Information Retrieval - Sommer 2006
Dipl.-Inform. Sascha Kriewel, LF 137
kriewel@is.informatik.uni-duisburg.de

Übungsblatt 12 — Zugriffspfade

Aufgabe 31: Pattern Matching

Sowohl Knuth-Morris-Pratt als auch Boyer-Moore gehen davon aus, dass ein einzelnes Suchwort im Text gefunden werden soll. Die Verfahren basieren darauf, dass aus dem Wort Information gewonnen wird, die die Suche im Vergleich zum naiven Algorithmus beschleunigt. Falls viele Suchworte in einem einzigen Text gefunden werden sollen, kann es sinnvoll sein, statt der Muster den Text für die Suche aufzubereiten. Wie könnte eine solche Aufbereitung aussehen?

Aufgabe 32: PAT-Bäume

Abgabe bis 6. Juli 2006, 14 Uhr:
iruebg-abgaben@is.informatik.uni-duisburg.de

(a) Bei der Suche in einem Trie (von *reTRIEval*) 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) Gegeben sei ein Text der Länge n aus dem Alphabet $\{0, 1\}$. Die Bereichssuche auf dem zugehörigen PAT-Tree liefert als Ergebnis eine Menge von Teilbäumen. Wie viele Teilbäume erhält man im *worst case* und warum?

(c) Baue aus dem Bitstring `1001100101` schrittweise einen PAT-Tree der ersten sieben Sistrings auf.

Aufgabe 33: Signaturen

(a) Was versteht man allgemein unter Superimposed Coding und wodurch können *false drops* entstehen?

(b) Eine Dokumentsammlung wird mit Hilfe von Signaturen indexiert. Dabei seien die Suchterme t_1 bis t_7 und die zugehörigen Signaturen gegeben:

t_1	100001
t_2	100100
t_3	011000
t_4	010001
t_5	001010
t_6	001100
t_7	000110

Seien weiterhin die folgenden Dokumente gegeben: Dokument d_1 mit den Termen t_2, t_4, t_7, Dokument d_2 mit den Termen t_3, t_4, t_7 und Dokument d_3 mit den Termen t_1, t_3.

Welche Dokumente werden für eine Anfrage nach t_4 zurückgeliefert? Sind dabei auch *false drops*?

(c) Erläutere das Verfahren der Bitscheibenorganisation zur Verwaltung von Signaturen. Welche Vorteile ergeben sich gegenüber dem Einsatz sequentieller Signaturdateien?