ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΗΛΕΚΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΔΟΜΕΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΡΧΕΙΩΝ

Ραφαήλ Τσιριβάκος 2013030199
LAB20243880

$3_η$ άσκηση
Τεκμηρίωση των αποτελεσμάτων

Στον πίνακα που φαίνεται δεξιά παρουσιάζονται οι ζητούμενες μετρήσεις. Δηλαδή, για κάθε εισαγωγή 100 στοιχείων στον πίνακα κατακερματισμού, ο μέσος όρος συγκρίσεων που προκύπτει. Αντίστοιχα, φαίνονται οι μέσοι όροι που προκύπτουν ανά αναζήτηση 50 στοιχείων και ανά διαγραφή 50 στοιχείων από τον πίνακα. Οι μέσοι όροι έχουν υπολογιστεί για κάθε 100 στοιχεία που εισάγονται έως ότου φτάσουν τα N=10000.

N	u>50% insert	u>50% search	u>50% delete	u>80% insert	u>80% search	u>80% delete	bst search
100	8	8	16	8	8	16	27
200	8	10	21	8	11	20	32
300	8	12	25	8	12	24	34
400	8	14	29	8	14	27	35
500	8	16	30	8	16	33	37
600	15	18	35	8	17	37	40
700	17	17	37	8	21	40	41
800	19	19	35	9	21	46	38
900	20	18	35	15	24	47	41
1000	19	16	34	17	25	53	40
1100	15	15	32	17	26	44	42
1200	16	17	37	18	26	53	43
1300	17	19	32	18	22	46	45
1400	17	18	33	17	26	50	43
1500	18	17	37	18	22	46	45
1600	19	16	35	16	20	39	45
1700	20	18	36	15	20	42	43
1800	20	18	39	15	20	44	42
1900	20	17	32	16	22	45	46
2000	19	15	32	16	22	42	43
2100	15	14	31	16	21	46	44
2200	16	14	33	16	22	45	44
2300	16	15	33	17	21	45	46
2400	16	16	36	16	21	43	45
2500	16	14	36	18	20	42	46
2600	18	14	39	17	20	45	50
2700	18	16	36	18	19	43	47
2800	18	15	36	17	20	47	49
2900	18	16	36	18	22	46	45
3000	17	17	39	17	20	42	47
3100	19	17	35	18	18	42	47
3200	19	16	40	16	16	40	47
3300	20	16	38	15	17	37	47
3400	19	16	40	14	16	40	47
3500	20	16	38	16	18	44	52
3600	20	15	40	14	17	40	48
3700	21	17	40	16	19	42	46
3800	20	14	39	15	20	39	48
3900	21	14	36	17	19	44	46
4000	20	14	37	16	20	42	48
4100	15	14	35	16	18	41	48
4200	15	14	34	16	20	44	48
4300	15	14	36	17	19	42	48
4400	15	13	35	16	19	42	45
4500	16	16	37	17	18	46	48
4600	16	14	37	17	20	46	51
4700	17	14	38	17	18	43	51
4800	16	14	36	16	19	45	45
4900	17	14	39	17	20	44	49
5000	17	14	37	17	20	43	46

Η δεύτερη, η τρίτη και η τέταρτη στήλη παρουσιάζουν τις μετρήσεις στην περίπτωση που το κριτήριο διάσπασης σελίδας είναι u>50%, ενώ οι επόμενες τρεις στήλες είναι για την περίπτωση που το κριτήριο είναι u>80%. Στην τελευταία στήλη παρουσιάζεται ο μέσος όρος συγκρίσεων ανά αναζήτηση 50 στοιχείων σε ένα δυαδικό δέντρο έρευνας, στο οποίο τα στοιχεία εισάγονται ανά 100 και είναι τα ίδια με αυτά που χρησιμοποιήθηκαν στον πίνακα κατακερματισμού. Τα κλειδιά που αναζητήθηκαν στο δέντρο είναι τα ίδια με αυτά που αναζητήθηκαν στον πίνακα κατακερματισμού με κριτήριο u>80%. Στην άλλη περίπτωση τα αποτελέσματα δεν είχαν διαφορά από αυτά που τελικά κράτησα.

