Übung zur Vorlesung

Webbasierte Informationssysteme

Heinz Schweppe, Annika Hinze, WS2000/2001

- Übung 5 -

1. Aufgabe Geben Sie eine Codierung an, sodass die Zuordnung einer Person zu einer Telefonnr. effizient hergestellt werden kann (Die Anfragen sind also von der Art: "Wer hat die Telefonnr. 4711?")
   Verwenden Sie die folgenden Beispieldaten: (4712, Abel), (5588, Bebel), (5834, Cebel), (7754, Debel), (8832, Ebel), (9912, Febel), (9999, Gebel)
   Zur Erinnerung: Präfixbäume enthalten in den entsprechenden inneren Knoten die Substrings s, die übersprungen werden, weil s in mehr als einer Zeichenkette an derselben Position vorkommt. Bitte geben sie in den inneren Knoten an (wie in der VL), wieviele Zeichen eines zu suchenden Schlüssel zu überspringen sind, nicht, welche Zeichenposition als nächstes zu überprüfen ist (wie es häufig in der Literatur zu finden ist).

   Verwenden Sie

   (a) einen Trie

   (b) einen Patricia-Baum,

   (c) einen Präfixbaum.

   (d) Geben Sie einen Pseudocode-Algorithmus an, der ein Wort in einem Patricia-Baum sucht.

2. Aufgabe

   (a) Bauen Sie einen Suffix-Baum für die folgende Zeichenfolge auf: TACCCGAGGTAGCCT

   (b) Angenommen das Alphabet besteht nur aus 4 Zeichen (plus ein Sonderzeichen, das wenn nötig eine Zeichenkette abschließt). Geben Sie den Speicheraufwand für diese Zeichenkette ab. Machen Sie dazu geeignete Annahmen.

   (c) Stellen Sie den Baum dar, wenn jedes vierte Zeichen "Indexpunkt" ist. Die obige Zeichenkette soll also eine Wortfolge von Wörtern sein, die alle die Länge 3 besitzen. Für diese Wortfolge soll ein Suffixbaum aufgebaut werden.

   (d) Kann man ohne weitere Annahmen allgemein etwas über den Speicherbedarf für eine Suffix-Baum - Darstellung einer Zeichenkette der Länge n sagen?

3. Projekt-Aufgabe
   Die verschiedenen Projekt-Aufgaben werden in der Übung besprochen.

4. Fortfuehren der auf Blatt 3 begonnenen Aufgabe (insgesamt 3 Wochen Bearbeitungszeit)
   Entwerfen, Implementieren und Testen Sie einen Index, der innerhalb der im Projekt entwickelten Suchmaschine eingesetzt werden koennte. Ziel der Implementierung ist eine schnelle Abarbeitung von Anfragen. Die Aufgabe wird ueber einen Zeitraum von drei Wochen bearbeitet, die einzelnen Zwischenschritte in den Uebungen besprochen:

   (a) Entwickeln Sie eine geeignete Datenstruktur zum Speichern aller fuer das Projekt benoetigten Daten; diese werden in der Uebung am 8.11. besprochen. Legen Sie den geplanten Ablauf von Indexaufbau und Abfrageabarbeitung auf Basis der von ihnen gewaehlten Datenstruktur dar.
   Abgabe: 15.11.2000

(b) Implementieren Sie die Indexstruktur und die Indexierung. Nutzen Sie dazu als Datengrundlage die Sammlung von Datenfiles, die aus den man-pages erstellt wurde (Link auf der Homepage der VL). Implementieren Sie in Java, ermoeglichen Sie die Einbindung ihres Indexierungsmoduls in einen groesseren Projektkontext durch einen modularen Aufbau der einzelnen Komponenten und eine gute Dokumentation des Quellcodes und der Schnittstellen. Die Inverslisten sollen mit einem in der VL vorgestellten Kompressionscode ihrer Wahl kodiert werden (z.B. Gamma oder Golomb-Code).
Abgabe: 22/29.11.2000

(c) Erarbeiten Sie geeignete Kriterien und Test-Strategien zur Evaluation des Index und testen Sie ihre Implementierung.
Abgabe: 22/29.11.2000