



Hochschule **RheinMain**



Exposé

im Studiengang
Informatik (M.Sc.)

Automatisierte Personen-Identifikation auf Basis von Adressdaten
von
Kevin Sapper

Referent: Prof. Dr.
Korreferent: Prof. Dr.
Betreuer: Thomas Strauß
Unternehmen: Detim Consulting GmbH / Universum Group

1 Einleitung

Die Masterarbeit soll in Zusammenarbeit mit der Firma Detim Consulting GmbH geschrieben werden. Dazu wird ein Problemfeld bei dem Kunden Universum Group, der Detim Consulting GmbH, in deren Geschäftsfeld dem Inkasso-, Liquiditäts-, und Risikomanagement, gewählt.

2 Problemfeld

Die Universum Group bietet Lösungen für Onlineshops zur Bonitätsprüfung, Adressprüfung und Forderungsankauf. Dabei wird dem Händler bei entsprechender Bonität seines Kunden das Angebot gemacht, die Forderung, nach Ablauf einer Zahlungsperiode, zu 100 % zu übernehmen. Sowohl bei der Risikobewertung, während der Bonitätsprüfung, als auch beim Inkassomanagement müssen Kunden eindeutig identifiziert werden. Dabei ist Hauptmerkmal zur Identifizierung von Kunden deren Adresse und eventuell das Geburtsdatum. Das Problem an dieser Stelle ist, dass der Kunde selbst die Daten erfasst und diese nicht anhand von Personalausweis oder Ähnlichem überprüft werden können. Fehler bei der Datenerhebung sind, beispielsweise unterschiedliche Schreibweisen, insbesondere bei Adressen, Tippfehler, welche bei Namen nicht auffallen, Abkürzungen, etwa Str. für Straße, oder akademische Titel und Adelstitel, welche in Onlineformularen nicht standardisiert sind. Bei der Bonitätsprüfung dient die Personenidentifizierung dazu, Kunden mit positiver oder negativer Zahlungsmoral zu erkennen und anzunehmen bzw. abzulehnen. Je genauer die Personenidentifikation ist, desto aussagekräftiger sind die Bonitätsauskünfte von externen Dienstleistern, beispielsweise der Schufa. Beim Inkassomanagement gilt das sog. Schadensminderungsprinzip. Das bedeutet, dass alle angekauften Forderungen eines Kunden nur einmalig abgemahnt werden dürfen. Daher müssen hier Personendubletten gefunden und zusammengeführt werden.

Das aktuelle System zur Personenidentifizierung ist durch einen externen Dienstleister realisiert. Dieser bereinigt und prüft Namen und Adressen. Allerdings skaliert das System dabei nur innerhalb eines vorgegebenen Kontingents.

3 Zielsetzung

Die Personenidentifizierung soll zukünftig schnell, zuverlässig und skalierbar sein. Personen sollen trotz abweichender Namen, Adressen und Geburtsdaten möglichst genau identifiziert werden. Bei Abweichungen soll ein Ähnlichkeitswert (Score) bestimmt werden, der mit Hilfe einer Toleranz für die Identifikation genutzt werden soll, um Personen zu erkennen. Dazu müssen die Eingabedaten bereinigt werden. Die Adressbereinigung und Adressvalidierung soll sich auf Adressen im deutschen Sprachraum konzentrieren. Als Primärziel ist dabei

Deutschland zu betrachten. Die Lösung muss allerdings auch vorsehen andere Länder mit überschaubarem Aufwand anzubinden. Bei den Namen sollen Vertauschungen bei Vor- und Nachname, sowie etwaige Titel im Namen erkannt werden. Die gewählten Algorithmen, die Architektur und das Design sollen in einem Proof of Concept evaluiert werden. Für das Proof of Concept soll, sofern möglich auf Open Source Lösungen gesetzt werden.

4 Methoden

Zur Umsetzung der in Abschnitt 3 beschriebenen Ziele muss zunächst eine Wissensbasis durch Literaturarbeit in folgenden Grundlagen geschaffen werden:

- Formate von Adressen im Euroraum
- Charakteristiken von persönlichen Namen
- Möglichkeiten von Eingabefehler persönlicher Daten bei unterschiedlichen Medien
- Duplikatserkennung (phonetisch, regelbasierend, ...)
- Fehlerkorrektur von Namen und Adressen

Weitere Methoden sind:

- OOP-Entwurf
- Schnittstellen-Entwurf
- Proof of Concept
- Funktionelle Leistungsbewertung anhand realer Anwendungsfälle

5 Erwartete Ergebnisse

Die erwarteten Ergebnisse der Masterarbeit sind:

- Analyse der Adressaufbauten im deutschen bzw. europäischen Sprachraum.
- Analyse der Möglichkeiten zur Namensbereinigung
- Analyse der Möglichkeiten zur Adressbereinigung und -validierung
- Analyse der Personenidentifikation anhand bereinigter und validierter Adressen
- Design eines Identifikationschecksystems zur Erkennung von Duplikaten
- Prototyp der wesentlichen Funktionen
- Evaluation des Prototypen, anhand der Erwartungen aus der Analyse

6 Vorbedingungen

- Adressdatenbank zur Validierung. Kann für Testzwecke aus OpenStreet-Map extrahiert werden.
- Namesregister zum Abgleich von Vor- und Nachnamen.
- Historie an fehlerhaften und korrigierten Personen- und Adressdaten.

7 Literatur