ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
Τμήμα Πληροφορικής και Τηλεπικοινωνιών
Κ23γ ¨Ανάπτυξη Λογισμικού για Αλγοριθμικά Προβλήματα' – Χειμερινό Εξάμηνο 2023
Εργασία από τους φοιτητές
Δημήτριος Σταύρος Κωστής Sdi1115201700304
Ορέστης Θεοδώρου Sdi1115202000058

2η Εργασία

ReadME

- Graph: Η συνάρτηση initializeGraph αρχικοποιεί τον γράφο.

  Η συνάρτηση addVertex προσθέτει έναν κόμβο στον γράφο που την κάλεσε.
  Η addVertexNeighbour συνδέει δύο κόμβους, αν ο κόμβος προορισμός υπάρχει, αλλιώς δημιουργεί τον κόμβο προορισμό και έπειτα τους συνδέει.
  Η συνάρτηση GNN υλοποιεί τον αλγόριθμο Graph Nearest Neighbor Search.

  Η συνάρτηση constructIndex υλοποιεί τον αλγόριθμο κατασκευής ευρετηρίου.
  Η connectMRNGEdges συνδέει τους κόμβος με τις ακμές που πήραμε από την constructIndex.
  Η συνάρτηση searchOnGraph υλοποιεί έναν γενικό αλγόριθμο αναζήτης πρώτα κατά πλάτος στον γράφο χρησιμοποιόντας τις ακμές MRNG.

  Η κλάση Vertex αναπαριστά έναν κόμβο μέσα στον γράφο.
  Με τη χρήση addNeighbour προσθέτουμε μια ακμή η οποία οδηγεί σε έναν κόμβο προορισμό.

  Η κλάση Edge αναπαριστά τις ακμές μέσα στον γράφο.

  Η LSHtoGraph χρησιμοποιεί τη δομή LSH Hash Tables και τα σημεία δεδομένων σε κάθε bucket για να δημιουργήσει έναν γράφο, δημιουργόντας ¨γειτονιές' όπου κάθε σημείο συνδέεται με τα k πλησιέστερα σημεία από το ίδιο bucket.

  Η συνάρτηση LSHGraphnn χρησιμοποιείται για την εύρεση των k πλησιέστερων γειτόνων ενός σημείου εντός του ίδιο bucket.

  Η FindNeighbours χρησιμοποιείται για την εύρεση γειτόνων ενός σημείου.

  Η συνάρτηση FindNavNode χρησιμοποιείται για την εύρεση του Navigation Node.

  Η calculateCentroid χρησιμοποιείται για την εύρεση του κέντροειδους του γράφου.

  Με την isLongestEdgeInTriangle γίνεται έλεγος αν μια ακμή είναι η μεγαλύτερη σε ένα δεδομένο τρίγωνο.

  Με τη χρήση initializeLp γίνεται η αρχικοποίηση στον πίνακα Lp στη δημιουργία του ευρετηρίου MRNG.

Η Υλοποίηση μας περίεχει MakeFile για το compile του κώδικα με 'make all'.
Οι αλγόριθμοι διάβασουν σύμφωνα με την εκφώνηση τις σημαίες στην γραμμή εντολής (εάν δίνονται), ενδεικτικά παραδείγματα εκτέλεσης των αλγορίθμων:
Για εκτέλεση του GNNS
`./graph_search -d input.dat -q query.dat -k 75 -E15 -m 1 -R 10 -o output.txt`
Για εκτέλεση του MRNG
`./graph_search -d input.dat -q query.dat -k 75 -E 15 -m 2 -R 10 -o output.txt`
Όπως αναφέρθηκε και στην εκφώνηση, μετά την ολοκλήρωση του το πρόγραμμα ρωτάει τον χρήστη αν θέλει να επαναλάβει με διαορετικά αρχεία ή όχι, στην περίπτωση που ο χρήστης θέλει να επαναλάβει, θα του ζητηθούν ξανά τα αρχεία input και query,τα αποτελέσματα όλων θα εκτυπωθούν το αρχικό ouput αρχείο.

