DS4DM

Schlussbericht der RapidMiner GmbH und der Universität Mannheim über das Verbundprojekt DS4DM

Datensuche für Data-Mining-Prozesse

Bericht vom 24.01.2019

Zuwendungsempfänger:	RapidMiner GmbH & Universität Mannheim
Vorhabensbezeichnung:	DS4DM – Datensuche für Data-Mining-Prozesse
Förderkennzeichen:	01IS15027A & 01IS15027A-B
Laufzeit des Vorhabens:	01.08.2015 - 31.07.2018
Berichtszeitraum:	01.08.2015 - 31.07.2018

Projektpartner

1. RapidMiner GmbH (Projektkoordination)
2. Universität Mannheim

Ansprechpartner bei der RapidMiner GmbH

- Ralf Klinkenberg (Projektleitung, Projektdurchführung, Administratives):
  E-Mail: rklinkenberg@rapidminer.com , research@rapidminer.com
- Dr. Edwin Yaqub (Projektdurchführung)
- David Arnu (Projektdurchführung)
- Dr. Fabian Temme (Projektdurchführung)
- Philipp Schlunder (Projektdurchführung)

Ansprechpartner bei der Universität Mannheim

- Prof. Dr. Christian Bizer (Projektleitung, Projektdurchführung)
  E-Mail: chris@informatik.uni-mannheim.de
- Prof. Dr. Heiko Paulheim (Projektleitung, Projektdurchführung)
  E-Mail: heiko@informatik.uni-mannheim.de
- Benedikt Kleppmann (Projektdurchführung)

Inhaltsverzeichnis des DS4DM-Schlussberichts

Motivation

Das DS4DM-Projekt (Datensuche für Data-Mining-Prozesse („Data Search for Data Mining“)) lief vom 01.08.2015 bis zum 31.07.2018. Während dieses Zeitraums wurden alle 6 Monate Zwischenberichte verfasst. In jedem Zwischenbericht berichteten die beiden Projektpartner RapidMiner GmbH und Universität Mannheim detailliert über die aktuellen Fortschritte. Die Berichte zum Ende der Projektjahre gaben zusätzlich einen jährlichen Überblick über die geleisteten Arbeiten. Zusätzlich verfasste RapidMiner einen Koordinatorbericht am Ende jedes Projektjahres. In dem vorliegenden Abschlussbericht erfolgt eine Gesamtübersicht über die geleisteten Arbeiten von RapidMiner. Die Arbeiten des Projektpartners Universität Mannheim sind in einen eigenen Abschlussbericht zusammengefasst.

Ziel des DS4DM-Projektes war die Erforschung und Entwicklung von Methoden für die Datensuche und -integration. Die Menge der verfügbaren Daten innerhalb von Unternehmen sowie im öffentlichen Internet (World Wide Web (WWW)) ist in den letzten Jahren sprunghaft gestiegen. Aufgrund der sehr großen Anzahl an Datenquellen sehen sich Analysten zunehmend mit der für sie sehr unbefriedigenden Situation konfrontiert, dass die Daten, die sie für ihr Data-Mining-Projekt benötigen, zwar irgendwo innerhalb des Unternehmens oder im Web vorliegen, sie aber nur schwer auffindbar sind und mir großem Aufwand integriert werden müssen. Der Fokus vieler Data-Mining-Projekte verschiebt sich somit zunehmend von der eigentlichen Datenanalyse auf die Suche nach geeigneten Daten zur Beantwortung der jeweiligen Fragestellung sowie auf die Integration dieser Daten.

Um von Rohdaten zu Erkenntnissen und schließlich zu Handlungsempfehlungen zu gelangen, durchlaufen Data-Mining-Projekte für gewöhnlich die folgenden Phasen [Fayyad, 1996]:

1. Auswahl der relevanten Daten
2. Vorverarbeitung (d.h. insbesondere: Integration) der Daten
3. Projektion der Daten (d.h. Auswahl der relevanten Variablen)
4. Finden von Mustern (Data Mining im engeren Sinne, z.B. überwachtes oder unüberwachtes maschinelles Lernen)
5. Interpretation und Evaluierung der Muster

Während in den vergangenen Jahrzehnten viel Forschung im Bereich des Data Mining betrieben wurde, die insbesondere die nachgelagerten Schritte adressiert (Projektion und Musterfindung), ist eine Werkzeugunterstützung und Teilautomatisierung der ersten Schritte, insbesondere der Auswahl der relevanten Daten, bislang wenig erforscht. So ist es häufig Alltag, dass Analysten viel Zeit damit verbringen, Daten zu suchen (sei es im Web oder im Unternehmen) und diese manuell (z.B. mit Werkzeugen wie Excel, Informatica, MapForce, Pentaho oder Talend) zu integrieren. Werkzeuge, die Analysten bei der Suche nach relevanten Daten unterstützen und die Integration der Daten teilweise automatisieren, haben somit das Potential, sowohl die Kosten und die Dauer von Data-Mining-Projekten relevant zu reduzieren, als auch die Qualität der Projektergebnisse zu verbessern, da bei der Suche automatisiert relevante Einflussfaktoren identifiziert werden könne, an die der Analyst selbst nicht gedacht hätte.

