Produktbericht: Dreifachboost für Datenkonsistenz und Anwenderfreundlichkeit (Teil 3 von 3)

Duplikaten-Erkennung mit Deep Qualicision KI

Bereits in den letzten beiden Ausgaben des Production manager wurden zwei Module für den Dreifachboost der Datenkonsistenz und Anwenderfreundlichkeit auf Basis des Deep Qualicision KI Framework thematisiert: Die Auto-Vervollständigung bei der Datenerfassung sowie die Eingabe-Validierung bei der Datenspeicherung. Beide führen – schon jeweils für sich und in Kombination verstärkend – zu einer messbaren Verbesserung der Datenkonsistenz sowie der Anwenderfreundlichkeit. Der Einsatz dieser Kombination bringt jedoch so lediglich einen Mehrwert für diejenigen Datensätze, die neu im Prozess erfasst werden. In einer bereits seit vielen Jahren bestehenden Datenbasis können somit noch existierende Duplikate einer Gesamtkonsistenz entgegenwirken. An dieser Stelle lässt sich mit der auf dem Deep Qualicision KI Framework basierenden Duplikaten-Erkennung ansetzen. Hierfür werden die aus historisierten Daten sowie die während der Dateneingabe und -kontrolle gelernte Syntax und Semantik der Datensätze unmittelbar für die Suche nach Duplikaten in bereits vorhandenen Datenbanken genutzt.

In nahezu jedem Geschäftsprozess sind Daten heutzutage die Basis für ein effizientes und effektives Handeln. Eine große Herausforderung sowohl an einen Bearbeiter als auch einen Administrator solcher Datenbasen ist das Aufrechthalten einer kontinuierlich hohen Datenqualität. Bei einer teilweise unüberwachten Datenerfassung – etwa ohne den Einsatz einer Auto-Vervollständigung oder einer automatisierten Dateneingabe-Validierung – entstehen im Zeitverlauf zunehmend Inkonsistenzen, die mitunter zu Störungen im Prozess selbst und dessen Nachfolgern führen können. Die Folgen sind häufig manuelle Nachbesserungen oder sogar Fehlplanungen.

Duplikaten-Erkennung mit Deep Qualicision KI.

Ein Kunden-Use-Case: Adressverwaltung von Lieferanten

Über viele Jahre wurden Adressen von Lieferanten, die weltweit operieren, in einer Datenbank gesammelt. Dabei sind die Eingaben stets manuell und durch viele verschiedene Bearbeiter erfolgt. Bei Adressen, die vermeintlich nicht gefunden wurden, erfolgte eine Neuanlage. Im Laufe der Zeit sind so durch verschiedene Schreibweisen Duplikate ein und derselben Lieferanten entstanden.

Ein Beispiel ist ein Lieferant in Italien. Hier kann der Straßenname auf viele verschiedene Weisen eingegeben werden: In der Landessprache als „Via delle Fabbriche“ oder als deutsche Übersetzung mit den Varianten „Fabrikstr.“, „Fabrikstrasse“ oder „Fabrikstraße“. Zudem kann der Firmenname ebenfalls in der Landessprache oder als deutsche Übersetzung eingetragen werden. Allein so ergeben sich acht Möglichkeiten für denselben Sachverhalt. Zudem können noch Varianten hiervon mit Groß- und Kleinbuchstaben entstehen. Die Konsistenz in der Adressverwaltung nimmt so stetig ab und verschlechtert damit auch die Anwenderfreundlichkeit sowie den Prozess selbst.

Suche nach Datenduplikaten auf Basis von Ähnlichkeitsmetriken

Da bei einer jahrelang bestehenden und stetig gewachsenen Datenbasis eine manuelle Suche nach Duplikaten zur Erhaltung der Konsistenz aufgrund des nicht zu bewältigenden Zeitaufwands außerfrage steht, ist ein erster Ansatz die Verwendung von Ähnlichkeitsmetriken. Hierbei werden die Inhalte von Datensätzen als Textobjekte mit einer Folge von Buchstaben interpretiert und anschließend Distanzen untereinander berechnet. Überschreitet diese Abweichung ein vorgegebenes Maß nicht, werden die beiden geprüften Objekte als Duplikate behandelt. Dies repräsentiert jedoch einen Methodenansatz mit der Suche nach fest definierten Anomalien, da es sich im Kern um

eine Schwellenwertprüfung für einen Ähnlichkeitsvergleich handelt, der zudem von der Wortlänge abhängig ist. Darüber hinaus weisen solche Verfahren ein schlechtes Laufzeitverhalten bei großen Datenmengen auf, was die Anwendbarkeit im Umfeld von Big Data einschränkt. Zusätzlich verhalten sich Ähnlichkeitsmetriken bei sich im Zeitverlauf ändernden Prozessen zum Teil instabil bzgl. der Semantik. Es bedarf also vielmehr eines Mechanismus, der Anomalien in den Strukturen eines Datensatzvergleichs selbsttätig erkennt, und sich jederzeit an aktuelle Rahmenbedingungen anpassen kann.

