Matching von Produktdaten

Wie Sie mit Hilfe lernender Match-Algorithmen eine saubere Datenbasis schaffen.

Einleitung

Ob Online-Shop oder Markenhersteller - Unternehmen zahlreicher Branchen stehen vor der Herausforderung, eine steigende Anzahl von Produktdaten zu managen. Ein Problem sind mehrfach vorhandene Produktdaten („Dubletten“, „Duplikate“). Wird beispielsweise ein Produkt von unterschiedlichen Lieferanten bezogen, kann das Produkt durch verschiedene Titel und Beschreibungen in der Produktdatenbank mehrfach vorhanden sein. Im ungünstigsten Fall wird dem Kunden und dem Produktmanager angezeigt, dass das Produkt gar nicht verfügbar ist. Der Kunde kauft in einem anderen Online-Shop und der Produktmanager löst eine Nachbestellung aus, obwohl noch genügend Produkte am Lager sind.

Den Prozess der automatisierten Identifizierung ähnlicher oder äquivalenter Datensätze nennt man Matching. Dabei bestimmen die Domäne und das Anwendungsszenario, wann zwei Datensätze als ähnlich oder äquivalent und damit als Match gelten.

Wird das Matching mit dem Ziel eingesetzt, Duplikate in den Produktstammdaten eines Online-Shops zu identifizieren, so sind zwei Datensätze ein Match, wenn sie dasselbe Produkt repräsentieren. Im Gegensatz dazu sind beim Einsatz des Matchings zur Konkurrenzbeobachtung für Markenhersteller zwei Datensätze ein Match, wenn sie sich auf vergleichbare Produkte beziehen.

Die Vorteile eines Matchings ergeben sich aus den jeweiligen Anwendungsfällen:

_Verbesserung der Datenqualität: Matching hilft, Duplikate zu identifizieren und zu entfernen und schafft damit eine saubere Datenbasis. In Online-Shops erhöht dies beispielsweise die Usability und sorgt für zufriedene Kunden, die hier gerne (wieder) einkaufen.

_Automatische Preisbeobachtung: Matching ermöglicht den Vergleich der eigenen Preise mit den Angeboten von Konkurrenten und erlaubt dadurch fundierte Analysen der Preisstrategien der Wettbewerber.

Dieses Whitepaper soll Ihnen einen Überblick über die neuen Möglichkeiten für das Matching von Produktdaten vermitteln. Sie erfahren, wie Sie mit neuen automatisierten Methoden eine saubere und zuverlässige Datenbasis schaffen.

Status quo

Auch wenn die meisten global agierenden Unternehmen die hohe Bedeutung einer sauberen und zuverlässigen Datenbasis für die Effizienz ihrer Geschäftsprozesse längst erkannt haben, haben nur wenige daraus bereits die erforderlichen Konsequenzen gezogen. Das ist das Kernergebnis der Studie „Strategisches Management von Stammdatenqualität", für das die Strategie- und Organisationsberatung Camelot Management Consultants 56 Entscheider aus global agierenden Unternehmen aller Branchen und Größen zu ihrem Ansatz zur Qualitätssicherung ihrer Unternehmensdaten befragt hat. Mehr als die Hälfte der befragten Unternehmen gibt an, dass unzureichende Stammdatenqualität sich nach wie vor massiv negativ auf die Prozesse entlang der gesamten Wertschöpfungskette auswirkt. Rund 60% der Befragten sehen enormen Nachholbedarf bei der Messbarkeit und Kontrolle von Datenqualität sowie beim Einsatz von automatisierten Tools.[1]

Eine unabhängige Anwenderbefragung von TDWI Europe kam zu einem ähnlichen Ergebnis. Nur 36% der Befragten waren mit der Datenqualität in ihrem Unternehmen zufrieden. Positive Auswirkungen bei einer Steigerung der Datenqualität sehen die meisten Befragten vor allem in den Bereichen Reporting, Kundenbindung und Prozessoptimierung. Besonders interessant ist, dass einige Teilnehmer davon ausgehen, dass die jährlichen Kosten mangelnder Datenqualität in ihrem Unternehmen über 200 TEUR jährlich betragen.[2]

Bislang werden häufig eigens Mitarbeiter eingesetzt, die manuell Produktdaten bereinigen oder die Preise von Wettbewerbern erfassen. Dies ist bei geringen Produktdatenmengen relativ unproblematisch möglich. Mit wachsender Datenmenge steigt allerdings der Aufwand überproportional an.