Im Projekt DS4DM (Datensuche für Data-Mining-Prozesse) wurde wie geplant die Software RapidMiner, eine der weltweit führenden Datenanalyse-Umgebungen, um die Funktionalität zur Suche und Integration von Daten aus unternehmensinternen Datenquellen sowie aus dem öffentlichen Web erweitert. Diese Funktionalität in RapidMiner einzubauen hat den Vorteil,

dass Analysten alle Schritte eines Datenanalyse-Projekts – von der Auswahl der relevanten Daten bis zur Interpretation der Muster – in einer einheitlichen Arbeitsumgebung durchführen und somit die einzelnen Arbeitsschritte eng miteinander verzahnen. Die im Projekt prototypisch entwickelte Funktionalität wird als RapidMiner Erweiterung nach Projektende weiter entwickelt und vermarktet.

Der Schwerpunkt dieser Arbeit lag also in der Entwicklung von Methoden für die Suche und Integration von Daten aus der stetig wachsenden Anzahl verfügbarer Daten in unternehmens-internen und externen Datenquellen wie dem World Wide Web (WWW), um die notwendige und bisher manuelle und zeitaufwendige Suche, Extraktion und Integration von Daten aus den verschiedenen heterogenen Datenquellen und -formaten so weit wie möglich zu automatisieren, um dann mit Data-Mining-Prozessen beispielsweise zum überwachten oder unüberwachten maschinellen Lernen Mehrwerte generieren zu können und beispielsweise mit angereicherten Daten höhere Klassifikationsgenauigkeiten oder Prognosegüten erreichen zu können also mit zuvor unangereicherten Daten aus nur einer Quelle. Somit zielte das DS4DM-Projekt darauf ab, neue Methoden zu entwickeln, um neue relevante Informationen aus einem heterogenen Datenkorpus zu extrahieren und diese möglichst automatisiert mit bestehenden Datensätzen zu verbinden. Außerdem wurde an der Extraktion von Daten aus nichttrivialen Datenformaten gearbeitet. Als Basis für die Entwicklung der neuen Methoden diente die Data-Science-Plattform von RapidMiner, für die neu entwickelten Methoden prototypisch in Form von Erweiterungen für die RapidMiner-Plattform implementiert und getestet wurden.

Erfolge

Das DS4DM-Projekt hat alle Meilensteine zum jeweils geplanten Zeitpunkt erreicht. Direkte Ergebnisse des Projektes umfassen eine prototypische grafische Benutzeroberfläche (DS4DM-Front-End (GUI)) für Datensuche und Datenintegration für RapidMiner, mehrere RapidMiner-Erweiterungen für die Datenextraktion aus verschiedenen Unternehmens- und WWW-Datenquellen sowie eine Schnittstelle zu Informatica, einer weitverbreitenden Daten-integrationsplattform. Die Universität Mannheim entwickelte das zugehörige BDS4DM-Back-End, um Daten zu speichern und zu indexieren und beschränkte, sowie unbeschränkte Daten-suchen durchzuführen. Neben der Integration von Back- und Front-End, entwickelte RapidMiner auch eine Variante der Such- und Integrationsverfahren als eigenständige RapidMiner-Erweiterung. Insgesamt wurden fünf RapidMiner-Erweiterungen entwickelt, welche frei verfügbar (kostenlos herunterladbar) sind und mit RapidMiner Studio (Desktop-Software) und RapidMiner Server angewendet werden können. Beide DS4DM-Projektpartner haben fortwährend die bestehenden Arbeiten verbessert und regelmäßige Aktualisierungen veröffentlicht. Im zweiten Projektjahr führte RapidMiner zwei Pilotstudien durch, um Rückmeldungen von unabhängigen Experten und Nutzern, zu den veröffentlichten RapidMiner-Erweiterungen zu sammeln. Während des Projektes nahmen beide Projektpartner an zahlreichen Veranstaltungen teil und gaben Präsentationen der jeweils aktuellen DS4DM-Ergebnisse. Zudem wurden die wissenschaftlichen und technischen Ergebnisse auf Konferenzen und Workshops präsentiert. RapidMiner nutzte zusätzlich das offizielle RapidMiner-Community-Portal, um das DS4DM-Projekt und seine Ergebnisse der RapidMiner-Community mit ihren mehr als 500.000 registrierten RapidMiner-Anwendern zu präsentieren.

Zur Koordination haben beide Projektpartner vier gemeinsame Arbeitstreffen veranstaltet, zwei in Dortmund und zwei in Mannheim. Beide Projektpartner standen im regen Kontakt mit regelmäßigen, wöchentlichen Telefonkonferenzen und per E-Mail. Die Zusammenarbeit verlief gut und beide Partner haben das Interesse an weiteren gemeinsamen Projekten bekundet.

