Aufbau und Pflege von Kunden-Ids im Finanzdienstleistungsgewerbe

Thomas Rüdiger
AXA Köln
Thomas.Ruediger@gmx.at

Zusammenfassung

Dass um 12 % einer Kundenbasis Dubletten sind, ist in Unternehmen mit Adressbeständen keine seltene Situation. In der direkten Folge von Dubletten entstehen auf der Unternehmensseite Mehrkosten fürs Unternehmen (z.B. Mailkampagnen, Management von Kundenbeschwerden, CRM-Analysen). Auf der Kundenseite wollen Kunden gezielt kontaktiert werden und nicht Produkte angeboten bekommen, die sie bereits mit dem Unternehmen abgeschlossen haben.
Probleme entstehen, wenn Unternehmen keine Kunden-Id im Datawarehouse pflegen oder im Fall einer im Datawarehouse vorhandenen Kunden-Id Dubletten sich nicht erkennen lassen (Undermapping) oder die Dublettenzusammenführung ungerechtfertigt ist (Overmapping).
SAS bietet als Lösung Funktionen im Bereich der Datenvorwäsche (Adressbereinigung) vor dem eigentlichen Deduplikationsverfahren und für die eigentliche Deduplikation den hier beschriebenen SAS-Spedis-Algorithmus an.

Keywords: Deduplikation, Dublettenbereinigung, Adressbereinigung, Hoax Data, Overmapping, Undermapping, Id-Journal, Spedis.

1 Der Kunde hat viele Gesichter, in der Regel eins

Deduplikation lässt sich beschreiben als eine hierarchische n:1-Abbildung unter Berücksichtigung möglicher textlicher Unschärfen in den untersuchten Feldern, z.B. N Verträge oder Partnernummern auf n Kunden-Ids (N > n).

Tabelle 1: Beispieldaten vor der Adressbereinigung

Name	Straße/ Hausnummer	PLZ/ Ort	Geburtsdatum
Schäfer, Manfred	Blumenstr. 7	50678 Köln-Deutz	01.01.1900
Schäfer, Manni	Blumenweg 7	50678 KOELN	05.01.1975
Schäfer,M.	Bluemenstr. 7a	5000 Keoln	05.01.1957

Die in Tabelle 1 enthaltenen Daten zeigen exemplarisch die Probleme, die sich bei der Zusammenführung von Adresseinheiten zu einer Kunden-Id ergeben. Textliche Unschärfen tragen einen wesentlichen Beitrag dazu bei, ob sich abhängig von der jeweiligen Datensortierung der Kunde eindeutig erkennen lässt oder nicht.

2 Deduplikation verläuft in Schritten

In der Sammelphase der Deduplikation ist es notwendig, die Datenarchitektur, die Feldstrukturen und die Wege zu einer Adresshistorisierung zu analysieren. In der Phase der Datenvorwäsche werden Adressen bereinigt und Kundenzuordnungsregeln des Unternehmens (z.B. Partnernummern) als Vorstufe der Deduplikation herangezogen. Unvollständige oder ungültige Adressen ('Hoax Data') wie 'Testperson Mustermann' oder 'Mickey Mouse' erhalten einen Sonderwert für die Kunden-Id (z.B. -1).
Die Deduplikation verwendet SAS-Funktionalitäten (spedis), hinterfragt vorgenommene Deduplikationen und hinterlegt Änderungen in einer Journal-Datei. Im 'CRM-Paradies' kommt es mit den zusätzlichen Methoden des Data Mining, Kampagnen Controlling, Lifetime Value und Share of Wallet – Analysen zu Kunden-Id-optimierten Responsequoten.

3 Daten-Vorwäsche erfolgt unter Nutzung von SAS/Base-Funktionen

Es gibt zahlreiche Textfunktionen aus SAS/BASE, die die Adressbereinigung unterstützen (siehe Tabelle 2).

Tabelle 2:

Waschgang	**SAS/Base-Funktionen**
standardisieren	`left, compbl`
Groß-/Kleinschreibung	`substr, upcase, lowcase`
Umlaute/Sonderzeichen	`tranwrd` (Ersetzen + Korrektur)
Textfelder splitten	`scan, index, substr`
Vollständigkeit von Feldern, Anzahl Vokale/Konsonanten	`substr`
Feldabhängigkeiten, z.B. Geschlecht–Vorname, PLZ–Ort (Länderformate)	`spedis`
Rechtschreibkorrektur, Fantasienamen, Text Mining	`spedis`

In Tabelle 3 sind die Beispieldaten aus Tabelle 1 am Ende der Datenbereinigungphase dargestellt.

