Povzetek

Učinkovito reševanje problema iskanja nizov v tekstu je odvisno od lastnosti danega teksta in iskanih nizov. Že število iskanih nizov grobo razdeli algoritme za iskanje. Za hitro iskanje vseh pojavitev enega niza si bomo pogledali algoritem, ki uporabi tabelo pripon in ima časovno zahtevnost iskanja $\mathcal{O}(m + \log N + k)$, kjer je m dolžina niza, N dolžina teksta in k število vseh pojavitev. Za primer algoritma, ki zna v enem iskanju hitro poiskati vse pojavitve nizov iz končne množice, si bomo pogledali algoritem s časovno zahtevnostjo $\mathcal{O}(N + k)$, ki si pri iskanju pomaga z avtomatom *Aho-Corasick*. Tudi tekst, v katerem iščemo nize, lahko vpliva na hitrost iskanja. Opisali bomo algoritem, ki bo iskal nize okoli izhodišč v tekstu. Časovna zahtevnost bo odvisna od števila izhodišč v tekstu in števila iskanj okoli njih. Pokazali bomo, da je v primeru, ko so znaki v tekstu porazdeljeni po Zipfovem zakonu, število iskanj majhno. Poleg časovnih zahtevnosti iskanj bo pomembna tudi časovna in prostorska zahtevnost predpriprav za strukture, ki jih posamezen algoritem potrebuje.

Math. Subj. Class. 2000: 68W05, 68P05, 68W40

Comput. Class. System 1998: F.2.2, E.1, G.2

Ključne besede: iskanje nizov, tabela pripon, avtomat Aho-Corasick, izhodišča v tekstu, minimalni prerez, Zipfov zakon

Keywords: string searching, suffix array, Aho-Corasick automaton, points of departures in text, minimal cut, Zipf's law

Literatura

[AC] A. V. Aho in M. J. Corasick, Efficient string matching: an aid to bibliographic search, Communications of the ACM 18 (1975), str. 333–340

[Ah] A.V. Aho, Algorithms for finding patterns in strings, v: J. V. Leeuwen, *Handbook of theoretical computer science* — algorithms and complexity, Elsevier, Amsterdam (1990), str. 255-300

[DL] S. Dori in G. M. Landau, Construction of Aho Corasick automaton in linear time for integer alphabets, Inf. Process. Lett. 98 (2006), str. 66–72

[HMU] J. E. Hopcroft, R. Motwani in J. D. Ullman, *Introduction to automata theory, languages, and computation*, 2. izdaja, Addison-Wesley 2001

[KS] J. Kärkkäinen in P. Sanders, Simple linear work suffix array construction, v zborniku konference ICALP '03, Lecture Notes in Comput. Sci. vol. 2719 (2003), str. 943–955

[MM] U. Manber in E. W. Myers, Suffix arrays: a new method for on-line string searches, SIAM J. Comput. 22 (1993), str. 935–948

[Se] J. Senellart, Fast pattern matching in indexed texts, Theoret. Comput. Sci. 273 (2000), str. 239–262

[WBc] Brownova zbirka dokumentov: http://en.wikipedia.org/wiki/Brown_Corpus

[WGT] Algoritem Goldberg-Tarjanov:
http://en.wikipedia.org/wiki/Relabel-to-front_algorithm

[WGz] Podatki o številu prebivalcev v večjih mestih:
http://people.few.eur.nl/vanmarrewijk/geography/zipf/index.htm

[WNb] Slovenska zbirka dokumentov Nova beseda: http://bos.zrc-sazu.si/s_beseda.html

[WSh] Podatki o številu besed v Shakespearovem delu Hamlet:
http://www.mta75.org/curriculum/english/Shakes/

[WUn] Program Unitex za proučevanje besedil naravnih jezikov:
http://www-igm.univ-mlv.fr/∼unitex/index.html

[WZl] Zipfov zakon: http://en.wikipedia.org/wiki/Zipf's_law