In den folgenden Abschnitten fasst dieser Abschlussbericht die Einzelarbeiten von RapidMiner und die gemeinsamen Arbeiten mit der Universität Mannheim zusammen. Zunächst werden die Hauptergebnisse aufgelistet und im Anschluss die einzelnen Arbeitspakete, Aufgaben und Meilensteine für jedes Jahr erläutert.

Kurzer inhaltlicher Bericht

In diesem Abschnitt erfolgt eine kurze Zusammenfassung der Ziele mit den jeweils dazu geleisteten Arbeiten und erzielten Ergebnissen:

- Ziel: Entwicklung von RapidMiner-Funktionsmodulen (RapidMiner-Operatoren) für die Exploration großer digitaler Datensammlungen (eDataset-Collections), die Repräsentation und manuelle Verfeinerung von Datenintegrations-Workflows und die Auflösung von Widersprüchen in Daten aus verschiedenen Quellen (Datenkonflikten).

    - Inhalt:
        - Eine umfassende Umgebung zur Datensuche und -integration wurde entwickelt. Sie bietet dem Nutzer umfangreiche Darstellungs-möglichkeiten und grafische Werkzeuge, um die Ergebnisse manuell zu verfeinern. Die Umgebung erlaubt die Suche nach einzelnen, aber auch nach mehreren Attributen (Datenspalten) mit Hilfe von Algorithmen für beschränkte (*constrained*) und unbeschränkte (*unconstrained*) Daten-suchen, welche mit Hilfe von Übereinstimmungen auf Daten-Schema- und Instanz-Ebene Ähnlichkeiten in heterogenen Datenmengen finden. Die gleichen Methoden ermöglichen auch die Konfliktbehebung und die semi-automatische Integration von Daten sowie die Verbesserung neuer Attribute um eine bestehende Datentabelle zu erweitern.

    - Ergebnisse und Aussagen zum konkreten Nutzen:
        - Die RapidMiner-Erweiterung „Data Search for Data Mining“ (DS4DM) ist ein konkretes Ergebnis dieses Projektes und ist kostenlos über den RapidMiner-Marktplatz (RapidMiner Marketplace) herunterladbar.

- Ziel: Evaluation der entwickelten Lösung anhand von Anwenderstudien (User Studies) und Pilotprojekten.

    - Inhalt:
        - Pilotstudien wurden mit zwei erfahrenen RapidMiner-Anwendern durchgeführt. Jeder der Nutzer testete anhand eigener Beispiele die RapidMiner-DS4DM-Erweiterung, welche zum Ende des zweiten Projekt-jahres entwickelt wurde.

    - Ergebnisse und Aussagen zum konkreten Nutzen:
        - Die Rückmeldungen der beiden Studien wurden genutzt, um weitere Verbesserungen in die beiden RapidMiner-Erweiterungen „Data Search for Data Mining” (DS4DM) und „Web Table Extraction” einzubauen. Der Schwerpunkt lag dabei darauf, einen Datenkorpus mit möglichst hoher

Qualität zu erstellen und den Nutzern mehr Kontrolle über die beiden Suchmethoden (constrained und unconstrained) zu ermöglichen.

- Ziel: Entwicklung von Methoden zur Extraktion von Web-Daten, um gezielt Web-Seiten mit nützlichen Daten aufzufinden und tabellarische Daten zu extrahieren.

    - Inhalt:
        - Eine generische, parameterfreie Methode wurde entwickelt, um Tabellen aus HTML-Dokumenten (Web-Seiten) zu extrahieren. Weitere Methoden befassen sich mit der Extraktion von Tabellen aus PDF-Dokumenten und Online-Kalkulationstabellen.

    - Ergebnisse und Aussagen zum konkreten Nutzen:
        - Verwertbare Ergebnisse sind die RapidMiner-Erweiterungen „Web Table Extraction“, „PDF Table Extraction” und „Spreadsheet Table Extraction”.

- Ziel: Entwicklung von Komponenten, die Daten aus Unternehmens-IT-Infrastrukturen extrahieren; Schnittstellen zu weitverbreiteten IT-System(en) (z.B. Datenintegrations-Werkzeugen); Konzept des Privacy-by-Design-Ansatzes.

    - Inhalt:
        - Um den Zugriff und die Extraktion von Daten in geschlossenen Unternehmensnetzwerken zu ermöglichen, wurden zwei Schnittstellen entwickelt. Die eine Schnittstelle erlaubt es, mit Hilfe von RapidMiner Dateien und Ordner einer Microsoft-SharePoint- oder Google-Spreadsheet-Seite anzuzeigen und herunterzuladen. Die andere Schnittstelle wurde für die Informatica-Datenintegrations-Plattform entwickelt. Diese Schnittstelle ermöglicht es, RapidMiner-Prozesse (als Web-Service) in einen Informatica-Workflow einzubetten. Das Privacy-by-Design-Prinzip wurde dabei auf die Dokumenten-Ebene übertragen. Dabei wird auf die Nutzer- und Zugriffsrechteverwaltung der Cloud-Anbieter wie Google und Microsoft zurückgegriffen.

