Краткое изложение заявки “Эффективные алгоритмы для некоторых задач обработки слов”, Стариковская Т.А.

В 1973 г. Питером Вайнером было предложено понятие суффиксного дерева. Суффиксное дерево для слова T длины n хранит информацию о всех его суффиксах (подсловах слова T, начинающихся в некоторой позиции i и заканчивающихся в позиции n). Суффиксные деревья позволяют эффективно решать множество задач обработки слов, но объем памяти, необходимый для хранения суффиксных деревьев, довольно большой. Одним из способов уменьшить объем занимаемой памяти является рассмотрение так называемых разреженных суффиксных деревьев, в которых представлены не все суффиксы, а только некоторые.

Например, если включать в дерево только каждый r-ый суффикс, где r — некоторое заранее заданное число, то объем требуемой памяти уменьшится в r раз. Такие деревья называются r-разреженными суффиксными деревьями, и они позволяют эффективно вычислять вхождения слова P в слово T [1], а также вычислять разложение Лемпеля — Зива слова T [2].

Тем не менее, r-разреженные суффиксные деревья не могут быть использованы для построения эффективных алгоритмов для многих других задач, например, для задачи о наибольшем общем подслове. Одна из целей настоящего проекта — исследовать возможность применения другого варианта разреженных суффиксных деревьев [3] для решения указанной задачи.

Другой целью данного проекта является разработка эффективного алгоритма для следующей задачи. Пусть дано множество слов $T_1, T_2, \ldots, T_m$, каждому из которых приписан некоторый вес. Задача состоит в том, чтобы по слову T, его весу w и числу d быстро найти все слова T_i, $i \in [1, m]$, вес которых меньше w и длина наибольшего общего подслова с T которых хотя бы d. Методы и структуры данных, полученные в совместной работе [4], позволяют надеяться получить эффективный по времени и памяти алгоритм. Идея заключается в использовании так называемого обобщенного суффиксного дерева, содержащего суффиксы слов $T_1, T_2, \ldots, T_m$, в сочетании со структурами данных, использующихся для решения задач вычислительной геометрии.

Список литературы

[1] R. Kolpakov, G. Kucherov, T.A. Starikovskaya. Pattern Matching on Sparse Suffix Trees. *Proceedings of the 1st International Conference on Data Compression, Communications and Processing.* New York, NY: IEEE Computer Society Press, 2011. — P. 92–97.

[2] T.A. Starikovskaya. Computing Lempel-Ziv factorization online. *Proceedings of the 37th International Symposium on Mathematical Foundations of Computer Science,* ed. by V. Sassone, P. Widmayer. *Lecture Notes in Computer Science,* Vol. **7464**. Berlin etc.: Springer, 2012. — P. 789-799.

[3] P. Bille, I. Gørtz, B. Sach, H. Vildhøj. Time-Space Trade-Offs for Longest Common Extensions. *Proceedings of the 23rd Symposium on Combinatorial Pattern Matching,* ed. by J. Kärkkäinen, J. Stoye. *Lecture Notes in Computer Science,* Vol. **7354**. Berlin etc.: Springer, 2012. — P. 293–305.

[4] G. Kucherov, Y. Nekrich, T.A. Starikovskaya. Cross-document pattern matching. *Proceedings of the 23rd Symposium on Combinatorial Pattern Matching,* ed. by J. Kärkkäinen, J. Stoye. *Lecture Notes in Computer Science,* Vol. **7354**. Berlin etc.: Springer, 2012. — P. 196–207.