Ludwig-Maximilians-Universität München
Department "Institut für Informatik"
Dr. Matthias Schubert
Arthur Zimek

München, den 26.6.2008

Knowledge Discovery in Databases II

SoSe 2008

Übungsblatt 10: Edit Distanz

Besprechung 27.6.2008 / 30.06.2008

Aufgabe 10-1 *Levensthein-Distanz auf Strings*

In der Vorlesung wurde die Edit-Distanz als mögliches Abstandsmaß für Graphen eingeführt. Betrachtet man einen Pfad in einem Graphen, kann man diesen als String zwischen den Knoten bzw. Kantelabels darstellen. Sequenzen können also auch als Teilmenge der Graphen betrachtet werden, die mittels Edit Distanz verglichen werden können. Die Levensthein-Distanz ist ein Distanzmaß auf Strings, das einen Vergeleich in quadratischer Zeit erlaubt.

Seien S_1, S_2 zwei String über dem Alphabetbet Σ. Wobei das Gap-Symbol " $\in \Sigma$ ist und eine Auslassung symbolisieren soll.

Sei K eine Kostenmatrix, die die Kosten des Vertauschens eines Elements aus Σ mit einem anderen Element beschreibt. Hierbei gilt, dass Löschungen oder Einfügungen eines Symbols als Vertauchen mit dem Gap-Symbol behandelt werden können.

Die Levensteindistanz besteht jetzt aus den minimalen Kosten aller Sequenzen von Operationen die S_1 in S_2 überführen. Hierbei verwendet man dynamische Programmierung um diese Distanz möglichst effizient zu berechnen.

Vergleichen Sie die folgenden zwei Strings "ABBBA" und "BBAB" über dem Alphabet $\Sigma = \{A, B, -\}$ mit der LevenSthein-Distanz. Verwenden sie dabei die folgenden Kostenmatritzen.

(a)

$$K_1 = \begin{bmatrix} & A & B & - \\ A & 0 & 1 & 1 \\ B & 1 & o & 1 \\ - & 1 & 1 & 0 \end{bmatrix}$$

(b)

$$K_2 = \begin{bmatrix} & A & B & - \\ A & 0 & 2 & 1 \\ B & 2 & o & 1 \\ - & 1 & 1 & 0 \end{bmatrix}$$