Αξιολόγιση Αλγορίθμων :

Για την αξιολόγηση των αλγορίθμων, τους τρέξαμε σε κάθε πιθανό συνδιασμό των μεταβλητών που υπάρχουν στους πίνακες.Έπειτα βγάλαμε τον μέσο όρο του χρόνου και του MAF για κάθε τιμή της εκάστοτε μεταβλητής.Τα αποτελέσματα είναι ενδεικτικά της υλοποίησης μας και πάνω σε αυτή γίνονται τα σχόλια παρακάτω.

LSH:

Variable	Value 1	Value 2	Value 3	Average tFraction, MAF	Average tFraction, MAF	Average tFraction, MAF
k	3	5	10	0.076, 4.556	0.070, 4.523	0.071, 4.524
L	5	10	20	0.027, 4.557	0.048, 4.523	0.137,4.523
N	1	5	10	0.075, 4.262	0.071, 4.655	0.071,4.654

1: Μέσος όρος κλάσματος (χρόνος εύρεσης approximate)/(Πραγματικός χρόνος με brutalForce)

Ανάλυση κατά μεταβλητή K:
Ο λόγος απόδοσης χρόνου αυξάνεται ελαφρώς καθώς αυξάνεται το k, υποδεικνύοντας μια μικρή μείωση στην απόδοση. Αυτό είναι αναμενόμενο καθώς περισσότερες συναρτήσεις κατακερματισμού μπορούν να οδηγήσουν σε περισσότερους υπολογισμούς. Η MAF εμφανίζει βελτίωση με την αύξηση του k, γεγονός που υποδηλώνει ότι η ποιότητα της προσέγγισης βελτιώνεται με περισσότερες συναρτήσεις κατακερματισμού. Αυτό θα μπορούσε να οφείλεται στο ότι περισσότερες συναρτήσεις κατακερματισμού οδηγούν σε μια πιο λεπτή κατάτμηση του χώρου δεδομένων, με αποτέλεσμα την ακριβέστερη αναγνώριση γειτόνων.

Αριθμός πινάκων κατακερματισμού L:
Ο λόγος απόδοσης χρόνου αυξάνεται σημαντικά με υψηλότερο L, ειδικά από 10 σε 20. Αυτό υποδηλώνει ότι περισσότεροι πίνακες κατακερματισμού επιβραδύνουν τον αλγόριθμο λόγω της αυξημένης πρόσβασης και διαχείρισης της μνήμης.Η MAF παραμένει σχετικά σταθερή, υποδεικνύοντας ότι ο αριθμός των πινάκων κατακερματισμού δεν επηρεάζει σημαντικά την ποιότητα προσέγγισης σε συνεπή κατεύθυνση.

Αριθμός πιο κοντινών γειτόνων N:

Ο λόγος απόδοσης χρόνου παραμένει σχετικά σταθερός σε διαφορετικά N, υποδηλώνοντας ότι ο αριθμός των γειτόνων δεν επηρεάζει σημαντικά την απόδοση του αλγορίθμου.Η MAF δείχνει βελτίωση με την αύξηση του N, υποδεικνύοντας ότι η αναζήτηση για περισσότερους γείτονες μπορεί να βελτιώσει την ποιότητα προσέγγισης. Αυτό μπορεί να οφείλεται στο ότι η εξέταση περισσότερων γειτόνων επιτρέπει περισσότερες πιθανότητες εύρεσης των αληθινών πλησιέστερων γειτόνων.

Συνολικές παρατηρήσεις: Ανταλλαγή αποτελεσματικότητας έναντι ποιότητας: Υπάρχει σαφής αντιστάθμιση μεταξύ της αποτελεσματικότητας (tFraction) και της ποιότητας προσέγγισης (MAF). Παράμετροι όπως το k θα πρέπει να επιλέγονται για να εξισορροπηθεί αυτή η ανταλλαγή ανάλογα με τις απαιτήσεις της εφαρμογής.
Ευαισθησία παραμέτρων: Η ευαισθησία των MAF και tFraction στις αλλαγές σε k, L και N ποικίλλει, υποδηλώνοντας ότι ο προσεκτικός συντονισμός αυτών των παραμέτρων είναι ζωτικής σημασίας για τη βελτιστοποίηση της απόδοσης.

