Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

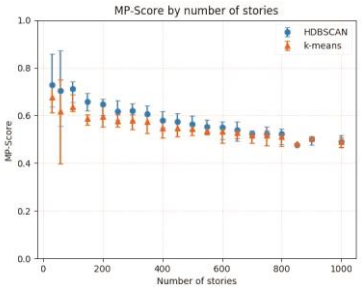
InIT Institut für angewandte Informationstechnologie

Dynamische Eventerkennung in Datenströmen

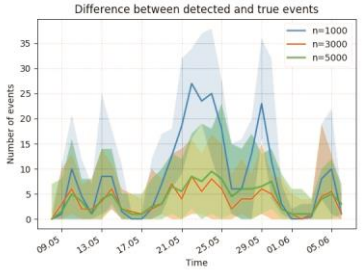
Diplomierende
Daniel Milenkovic
David Pacassi Torrico

Dozierende
Andreas Weiler
Kurt Stockinger

Die Erkennung von Ereignissen in Datenströmen kann sich als schwierig erweisen, vor allem dann, wenn sich die Definition, Inhalte oder Eigenschaften eines Ereignisses im Laufe der Zeit verändern können.
Diese Bachelorarbeit fokussiert auf die Entwicklung und Evaluation einer Online-Clustering-Lösung, in welcher Ereignisse entweder als Veränderungen bestehender Cluster oder aber als Bildung neuer Cluster definiert sind. Die Lösung ist eine Text-Mining-Software, welche über einen Datenstrom neue News-Artikel erhält und diese verarbeitet. Dabei werden Artikel aufgrund ihrer Ähnlichkeit zu anderen Artikeln verschiedenen Clustern zugewiesen. Die Annahme ist, dass sehr ähnliche Artikel über dasselbe Thema schreiben. Zusätzlich wurde für die Evaluation der Clusteringalgorithmen eine eigene Bewertungsfunktion entwickelt.
Im ersten Teil dieser Arbeit wurde nach einem geeigneten Datensatz gesucht, in welchem Inhalte gemäss ihrer Ähnlichkeit gruppiert werden. Die umgesetzte Lösung verwendet HDBSCAN als Clustering-Methode und vergleicht diese mit dem State-of-the-Art-Verfahren k-means. Dabei stellte sich heraus, dass die Verwendung von HDBSCAN Vorteile bei der Performanz, wie aber auch bei der Präzision gegenüber k-means auf-weist. Des Weiteren wurden auch verschiedene Textvorverarbeitungsmethoden evaluiert. Der Einsatz von Text-Lemmatisierung und des Tf-idf-Masses verbesserten das Clustering in Hinsicht auf Präzision. Bei der abschliessenden Evaluation stellte man fest, dass nicht zugewiesene News-Artikel die Präzision des Clusteringverfahrens reduzieren.
Die resultierende Präzision des Clusteringverfahrens beträgt 72 % bei einer Standardabweichung von 12 %. Die Präzision zur Erkennung neuer Ereignisse beträgt 62 % bei einer Standardabweichung von 43 %. Die Erkennung von Änderungen bestehender Ereignisse ergibt eine Präzision von 69 % bei einer Standardabweichung von 16 %. Eine Fortsetzung dieser Arbeit sollte die Verbesserung des Clusterings sein, um die Präzision bei der Erkennung von Ereignissen zu erhöhen.



Der Vergleich zwischen HDBSCAN und k-means anhand der MP-Score und der Anzahl von Stories.



Die Differenz zwischen erkannten Ereignissen und realen Ereignissen in einem dynamischen Datenstrom über 30 Tage. Die Grafik zeigt drei Durchläufe mit unterschiedlichen Batch-Grössen.