Die überwiegende Zahl der existierenden automatisierten Matching-Tools bieten bisher keine spezialisierten Lösungen für das Matching von Produktdaten. Die spezifischen Herausforderungen, die für Produktdaten zu meistern sind, werden im Folgenden näher erläutert.

Herausforderungen

Die größten Herausforderungen beim Matching von Produktdaten liegen im Umgang mit sehr heterogenen Produktbeschreibungen sowie fehlenden einheitlichen Nummern zur eindeutigen Produktidentifizierung (wie zum Beispiel EAN/GTIN). Uneinheitliche Produktbeschreibungen entstehen unter anderem durch die Verwendung von:

_domänenspezifischen Bezeichnungen und Abkürzungen: Im Bereich Fashion ist zum Beispiel „lg. A." eine Abkürzung für „lang Arm". Im Bereich Elektronik ist zudem LG ein Markenhersteller.

_heterogenen Bezeichnungen für Größen und Stückangaben: 168 Stück vs. 3x56 Stück.

_heterogenen Schreibweisen von Modellbezeichnungen: DHI655FX vs. DHI 655 FX vs. DHI-655 FX.

_Synonymen: Im Bereich Fashion bezeichnen zum Beispiel Hoody, Hoodie und Kapuzenpullover einen Pullover mit Kapuze.

Eine generelle Herausforderung für das Matching ist die Effektivität. Um diese zu bewerten, werden üblicherweise zwei Maße verwendet: Precision und Recall.

[1] https://www.hs-heilbronn.de/6691397/Trendstudie_SDQ_2013_Management_Summary.pdf

[2] http://www.emagixx.de/images/Unternehmen/Studie_Datenqualitaet_in_Unternehmen.pdf

Die **Precision** misst den Anteil der tatsächlichen Matches an den erkannten Matches, ist also ein Maß für die Treffgenauigkeit. Eine hohe Precision besagt, dass erkannte Matches unbesorgt als solche angenommen werden können.

Der **Recall** hingegen misst den Anteil der erkannten Matches an allen tatsächlichen Matches, ist also ein Maß für die Vollständigkeit des Matchings. Ein hoher Recall bezeugt also, dass viele der tatsächlich vorhandenen Matches gefunden wurden.

Je nach Anwendungsfall und Qualität der Ausgangsdaten ist es schwierig, sowohl die Precision als auch den Recall zu maximieren. Es ist deshalb in der Regel erforderlich, beide Ziele anwendungsspezifisch zu gewichten. Liegt der Fokus auf der Genauigkeit (Precision), muss in Kauf genommen werden, dass möglicherweise einige korrekte Matches nicht erkannt werden. Liegt der Fokus dagegen auf der Vollständigkeit (Recall) enthält das Ergebnis unter Umständen einige falsche Zuordnungen.

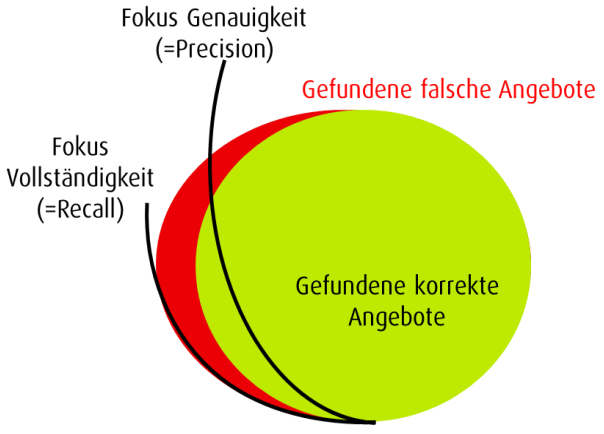

Abbildung 1: Vollständigkeit versus Genauigkeit

Weitere Herausforderungen für das Matching von Produktdaten ergeben sich je nach Anwendungsfall aus den folgenden Faktoren:

_Datenmenge: Je nach Situation muss mit einer riesigen Datenmenge umgegangen werden, deren Umfang die Laufzeit des Matchings direkt beeinflusst.

