Zusammenfassung

Entity Resolution ist der Prozess, in einer oder mehreren Datenquellen Gruppen von Datensätzen zu identifizieren, die derselben realen Entität entsprechen. Dabei gibt es kein einzigartiges Attribut, welches zur Zuordnung genutzt werden kann. Die Unterschiede in den Datensätzen entstehen, beispielsweise durch Rechtschreibfehler oder fehlende und vertauschte Attribute, welche eine Vieldeutigkeit erzeugen, die bei manueller Betrachtung durch einen Menschen meist nur mit viel Zeitaufwand aufzulösen sind. Damit ein Entity Resolution Workflow diese Vieldeutigkeiten auslösen kann, muss dieser abhängig von der Domäne der Daten konfiguriert werden. Diese Konfiguration besteht aus einer Vielzahl von Parametern, die auch von einem Domänexperten nur aufwändig zu bestimmen sind. Erster Beitrag dieser Arbeits ist deshalb die Analyse und Entwicklung von Verfahren, die eine Selbstkonfiguration der Parameter in Abhängigkeit der Datenquelle ermöglichen. Dabei liegt der Fokus dieser Arbeit auf Entity Resolution Verfahren für Event Stream Processing Systeme. Hierbei ist neben der Qualität der Ergebnisse auch die Antwortzeit von Bedeutung, welche oft im Subsekundenbereich liegen muss. Die Suche nach Duplikaten ist jedoch mit enormen Kosten verbunden, die beim vollständigen Durchsuchen aller Bestandsdatensätze in einer quadratischen Komplexität resultiert. Der Zweite Beitrag dieser Arbeit ist daher ein sog. Blocking-Verfahren zur Reduzierung der Komplexität, welches für Event Stream Processing tauglich ist. Für die Selbstkonfiguration bedeutet dies, dass neben der Qualität auch die Effizienz berücksichtigen muss. Die analysierten und entwickelten Verfahren wurden in einem prototypischen System implementiert, dass sich unüberwacht (ohne Eingreifen des Benutzers) vor der Laufzeit selbst konfiguriert und anschließend Anfragen aus einem Ereignisstrom beantwortet. Die Auswertung dieses Systems zeigt, dass die Selbstkonfiguration auf einem Datensatz mit 4 Mio Einträgen ein F-Measure von bis zu 70 % erreicht und bei 1.3 Mio Anfragen im Durchschnitt über 500 pro Sekunde beantwortet.