

Πολυτεχνική Σχολή

Τμήμα Μηχανικών Η/Υ & Πληροφορικής

**ΕΙΣΑΓΩΓΗ ΣΤΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ**

**ΠΡΩΤΟ ΣΥΝΟΛΟ ΕΡΓΑΣΤΗΡΙΑΚΩΝ ΑΣΚΗΣΕΩΝ 2023-2024**

<Όνομα Φοιτητή>

Α.Μ. <Φοιτητή>

Ασπασία Τζερμιά

Α.Μ. 1067455

Πάτρα, 2024

**Ερώτημα 3**

Οι k συμβολοσειρές μπορούν να αποθηκευτούν σε ένα γενικευμένο δέντρο επιθεμάτων σαν μια ενιαία συμβολοσειρά που θα είναι η συνένωση όλων των συμβολοσειρών, όπου ανάμεσα σε κάθε μία από τις συμβολοσειρές μπορεί να παρεμβάλλεται κάποιος ειδικός χαρακτήρας που δεν ανήκει στο αλφάβητό τους (για παράδειγμα ο χαρακτήρας “$”). Δηλαδή για δεδομένο σύνολο με τις συμβολοσειρές x1, x2, …, xk, το δέντρο θα κατασκευαστεί από τη συμβολοσειρά x1$x2$...xk$. Για να μπορούν να προστίθενται ή να αφαιρούνται από αυτό συμβολοσειρές πρέπει να μπορεί να διατηρεί και να προσαρμόζει «δυναμικά» τη δομή του, δηλαδή να είναι δυναμικό δέντρο επιθεμάτων (dynamic suffix tree). Πρέπει επίσης να μπορεί να γίνεται αναζήτηση για ταιριάσματα συμβολοσειρών στο δέντρο που δεν έχει σταθερή δομή (dynamic string matching).

Για τις παραπάνω λειτουργίες απαιτείται όμως περισσότερος χρόνος (για αναδιοργάνωση της δομής μετά από κάθε εισαγωγή ή διαγραφή και για αναζήτηση στο δέντρο) και χώρος (γιατί η κατάσταση του δέντρου πριν και μετά από κάθε αλλαγή πρέπει να διατηρείται στη μνήμη του υπολογιστή). Η αναζήτηση για ταιριάσματα μπορεί να γίνει πολύ χρονοβόρα, ειδικά όταν πρόκειται για βιολογικές ακολουθίες, που είναι μεγάλες σε μέγεθος. Επιπλέον, η διαγραφή μιας συμβολοσειράς xj γίνεται ουσιαστικά με διαγραφή των φύλλων στο δέντρο που αντιστοιχούν στις υποσυμβολοσειρές της xj, όμως η κατάσταση των εσωτερικών κόμβων του δέντρου δεν ανανεώνεται αμέσως στη μνήμη και έτσι ο συνολικός χώρος για την αποθήκευση της δομής αυξάνεται.

Μια λύση για βελτίωση της χρονικής και χωρικής πολυπλοκότητας είναι να αποθηκεύονται οι k συμβολοσειρές ως γραμμές ή στήλες ενός μητρώου που θα αποθηκεύεται στη μνήμη. Οι εισαγωγές/διαγραφές συμβλοσειρών μπορούν να γίνονται με εισαγωγή/διαγραφή γραμμών ή στηλών. Κάθε υποσυμβολοσειρά, που κατά την αναζήτηση ταιριάσματος θα είναι ουσιαστικά ένα υπομητρώο του μητρώου αυτού, θα αποθηκεύεται σε ένα ξεχωριστό dynamic suffix tree.

Μια άλλη λύση είναι τα dynamic Fully Compressed Suffix Trees (FCSTs). Αποτελούν βελτίωση των στατικών Compresses Suffix Trees για δυναμικά μεταβαλλόμενα δεδομένα. Ένα FCST αποτελείται από έναν πίνακα Compressed Suffix Array (που χρησιμοποιείται ως ευρετήριο για τις συμβολοσειρές που είναι αποθηκευμένες σε αυτό), από ένα δ-Sampled tree (που εκμεταλλεύεται τις ιδιότητες των suffix trees ώστε να μπορούν να αποθηκεύονται μόνο κάποιοι κρίσιμοι κόμβοι και να μη χρειάζεται αποθήκευση ολόκληρης της δενδρικής δομής) και από αντιστοιχίσεις ανάμεσα στις δύο αυτές δομές. Το FCST μεταβάλλεται με χρήση ενός κανόνα που τοποθετεί νέα στοιχεία από δεξιά προς τα αριστερά δημιουργώντας ένα suffix tree ως «ενδιάμεση» κατασκευή. Έτσι μπορούν να αναπαρασταθούν όλο και μεγαλύτερες συμβολοσειρές. Κάθε αλλαγή στο δέντρο γίνεται μόνο στα φύλλα και αλλαγές στους εσωτερικούς κόμβους γίνονται μόνο αν χρειάζεται. Τα δέντρα αυτά πετυχαίνουν βέλτιστη χωρική πολυπλοκότητα.

Πηγές:

1. Y. Choi, T.W. Lam, Dynamic suffix tree and two-dimensional texts management, Information Processing Letters, Volume 61, Issue 4, 1997, Pages 213-220, ISSN 0020-0190, <https://doi.org/10.1016/S0020-0190(97)00018-5> .
2. Russo, L. M. S., Navarro, G., & Oliveira, A. L. (2008). Dynamic Fully-Compressed Suffix Trees.. CPM. (pp. 191-203).