

Wie man Dubletten im Datenbestand findet und vermeidet

Über den Autor:

Der Informatiker Dipl.-Math. Klaus-Dieter Sedlacek studierte an der Universität Stuttgart Mathematik und Informatik und beschäftigte sich seit dem Studium mit Fragen der künstlichen Intelligenz. Heute arbeitet er als System Analyst für die Firma TOLERANT Software in Stuttgart.

Über das Buch:

In einer Welt, in der Daten das neue Gold sind, ist die Vermeidung von Dubletten in Datenbeständen entscheidend für die Effizienz und Genauigkeit von Datenanalysen und Geschäftsentscheidungen. Klaus-Dieter Sedlacek, ein Experte auf dem Gebiet der Informatik und künstlichen Intelligenz, bietet im Werk "Wie man Dubletten im Datenbestand findet und vermeidet" ein umfassendes Kompendium an Strategien und Methoden zur Datenpräparation. Von der Normalisierung von Namen und Adressen bis hin zu fortgeschrittenen Algorithmen zur Duplikaterkennung, deckt Sedlacek alle Aspekte ab, die für das moderne Datenmanagement unerlässlich sind. Dieses Buch ist ein unverzichtbarer Leitfaden für jeden, der in der Datenverarbeitung tätig ist, und bietet praktische Lösungen für ein Problem, das in der digitalen Ära zunehmend an Bedeutung gewinnt.

Wie man Dubletten im Datenbestand findet und vermeidet

Personen-Suchfelder passend
präparieren

Von

Klaus-Dieter Sedlacek

TOLERANT Software

TOLERANT Software Fachbuch Bd. 3

TOLERANT Software

© 2024 Klaus-Dieter Sedlacek

Sprache der Originalausgabe: Deutsch

Druck und Distribution im Auftrag des Autors/der Autorin:
tredition GmbH, Halenreihe 40-44, 22359 Hamburg, Deutschland

Softcover ISBN 978-3-384-15178-0

Das Werk, einschließlich seiner Teile, ist urheberrechtlich geschützt. Für die Inhalte ist der Autor/die Autorin verantwortlich. Jede Verwertung ist ohne seine/ihre Zustimmung unzulässig. Die Publikation und Verbreitung erfolgen im Auftrag des Autors/der Autorin, zu erreichen unter: tredition GmbH, Abteilung "Impressumservice", Halenreihe 40-44, 22359 Hamburg, Deutschland.

Inhaltsverzeichnis

1. Einführung in die Präparation von Personen-Suchfeldern.....	7
1.1 Bedeutung der Suchfeldpräparation zur Dublettenvermeidung.....	7
1.2 Überblick über gängige Herausforderungen und Ziele.....	9
2. Grundlagen der Datenpräparation.....	11
2.1 Definition und Relevanz von Personen-Suchfeldern.....	11
2.2 Typische Struktur und Inhalte von Personen-Suchfeldern.....	12
2.3 Allgemeine Prinzipien der Datenbereinigung.....	14
3. Spezifische Methoden zur Präparation von Personen-Suchfeldern.....	18
3.1 Namensfelder.....	18
3.1.1 Normalisierung von Vor- und Nachnamen (z.B. Entfernung von Titeln und Suffixen).....	18
3.1.2 Umgang mit internationalen Namenskonventionen.....	20
3.1.3 Behandlung von Sonderzeichen und Diakritika.....	22
3.2 Adressfelder.....	24
3.2.1 Standardisierung von Adressformaten.....	24
3.2.2 Bereinigung und Konsolidierung von Adresszusätzen.....	27
3.2.3 Umgang mit abweichenden Schreibweisen von Ortsnamen.....	29
3.3 Kontaktdaten.....	31
3.3.1 Einheitliche Formate für Telefonnummern.....	31

3.3.2 Normalisierung von E-Mail-Adressen.....	33
3.4 Geburtsdaten.....	35
3.4.1 Standardformate und deren Einfluss auf die Eindeutigkeit...	35
4. Phonetische Algorithmen zur Dublettenvermeidung.....	38
4.1 Überblick und Einsatzgebiete phonetischer Algorithmen.....	38
4.2 Anwendungsbeispiele und Effektivität im Kontext von Personen-Suchfeldern.....	40
5. Best Practices und Richtlinien.....	43
5.1 Strategien zur laufenden Pflege und Aktualisierung von Personen-Suchfeldern.....	43
5.2 Qualitätskontrolle und Testing von Präparationsmethoden.....	45
5.3 Datenschutzaspekte bei der Bearbeitung von Personenbezogenen Daten.....	48
6. Fallstudien und Anwendungsbeispiele.....	51
6.1 Erfolgreiche Projekte zur Dublettenvermeidung durch geeignete Präparation.....	51
6.2 Analyse von Problemfällen und deren Lösungsansätze.....	53
7. Fazit und Ausblick.....	55
7.1 Zusammenfassung der Kernpunkte.....	55
7.2 Empfehlungen für die Implementierung in der Praxis.....	57