_Frequenz: Je häufiger das Matching einer (großen) Datenbasis wiederholt werden muss, desto umfangreicher gestaltet sich der Prozess.

_Änderung der Datenbasis: Je häufiger sich die Datenbasis, für die ein Matching erfolgen soll, ändert, desto häufiger müssen neue (umfangreiche) Matchingprozesse erfolgen.

Nutzen

Unternehmen, die automatisierte Softwarelösungen für das Matching von Produktdaten einsetzen, profitieren in mehrfacher Hinsicht. So entfällt der interne Aufwand für eine manuelle Überprüfung und Zuordnung der Daten. Besonders in Bereichen mit hoher Dynamik kann der Aufwand für eine Anpassung an sich ändernde Bedingungen drastisch reduziert werden, sodass weniger interne Ressourcen gebunden werden.

Darüber hinaus ergibt sich je nach Anwendungsfall des Matchings ein zusätzlicher individueller Nutzen.

Beispiele:

Nach der Optimierung der Produktdaten finden Besucher eines Online-Shops die gewünschten Produktinformationen bedeutend leichter, weil die Daten klar strukturiert vorliegen und damit auch die Suche bessere Ergebnisse liefert. Gut gepflegte und einheitliche Produktinformationen schaffen mehr Vertrauen beim Kunden. Der Nutzen für den Online-Shop zeigt sich in einer höheren Kundenzufriedenheit und steigenden Verkaufszahlen.

Außerdem sorgen einheitliche Produktdaten für mehr Transparenz beim Einkauf. Einkäufer erhalten einen schnellen Überblick über vorhandene Lagerbestände und können die Einkaufspreise verschiedener Lieferanten trotz unterschiedlicher Bezeichnungen einfach vergleichen. Dies schafft Potenzial für höhere Margen und vermeidet Doppelbestände im Lager.

Neben dem internen Matching der eigenen Produkte verschiedener Lieferanten haben Shopbetreiber die Möglichkeit, das eigene Sortiment direkt mit den Angeboten verschiedener Konkurrenten zu vergleichen. Dabei ist insbesondere die Beobachtung des Preises für übereinstimmende Produkte im Sortiment der Wettbewerber von Interesse. Mit Hilfe genauer und verlässlicher Marktdaten können Händler so die eigene Preissetzungsstrategie auf die Wettbewerbssituation anpassen und ihre Konkurrenzfähigkeit steigern. Dieses Verfahren ist im Whitepaper „Preisbeobachtung im Internet - wie Sie mithilfe strukturierter Wettbewerbsdaten Ihre Marge optimieren können“ näher beschrieben.

Matching Process

Der Matching Prozess kann grob in die vier Phasen Data Extraction, Preprocessing, Matching und Reporting unterteilt werden. Im Folgenden werden die vier Phasen exemplarisch anhand der Software blackbee erläutert.

1. Data Extraction

Zunächst werden die Daten zur Verfügung gestellt, für die ein Matching durchgeführt werden soll. Die zu matchenden Daten können in unterschiedlichen Quellsystemen vorliegen, beispielsweise in verschiedenen Datenbanken oder in Textdateien unterschiedlicher Formatierung. blackbee unterstützt verschiedene Quellsysteme und Dateiformate. Darüber hinaus sind die Einbindung von Produktfeeds und die Abfrage von diversen Webquellen wie Webservices, Preisvergleichsportalen und Online-Shops möglich.

2. Preprocessing

Typischerweise werden die Daten zunächst vorverarbeitet, um die Durchführung des eigentlichen Matchings zu erleichtern. Zur Vorverarbeitung zählen unter anderem die Standardisierung von Attributwerten, die Ergänzung fehlender Attributwerte sowie die Ergänzung zusätzlicher Attribute.