HyperCube:

Variable	Value 1	Value 2	Value 3	Average tFraction, MAF	Average tFraction, MAF	Average tFraction, MAF
k	3	5	10	1.936, 3.712	3.555,3.914	6.620,3.851
M	100	500	1000	3.489, 3.965	4.604, 3.656	4.679,3.729
probes	2	5	10	4.548, 3.876	4.186,3.729	3.920,3.757
N	1	5	10	4.398, 3.618	4.142,3.849	4.166,3.901

2: Μέσος όρος κλάσματος (χρόνος εύρεσης approximate)/(Πραγματικός χρόνος με brutalForce) και η μεταβλητή MAF

Διάσταση k:
Καθώς αυξάνεται το k, ο λόγος απόδοσης χρόνου αυξάνεται σημαντικά από 1,936 σε 6,620, υποδεικνύοντας ότι οι υψηλότερες διαστάσεις ενδέχεται να απαιτούν περισσότερο χρόνο σε σχέση με τη μέθοδο της Brutal Force. Ωστόσο, το MAF παραμένει σχετικά σταθερό γύρω στο 3,7 έως το 3,9, υποδηλώνοντας ότι η ακρίβεια εύρεσης του κατά προσέγγιση πλησιέστερου γείτονα σε σύγκριση με τον πραγματικό πλησιέστερο γείτονα είναι συνεπής σε διαφορετικές διαστάσεις.

Μέγιστος αριθμός θέσεων M:
Για το M, υπάρχει μια σταδιακή αύξηση του λόγου απόδοσης χρόνου από 3.489 σε 4.679 καθώς το M πηγαίνει από 100 σε 1000. Αυτό θα μπορούσε να σημαίνει ότι οι υψηλότερες τιμές του M οδηγούν σε μεγαλύτερους χρόνους αναζήτησης.Το MAF μειώνεται ελαφρώς από 3,965 σε 3,729 καθώς αυξάνεται το M. Αυτό μπορεί να υποδηλώνει μια μικρή μείωση στην ακρίβεια με περισσότερες θέσεις προς έλεγχο.

Αριθμός probes:
Ο λόγος απόδοσης χρόνου μειώνεται καθώς αυξάνεται ο αριθμός των ανιχνευτών. Αυτό υποδηλώνει ότι ο έλεγχος περισσότερων άκρων μπορεί να οδηγήσει σε πιο αποδοτικές αναζητήσεις από άποψη χρόνου.Το MAF μειώνεται επίσης ελαφρώς από 3.876 σε 3.757 καθώς αυξάνεται ο αριθμός των ανιχνευτών, υποδεικνύοντας οριακή αύξηση της ακρίβειας με αυξημένο αριθμό ανιχνευτών.

Αριθμός πιο κοντινών γειτόνων N:
Ο λόγος απόδοσης χρόνου αυξάνεται ελαφρά από 4.398 σε 4.166 καθώς το N πηγαίνει από το 1 στο 10.

Η αύξηση δεν είναι τόσο σημαντική όσο σε άλλες μεταβλητές, υποδηλώνοντας ότι η επίδραση του N στην απόδοση χρόνου είναι μέτρια.Το MAF δείχνει μια σταθερή αύξηση από 3.618 σε 3.901, υποδεικνύοντας μειωμένη ακρίβεια στην εύρεση πλησιέστερων γειτόνων καθώς μεγαλώνει ο εξεταζόμενος αριθμός.

