Average-Case Analysis of Pattern-Matching

Mireille Régnier
INRIA-Rocquencourt

November 29, 1993

[summary by Jean-Marc Steyaert]

Knuth-Morris-Pratt algorithm for pattern-matching has been long studied as well as its variants. Its worst-case behaviour has been analyzed in several implementations and upper-bounds on the delay have been given. This talk gives precise estimates for the average-case behaviour under a variety of probabilistic models: uniform, Bernoulli, Markov, both for patterns and texts. The average complexity happens to be linear in all cases and the linearity constant K can be precisely computed thus allowing full comparison with other algorithms.

The results are obtained by an algebraic and language theoretic approach. Basically variations around the average-case behaviour are due to overlapping subpatterns. These overlaps are formally described by means of formal language theory and in this particular case of context-free grammars.

Average costs can then be expressed in terms of formal power series that satisfy quasi-algebraic equations: perturbative terms happen to be almost neglectible. With some use of computer algebra it is then possible to determine precisely expectation and variance for a number of strategies.

Bibliography

[1] Régnier (Mireille). – Average performance of Morris-Pratt-like algorithms, 1993.