Information Mining - Wintersemester 2014
Dipl.-Inform. Vu Tran, LF 139
vtran@is.inf.uni-due.de

Übungsblatt 11

Aufgabe 24: Praktische Übung: Clustering

Die Websuchmaschine Carrot[1] führt automatisch ein Clustering der Suchergebnisse durch und zeigt dieses an.

Stelle einige Suchanfragen und schaue dir die resultierenden Clusterings an.

(a) Wonach werden die Suchergebnisse strukturiert?

(b) Findest du die Strukturierung hilfreich? Warum bzw. warum nicht? Welche Schwierigkeiten könnten auftreten?

(c) Bei welchen Arten von Suchaufgaben und -anfragen könnte die Strukturierung besonders hilfreich sein und warum?

Aufgabe 25: Clustering: k-Means

Im Information Retrieval spielt das Vektorraummodell eine wichtige Rolle. Dabei werden alle Dokumente als Vektoren von Termen aufgefasst, deren einzelne Elemente das Gewicht eines jeweiligen Terms im Dokument wiedergibt.

Dieses Modell kann auch zum Clustering verwendet werden. Angenommen, wir haben die Termmenge T = {`haus, auto`}; der dazugehörige Vektorraum ist also zweidimensional und wird von den beiden Termen `haus` und `auto` aufgespannt. Wir haben nun folgende fünf Dokumente, die sich als Vektoren beschreiben lassen (dabei sei $\vec{d}$ der zu einem Dokument d zugehörige Vektor):

$$\vec{d_1} = \begin{pmatrix} 0,8 \\ 0,9 \end{pmatrix}, \vec{d_2} = \begin{pmatrix} 0,9 \\ 0,65 \end{pmatrix}, \vec{d_3} = \begin{pmatrix} 0,5 \\ 0,5 \end{pmatrix}, \vec{d_4} = \begin{pmatrix} 0,2 \\ 0,25 \end{pmatrix}, \vec{d_5} = \begin{pmatrix} 0,25 \\ 0,1 \end{pmatrix}$$

`haus` hat also im Dokument d_1 das Gewicht 0,8, während `auto` dort das Gewicht 0,9 hat, usw.

(a) Fasse die Dokumente mittels k-Means-Clustering zusammen. Es sollen dabei zwei Cluster gebildet werden. Als initiale Seeds sollen d_4 und d_5 verwendet werden.

(b) Skizziere graphisch den Dokumentenraum und die Bewegung der Zentroiden. Was fällt auf?

(c) Berechne für jeden Iterationsschritt die *Purity* und den *Rand-Index*. Gehe dabei von folgender manuellen Klassifikation aus: $C_1 = \{d_1, d_2, d_3\}$ und $C_2 = \{d_4, d_5\}$.

[1] `http://search.carrot2.org/stable/search`