Συνολικά, ο αλγόριθμος δείχνει μια ισορροπία μεταξύ της απόδοσης χρόνου και της ακρίβειας, με συγκεκριμένες τάσεις που παρατηρούνται για κάθε μεταβλητή. Οι υψηλότερες διαστάσεις και οι περισσότεροι ανιχνευτές τείνουν να βελτιώνουν την απόδοση του χρόνου, αλλά μειώνουν ελαφρώς την ακρίβεια. Η αύξηση του μέγιστου αριθμού θέσεων M ή του αριθμού των πλησιέστερων γειτόνων N τείνει να αυξάνει τον χρόνο που απαιτείται, αλλά βελτιώνει επίσης την ακρίβεια εύρεσης του πλησιέστερου γείτονα. Η απόδοση του αλγόριθμου φαίνεται να είναι ισχυρή σε διαφορετικές ρυθμίσεις, διατηρώντας μια λογική ισορροπία μεταξύ ταχύτητας και ακρίβειας.Παρόλα αυτά ο αλγόριθμος απέχει πολύ από βέλτιστος και οι τιμές τους μας δείχνουν ότι δεν είναι ακριβής και γρήγορος, όπως θα έπρεπε (σε σίγκριση με την εξαντλητική μέθοδο).

GNNS:

Variable	Value 1	Value 2	Value 3	Average tFraction
k	25	50	75	0.00202, 0.001997, 0.002173
E	15	30	45	0.001994, 0.002161, 0.002034
R	1	5	10	0.000398, 0.001871, 0.003973

3: Μέσος όρος κλάσματος (χρόνος εύρεσης approximate)/(Πραγματικός χρόνος με brutalForce)

Μεταβλητή k:
Το tFraction αυξάνεται καθώς αυξάνεται το k, αν και όχι σημαντικά (0,00202 έως 0,002173).Αυτό υποδηλώνει ότι καθώς ο αριθμός των πλησιέστερων προς αναζήτηση γειτόνων αυξάνεται, η αποτελεσματικότητα του αλγορίθμου GNNS σε σύγκριση με τη μέθοδο της Brutal force βελτιώνεται ελαφρώς.Η διαφορά στην απόδοση μεταξύ $k = 25$ και $k = 75$ είναι ελάχιστη, υποδεικνύοντας ότι ο αλγόριθμος κλιμακώνεται αρκετά καλά με αύξηση του k.

Μεταβλητή E:
Το tFraction αυξάνεται επίσης καθώς το E αυξάνεται, από 0,001994 σε 0,002161, και στη συνέχεια μειώνεται ελαφρώς στο 0,002034.Αυτό το μοτίβο μπορεί να υποδεικνύει ότι η αποτελεσματικότητα του αλγορίθμου βελτιώνεται μέχρι ένα ορισμένο σημείο καθώς το E αυξάνεται, αλλά στη συνέχεια τα επίπεδα μειώνονται ή μειώνονται ελαφρώς.Ο αλγόριθμος φαίνεται να χειρίζεται την αυξημένη πολυπλοκότητα μέχρι ένα συγκεκριμένο όριο.

Μεταβλητή R:
Υπάρχει μια πιο αισθητή αύξηση στο tFraction με υψηλότερες τιμές του R.Αυτό θα μπορούσε να σημαίνει ότι η σχετική αποτελεσματικότητα του αλγορίθμου GNNS βελτιώνεται σημαντικά καθώς το R αυξάνεται.Ο αλγόριθμος φαίνεται να γίνεται πιο αποτελεσματικός σε σύγκριση με τη μέθοδο της Brutal force όταν ασχολείται με ευρύτερες παραμέτρους αναζήτησης ή μεγαλύτερα πεδία αναζήτησης.

MAF μεταβλητή:
Ο αλγόριθμος παρουσίασε μεγάλη σταθερότητα στην τιμή MAF με όλες δοκιμές να επιστρέφουν την τιμή $MAF = 0.616748$.Αυτό δείχνει ότι παρά τις διάφορες τιμές, ο αλγόριθμος φτάνει σε αντίστοιχα προσεγγιστικά σημεία πολύ κοντά στο πραγματικό σημείο.

