Explorative Analyse von Textkorpora mit Clusterverfahren

Daniel Pullwitt

Die Handhabung großer Sammlungen von Textdokumenten erfordert Methoden zur automatischen Analyse der inhaltlichen Struktur. Eine geeignete Technik zur explorativen Analyse ist die Clusterbildung über geeigneten Vektorrepräsentationen. Die vorliegende Arbeit behandelt verschiedene Teilaufgaben der Clusteranalyse.

Die Modellierung mit wortformbasierten Merkmalen wird inklusive Methoden zur Reduktion der Komplexität beschrieben. Eine Analyse der Struktur und Dimension von Korpora in wortformbasierten Merkmalsräumen wird durchgeführt und ein zweistufiges Modell zur Ableitung von Merkmalen über Dokumentteilen (z. B. Sätzen) entwickelt, dass Dokumentvektoren einer Dimension nahe der ermittelten intrinsischen Dimension erlaubt.

Es wird ein Überblick gegeben über verschiedene gebräuchliche Clusterverfahren sowie Methoden zur Visualisierung und eine neue, auf Überlegungen der Informationstheorie basierende, k-means-Variante vorgestellt. Methoden zur Messung der Qualität von Clustereinteilungen sowohl im Merkmalsraum als auch in Bezug auf eine vorgegebene Referenzeinteilung werden entwickelt bzw. beschrieben und zum Vergleich einer Reihe von Verfahren und Modellierungen anhand von vier verschiedenen Testkorpora benutzt.

Die Experimente zeigen eine geringfügige Überlegenheit informationstheoretischer Verfahren gegenüber klassischen auf Euklidischem Abstand bzw. innerem Produkt basierenden Abständen sowie Vorteile der satzbasierten Modellierung für lange Texte. Ebenso wird demonstriert, dass die zur Visualisierung geeigneten selbst organisierenden Karten nur mit geringen Verlusten der Clustereffektivität verbunden sind und auf Grund von Komplexität und Eignung für große Datenmengen daher eine bevorzugte Technik zur Darstellung von Korpusstruktur sind.

Zusätzlich werden Optimierungen auf Modell- und Implementierungsebene beschrieben, die eine effiziente Anwendung der Algorithmen bei größeren Datenmengen ermöglichen.