Duplikate Daten-basiert erkennen mittels Qualitativem Labeln kombiniert mit maschinellem Lernen

In den meisten Geschäftsprozessen existiert ohnehin bereits eine breite Basis historisierter Daten. Durch Qualitatives Labeln vereint mit maschinellem Lernen basierend auf dem Deep Qualicision KI Framework lassen sich aus Daten der Vergangenheit prozessspezifisch die Strukturen einer Datenbasis erlernen. Insbesondere zur Erkennung mehrstufiger Zusammenhänge und komplexer Ähnlichkeiten in Daten – wie beispielsweise das Auffinden eines Lieferanten, der mit mehreren Einträgen in der Adressverwaltung geführt wird – bieten datengetriebene Methoden mannigfaltige Vorteile.

KPI-basierte selbstlernende Duplikaten-Erkennung als Bestandteil eines Deep Qualicision KI-Gesamtsystems

Grundlage einer Duplikaten-Erkennung auf Basis des Deep Qualicision KI Framework ist die Vereinigung von Qualitativem Labeln mit einer mittels maschinellem Lernen trainierten Wissensbasis aus historisierten Daten. Darüber hinaus kommen Ähnlichkeitsmetriken zum Einsatz, um die Vergleiche zwischen Textobjekten zu realisieren. Allerdings wird mit dem Framework zusätzlich eine Entscheidungsunterstützung durch einfaches Präferieren verschiedener Bewertungs-KPIs ermöglicht. Auf diese Weise können nicht nur syntaktische Ähnlichkeiten, sondern auch semantische Analogien – wie bei unterschiedlichen Schreibweisen von Straßen- oder Firmennamen – für das Auffinden von Duplikaten einbezogen werden. Ein solcher auf KPI-Basis selbstlernender Prüfmechanismus kann so einen Automatismus zur kontinuierlichen Erkennung von Datenduplikaten bereitstellen, der auf einer Daten-Historie aufsetzt und eine im Prozess stetig anwachsende Wissensbasis umfasst. Für den Prozess selbst und seine Nachfolger ist als Konsequenz damit sichergestellt, dass die Planung mit konsistenten Daten vollzogen werden kann, um manuelle Nacharbeit zu reduzieren und Fehler zu vermeiden.

Deep Qualicision-basierte Duplikaten-Erkennung als Erweiterung der Auto-Vervollständigung und Dateneingabe-Validierung

Ein bereits im Betrieb befindliches System mit Auto-Vervollständigung und Dateneingabe-Validierung ist durch Nutzung des gemeinsamen Deep Qualicision KI Framework modular um die Duplikaten-Erkennung erweiterbar. So lässt sich eine weitere messbare Steigerung für die Anwenderfreundlichkeit sowie die Datenkonsistenz erreichen.

Nutzen der Duplikaten-Erkennung

+ Erkennung von Duplikaten als Anomalien in der gesamten Datenbank
+ Automatisierte Erkennung von duplizierten Datensätzen
+ Signifikante Zeitersparnis und Planungssicherheit in nachgelagerten Prozessen
+ Konsistenz der gesamten Datenbasis
+ Qualitative Standardisierungs- und Plausibilitätsanalysen
+ Permanentes Nachlernen der Wissensbasis zur Erhaltung eines aktuellen Datenstands

Ein KI-Gesamtsystem mit Dreifachboost für Datenkonsistenz und Anwenderfreundlichkeit

Durch das modulare Verknüpfen der Bausteine Auto-Vervollständigung, Dateneingabe-Validierung und Duplikaten-Erkennung – die jeweils für sich auch einzeln betrieben werden können – entsteht eine sich stetig maschinell selbstlernend erweiternde Wissensbasis zur automatisierten Unterstützung bei der Datenerfassung, -überprüfung und -haltung. Dies liefert in Summe den Dreifachboost für Datenkonsistenz und Anwenderfreundlichkeit auf Basis des Deep Qualicision KI Framework.

PSI FLS
Fuzzy Logik & Neuro Systeme GmbH
Dr. Jonas Ostmeyer
Consultant Supply Chain Optimization
ostmeyer@psi.de
www.deepqualicision.ai