    - Ergebnisse und Aussagen zum konkreten Nutzen:
        - Direkte Ergebnisse sind die beiden RapidMiner-Erweiterungen „SharePoint Connector“ und „Spreadsheet Table Extraction”, um auf Excel-Online- und Google-Spreadsheet-Dateien zugreifen zu können. Weiterhin wurde der „RapidMiner Connector for Informatica” entwickelt und über den Informatica-Cloud-Marktplatz veröffentlicht.

- Ziel: Das DS4DM-Projekt und seine Ergebnisse einer breiten Nutzergruppe bekannt zu machen und so deren Weiterbestehen nach Projektende zu gewährleisten sowie eine nachhaltige Verwendung als Produkt vorzubereiten.

    - Inhalt:
        - Die meisten oben erwähnten RapidMiner-Erweiterungen wurden im März 2017 auf dem RapidMiner-Marktplatz veröffentlicht, um frühzeitig Nutzerfeedback zu sammeln. Diese RapidMiner-Erweiterungen wurden kontinuierlich aktualisiert und verbessert. Sämtliche Veröffentlichungen

wurden von Blog-Postings in der RapidMiner-Community begleitet, um die über 500.000 registrierten RapidMiner-Nutzer zu informieren.

    - Ergebnisse und Aussagen zum konkreten Nutzen:
        - Das offizielle RapidMiner-Community-Portal wurde mit Beginn des zweiten Projektjahres rege genutzt, um über das DS4DM-Projekt zu informieren und eine aktive Nutzergemeinschaft für die im DS4DM-Projekt entwickelten RapidMiner-Erweiterungen aufzubauen.

- Ziel: Veröffentlichung der wissenschaftlichen Ergebnisse des Projekts
    - Inhalt:
        - Zahlreiche Veröffentlichungen wurden von beiden DS4DM-Projekt-partnern erstellt, auch als gemeinsame Publikationen, um Kenntnis über die Arbeiten und Ergebnisse des DS4DM-Projekts zu verbreiten.
    - Ergebnisse und Aussagen zum konkreten Nutzen:
        - Die Veröffentlichungen beinhalten Beiträge auf Fachkonferenzen, Workshops, Posterbeiträge auf Fachkonferenzen, Beiträge in Diskussionsrunden und Demonstrationen der RapidMiner-DS4DM-Erweiterungen auf einigen Veranstaltungen.

Wichtigste wissenschaftlich-technische, sowie weitere wesentliche Ergebnisse

Die wichtigsten Kernpunkte der wissenschaftlichen und technischen Ergebnisse sind im Folgenden aufgeführt.

1.1 RapidMiner-Erweiterung „Data Search for Data Mining" (DS4DM)

- Diese RapidMiner-Erweiterung bietet eine interaktive Umgebung innerhalb der grafischen Bedieneroberfläche (*Graphical User Interface (GUI)*) der Desktop-Software *RapidMiner Studio*, um eine *Constrained* oder *Unconstrained* Datensuche durchzuführen. Diese Suche erfolgt entweder auf dem Datenserver der Universität Mannheim oder innerhalb der RapidMiner-Erweiterung mit Datensammlungen, die in einem RapidMiner-Verzeichnis verfügbar sind. Bei der *Constrained* Suche werden die Namen der Tabellenspalten vorgegeben, die man mit passenden zu findenden Daten befüllt haben möchte. Bei der *Unconstrained* Suche werden keine solchen Vorgaben gemacht, sondern frei nach allen potentiell passenden neuen Spalten für die zu erweiternde Datentabelle gesucht.
- Interaktive Prozesse: Die Parameter der Suche und Datenintegration können interaktiv verändert und verbessert werden, um die Genauigkeit und die Abdeckung der Ergebnisse zu verbessern.
- *Correspondence* oder auch *Constrained Search*: Diese erweitert eine Tabelle mit relevanten Attributen (Spalten), welche Übereinstimmungen auf Schema- oder Instanz-Ebene mit der ursprünglichen Tabelle (*Query Table* genannt) haben. Der Name des gewünschten neuen Attributes (Spalte) wird vom Nutzer vorgegeben.
- *Unconstrained Search*: Erweitert eine Tabelle mit allen relevanten Attributen, welche eine Übereinstimmung auf Schema- oder Instanz-Ebene mit der Ursprungstabelle haben.
- Es ist möglich, eigene Datenverzeichnisse zu erstellen und diese dem Datenserver hinzuzufügen.

