

Information Retrieval
Prof. Dr. N. Fuhr

Sascha Kriewel
sascha.kriewel@uni-duisburg.de

Abgabe bis 14. Juli 2004

Übungsblatt Nr. 10

Aufgabe 1: String Matching

Gegeben seien der folgende Text (eine Basenkette) und das gesuchte Pattern:

Text: GCATCGCAGAGATACAGTACG
Pattern: GCAGAGAG

- (a) Führe in allen Einzelschritten ein naives String Matching durch.
- (b) Führe in allen Einzelschritten eine Textsuche nach Knuth-Morris-Pratt durch. Wie sieht insbesondere die Präfix-Funktion *next* aus?
- (c) Führe in allen Einzelschritten eine Textsuche nach Boyer-Moore durch. Betrachte sowohl den Standardalgorithmus als auch den vereinfachten Boyer-Moore-Algorithmus (BMH), der nur eine Vorkommensheuristik verwendet. Wie sehen insbesondere die Felder *dd* und *d* aus?
- (d) Verwende schließlich den Shift-Or-Algorithmus für eine Suche nach dem Pattern **GCAGAGAG**. Betrachte insbesondere, wie sich der jeweils der neue Statusvektor ergibt.
- (e) Vergleiche die Laufzeit des naiven, des Boyer-Moore- und des Shift-Or-Algorithmus. Decken sich die Ergebnisse mit den experimentellen Befunden für englischsprachigen Text?

20 Punkte

Aufgabe 2: PAT-Bäume

- (a) Bei der Suche in einem Trie (von *retrieval* aber wie *try* gesprochen) oder digitalen Suchbaum, kommt es durch „Einweg-Verzweigen“ zu zusätzlichen Knoten im Baum. Erkläre, wie Patricia-Bäume dieses Problem lösen. Was bedeutet das für die maximale Anzahl innerer Knoten im Baum?
- (b) Baue aus dem Bitstring 1001100101 schrittweise einen PAT-Tree der ersten sieben Sistrings auf.

10 Punkte