Die Standardisierung bringt Attributwerte in ein einheitliches Format, um heterogene Herstellerbezeichnungen wie „HP“ und „Hewlett-Packard“ vergleichbar zu machen. Die Ergänzung fehlender Attributwerte kann mit Hilfe einer Referenzliste erfolgen. Eine solche Liste wird beispielsweise für das Attribut „Marke“ aus vorhandenen Angaben aufgebaut, sodass die Marke für Produkte mit dem fehlenden Attribut „Marke“ aus dem Titel oder der Beschreibung ergänzt werden kann. Durch die Ergänzung um zusätzliche Attribute können Produktdaten mit klar zu identifizierenden Merkmalen angereichert werden. So kann der Produktcode aus dem Titel oder der Beschreibung eines Produktes als zusätzliches Attribut extrahiert werden. Ein Produktcode ist eine vom Hersteller gewählte eindeutige Zeichenfolge zum Zweck der Produktidentifikation. Für ein Produkt mit dem Titel „Bosch DHI 655FX grau-metallic Dunstabzugshaube Einbau“ lautet der Produktcode „DHI 655FX“.

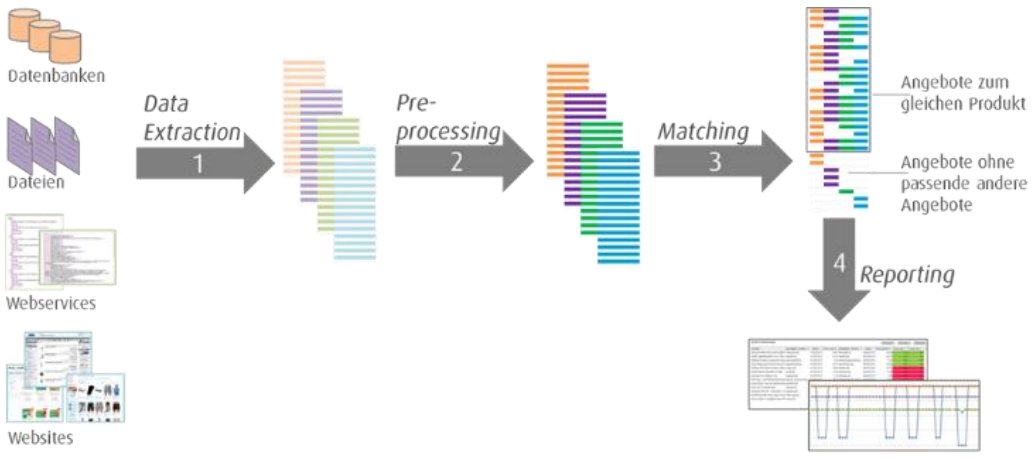

Abbildung 2: Matching Process blackbee

3. Matching

In diesem Schritt werden die einzelnen Datensätze miteinander verglichen. Für große Datenmengen ist es erforderlich, zuvor den Suchraum einzugrenzen, um nicht jeden Datensatz mit jedem anderen vergleichen zu müssen. Dazu kommen sogenannte Blockingverfahren zum Einsatz, wobei die Produktdaten in Blöcke aufgeteilt werden. Das kann auf unterschiedliche Weise erfolgen. Eine Möglichkeit ist die Generierung oder Nutzung eines Blockschlüssels. Alle Produktdaten, die denselben Blockschlüssel (zum Beispiel die Produktmarke) haben, werden in einem Block zusammengefasst. Es werden dann nur noch Produktdaten im selben Block (also Produkte derselben Marke) miteinander verglichen.