- Es können auch manuelle Verbesserungen für die Suchergebnisse durchgeführt werden: Bei der *Constrained* Suche können relevante Ergebnisse vom Nutzer weiter verbessert werden, in dem er mit Hilfe von eigenem Domänenwissen Ergebnisse entfernt, bevor sie der Tabelle hinzugefügt werden.
- Verschiedene grafische Darstellungsmöglichkeiten: Eine interaktive Dokumentenkarte (basierend auf dem Prinzip einer *Self-Organizing-Map (SOM)*) zeigt an, wie verschiedene Tabellen sich einander ähneln. Mit Hilfe von Zoomen und Auswahlmöglichkeiten können ganze Tabellen so einfach betrachtet werden. Außerdem gibt es einen 3D-Scatterplot, welcher Pareto-Fronten auf der Schema- und Instanz-Ebene darstellen kann.
- Mit Hilfe der Google-Tabellen-Suche können weitere Tabellen gefunden werden, welche den angegebenen Suchwörtern entsprechen. Diese können dann mit der RapidMiner-Erweiterung „Web Table Extraction“ heruntergeladen und in die Suche integriert werden.

1.2 RapidMiner-Erweiterung „Web Table Extraction“

- Die Erweiterung erlaubt eine einfache Extraktion von Tabellen aus HTML Dokumenten. Diese können entweder als Dateipfad vorliegen oder als URL, oder einer Sammlung von mehreren Pfaden oder URL.
- Parameterfreier Ansatz: Im Gegensatz zu üblichen Web-Scraping Ansätzen, welche Dokumentenspezifische reguläre Ausdrücke, XPath Regeln oder andere komplexe Anfragen benötigen, benötigt die Erweiterung keine solchen Parameter.
- Smart extraction: Die Methode verwendet ein von der Universität Mannheim entwickeltes Klassifikationsmodel um zwischen relationalen Datentabellen und Layouttabellen welche ebenfalls den <table> HTML-tag benutzen.

1.3 RapidMiner-Erweiterung „PDF Table Extraction“

- Die Erweiterung erlaubt die Extraktion von Tabellen aus PDF-Dokumenten, welche entweder als Dateipfad vorliegen oder als URL, oder einer Sammlung von mehreren Pfaden oder URL.
- Die Tabellen werden mit Hilfe der Tabula Java Bibliothek erkannt. Die genauen Parameter der Bilderkennungsalgorithmen und Extraktionsalgorithmen können vom Nutzer verändert werden um die Ergebnisse zu verbessern.

1.4 RapidMiner-Erweiterung „Spreadsheet Table Extraction”

- Ermöglicht die Extraktion von Google-Online und Excel Online Tabellen, welche als Standard bereits Privacy-by-Design Lösungen integriert haben. Die online Kalkulationstabellen sind in vielen Organisationen weit verbreitet, da:
    - Der Zugriff zu den Tabellen vom Anbieter verwaltet werden kann
    - Der Speicherplatz ebenfalls vom Anbieter in der Cloud verwaltet wird und bei Bedarf einfach erweitert werden kann.

1.5 RapidMiner-Erweiterung „SharePoint Connector“

- Sharepoint ist ein verbreitetes online Speichersystem im Unternehmensumfeld. Mit Hilfe der Erweiterung kann leicht auf Dokumente aus dem Unternehmensintranet zugegriffen werden.
- Die beiden wichtigsten Funktionen sind die Auflistung von vorhandenen Dateien und das Herunterladen von Dateien.

1.6 Informatica Connector für RapidMiner

- Eine strategische Partnerschaft wurde zwischen RapidMiner und Informatica, einem der führenden Anbieter für Datenintegration, vereinbart. Ziel dabei ist, RapidMiner Prozesse für Informatica Anwender verfügbar zu machen.
- Diese Anbindung ist ein weiterer Adapter für Unternehmensdaten, welcher sich an Nutzer von Informatica richtet. Er wurde für die Codebasis von Informatica entwickelt und ist über den Informatica Cloud Marketplace verfügbar.
- Mit Hilfe dieses Adapters kann innerhalb eines Informatica Cloud Workflows ein zuvor definierter RapidMiner Prozess aufgerufen werden. Damit ist es Informatica Nutzern möglich, die zahlreichen Data-Mining Methoden von RapidMiner zu verwenden.

Zusammenfassend hat das Projekt Möglichkeiten für eine ganzheitliche Datensuche und halbautomatisierte Datenverbesserung erforscht und entwickelt. Es ist möglich neue Anfragen an bestehende Datensammlungen zu stellen, Daten aus nicht trivialen Dokumentenformaten zu extrahieren und somit bestehende Datensätze mit neuen Attributen anzureichern. Außerdem wurden Anbindungen an weitverbreitete Plattformen geschaffen um einen verbesserten Datenaustausch zu ermöglichen.