Απόδοση Γραμμικού Κατακερματισμού

Ανεξάρτητα ποιο δείκτη διάσπασης εφαρμόζεται, παρουσιάζονται συγκεκριμένα χαρακτηριστικά στα αποτελέσματα.

εισαγωγή

Στην εισαγωγή των στοιχείων παρατηρώ μια περιοδικότητα. Αυτό συμβαίνει καθώς ο αριθμός των συγκρίσεων είναι μεγαλύτερος όταν γίνεται η εισαγωγή κλειδιού στον πίνακα και γίνεται διάσπαση. Αφού γίνει η διάσπαση, ο αριθμός συγκρίσεων μειώνεται μέχρι και την επόμενη διάσπαση. Όσο μεγαλώνει το N, ο αριθμός των κλειδιών που έχουν εισαχθεί στον πίνακα δηλαδή, μεγαλώνει και η περίοδος που έχουμε ψηλές ή χαμηλές τιμές. Αυτό συμβαίνει διότι εισάγουμε πάντα 100 κλειδιά σε ένα πίνακα που μεγαλώνει, κι έτσι καθυστερεί να ξεπεραστεί ο δείκτης διάσπασης. Αυτό βέβαια είναι εντονότερο όταν έχουμε χαμηλό δείκτη διάσπασης. Αυτό γιατί όταν είναι υψηλός, έχουμε λιγότερες διασπάσεις, οπότε οι τιμές του μέσου

όρου συγκρίσεων κυμαίνονται σε μια πιο περιορισμένη περιοχή. Τέλος, όσων αφορά την πράξη της εισαγωγής, οι μετρήσεις είναι πάρα πολύ χαμηλές για τις πρώτες εισαγωγές, μέχρι να γίνει η πρώτη διάσπαση. Πέρα αυτού κοιτώντας μακροσκοπικά τις τιμές, έχουν μια γραμμική μεταβολή, καθώς μεταξύ τους δεν έχουν μεγάλες διαφορές και σχεδόν επαναλαμβάνονται. Αυτό φαίνεται και από το διάγραμμα παρακάτω, στο οποίο φαίνεται και η σταδιακή μείωσή τους.

5100	17	15	37	17	19	44	49
5200	17	14	37	17	18	47	48
5300	17	15	34	18	20	46	48
5400	18	14	41	17	18	47	50
5500	18	16	38	18	19	41	47
5600	18	14	43	17	17	46	47
5700	17	15	40	18	17	46	51
5800	17	14	38	17	16	42	47
5900	18	14	42	18	16	44	49
6000	19	14	39	17	17	42	48
6100	19	13	38	18	19	45	51
6200	19	15	41	17	16	39	50
6300	19	14	41	18	19	42	48
6400	19	14	41	16	17	37	50
6500	19	15	40	15	16	42	51
6600	20	14	41	14	18	41	49
6700	19	13	42	16	15	38	49
6800	20	14	41	14	16	41	52
6900	20	13	43	15	15	42	51
7000	21	13	41	14	14	40	51
7100	20	14	36	17	17	42	48
7200	19	13	40	15	17	43	49
7300	21	15	42	15	15	44	50
7400	20	13	40	16	16	44	55
7500	20	13	42	16	15	41	48
7600	20	12	46	15	13	43	51
7700	21	13	39	16	16	41	49
7800	21	13	43	16	17	46	49
7900	21	12	40	16	16	46	48
8000	19	12	43	15	16	43	50
8100	16	11	38	16	16	44	49
8200	14	11	39	15	16	46	49
8300	16	11	39	17	17	44	52
8400	16	11	39	16	17	46	51
8500	16	12	39	17	15	47	50
8600	15	12	40	15	16	46	53
8700	17	13	38	17	17	48	51
8800	15	12	39	16	17	44	50
8900	16	12	41	17	15	45	49
9000	15	12	38	16	14	44	50
9100	15	12	40	17	15	46	50
9200	17	11	43	16	16	48	52
9300	16	11	39	17	15	43	48
9400	17	11	43	16	15	48	53
9500	16	12	43	18	15	44	52
9600	17	11	41	16	15	46	53
9700	17	11	39	18	16	50	53
9800	16	11	42	17	17	46	48
9900	16	11	43	18	14	45	50
10000	16	10	42	17	13	48	49

