ανεπιτυχή αναζήτηση, ii) το μέσο πλήθος probes για εισαγωγή ενός στοιχείου και iii) το μέσο πλήθος probes για επιτυχή αναζήτηση.

Θα κάνουμε την υπόθεση ότι κατά την αναζήτηση ή την εισαγωγή ενός στοιχείου, κάθε μία από τις $m!$ διαφορετικές ακολουθίες ελέγχων μπορεί να συμβεί με την ίδια πιθανότητα. Έχουμε δηλαδή *ομοιόμορφο κατακερματισμό* (uniform hashing). Επίσης, για τις ανάγκες της ανάλυσης θα υποθέσουμε ότι οι τιμές n και m είναι αρκετά μεγάλες (πηγαίνουν στο άπειρο).

Θεώρημα 2.1. *Με την υπόθεση ότι έχουμε ομοιόμορφο κατακερματισμό, το μέσο πλήθος των probes που απαιτούνται για μια ανεπιτυχή αναζήτηση σε έναν πίνακα κατακερματισμού με ακοικτή διεύθυνση είναι το πολύ* $1/(1-\alpha)$*, όπου* $\alpha = n/m < 1$ *ο παράγοντας φόρτωσης του πίνακα.*

Απόδειξη. Έστω ότι αναζητούμε ένα στοιχείο x που δεν υπάρχει στον πίνακα κατακερματισμού. Αυτό σημαίνει ότι όλες οι προσπελάσεις στον πίνακα (εκτός από την τελευταία) μας οδηγούν σε θέση που περιέχει ένα στοιχείο διαφορετικό από το x, ενώ ο τελευταίος έλεγχος μας οδηγεί σε άδειο κελί. Ορίζουμε ως p_i την πιθανότητα **ακριβώς** i θέσεις του πίνακα να είναι γεμάτες κατά τη διαδικασία των ελέγχων για την αναζήτηση ενός στοιχείου. Πιο συγκεκριμένα:

$$p_i = Prob\{\text{ακριβώς } i \text{ έλεγχοι οδηγούν σε γεμάτες θέσεις}\}$$

Είναι προφανές ότι όταν $i > n$ τότε $p_i = 0$. Από την προηγούμενη συζήτηση προκύπτει ότι αν συμβολίσουμε με $\overline{probes}$ το μέσο αριθμό των ελέγχων, σύμφωνα με τον ορισμό της μέσης τιμής θα έχουμε:

$$\overline{probes} = 1 + \sum_{i=0}^{\infty} i \cdot p_i \tag{1}$$

Για τον υπολογισμό του αθροίσματος θα χρησιμοποιήσουμε μία από τις σημαντικές ιδιότητες της μέσης τιμής. Ορίζουμε ως q_i την πιθανότητα **τουλάχιστον** i θέσεις του πίνακα να είναι γεμάτες κατά τη διαδικασία των ελέγχων για την αναζήτηση ενός στοιχείου. Τότε έχουμε:

$$\overline{probes} = 1 + \sum_{i=0}^{\infty} i \cdot p_i = 1 + \sum_{i=0}^{\infty} q_i \tag{2}$$

Ας προσδιορίσουμε την τιμή του q_i για $i >= 1$. Η πιθανότητα ο πρώτος έλεγχος να οδηγήσει σε γεμάτη θέση του πίνακα κατακερματισμού ισούται με n/m, αφού οι συνολικές θέσεις είναι m από τις οποίες οι n είναι γεμάτες. Άρα με πιθανότητα n/m θα χρειαστεί δεύτερος έλεγχος, ο οποίος με πιθανότητα $(n-1)/(m-1)$ θα οδηγήσει πάλι σε γεμάτη θέση και θα χρειαστεί τρίτος έλεγχος. Επομένως έχουμε έως τώρα: $q_1 = n/m$, $q_2 = n/m \cdot (n-1)/(m-1)$. Ακολουθώντας το σκεπτικό αυτό, η πιθανότητα να χρειαστούν τουλάχιστον i έλεγχοι είναι:

$$q_i = \frac{n}{m} \cdot \frac{n-1}{m-1} \cdot ... \cdot \frac{n-i+1}{m-i+1} \tag{3}$$

Εδώ θα χρησιμοποιήσουμε μία γνωστή ιδιότητα των κλασμάτων, που μας λέει ότι $n/m \leq (n-j)/(m-j)$ όταν $n < m$ και $j \geq 0$. Σύμφωνα με αυτήν την ιδιότητα και με αντικατάσταση στην παραπάνω σχέση παίρνουμε την ανισότητα:

$$q_i \leq \left(\frac{n}{m}\right)^i = \alpha^i \tag{4}$$

Σύμφωνα με τη σχέση 4 η σχέση 2 γίνεται:

$$\overline{probes} = 1 + \sum_{i=0}^{\infty} q_i \leq 1 + \alpha + \alpha^2 + \alpha^3 + ... = \frac{1}{1-\alpha}$$

