Ανάπτυξη Λογισμικού για Αλγοριθμικά Προβλήματα

1η Προγραμματιστική Εργασία
Αναζήτηση και Συσταδοποίηση Διανυσμάτων στη C++

Μηνάς Διολέτης (Α.Μ. 1115201400272)
Απόστολος Θεοδώρου (Α.Μ.: 1115201500046)

Χειμερινό Εξάμηνο 2021-2022
Τμήμα Πληροφορικής και Τηλεπικοινωνιών
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Contents

1 Περιγραφή της εργασίας

Σκοπός της παρούσας εργασίας είναι η υλοποίηση αλγορίθμων εύρεσης πλησιέστερου γείτονα και συσταδοποίησης σε γλώσσα C++.

Τα προγράμματα εύρεσης κοντινότερου γείτονα λαμβάνουν ως είσοδο αρχείο που περιέχει διανύσματα (ή σημεία) ενός χώρου διάστασης d (input_file) και ένα δεύτερο αρχείο που επίσης περιέχει διανύσμα του ίδιου χώρου (query_file). Τα διανύσματα του πρώτου αρχείου αποτελούν το σύνολο δεδομένων (dataset), ενώ αυτά του δεύτερου το σύνολο αναζήτησης. Στόχος είναι να παραχθεί ως έξοδος ένα αρχείο, στο οποίο, ανάλογα με τις δοθείσες παραμέτρους, να επιστρέφεται το κοντινότερο ή τα κοντινότερα διανύσματα του dataset για καθένα από τα διανύσματα του συνόλου αναζήτησης.

Το πρόγραμμα της συσταδοποίησης λαμβάνει, όπως και τα προηγούμενα προγράμματα, ένα αρχείο που περιέχει το dataset. Επιπλέον δέχεται και ένα αρχείο (configuration_file) στο οποίο περιέχονται πληροφορίες για την παραμετροποίηση της συσταδοποίησης. Η έξοδός του είναι ένα αρχείο που περιλαμβάνει πληροφορίες για τις συστάδες σημείων που δημιουργήθηκαν.

Οι αλγόριθμοι που υλοποιούν τα προγράμματα της εύρεσης κοντινότερων γειτόνων είναι ο lsh (Locality Sensitive Hashing) και ο αλγόριθμος τυχαίας προβολής στον υπερκύβο. Το πρόγραμμα της συσταδοποίησης δύναται να εκτελεστεί με τις εξής εναλλακτικές μεθόδους:

- τον ακριβή αλγόριθμο του Lloyds
- την αντίστροφη ανάθεση(reverse) μέσω Range Search
  - με lsh.
  - με τυχαία προβολή

Δείτε αναλυτικά την εκφώνηση της εργασίας

2 Αρχεία

Τα προγράμματα αναπτύχθηκαν σύμφωνα με τις αρχές του modular programming και ο κώδικας, αν και εκτενής, είναι διαρθρωμένος σε ξεχωριστά αρχεία, ανάλογα με τη λειτουργικότητα που επιτελεί η κάθε συνάρτηση. Ακολουθεί λίστα των αρχείων και περιγραφή τους.

Η main και οι συναρτήσεις της:

- main.cpp: περιέχει τη συνάρτηση main που υλοποιεί τη λειτουργικότητα κάθε προγράμματος καλώντας τις κατάλληλες συναρτήσεις, ανάλογα με τις παραμέτρους της γραμμής εντολών ή/και του configuration file.
- cmd_line_args.cpp: υλοποιεί την ανάθεση των παραμέτρων της γραμμής εντολών στις αντίστοιχες μεταβλητές.
- conf_file.cpp: υλοποιεί το διάβασμα του configuration file του cluster και την αρχικοποίηση των αντίστοιχων μεταβλητών.
- file_functions.cpp: χρησιμοποιείται για το άνοιγμα αρχείων (input, query, output και configuration) , το διάβασμα και το γράψιμο και από και προς αυτά, το κλείσιμο του μετά το πέρας του προγράμματος.
- user_input.cpp: διαχειρίζεται την ανάδραση του τον χρήστη στην περίπτωση του ή/και hypercube όταν ζητείται η επανάληψη εκτέλεσης του προγράμματος για διαφορετικό σύνολο (αρχείο) αναζήτησης.

Συναρτήσεις γενικού σκοπού:

- hamming_distance.cpp: υλοποιεί την εύρεση της Χάμινγκ απόστασης μεταξύ δυαδικών συμβολοσειρών, χρησιμοποιείται στην μέθοδο του υπερκύβου.
- mod.cpp: υλοποιεί την εύρεση της ομώνυμης πράξης (επιστρέφοντας μόνο θετικά υπόλοιπα).

- vector_ops.cpp: περιέχει όλες τις συναρτήσεις που αφορούν πίνακες (vectors)και χρησιμοποιούνται από πολλές συναρτήσεις τόσο της εύρεσης γειτόνων, όσο και τις συσταδοποίησης. Λεπτομέρειες για τον σκοπό της κάθε συνάρτησης θα βρείτε σε σχόλιο πάνω από τον ορισμό της.

Για τα προγράμματα lsh και cube:

- lsh.cpp: περιέχει τις υλοποιήσεις των αλγορίθμων lsh, range_search με lsh και brute force knn.
- cube.cpp: περιέχει τις υλοποιήσεις της τυχαίας προβολής στον υπερκύβο και range search με hypercube.
- hashTable.cpp:υλοποιεί τους πίνακες κατακερματισμού των μεθόδων hypercube και lsh.
- hash_functions.cpp: υλοποιεί τις συναρτήσεις κατακερματισμού h, g που-χρησιμοποιούνται στην καταχώρηση σημείων στους πίνακες και στην εύρεση των id στις περιπτώσεις που απαιτείται.
- knn_table_functions.cpp: περιέχει χρήσιμες βοηθητικές συναρτήσεις (ταξινόμησης, αναζήτησης κ.α.) για τον αλγόριθμο της εύρεσης κοντινότερων γειτόνων.

Για το πρόγραμμα cluster:

- cluster.cpp: περιέχει τις υλοποιήσεις των αλγορίθμων Lloyds, reverse range search with LSH, reverse range search with hypercube.
- lloyds_auxiliary.cpp: περιέχει βοηθητικές συναρτήσης για τον αλγόριθμο του Lloyds.
- silhouette.cpp: υλοποιεί τον υπολογισμό και την εμφάνιση της μετρικής της σιλουέτας.

3 Μεταγλώττιση

Και τα τρία προγράμματα μεταγλωττίζονται και παράγονται ταυτόχρονα μέσω της εντολής make από το command line.

Για την εκτέλεση του κάθε προγράμματος τρέξτε τις εντολές που δίνονται στην εκφώνηση. Σε περίπτωση που δεν δώσετε κάποιο από τα μη υποχρεωτικά ορίσματα, αυτό θα αρχικοποιηθεί από τις default τιμές.

Για την ανάπτυξη του λογισμικού χρησιμοποιήθηκε το εργαλείο git και η πλατφόρμα github.

Github repository