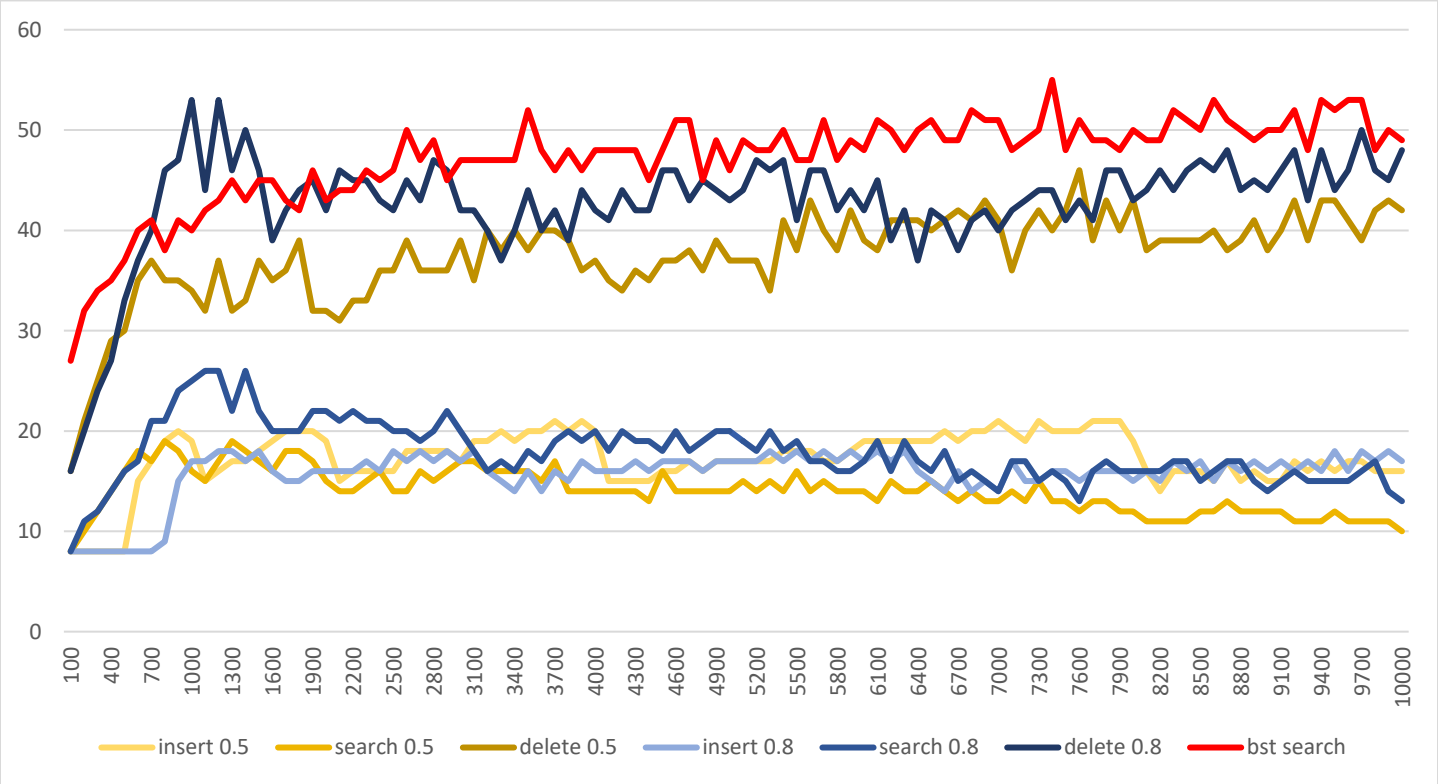
αναζήτηση

Αντίθετα με την πράξη της εισαγωγής, στην αναζήτηση εμφανίζονται υψηλές μετρήσεις από τις πρώτες εισαγωγές. Μάλιστα, οι υψηλότερες είναι λίγο μετά τις πρώτες εισαγωγές, όπου έχουμε τις πρώτες διασπάσεις αλλά ακόμα πολύ γεμάτες θέσεις. Αυτό συμβαίνει διότι ο αριθμός των συγκρίσεων που γίνονται είναι μεγαλύτερος όταν έχουμε θέσεις με πολλά κλειδιά ή ακόμα και με υπερχείλιση. Αυτές οι περιπτώσεις μειώνονται με τις διασπάσεις, όπου τα κλειδιά μοιράζονται. Όσο λοιπόν αυξάνεται ο αριθμός των στοιχείων που είναι μέσα στον πίνακα, μεγαλώνει και το μέγεθος του πίνακα, με βάση το κριτήριο διάσπασης που λειτουργεί. Έτσι είναι λογικό αρχικά να έχουμε υψηλές μετρήσεις που σταδιακά μειώνονται. Όσο αφορά το κριτήριο διάσπασης, όταν είναι ψηλό το όριο, όπως στη συγκεκριμένη υλοποίηση για u>80%, αργεί ο πίνακας να μεγαλώσει, σε σχέση με την υλοποίηση όπου u>50%. Οπότε τοποθετούνται πολλά στοιχεία σε κάθε θέση και δημιουργούνται πολλές θέσεις υπερχείλισης. Αυτό έχει σαν αποτέλεσμα να έχουμε μεγάλες τιμές στο μέσο όρο συγκρίσεων ανά αναζήτηση όταν το όριο διάσπασης είναι μεγάλο και μικρότερες όταν το όριο διάσπασης είναι μικρό. Και στις δύο περιπτώσεις όμως, σημειώνεται σταδιακή μείωση και φαίνεται καλύτερα στο παρακάτω διάγραμμα, καθώς επίσης φαίνεται η γραμμική μεταβολή των τιμών όσο αυξάνεται ο αριθμός των κλειδιών στον πίνακα, όπως και στην πράξη της εισαγωγής.

διαγραφή