□

Παρατηρούμε το σημαντικό ρόλο που παίζει ο παράγοντας φόρτωσης στην απόδοση του πίνακα κατακερματισμού. Για παράδειγμα, αν ο πίνακας είναι γεμάτος κατά 50%, άρα $\alpha = 0.5$, τότε με απλή αντικατάσταση θα έχουμε 2 ελέγχους στη μέση περίπτωση, ενώ αν είναι γεματός κατά 90%, άρα $\alpha = 0.9$, τότε ο μέσος αριθμός ελέγχων γίνεται 10 (πάντα για την περίπτωση που αναζητούμε στοιχείο που δεν υπάρχει στον πίνακα).

Εύκολα μπορούμε να δείξουμε ότι ο μέσος αριθμός ελέγχων για την εισαγωγή ενός νέου στοιχείου σε έναν πίνακα που έχει ήδη n στοιχεία και συνολικά m θέσεις είναι πάλι $1/(1-\alpha)$.

Θα προχωρήσουμε στη συνέχεια στη μελέτη του μέσου αριθμού των ελέγχων στην περίπτωση της επιτυχούς αναζήτησης, όταν δηλαδή το στοιχείο που αναζητούμε υπάρχει στον πίνακα κατακερματισμού. Θεωρούμε, ότι το κάθε ένα από τα $n < m$ αποθηκευμένα στοιχεία μπορεί να ζητηθεί με την ίδια πιθανότητα. Ισχύει το ακόλουθο θεώρημα:

Θεώρημα 2.2. *Σε έναν πίνακα κατακερματισμού με ανοικτή διεύθυνση όπου $\alpha = n/m < 1$ ο παράγοντας φόρτωσης, ο μέσος αριθμός ελέγχων για επιτυχή αναζήτηση, θεωρώντας ομοιόμορφο κατακερματισμό και ότι κάθε στοιχείο μπορεί να ζητηθεί με την ίδια πιθανότητα, είναι:*

$$\overline{probes} = \frac{1}{\alpha} \cdot \ln \frac{1}{1-\alpha} + \frac{1}{\alpha}$$

Απόδειξη. Για την αναζήτηση ενός στοιχείου k, στην ουσία εφαρμόζεται η ακολουθία ελέγχων που πραγματοποιήθηκε κατά την εισαγωγή του στοιχείου. Εάν το k ήταν το $(i+1)$-οστό στοιχείο που μπήκε στον πίνακα, ο μέσος αριθμός ελέγχων που απαιτείται για την εύρεση του k στον πίνακα είναι $1/(1-i/m) = m/(m-i)$. Λαμβάνοντας το μέσο όρο για όλα τα n στοιχεία που βρίσκονται στον πίνακα (άρα από $i=0$ έως $n-1$) έχουμε ότι:

$$\overline{probes} = \frac{1}{n}\sum_{i=0}^{n-1}\frac{m}{m-i} = \frac{m}{n}\sum_{i=0}^{n-1}\frac{1}{m-i}$$

Είναι γνωστό ότι το άθροισμα $\sum_{j=1}^{k}(1/j)$ καλείται και k-οστός αρμονικός αριθμός. Αν λοιπόν συμβολίσουμε με H_k το άθροισμα αυτό, τότε από την παραπάνω σχέση παίρνουμε:

$$\overline{probes} = \frac{1}{\alpha}(H_m - H_{m-n})$$

Από τις ιδιότητες των αρμονικών αριθμών είναι γνωστό ότι η τιμή του H_k είναι μεγαλύτερη από $\ln k$ και μικρότερη από $\ln k + 1$, δηλαδή:

$$\ln k \leq H_k \leq \ln k + 1$$

Με αντικατάσταση $H_m \leq \ln m + 1$ και $H_{m-n} \geq \ln(m-n)$ (επειδή έχουμε αφαίρεση!) παίρνουμε την παρακάτω ανισότητα:

$$\overline{probes} = \frac{1}{\alpha}(H_m - H_{m-n}) \leq \frac{1}{\alpha}(\ln m + 1 - \ln(m-n)) = \frac{1}{\alpha}\ln\frac{m}{m-n} = \frac{1}{\alpha}\ln\frac{1}{1-\alpha} + \frac{1}{\alpha} \qquad (5)$$

□

Βιβλιογραφία

[1] T. H. Cormen, C. E. Leiserson, R. L. Rivest, "*Introduction to algorithms*", MIT Press, 1990.