Die Arbeitspakete

Wie im Vollantrag festgehalten, wurde die Arbeit in die folgenden 6 Arbeitspakete aufgeteilt:

- **Arbeitspaket 1: Datensuchmaschine.** Das Arbeitspaket beinhaltet die Entwicklung, Implementierung und Evaluation der Methoden für die Datenindexierung und Datensuche sowie die Implementierung von Back-End-Diensten für die Datenintegration.
- **Arbeitspaket 2: Interaktive Datenintegrationsumgebung.** Dieses Arbeitspaket beinhaltet die Entwicklung der Benutzerschnittstelle für die iterative Suche und Exploration der Daten sowie zur interaktiven Verfeinerung der vorgeschlagenen Datenintegrations-Workflows.
- **Arbeitspaket 3: Zugriff auf Webdaten.** Ziel des Arbeitspakets ist die Entwicklung von Methoden zur Extraktion von Daten aus Web-Datenquellen (inkl. Excel-CSV, XML, HTML, Microdata und Microformats, RDF/RDFa und PDF) und die Integration dieser Methoden in die Datensuchmaschine (AP1) sowie die interaktive Datenintegrationsumgebung (AP2).
- **Arbeitspaket 4: Zugriff auf Intranet-Daten.** Das Arbeitspaket entwickelt Komponenten zur Extraktion von Daten aus Unternehmens-IT-Infrastrukturen sowie Schnittstellen zu Werkzeugen wie Microsoft SharePoint, IBM Content Manager oder SAP Enterprise Suite.
- **Arbeitspaket 5: Dissemination und Verwertung.** Das Arbeitspaket legt die Grundlage für die kommerzielle Verwertung der Projektergebnisse nach Projektende indem in

Pilotprojekten gemeinsam mit Industriepartnern branchen-spezifische Showcases erstellt werden und eine Nutzer-Community um den DS4DM Demonstrator aufgebaut wird. Darüber hinaus werden die DS4DM-Projektergebnisse auf Messen und wissenschaftlichen Konferenzen beworben.

- **Arbeitspaket 6: Management.** Das Arbeitspaket Management koordiniert die Aktivitäten der Projektpartner, überwacht den Projektfortschritt und ist für die Bereitstellung der gemeinsam genutzten Projektinfrastruktur verantwortlich.

In diesem Abschlussbericht befindet sich ein Kapitel für jedes Arbeitspaket indem diese genauer geschrieben wird.

RapidMiner GmbH ist für die Arbeitspakete **AP2**, **AP4**, **AP5** und **AP6** zuständig. Die Universität Mannheim ist für die Bearbeitung der Arbeitspakete **AP1** und **AP3** zuständig.
Alle in diesen Arbeitspaketen enthaltenen Ziele und Meilenstein wurden planmäßig und pünktlich erreicht.

AP1 Datensuchmaschine

Die Entwicklung der Datensuchmaschine wurde wie geplant im August 2015 begonnen und im Juli 2017 abgeschlossen. Gemäß Arbeitsplan wurde das Arbeitspaket bis mittels mehrerer, parallel-verlaufender Aufgaben (Tasks **T1.1** bis **T1.5**) mit mehreren Meilensteinen (**M1.1** bis **M1.3**) bearbeitet.

Dabei wurden alle Meilensteine planmäßig erreicht:

- Der Meilenstein **M1.1** „Initialer Prototyp der Datensuchmaschine fertiggestellt“ wurde im Juli 2016 vollendet.
- Der Meilenstein **M1.2** „Verbesserter Prototyp der Datensuchmaschine inklusive korrespondenz-basierter Suche fertiggestellt“ wurde im Juli 2017 erreicht
- Der Meilenstein **M1.3** „Finaler Prototyp der Datensuchmaschine inklusive korrelations-basierter Suche fertiggestellt“ wurde planmäßig im Juli 2018 erreicht.

Anmerkung:
RapidMiner beteiligte sich am Design, der Integration und dem Testen der Aufgaben **T1.1** bis **T1.5**. Das Arbeitspaket **AP1** wurde von der Universität Mannheim geleitet und so sind weitere Details über die Arbeit sind im Bericht der Universität Mannheim zu finden.

AP2 Interaktive Datenintegrationsumgebung

Dieses Arbeitspaket wurde planmäßig im August 2015 begonnen und im April 2018 fertiggestellt. Es besteht aus den Meilensteinen **M2.1**, **M2.2**, **M2.3** und **M2.4**. Diese wurden jeweils planmäßig erreicht.

M2.1 Initialer Prototyp der Integrationsumgebung fertiggestellt

Der Initiale Prototyp der Integrationsumgebung wurde planmäßig im Juli 2016 fertiggestellt.

Der Prototyp nutzte die erste Fassung der Data Search API, welche von der Universität Mannheim entwickelt wurde. Er umfasste eine Such-Methode, welche passende Tabellen aus dem bereits indizierten Corpus liefert. Auch liefert er eine erste Version der Schema- und Instanz-Übereinstimmung. Eine weitere API Methode, „fetch table“, erlaubt es einzelne Tabellen zu laden. Basierend auf dieser API konnte RapidMiner einen ersten Prototyp der „Data Search for Data Mining” Erweiterung (im Folgenden kurz als „Data Search“ Erweiterung bezeichnet) veröffentlichen. Sie enthielt die Operatoren „Data Search“, „Translate“ und „Fuse“.

M2.2 Verbesserter Prototyp der Integrationsumgebung fertiggestellt

Der verbesserte Prototyp der Integrationsumgebung wurde planmäßig im Juli 2016 fertiggestellt.

