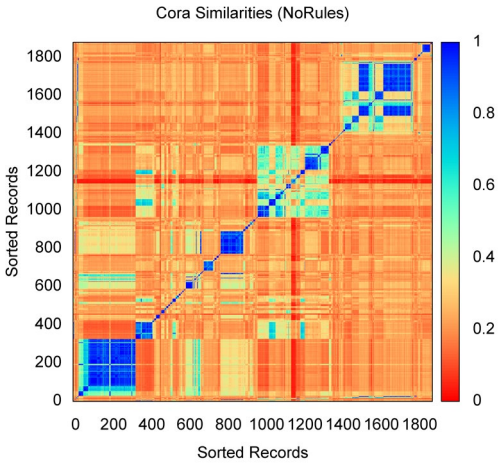
Data Matching Benchmark

Bachelorprojekt für 2020/2021
Fachgebiet Informationssysteme
Prof. Dr. Felix Naumann

Data Matching Verfahren, auch bekannt als **Duplikaterkennung**, Record Linkage oder Entity Resolution, identifizieren mehrfache Datensätze, die das gleiche Realweltobjekt repräsentieren, z.B. Produkte in einem Katalog oder Personen in Kundenlisten. Duplikaterkennung ist ein altes, viel behandeltes, aber noch immer nicht ausreichend gelöstes Problem der Informationsintegration. Data Matching Verfahren haben zwei wesentliche Zutaten: Erstens ein Maß, das die Ähnlichkeit zweier Datensätze bestimmt. Zweitens ein Algorithmus, der Datensätze zum Vergleich auswählt, aber den paarweisen Vergleich aller Datensätze miteinander vermeidet (die Abbildung zeigt übrigens solchen einen paarweisen Vergleich, gefärbt nach Ähnlichkeit). Nur so kann man den inakzeptabel hohen quadratischen Aufwand umgehen. Viele modernere Verfahren basieren auf **maschinellem Lernen**, um zwischen Duplikaten und Nichtduplikaten zu unterscheiden.

Aufgrund der hohen Relevanz des Problems existieren in Forschung und Industrie viele verschiedene Verfahren und Ansätze mit unterschiedlichsten Eigenschaften. Unser **Industriepartner SAP** zum Beispiel bietet verschiedene Lösungen im Bereich **Master Data Management** an. Diese wiederum verwenden jeweils unterschiedliche Technologien und Algorithmen für das Matching, welche über SAP HANA Fuzzy Search, SAP Data Services Match Transform und SAP HANA Data Hub Matching zur Verfügung gestellt werden.

Data Matching Verfahren können auf vielfältige Weise bewertet werden. Neben der **Laufzeit** sind insbesondere **Precision** (Korrektheit) und **Recall** (Vollständigkeit) relevante Maße für die Fähigkeit der Verfahren, alle und nur alle tatsächlichen Duplikate zu finden. Die faire Berechnung dieser Maße ist jedoch nicht leicht, angesichts unvollständiger Goldstandards, verschiedener Konfgurationsparameter der Verfahren, nicht-binäre Ergebnismengen, usw. Aus Sicht der Industrie ist zudem auch der Konfigurations- und Wartungsaufwand, die „total cost of ownership“ relevant: Ähnlichkeitsmaße müssen entwickelt und gewichtet werden, **Trainingsdaten** müssen erzeugt werden, Daten müssen vor einem Duplikaterkennungslauf aufbereitet werden, usw.



Das Ziel dieses Bachelorprojektes ist die Entwicklung eines **Benchmarks** mittels dessen verschiedene Verfahren in verschiedenen Dimensionen leicht miteinander verglichen werden können. So kann unser Projektpartner SAP (i) neue Verfahren aus der Forschung und Entwicklung evaluieren, (ii) einschätzen wie gut sich aktuelle SAP Matchingsysteme für neue Anwendungsgebiete eignen und nicht zuletzt (iii) aktuelle Verfahren optimal konfigurieren.

Projektbeschreibung

Zu einem Benchmark gehören traditionell drei Komponenten: Eine Datenmenge, eine Workload, sowie ein Erfolgsmaß. Um einen Benchmark erfolgreich zu verbreiten, muss zudem seine Anwendung leicht sein. Wir wollen gemeinsam, basierend auf wissenschaftlichen Grundlagen und viel Praxiserfahrung, die folgenden Aufgaben angehen:

- **Datensammlung**: Wir wollen öffentliche und interne Goldstandard-Daten einsammeln und standardisiert speichern. Viele dieser Datenquellen sind bereits mit bekannten Treffern annotiert. Oft ist diese Annotation jedoch nicht vollständig und nicht konsistent – das wollen wir automatisiert untersuchen und bei der Bewertung von Matching-Verfahren berücksichtigen.

- **Goldstandard/Silberstandard**: Wir wollen *nicht* selbst mühsam Daten als Duplikate bzw. Nicht-Duplikate annotieren, sondern vielmehr Methoden entwickeln, die es erlauben, trotz unvollständiger Goldstandards, also sogenannter Silberstandards, Duplikatserkennungsansätze zu bewerten und zu vergleichen.

- **Entwicklung von Erfolgsmaßen**: Basierend auf wissenschaftlicher Literatur wollen wir eine Auswahl an Erfolgsmaßen implementieren und insbesondere so anpassen, dass sie auch bei unvollständigem Goldstandard aussagekräftig bleiben.

- **Vergleich zwischen Verfahren**: In der Regel genügt es nicht, nur die Kennzahlen zweier Verfahren miteinander zu vergleichen – auch deren tatsächliche Ergebnisse sollen verglichen werden: Welche Duplikate findet das eine Verfahren, das andere jedoch nicht? Welcher manuelle Aufwand ist nötig, um die Ergebnisse zu erzielen? Wir wollen eine intuitive Darstellung der Ergebnisse auf beiden Ebenen entwickeln.

- **Einsatz**: Der Benchmark soll beispielhaft an mehreren SAP Produkten, SAP-internen Verfahren, sowie weiteren öffentlich verfügbaren Systeme ausgeführt werden.

Die Entwicklung eines Data Matching Benchmark verbindet grundlegende **Forschungskonzepte** mit höchstrelevanter **Praxis**. Ein schlagkräftiges Projektteam kann diese Verbindung entscheidend beeinflussen und erlernt dabei zugleich grundlegende Kenntnisse und Fähigkeiten der Datenaufbereitung, Datenreinigung und Informationsintegration. Vorkenntnisse im Bereich Datenbanksysteme (DBS I) und SQL sind erforderlich.

Projektpartner

SAP nimmt bei der digitalen Transformation eine zentrale Rolle ein. Als Marktführer bei Unternehmenssoftware hilft SAP Unternehmen und Organisationen dabei, die steigende Komplexität zu minimieren, neue Möglichkeiten für Innovation und Wachstum zu schaffen und im Wettbewerb erfolgreich zu sein. SAP Master Data Governance (MDG) ist die Master Data Management Lösung von SAP. Anwendungsgebiete sind neben der zentralen Stammdatenpflege die Konsolidierung und das Qualitätsmanagement.

Das Projekt beginnt im Oktober 2020 und wird durch Prof. Dr. Felix Naumann, sein Team und Partner bei SAP betreut. Fragen können gerne an felix.naumann@hpi.de gerichtet werden.