6. Übungsblatt

Ausgabe: 3. Kalenderwoche · Besprechung: 4/5. Kalenderwoche

Aufgabe 0

- Wie viel Bits benötigt ein Bitmap Index für N Spalten einer Tabelle mit jeweils n unterschiedlichen Werten?
- Wie kann man einen Index für eine Spalten mit Werten von 0 bis 999.999 zerlegen?
- Welche Vorteile und Nachteile bringen Bitmaps Indices? (Tipp: Join Operationen und Datenzugriff)
- Warum verwendet man Kompressionstechniken für die Bitmaps und was ist dabei wichtig?

Aufgabe 1

Es gibt viele unterschiedliche Möglichkeiten Folgen von Bits zu kompremieren. Ein mögliches Verfahren ist das Word-Aligned Bitmap Compression Verfahren. Das Verfahren wird in folgendem Paper beschrieben: Optimizing Bitmap Indices with Efficient Compression, Wu et al. 2006[1] Lesen Sie das Paper im Anhang gründlich durch und beantworten Sie folgende Fragen:

1. Wie funktioniert Run-Length Encoding im Algemeinen und speziel by Bitmaps?
2. Was versteht man unter Word-Aligned Hybrid Code (WAH)? Erklären Sie WAH am Beispiel in Figure 2 im Paper! Was ist der Vorteil von WAH zu anderen Kompressionsverfahren vie GZip zum Beispiel?
3. Wie lassen sich einfache logische Operationen wie AND und OR effizient auf WAH kodierten Daten durchführen?

[1] http://dl.acm.org/citation.cfm?id=1132864

Aufgabe 2

In der Vorlesung haben Sie den ETL Prozess kennen gelernt als wichtiger Schritt in einer DWH Anwendung. Eine wichtige Aufgabe dabei ist das Erkennen von Duplikaten. Dies ist wichtig um konsistente und nicht redundante Daten zu haben. Bei Zeichketten nimmt man häufig die Edit Distanz um ganze Strings miteinander zu vergleichen. Je geringer die Edit Distanz ist, desto wahrscheinlicher handelt es sich um ein Duplikat. Implementieren Sie die Edit Distanz in Java! Auf der Webseite finden Sie unter sonstiges eine Datei mit Restaurantname. Lesen Sie diese Daten ein und erzeugen Sie eine Ähnlichkeitsmatrix wie sie in den Vorlesungfolien angedeutet wurde. Der Eintrag in der Zeile i und der Spalte j der Ähnlichkeitsmatrix ist die Edit Distanz vom Beispielstring i und j. Überlegen Sie sich einen sinnvollen Schwellwert ab welcher Edit Distanz ein String das Duplikat des anderen Strings ist.