1. EINFÜHRUNG IN DIE PRÄPARATION VON PERSONEN-SUCHFELDERN

1.1 Bedeutung der Suchfeldpräparation zur Dublettenvermeidung

In der digitalen Datenwelt, wo jedes Byte zählt und die Reinheit der Information Gold wert ist, erhebt sich die Suchfeldpräparation wie ein Leuchtturm der Ordnung aus dem stürmischen Meer der Datenchaos. Ein Schlüsselaspekt dieser strahlenden Wacht ist ihre unverzichtbare Rolle bei der Vermeidung von Dubletten, jenen heimtückischen Doppelgängern der Datenwelt, die nicht nur die Effizienz untergraben, sondern auch die Integrität und Zuverlässigkeit von Datenbanken gefährden.

Dubletten – flüchtige Schatten in der Datenlandschaft, entstehen oft durch kleinste Inkonsistenzen in den Suchfeldern, seien es Tippfehler, unterschiedliche Namenskonventionen oder variierende Adressformate. Diese scheinbar harmlosen Unterschiede sind wie Fallen, die unvorsichtige Daten in die Irre führen, sie in die Fänge von Redundanz und Verwirrung treiben.

Hier kommt die Kunst der Suchfeldpräparation ins Spiel. Sie gleicht einem sorgfältigen Gärtner, der sein Feld bestellt, Unkraut jätet und sicherstellt, dass jede Pflanze – oder in diesem Fall jedes Datenstück – in reinster Form wächst und gedeiht. Durch die Standardisierung und Bereinigung von Namen, Adressen und anderen kritischen Suchfeldern verwandelt sie potenzielle Duplikate

in eindeutig identifizierbare Einheiten, bereit, ihren einzigartigen Platz in der Datenbank einzunehmen.

Die Präparation von Suchfeldern ist daher nicht nur eine Frage der Ästhetik oder Ordnung. Sie ist eine strategische Notwendigkeit, die das Fundament einer jeden Datenverarbeitungsoperation stärkt. Durch die Eliminierung von Dubletten verbessert sie nicht nur die Datenqualität, sondern optimiert auch Suchabfragen, erhöht die Effizienz von Marketingkampagnen und stärkt die Entscheidungsfindung – ein echtes Kraftwerk im Motor der Datenverarbeitung.

In der Praxis bedeutet dies, dass jeder Dateneintrag vor seiner Eingliederung in die Datenbank einer gründlichen Inspektion und Bearbeitung unterzogen wird. Diese Vorbereitung ist ein Bollwerk gegen die Flut von Fehlern und Inkonsistenzen, ein Schutzschild, das die Datenintegrität wahrt und das Risiko von Fehlinterpretationen minimiert.

Die Bedeutung der Suchfeldpräparation zur Dublettenvermeidung kann daher nicht hoch genug eingeschätzt werden. Sie ist ein wesentliches Element im Kampf gegen die Unordnung der Daten, ein Leitstern, der den Weg zu Klarheit, Effizienz und Verlässlichkeit weist. In einer Welt, die zunehmend von Daten angetrieben wird, ist die Suchfeldpräparation nicht nur eine Tugend, sondern eine unabdingbare Notwendigkeit.

1.2 Überblick über gängige Herausforderungen und Ziele

Im Herzen der digitalen Transformation, wo Daten das neue Öl sind, stehen Unternehmen vor der monumentalen Herausforderung, ihre Datenbestände nicht nur zu verwalten, sondern sie in eine Quelle nachhaltiger Wertschöpfung zu verwandeln. Der Gliederungspunkt "Überblick über gängige Herausforderungen und Ziele" wirft ein Schlaglicht auf die Sisyphusarbeit, die hinter der scheinbaren Ruhe der Datenlandschaften lauert, und skizziert die Ziele, die Unternehmen anstreben, um ihre Daten effektiver zu nutzen.

Die Herausforderungen sind vielfältig und komplex. An erster Stelle steht die Duplikaterkennung und -vermeidung, ein leidiges Problem, das nicht nur Speicherplatz verschwendet, sondern auch zu Fehlentscheidungen führen kann. Die Ursachen sind mannigfaltig: Von inkonsistenten Dateneingaben über unterschiedliche Schreibweisen bis hin zu fehlenden Standards. Jedes Duplikat ist ein Stolperstein auf dem Weg zu einer sauberen, effizienten Datenbank.

Ein weiteres großes Hindernis ist die Datenaktualität und -qualität. Veraltete Informationen können zu ineffizienten Marketingstrategien, fehlerhaften Analysen und letztendlich zu finanziellen Verlusten führen. Die kontinuierliche Pflege und Aktualisierung des Datenbestands ist daher kein Luxus, sondern eine Notwendigkeit.