Αντίστοιχη συμπεριφορά με αυτή της πράξης της αναζήτησης παρουσιάζεται στην πράξη της διαγραφής. Αυτό γιατί η διαγραφή βασίζεται πάνω στην αναζήτηση, οπότε η εξάρτηση στις διασπάσεις του πίνακα είναι αντίστοιχη, άρα η λογική σε πρώτο στάδιο είναι ίδια. Φαίνεται και από το παρακάτω διάγραμμα πιο έντονα, αφού η καμπύλη της διαγραφής ακολουθεί ίδια περιοδικότητα στις αυξομειώσεις των τιμών με αυτή της αναζήτησης, την οποία ακολουθεί και η καμπύλη της εισαγωγής. Στο επόμενο στάδιο όμως, στο οποίο οφείλεται και η μεγάλη διαφορά των τιμών των μετρήσεων της διαγραφής από τις άλλες δύο πράξεις, ο αριθμός των συγκρίσεων μεγαλώνει αρκετά, καθώς, στην περίπτωση που γίνει διαγραφή κάποιου στοιχείου έχουμε μετακίνηση όλων των ακόλουθών του στοιχείων ώστε να μην υπάρχει κενό. Πέρα αυτού, έχουμε πολλές συγκρίσεις όταν ικανοποιείται το κριτήριο συγχώνευσης. Όλα αυτά έχουν σαν αποτέλεσμα όταν έχουμε κριτήριο διάσπασης u>80% να έχουμε περισσότερες συγκρίσεις αρχικά λόγω των αναζητήσεων σε θέσεις με πολλά κλειδιά ή με υπερχείλιση ακόμα, σε σχέση με την περίπτωση που έχει κριτήριο το u>50%, όμως οι τιμές αυτές κυμαίνονται στην ίδια περιοχή για κάθε αριθμό κλειδιών στον πίνακα. Το τελευταίο συμβαίνει διότι συγχώνευση θέσεων σε αυτή την περίπτωση ίσως και να μη συμβαίνει ποτέ, καθώς ο δείκτης διάσπασης είναι πολύ μεγαλύτερος του δείκτη συγχώνευσης. Δηλαδή ο πίνακας κατακερματισμού δεν θα είναι πότε αρκετά άδειος για να ικανοποιηθεί το τελευταίο κριτήριο. Το αντίθετο συμβαίνει στην περίπτωση που το κριτήριο διάσπασης είναι u>50%, όπου έχουμε συνεχώς