Tabelle 3:

Name	Straße/ Hausnummer	PLZ/ Ort	Geburtsdatum
Schäfer, Manfred	Blumenstr. 7	50678 Köln	.
Schäfer, Manfred	Blumenweg 7	50678 Köln	05.01.1975
Schäfer,M.	Bluemenstr. 7a	5000 Köln	05.01.1957

4 Spedis-Algorithmus klassifiziert nach Ähnlichkeit

Die Spedis-Funktion liefert nach Konkattenieren von zu vergleichenden Datenfeldern einen Datenvergleichswert, der zu drei möglichen Entscheidungen im Deduplikationschritt führt (Tabelle 4). Die Entscheidungsgruppen (I) und (III) führen zu schnellen Ergebnissen (Zeitgewinn) und Mitarbeiter-Entlastung. Die Gruppe (II) benötigt einen zusätzlichen Verifikationsschritt durch die Mitarbeiter aus den Geschäftsstellen des Unternehmens. Der Cut-Off-Wert S1 zwischen (I) und (II) sowie der Cut-Off-Wert S2 zwischen (II) und (III) lassen sich mit Erfahrungswerten feinjustieren.

Tabelle 4:

Ähnlichkeits-Klasse	Spedis-Intervall (0<S1<S2<200)	Aktion	Vorteil
(I) Match-Group gleich oder sehr ähnlich	<S1	„Dublette!“	Zeitgewinn, Entlastung
(II) Match-Control-Group Ähnlichkeit unsicher	≥S1, ≤S2	Verifizieren	Risikominderung von Fehl-deduplikation
(III) Non-Match-Group ungleich oder sehr unähnlich	>S2	„Neukunde!“	Zeitgewinn, Entlastung

5 Die Spedis-Funktion von SAS bringt schnellen Nutzen statt langfristige Kosten!

Der SAS-Help-Kontext zur Spedis-Funktion zeigt die Text-untersuchenden Eigenschaften von Spedis auf. Der Wertebereich liegt zwischen 0 und 200 (0=match).
Datumsformate sind vorab in Datumstext, sonstige numerische Formate in Text zu konvertieren und danach die notwendigen Felder für den Textvergleich zu konkattenieren.
Werden zwei Adresseinheiten aus einem Datenbestand miteinander verglichen, sollte die Asymmetrie der Spedis-Funktion durch Mittelwertbildung aufgehoben werden. Je nach Performance, Arbeitsspeicher und Anzahl der Datensätze empfiehlt es sich, in übersichtlichen BY-Gruppen der Ausgabedateien nach einem PROC SORT oder nach INDEX CREATE zu operieren.
Um die durch Overmapping (Dublettenzusammenführung ungerechtfertigt) und Undermapping (nicht erkannte Dubletten) im laufenden Prozess zu verwalten, empfiehlt sich die Verwendung einer ID-Journal-Tabelle (Tabelle 5).
Zum Vertrag/Partnernummer V wurde am Tag D die Kunden-ID I1 auf I2 abgeändert (optional: von User U).

Tabelle 5:

Check-Point	Query	Keyword	SPEDIS-Wert
Match	Fuzzy	Fuzzy	0
Singlet	Fuzy	Fuzzy	6
Doublet	Fuzzy	Fuzzy	8
Swap	Fuzzy	Fuzzy	10
Truncate	Fuzz	Fuzzy	12
Append	Fuzzys	Fuzzy	5
Dolete	Fzzy	Fuzzy	12
Insert	Fluzzy	Fuzzy	16
Replace	Fizzy	Fuzzy	20
Firstdel	Uzzy	fuzzy	25
Firstins	Pfuzzy	fuzzy	33
Firstrep	Wuzzy	fuzzy	40
Several	floozy	Fuzzy	50

Referenzen

02/2002 wurde die SAS-Deduplikation im CRM-Mart von AXA Deutschland eingeführt.

Literatur

Zur SAS-Funktion spedis erscheint anstelle eines Literaturhinweises ein Verweis auf die SAS-Help-Kontext angebracht.