

Research on some pattern matching techniques and their application in approximate search problems

Phạm Văn Quyết

Abstract: This thesis investigates several pattern matching techniques and their application to the problem of approximate string search. While exact pattern matching is well-established, real-world data often contains errors, variations, or noise, necessitating robust approximate solutions. The research systematically analyzes key algorithms, including the Levenshtein automaton, bit-parallel methods, and filtering-based approaches such as the q-gram technique. A comparative evaluation is conducted to assess their efficiency and accuracy across different error thresholds and dataset characteristics. Furthermore, the study proposes a hybrid model that integrates the strengths of these methods to optimize performance for large-scale text corpora. The proposed approach is then applied to practical tasks, including spell-checking and fuzzy retrieval in Vietnamese text, where diacritic sensitivity poses unique challenges. Experimental results demonstrate that the hybrid method significantly reduces search time while maintaining high recall rates. This work contributes a comprehensive framework for selecting and combining pattern matching techniques, offering a practical solution for approximate search in noisy environments.

Keywords: pattern matching, approximate search, string matching, fuzzy search