Die Integration von Daten aus verschiedenen Quellen stellt Unternehmen vor die Herausforderung, eine kohärente, einheitliche

Datenbasis zu schaffen. Die Harmonisierung der Datenformate und -strukturen ist ein wesentlicher Schritt, um Silos aufzubrechen und die Zusammenarbeit sowie den Informationsfluss innerhalb des Unternehmens zu verbessern.

Das Ziel ist klar: Die Schaffung einer hochwertigen, zuverlässigen Datenbasis, die als Grundlage für geschäftliche Entscheidungen dient. Dies erfordert nicht nur technische Lösungen, sondern auch ein Umdenken in der Unternehmenskultur: Daten müssen als wertvolles Gut betrachtet werden, dessen Pflege und Entwicklung im Mittelpunkt der Unternehmensstrategie steht.

Um diese Herausforderungen zu meistern und die gesetzten Ziele zu erreichen, setzen Unternehmen auf fortschrittliche Technologien und Methoden der Datenbereinigung, wie etwa die Präparation von Suchfeldern, die Verwendung phonetischer Algorithmen zur Duplikaterkennung und die Implementierung von Datenqualitätsmanagementsystemen. Diese Werkzeuge sind unerlässlich, um die Integrität, Sicherheit und Nutzbarkeit der Daten zu gewährleisten.

Der "Überblick über gängige Herausforderungen und Ziele" verdeutlicht, dass die Datenpflege ein dynamischer, fortlaufender Prozess ist, der Flexibilität, Innovationsbereitschaft und eine klare Strategie erfordert. In diesem Sinne ist die Datenpflege weniger eine Last als vielmehr eine Chance: eine Gelegenheit, die Leistungsfähigkeit der Unternehmen im digitalen Zeitalter grundlegend zu steigern.

2. GRUNDLAGEN DER DATENPRÄPARATION

2.1 Definition und Relevanz von Personen-Suchfeldern

In der digitalen Ära, in der Daten nicht nur eine Ressource, sondern das Rückgrat von Geschäftsentscheidungen, Marketingstrategien und Kundenbeziehungen sind, nehmen Personen-Suchfelder eine zentrale Stellung ein. Diese Suchfelder sind die Navigationssterne in der endlosen Galaxie von Datensätzen, die es ermöglichen, Personen in Datenbanken präzise zu identifizieren und zu lokalisieren. Die Definition und Relevanz dieser Suchfelder sind daher von unschätzbarem Wert für jedes datengetriebene Unternehmen.

Personen-Suchfelder umfassen typischerweise Informationen wie Namen, Adressen, Telefonnummern, E-Mail-Adressen und Geburtsdaten. Sie sind die Schlüsselattribute, die es ermöglichen, individuelle Datensätze in einer Vielzahl von Anwendungen und Kontexten eindeutig zu identifizieren. Die korrekte und effiziente Handhabung dieser Suchfelder ist entscheidend für die Qualität der Datenpflege, die Vermeidung von Dubletten und letztlich für die Zuverlässigkeit der gesamten Datenbank.

Die Relevanz von Personen-Suchfeldern erstreckt sich über diverse Bereiche: Vom Customer Relationship Management (CRM) über gezielte Marketingkampagnen bis hin zur Compliance mit Datenschutzvorschriften. In CRM-Systemen ermöglichen sie eine personalisierte Kundenansprache, indem sie eine 360-Grad-Sicht

auf den Kunden liefern. Im Marketing ermöglichen sie die Segmentierung von Zielgruppen und die Durchführung präziser, maßgeschneiderter Kampagnen. Im Hinblick auf den Datenschutz gewährleisten sie, dass personenbezogene Daten korrekt gehandhabt und geschützt werden.

Die Herausforderung liegt in der Komplexität und Dynamik der Daten. Menschen ziehen um, ändern ihre Namen oder Kontaktinformationen, was die Aktualität und Genauigkeit der Datenbanken kontinuierlich bedroht. Hinzu kommt die Vielfalt der Datenquellen und -formate, die eine konsistente Datenpflege erschweren. Die Präparation von Personen-Suchfeldern – die Standardisierung und Bereinigung der Daten – ist daher ein kritischer Schritt, um diese Herausforderungen zu meistern.

Die Investition in die sorgfältige Verwaltung von Personen-Suchfeldern zahlt sich aus. Sie minimiert das Risiko von Fehlern, verbessert die Kundeninteraktion und -zufriedenheit und steigert die Effizienz von Geschäftsprozessen. In einer Zeit, in der der Wettbewerb zunehmend über die Qualität der Daten und die Fähigkeit, diese intelligent zu nutzen, entschieden wird, sind Personen-Suchfelder mehr als nur Datenpunkte; sie sind strategische Vermögenswerte, die gepflegt und geschützt werden müssen.

2.2 Typische Struktur und Inhalte von Personen-Suchfeldern

Die typische Struktur und Inhalte von Personen-Suchfeldern bilden das Fundament für effiziente Datenbankoperationen und