

Datenqualität und Datenreinigung

Felix Naumann

Datenqualität bezeichnet die Eigenschaft von Daten, einem bestimmten Zweck zu dienen. Kundendaten sind für Unternehmen besonders wertvoll und dem gezielten Kontakt mit Menschen. Leider sind sie oft verschmutzt und somit von niedriger Qualität: Hausnummern werden durch Zahlendreher verfälscht, Vornamen und Nachnamen werden vertauscht und enthalten Tippfehler, Postleitzahlen passen nicht zu den Ortsnamen, Telefonnummern fehlen, usw. Aufgrund solcher Fehler werden Kunden falsch oder gar doppelt angeschrieben, Briefe können nicht zugestellt werden, Kreditlimits werden mehrfach vergeben usw. Diese Effekte senken die Kundenzufriedenheit und verursachen im Unternehmen Kosten.

Methoden der Datenreinigung säubern und korrigieren solche verschmutzten Daten. Insbesondere betrachten wir Verfahren, welche Kundendubletten - also doppelte Einträge in der Kundendatenbank für eine einzelne Person - automatisch aufspüren. Dieses Problem ist aus zweierlei Hinsicht besonders schwierig: Erstens sind doppelte Kundendatensätzen selten identisch - so sollte beispielsweise automatisch erkannt werden, dass <P. Maier, Hauptstraße 9a> dieselbe Person ist wie <Paul Meier, Hauptstr. 9>. Zweitens ist die Suche nach solchen Dubletten sehr aufwändig; bei nur 10.000 Kunden müssen bereits 50 Millionen Vergleiche durchgeführt werden. Im Vortrag werden geeignete Ideen zur Lösung beider Probleme vorgestellt.