Γενική Ανάλυση:
Η απόδοση του αλγόριθμου GNNS σε σχέση με την προσέγγιση της ωμής δύναμης βελτιώνεται με την αυξημένη πολυπλοκότητα των παραμέτρων αναζήτησης (k, E και ειδικά R). Ο αλγόριθμος κλιμακώνεται αρκετά καλά με μια αύξηση σε αυτές τις παραμέτρους, υποδεικνύοντας καλή απόδοση σε πιο σύνθετα σενάρια αναζήτησης.Η αύξηση της αποτελεσματικότητας με το R είναι ιδιαίτερα αξιοσημείωτη, υποδηλώνοντας ότι τα δυνατά σημεία του αλγορίθμου βρίσκονται σε ευρύτερα ή πιο εκτεταμένα σενάρια αναζήτησης.

MRNG:

Variable	Value 1	Value 2	Value 3	Average tFraction
k	25	50	75	0.360252, 0.466894, 0.458460
E	15	30	45	0.558456, 0.349043, 0.461843
R	1	5	10	0.348694, 0.585424, 0.458759
I	10	20	30	0.350922, 0.676657, 0.349050

4: Μέσος όρος κλάσματος (χρόνος εύρεσης approximate)/(Πραγματικός χρόνος με brutalForce)

Αναλύοντας τα αποτελέσματα του πίνακα παρατηρούμε τα εξής για την ταχύτητα εκτέλεσης του αλγορίθμου.

Παραλλαγή σε τιμές k:
Υπάρχει μια αξιοσημείωτη διαφορά στο μέσο tFraction σε διαφορετικές τιμές k (25, 50, 75). Η αύξηση του tFraction από $k = 25$ σε $k = 50$ και μια ελαφρά μείωση στο $k = 75$ μπορεί να υποδηλώνει ότι η απόδοση ή η συμπεριφορά του αλγορίθμου αλλάζει σημαντικά με διαφορετικές ρυθμίσεις k. Αυτό θα μπορούσε να σημαίνει ότι ο αλγόριθμος είναι ευαίσθητος στην παράμετρο k.

Επίδραση της παραμέτρου E:
Η παράμετρος E δείχνει μια ποικίλη επίδραση στο tFraction, με τον υψηλότερο μέσο όρο tFraction στο E=15 και το χαμηλότερο στο E=30. Αυτή η παραλλαγή υποδηλώνει ότι η παράμετρος E παίζει κρίσιμο ρόλο στην απόδοση του αλγορίθμου. Μπορεί να είναι ωφέλιμο να διερευνήσουμε γιατί το E=15 οδηγεί σε υψηλότερο tFraction.

Επίδραση της παραμέτρου R:
Το μέσο tFraction για $R = 5$ είναι σημαντικά υψηλότερο από το $R = 1$ και $R = 10$. Αυτό θα μπορούσε να σημαίνει ότι η έξοδος του αλγορίθμου είναι πολύ ευαίσθητη σε αυτήν την παράμετρο και ότι υπάρχει μια βέλτιστη ρύθμιση γύρω από το $R = 5$ που αποδίδει σημαντικά διαφορετικά αποτελέσματα. Η κατανόηση του γιατί το $R = 5$ οδηγεί σε αυτή τη συμπεριφορά θα μπορούσε να είναι το κλειδί για την αποτελεσματική αξιοποίηση του αλγόριθμου.

Ανάλυση παραμέτρου I:
Η παράμετρος I δείχνει ένα ενδιαφέρον μοτίβο, με σημαντικά υψηλότερο μέσο όρο tFraction στο $I = 20$. Αυτή η ακραία τιμή υποδηλώνει ότι ο αλγόριθμος συμπεριφέρεται διαφορετικά κάτω από αυτή τη συγκεκριμένη συνθήκη. Είναι σημαντικό να κατανοήσουμε ότι με μικρότερες τιμές I έχουμε μικρότερη δεξαμενή τιμών,οπότε ο αλγόριθμος δεν θα έπρεπε να παρουσιάζει τέτοια αύσηξη στην μέση τιμή.