Dieser Meilenstein umfasst eine Verbesserung des Prototyps der RapidMiner-Erweiterung zur Datensuche (Aufgaben aus Tasks **T2.1**, **T2.2** und **T2.3**). Diese Erweiterung bietet eine komplette Implementierung des Search-Join-Verfahrens und besteht aus drei verbesserten RapidMiner-Operatoren: „Data Search“ (Datensuche), „Translate“ (Übersetzung), „Fuse“ (Zusammenführung). Zusammen erlauben diese Operatoren dem Nutzer das Auffinden neuer kontextuell relevanter Attribute und eine automatische Integration dieser in einen gegebenen Eingangsdatensatz.

Eine hervorzuhebende Funktion der Erweiterung ist die Möglichkeit zur Wahl zwischen einer automatischen und einer semi-automatischen Abarbeitung der Teilschritte. Bei der semi-automatischen Variante hat der Nutzer die Möglichkeit über einen Editor manuell mit visueller Unterstützung Verfeinerungsschritte vorzunehmen. Diese Funktion greift zum einen nach dem „Data Search“, als auch nach dem „Translate“-Operator. Durch diese Option entsteht eine Synergie aus menschlicher Intelligenz, bei der Datenidentifikation und -bereinigung, und der maschinellen Effizient bei der automatischen Datenintegration, welche für den Menschen fehleranfällig und sehr aufwendig wäre.

Die Erweiterung wurde inklusive graphischer Oberfläche entwickelt, die nahtlos in RapidMiner Studio integriert ist. Dadurch wird die Integration der Datensuche in den gesamten Data-Mining-Prozess stark vereinfacht.

Auch im dritten Jahr, neue Entwicklungen am Data Search Back-End (Tasks **T1.1**, **T1.3** and **T1.4**) führten zu zwei neuen Operatoren in der Data Search Extension. Diese sind „Unconstrained Search“ und „Correlation Based Search“, welche dem Nutzer die Erweiterung der Datentabelle mit mehreren relevanten Attributen. Diese Operatoren arbeiten auf Seite des Clients und sind ideal geeignet, wenn große Sammlungen an Tabellen bereits vorliegen oder zum Back-End hochgeladen werden können.

Um neue Repositories zu befüllen und um bestehende mit öffentlich verfügbaren Daten, wurde der „Google Table Search“ Operator der Data Search Erweiterung hinzugefügt. Dieser stellt Anfragen an den Google Tabellensuchdienst um Webseiten URLs zu finden, welche Datentabellen enthalten. Die Liste der URLs kann an den Read HTML Operator (aus der Web

Table Extraction Erweiterung) übergeben werden, welche diese Tabellen dann extrahiert. (Task **T3.2**).

In Fällen, in denen es nicht erwünscht oder möglich ist Daten in das Back-End der Suchmaschine zu laden, wurde der neue Operator „Enrich Table by Data Fusion“ zur Data Search Extension hinzugefügt. Dieser Operator nimmt eine Kollektion von Datentabellen als Eingabe und verarbeitet diese um eine ausgewählte Tabelle mit mehreren relevanten Attributen anzureichern. Dieser Operator implementiert einen Datenintegrationsalgorithmus, mit der Schema Matching, Instance Matching und Data Fusion und benutzt die gleichen Konzepte wie die Back-End Implementierung aus Task **T1.4**. Dieser Operator erzeugt eine ähnliche Ausgabe wie die „Unconstraint Search“ und „Correlation-Based Search“ Operatoren, aber ohne die Notwendigkeit eines Back-Ends.

Die „Data Search“ Erweiterung wurde am 23.03.2017 veröffentlicht und am 27.07.2017, 19.10.2017, 30.01.2018, 30.07.2018 und zuletzt am 16.11.2018 aktualisiert. Die Erweiterung ist auf dem RapidMiner Marketplace zur frei verfügbar.

M2.3 Evaluation des Prototypens innerhalb von Pilotprojekten abgeschlossen

Zur Beurteilung der Nutzbarkeit wurden zwei Evaluationen mit Externen durchgeführt. Dabei wurden die Funktionalität und die Anwendbarkeit abgefragt. Im Rahmen der Evaluation

wurden alle Erweiterungen, die zum Zeitpunkt der Durchführung fertig waren, getestet. Dazu zählen die „Data Search“-Erweiterung, welche im Rahmen des Arbeitspaketes **AP2** entwickelt wurde und die Arbeiten den Paketes AP1 beinhaltet, da auf der Back-End zurückgegriffen wird. Weiterhin wurden die „Web Table Extraction“-Erweiterung (aus **AP3**), die „PDF Table Extraction“-Erweiterung (aus **AP3**), sowie die „Spreadsheet Table Extraction“-Erweiterung (aus **AP4**) bereitgestellt. Die Evaluation wurde dabei von zwei professionellen Beratern (Lindon Ventures und Genzer Consulting) durchgeführt. Beide sind langjährige RapidMiner-Nutzer und haben Erfahrungen im Bereich Web- und Data-Mining. Im Rahmen der Evaluationen wurde wichtiges Feedback erhoben, welches für zukünftige Arbeiten berücksichtigt wird.

