Übung zur Vorlesung

Webbasierte Informationssysteme

Heinz Schweppe, Annika Hinze, WS2000/2001

- Übung 4 -

1. Aufgabe
   Gesucht sind (Pseudocode-) Algorithmen zur Auswertung von speziellen Anfrageformen über einem invertierten Dokumentbestand (Posting-Listen mit Wortposition im jeweiligen Dokument, Stoppwörter eliminiert)

   (a) Nahbereichsoperator (proximity operator): "w1 near k w2 " heisst: w1 und w2 befinden sich in einem Dokument und sind höchstens durch k Wörter voneinander getrennt. Beschreiben Sie einen moeglichen Algorithmus.

   (b) Phrasensuche: Schlagen Sie einen Algorithmus zur Auswertung von Anfragen vor, die als Ganze Sätze formuliert werden. Diskutieren Sie ggf. Vor- und Nachteile verschiedener Verfahren.

   (c) Beschreiben Sie ein Verfahren zur Suche anhand von Beispieldokumenten: "Dokumente wie dieses" .

2. Projekt-Aufgabe

   (a) Untersuchen Sie die Moeglichkeit, Invers- (Posting-) listen mit Hilfe eines relationalen DBS zu gespeichern. Machen Sie dazu einen Vorschlag bezueglich eines moeglichen Schemas und der SQL-Anfragen, mit denen Vektorraumanfragen mit zwei Anfrageterme emuliert werden können.

   (b) Setzen Sie sich kritisch mit diesem Ansatz auseinander.

   (c) Prüfen Sie anhand der Systemlisteratur, wie Oracle 8i / Intermedia die Inverslisten speichert.

3. 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