Übung zur Vorlesung

Webbasierte Informationssysteme

Heinz Schweppe, Annika Hinze, WS2000/2001

- Übung 6 -

1. Aufgabe

   (a) Geben Sie Suffixbäume für die Zeichenketten (a) cccc und (b) acacgat an.

   (b) Gesucht das Suffixfeld für die Zeichenkette S der Länge n=18 über dem Alphabet $\Sigma = \{a, c, g, t\}$ die mit dem Sonderzeichen # (Position 19) abgeschlossen ist. Die auf Σ definierte Ordnung ist$a < c < g < t$, außerdem ist # kleiner als jedes Zeichen.
   Eine Sortierung nach dem ersten Zeichen hat folgende Mengen ergeben:

   - #, Position: 19
   - a, Position: 1,4,5,11
   - c, Position: 2,6,7,12,13,14,18
   - g, Position: 3,9,10,17
   - t, Position: 8,15,16

   Bestimmen Sie die Zeichenkette direkt ohne Aufbau des entsprechenden Suffixbaumes oder durch Bestimmen der lexikographischen Ordnung der Suffixe per Hand.

   (c) Geben Sie die Größenordnung des benötigeten Platzes für einen Suffixbaum (Suffix-Feld) eines TREC-Datenbestand mit 333 Millionen Wörtern an. Indexpunkte sind die Wortanfänge, ein Wort soll im Mittel 10 Zeichen lang sein, das Alphabet hat 96 Zeichen. Machen Sie ggf. zusätzliche Annahmen.

2. Projekt-Aufgabe

   (a) Testen Sie den Teil ihrer Suchmaschine, der fuer das Einlesen der Daten bis zum Indexaufbau verantwortlich ist. Geben Sie dazu geeignete Startadressen vor. Messen Sie die Zeit fuer die einzelnen Komponenten, wieviele Dokumente/Begriffe werden pro Zeiteinheit verarbeitet? Erstellen Sie eine statistische Uebersicht ueber den erstellten Index (wieviele Begriffe, URLs etc.).

   (b) Welche Problembereiche konnten sie durch die Tests (siehe a) feststellen, sind die Gruende entwurfsbedingt oder durch eine unguenstige Implementation hervorgerufen? Schlagen Sie Alternativen vor.

   (c) Untersuchen Sie die moegliche Einbindung eines Thesaurus. Nach welchen Gesichtspunkten sollte der Thesaurus gewaehlt werden? Eine Liste von Thesauri ist ueber die Homepage der VL zu erreichen.

   (d) Stellen Sie die fuer Ihre Suchmaschine geplante Rankingfunktion vor. Beziehen Sie sich auf die Sprachdefinitin und arbeiten sie mit aussagekraeftigen Beispielen.