Συνολική συμπεριφορά αλγορίθμου: Η διακύμανση του tFraction σε διαφορετικές ρυθμίσεις των k, E,

R και I υποδηλώνει ότι ο αλγόριθμος MRNG είναι ευαίσθητος στις ρυθμίσεις παραμέτρων του. Αυτό συνεπάγεται την ανάγκη για προσεκτική βαθμονόμηση αυτών των παραμέτρων με βάση τη συγκεκριμένη περίπτωση χρήσης ή τα επιθυμητά αποτελέσματα. Επιπλέον, ο λόγος πίσω από την υψηλή μεταβλητότητα, ειδικά στην περίπτωση των E=15 και $R = 5$, πρέπει να διερευνηθεί. Είναι σημαντικό να διασφαλιστεί ότι ο αλγόριθμος είναι ισχυρός και ότι η απόδοσή του είναι συνεπής και αξιόπιστη σε διάφορες ρυθμίσεις παραμέτρων, στην προκυμένη περίπτωση δεν παρουσιάζεται κάτι τέτοιο με τις σιγκρίσεις του χρόνου.

MAF μεταβλητή:
Ο αλγόριθμος παρουσίασε μεγάλη σταθερότητα στην τιμή MAF με τις περισσότερες δοκιμές να επιστρέφουν την τιμή $MAF = 0.456288$.Αυτό δείχνει ότι παρά τις διάφορες τιμές, ο αλγόριθμος φτάνει σε αντίστοιχα προσεγγιστικά σημεία πολύ κοντά στο πραγματικό σημείο.

Συγκριτική Ανάλυση

Ανταλλαγή απόδοσης έναντι ακρίβειας: Όλοι οι αλγόριθμοι παρουσιάζουν μια αντιστάθμιση μεταξύ της απόδοσης χρόνου και της ακρίβειας προσέγγισης. Το LSH και το HyperCube, ειδικότερα, δείχνουν αυτή την ισορροπία ευδιάκριτα,αυτό πιθανόν οφείλεται στην υλοποίηση και όχι εξ-ολοκλήρου στον αλγόριθμο.
Επεκτασιμότητα: Το GNNS ξεχωρίζει για την επεκτασιμότητα και τη σταθερή του ακρίβεια σε διάφορες ρυθμίσεις παραμέτρων, καθιστώντας το δυνητικά πιο προσαρμόσιμο σε πολύπλοκα σενάρια.
Ευαισθησία παραμέτρων: Το MRNG και το LSH είναι ιδιαίτερα ευαίσθητα στις αντίστοιχες παραμέτρους τους, κάτι που απαιτεί προσεκτικό συντονισμό για βέλτιστη απόδοση.
Συνέπεια: Το GNNS και το HyperCube διατηρούν ένα σχετικά σταθερό επίπεδο ακρίβειας (MAF) σε διαφορετικές ρυθμίσεις, υποδεικνύοντας μια σταθερή ποιότητα προσέγγισης.

Συμπεράσματα

Όταν επιλέγουμε έναν αλγόριθμο για μια συγκεκριμένη εφαρμογή, πρέπει να λάβουμε υπόψη τις συγκεκριμένες απαιτήσεις της εργασίας, όπως η ανάγκη για ακρίβεια σε σχέση με την απόδοση χρόνου και την επεκτασιμότητα του αλγορίθμου κάτω από διάφορες ρυθμίσεις παραμέτρων. Το GNNS φαίνεται να είναι μια ισχυρή επιλογή για πολύπλοκα σενάρια λόγω της επεκτασιμότητας και της σταθερής ακρίβειάς του, ενώ το LSH και το HyperCube προσφέρουν μια ισορροπημένη αντιστάθμιση μεταξύ αποτελεσματικότητας και ακρίβειας, κατάλληλη για λιγότερο σύνθετες εργασίες. Η μεταβλητότητα απόδοσης του MRNG υποδηλώνει ότι μπορεί να απαιτεί πιο προσεκτική ρύθμιση παραμέτρων για να επιτευχθούν τα επιθυμητά αποτελέσματα.