Ludwig-Maximilians-Universität München
Institut für Informatik
Priv. Doz. Dr. Matthias Schubert

München, den 6.7.2010

Knowledge Discovery in Databases II

SoSe 2010

Übungsblatt 9: Edit Distanz und Graphlets

Besprechung 8.7.2010

Aufgabe 9-1 *Levenshtein-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. Kantenlabels darstellen. Sequenzen können also auch als Teilmenge der Graphen betrachtet werden, die mittels Edit Distanz verglichen werden können. Die Levenshtein-Distanz ist ein Distanzmaß auf Strings, das einen Vergeleich in quadratischer Zeit erlaubt.

Seien S_1, S_2 zwei Strings über dem Alphabet $\Sigma \setminus -$. Um eine Auslassung zu symbolisieren verwenden wir $-$.

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

Die Levenshteindistanz besteht jetzt aus den minimalen Kosten aller Sequenzen von Operationen die S_1 in S_2 überführen. Mit Hilfe von dynamischer Programmierung geht dies effizient in $O(n^2)$.

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

(a)

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

(b)

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

Aufgabe 9-2 *Graphlet Kernels*

Wir suchen nach Graphlets der Größe 4, d.h. Subgraphen mit 4 Knoten, in einem Graphen.

(a) Wie viele Graphlets der Größe 4 gibt es? Wie groß ist der Zeitaufwand, um sie zu bestimmen?

(b) Wie teuer ist der Vergleich zweier Graphen mittels dieser Graphlets?

(c) Wie kann man den Vergleich der Graphlets beschleunigen?

(d) Bildet man die Graphlets explizit in den Feature Space ab? Ist die Antwort stets ja oder nein?