Der Vergleich von zwei Produktdaten geschieht im Allgemeinen anhand des Vergleiches ihrer Attribute. Um Attribute zu vergleichen, existieren eine Vielzahl unterschiedlicher Ähnlichkeitsmaße. Gebräuchliche Ähnlichkeitsmaße sind EditDistance, Jaccard, TFIDF und Trigram. Für die Qualität des Matchings ist es entscheidend, passende Ähnlichkeitsmaße auszuwählen. In den meisten Fällen führt ein einzelnes Maß nicht zu einem optimalen Ergebnis. Erfolgversprechender ist es, unterschiedliche Maße für verschiedene Attributwerte (zum Beispiel für die Attribute „Artikelbezeichnung“ und „Marke“) zu kombinieren. Die Bestimmung einer effektiven Matchingstrategie ist aufgrund der Vielzahl existierender Verfahren oft selbst für einen Experten eine Herausforderung. Hier bieten lernende Verfahren, wie blackbee sie anbietet, entscheidende Vorteile. Maschinelle Lernverfahren reduzieren den erforderlichen manuellen Tuning-Aufwand, indem sie eine optimierte Matchingstrategie semi-automatisch bestimmen („lernen“). Dazu benötigen sie positive und negative Trainingsdaten in Form von Beispielen für Matches und Nicht-Matches. blackbee unterstützt die Generierung und Verwaltung solcher Trainingsdaten und den Aufbau eines Feedbackmechanismus. Durch die Berücksichtigung von Feedback können Zuordnungsfehler korrigiert werden. Das System lernt aus diesen Korrekturen und erhöht somit von Durchlauf zu Durchlauf seine Treffgenauigkeit.

4. Reporting

Das Ergebnis des Matching ist eine Menge von Korrespondenzen. Eine Korrespondenz ist dabei ein Paar aus zwei Produktdaten, die vom System als Matches identifiziert wurden, und einem Ähnlichkeitswert, der angibt, mit welcher Wahrscheinlichkeit es sich um einen tatsächlichen Match handelt. Je nach Anwendungsfall kann dieses Ergebnis für weitergehende Analysen aufbereitet und für verschiedenste Reports genutzt werden. Für Preisbeobachtungen lässt sich mit der Reportfunktion beispielsweise eine Übersicht über die Top 5 Anbieter für jedes Produkt generieren. blackbee unterstützt den Export der Ergebnisse und generierten Reports in unterschiedliche Formate (zum Beispiel Excel, CSV).

Zusammenfassung

Unternehmen sehen im Bereich Datenqualität einen großen Nachholbedarf. Während die Mehrzahl der Anbieter von Datenqualitätslösungen keine Matching-Funktionalitäten anbietet oder auf Kundendaten spezialisiert ist, wird mit blackbee ein Service zum Matching und zur Bereinigung von Produktdaten angeboten. Der Einsatz lernender Matching-Algorithmen zur Qualitätsverbesserung von Produktdaten bietet Unternehmen einen strategischen Vorteil im Umgang mit großen Datenmengen. Dies trifft in besonderem Maße auf Unternehmen zu, deren Daten sich häufig ändern. Mit der Verwendung der Softwarelösung blackbee sind Unternehmen in der Lage, große Mengen von Produktdaten in kurzer Zeit und wiederholt um Duplikate zu bereinigen oder auch Daten zu einem Preisvergleich von Angeboten unterschiedlicher Anbieter zusammen zu führen. Die Reduzierung von manuellem Aufwand und die Vermeidung von Fehlentscheidungen aufgrund fehlerhafter Daten sind zwei entscheidende Vorteile, die sich aus dem Einsatz der Software blackbee ergeben.

Über Webdata Solutions GmbH

Der E-Commerce-Dienstleister ist im Jahr 2012 als Ausgründung aus einem Forschungsprojekt der Universität Leipzig entstanden und gehört heute zu den weltweiten Marktführern im Bereich **Online-Marktanalyse**. Die Lösungen, die auf der innovativen Plattformtechnologie **blackbee** basieren, werden von führenden online-Händlern und Herstellern erfolgreich in der Praxis eingesetzt. Webdata Solutions reduziert die Komplexität, die aus einer Vielzahl an produkt- und produktbezogenen Daten im Internet entsteht und generiert auf den **geschäftsnutzen fokussierte Informationen**. Das Unternehmen hat es sich zum Ziel gesetzt, mit blackbee das Potenzial von Webdaten umfassend zu heben und mit den **richtigen Kerninformationen zur richtigen Zeit** dazu beizutragen, dass E-Commerce zu einem **transparenten Markt** wird.

Hanna Köpcke
CTO

Kontakt

Telefon +49 341 351 361 70

E-Mail info@webdata-solutions.com

Internet www.webdata-solutions.com

Autoren: Hanna Köpcke
Carina Röllig

Jacobstraße 5
04105 Leipzig