συγχώνευση στις διαγραφές, οπότε αυξάνοντας τον αριθμό των κλειδιών στον πίνακα, αυξάνονται οι πιθανές διαγραφές, άρα και οι συγχωνεύσεις, άρα και ο μέσος αριθμός συγκρίσεων ανά διαγραφή. Τελικώς, δηλαδή για μεγάλο N, οι δύο περιπτώσεις για μεγάλο αριθμό κλειδιών δεν παρουσιάζουν μεγάλες διαφορές στις τιμές των μετρήσεων, οι οποίες κοιτάζοντάς τες μακροσκοπικά, μεταβάλλονται γραμμικά, όπως και στις δύο παραπάνω πράξεις.



Απόδοση δυαδικού δέντρου έρευνας στην αναζήτηση, σε σχέση με το γραμμικό κατακερματισμό

Στον πίνακα των μετρήσεων και στο διάγραμμα παρουσιάζεται ο μέσος όρος συγκρίσεων ανά αναζήτηση σε ένα δυαδικό δέντρο έρευνας με κλειδιά που αυξάνονται ανά 100 μέχρι να φτάσουν τα N=10000. Όπως ήταν αναμενόμενο, για τη συγκεκριμένη πράξη είναι αποδοτικότερη η μέθοδος του γραμμικού κατακερματισμού. Ο μεγάλος αριθμός συγκρίσεων στο δυαδικό δέντρο έρευνας οφείλεται στο ότι τα στοιχεία είναι ταξινομημένα μεταξύ τους με βάση το μέγεθός τους κι έτσι υπάρχουν πολλά πιθανά σημεία στα οποία βρίσκεται το κάθε κλειδί, έναντι του γραμμικού κατακερματισμού που οι πιθανές θέσει ενός κλειδιού είναι από μια μέχρι δύο. Στην αναζήτηση όμως εύρους τιμών το δυαδικό δέντρο έρευνας είναι αποδοτικότερο γι’ αυτό τον λόγο, και για το ότι στον πίνακα κατακερματισμού τα κοντινά κλειδιά αποθηκεύονται σε μακρινές θέσεις.

συμπέρασμα

Κοιτάζοντας το διάγραμμα συμπεραίνω ότι όταν έχουμε χαμηλό όριο διάσπασης η μέθοδος του γραμμικού κατακερματισμού είναι πιο αποδοτική στην αναζήτηση και στη διαγραφή στοιχείων. Όμως στην εισαγωγή είναι λιγότερο αποδοτική από όταν γίνεται χρήση μεγαλύτερου ορίου. Η διαφορά βέβαια δεν είναι μεγάλη, όμως θα μπορούσε να βελτιωθεί αν η διάσπαση γινόταν σε περισσότερες σελίδες κάθε φορά, αντί για μία σελίδα που υλοποίησα σε αυτή την εργασία. Έτσι δεν θα είχαμε τόσο συχνά διάσπαση, άρα και τόσο συχνή αύξηση του αριθμού των συγκρίσεων. Κάτι ακόμα που βοηθάει σε κάθε περίπτωση να κρατήσει υψηλά την απόδοση της μεθόδου είναι να έχουμε κατασκευάσει αρχικά έναν πίνακα μεγέθους ανάλογο με τον αριθμό των κλειδιών που θα εισάγουμε. Έτσι ο αριθμός των συγκρίσεων σε κάθε πράξη θα είναι αρκετά χαμηλός. Αλλά δεν μπορούμε να γνωρίζουμε το τελικό μέγεθος πάντα. Κι αυτό είναι που μας κάνει να χρησιμοποιούμε ένα πίνακα κατακερματισμού έναντι ενός απλού πίνακα. Χρησιμοποιούμε δηλαδή το πλεονέκτημα της δυναμικότητας της μεθόδου, κάτι που χρειαζόμαστε όταν δεν γνωρίζουμε τον τελικό αριθμό των κλειδιών.