Kopelowitz, Tsvi; Porat, Ely; Rozen, Yaron

Succinct online dictionary matching with improved worst-case guarantees. (English)

Zbl 1380.68154


Summary: In the online dictionary matching problem the goal is to preprocess a set of patterns \( D = \{P_1, \ldots, P_d\} \) over alphabet \( \Sigma \), so that given an online text (one character at a time) we report all of the occurrences of patterns that are a suffix of the current text before the following character arrives. We introduce a succinct Aho-Corasick like data structure for the online dictionary matching problem. Our solution uses a new succinct representation for multi-labeled trees, in which each node has a set of labels from a universe of size \( \lambda \). We consider lowest labeled ancestor (LLA) queries on multi-labeled trees, where given a node and a label we return the lowest proper ancestor of the node that has the queried label.

In this paper we introduce a succinct representation of multi-labeled trees for \( \lambda = \omega(1) \) that support LLA queries in \( O(\log \log \lambda) \) time. Using this representation of multi-labeled trees, we introduce a succinct data structure for the online dictionary matching problem when \( \sigma = \omega(1) \). In this solution the worst case cost per character is \( O(\log \log \sigma + \text{occ}) \) time, where \( \text{occ} \) is the size of the current output. Moreover, the amortized cost per character is \( O(1 + \text{occ}) \) time.

For the entire collection see [Zbl 1351.68018].

MSC:

68P05 Data structures
68Q25 Analysis of algorithms and problem complexity
68U15 Computing methodologies for text processing; mathematical typography
68W32 Algorithms on strings

Keywords:

succinct indexing; dictionary matching; Aho-Corasick; labeled trees

Full Text: DOI