Efficient tree-structured categorical retrieval

Djamal Belazzougui*[1] and Gregory Kucherov†[2,3]

[1]CAPA, DTISI, Centre de Recherche sur l'Information Scientifique et Technique, Algiers, Algeria.
[2]CNRS and LIGM/Univ Gustave Eiffel, Marne-la-Vallée, France.
[3]Skolkovo Institute of Science and Technology, Moscow, Russia.

October 11, 2023

Abstract

We study a document retrieval problem in the new framework where D text documents are organized in a *category tree* with a predefined number h of categories. This situation occurs e.g. with taxomonic trees in biology or subject classification systems for scientific literature. Given a string pattern p and a category (level in the category tree), we wish to efficiently retrieve the t *categorical units* containing this pattern and belonging to the category. We propose several efficient solutions for this problem. One of them uses $n(\log\sigma(1+o(1))+\log D+O(h))+O(\Delta)$ bits of space and $O(|p|+t)$ query time, where n is the total length of the documents, σ the size of the alphabet used in the documents and Δ is the total number of nodes in the category tree. Another solution uses $n(\log\sigma(1+o(1))+O(\log D))+O(\Delta)+O(D\log n)$ bits of space and $O(|p|+t\log D)$ query time. We finally propose other solutions which are more space-efficient at the expense of a slight increase in query time.

Index terms— pattern matching, document retrieval, category tree, space-efficient data structures

*Corresponding Author: dbelazzougui@cerist.dz

†Gregory.Kucherov@univ-mlv.fr