M2.4 Finaler Prototyp der Integrationsumgebung fertiggestellt

Unter diesem Meilenstein wurden verschiedene Verbesserungen in Hinblick auf die Search-Join Operatoren umgesetzt, welche die „Data Search“, „Translate“ and „Fuse“ Operatoren umfasst. Zusammen implementieren diese Operatoren das Frontend der *Correspondence Based* Suche, welche eines der Hauptergebnisse im zweiten Projektjahr war. Die Correspondence Suche verbessert die Schlagwort-basierte Suche, welche als initialer Prototyp im ersten Projektjahr entwickelt wurde.

Im dritten Jahr wurde der „Translate“ Operator verbessert um zusätzliche Informationen über die Herkunft der Kandidatentabelle anzuzeigen, z.B. Metadaten über die Quelle. Das hilft dem Benutzer bei der grafischen Bearbeitung und Auswahl von Suchergebnissen, um zum Beispiel Daten einer gewissen Quelle nicht zu berücksichtigen. Dies ergänzt die statistischen Metadaten zur Ähnlichkeit der Attribute, welche bereits angezeigt werden.

Darüber hinaus wurde im dritten Jahr ein „Advanced Fuse“ Operator implementiert, welcher dem Benutzer mehr Kontrolle bei der Konfliktauflösung gibt. Der Operator stellt eine fortgeschrittene Zusammenführungsstrategie bereit, die dem Benutzer es erlaubt Präferenzen zu mehreren Kriterien anzugeben. Diese werden dann gemeinsam verwendet um Konflikte beim Zusammenführen der erweiterten Attribute zu lösen. Auch hier kann jetzt optional auf die Herkunft der Tabellen zugegriffen werden. Da die erweiterte Tabelle Werte aus

verschiedenen Quellen beinhalten kann, kann die Information über die Herkunft durchaus relevant sein.

Die grafische Komponente der Verbindungsverwaltung, welche ein Teil der Data Search Extension ist, wurde ebenfalls verbessert und gibt dem Benutzer mehr Kontrolle über den Verbindungsaufbau und zeigt Informationen bei Timeouts. Die Verbindungsverwaltung vereinfacht das Management von mehreren Verbindungen, falls mehrere Back-End s zur Verfügung stehen.

Die aktuellste Version von „Data Search for Data Mining“ Erweiterung beinhaltet jetzt auch zusätzliche Beispieldatensätzen, damit die Nutzer verschiedene Datenanreicherungsmethoden direkt testen können.

T2.1 Design und Implementierung von Operationen zur Exploration großer Dataset-Collections

Die Entwicklung begann planmäßig im August 2015. Diese Aufgabe umfasst Arbeiten über alle drei Projektjahre hinweg.

Das erste Jahr

„Data Search for Data Mining”-Erweiterung für RapidMiner

Passend zur entwickelten Data Search API wurde der Operator „Data Search“ entwickelt. Dieser bekommt eine Tabelle und Parameter übergeben, ruft die Web Services mit Parametern auf und wandelt die Resultate in RapidMiner Datenstrukturen um. Der Nutzer erhält eine Sammlung von Tabellen, die das gesuchte Attribut beinhalten. Außerdem erhält der Benutzer das Zielschema der neuen Tabelle und die Korrespondenzen auf Schema- und Instanzlevel. Der Operator ist bereits mit den Web Services der Universität Mannheim verknüpft.

Außerdem wurden bereits die Operatoren „Translate“ und „Fuse“ von RapidMiner entwickelt, die bisher allerdings noch keine Funktionalität beinhalten. Im Rahmen der Entwicklung einer eigenen Extension für die genannten Operatoren wurde parallel die Dokumentation zum Thema „How to extend RapidMiner“ aktualisiert.

Das zweite Jahr

Im zweiten Jahr wurde die “Data Search“ Erweiterung maßgeblich verbessert und neue Möglichkeiten hinzugefügt. Insbesondere wurden die „Data Search“ und „Translate“ Operatoren erweitert um eine manuelle Verbesserung der gefunden Daten und ihre Schema- und Instanz-Übereinstimmungen zu ermöglichen.

Graphischen Anwendungsoberfläche:

Durch die, während des zweiten Jahres, erzielten Fortschritte kann der Search-Join-Prozess vom Nutzer manuell zur Laufzeit angepasst werden. Dabei kann die Analyseprozess-ausführung an zwei Stellen angehalten werden. Zum einen bei der Datensuche (nach dem „Data Search“ Operator) und zum anderen bei der Übersetzung (nach dem „Translate“ Operator). Dabei werden die Ergebnisse in einer legendenartigen Baumansicht dargestellt.

Diese Art der Ansicht erlaubt einen leichten Zugang zur Schema- und Instanz-basierten Übereinstimmungsmethodik der Suche. Jede vorgeschlagene Tabelle kann dabei einzeln betrachtet werden. Als Auswahlhilfe werden zudem Übereinstimmungsmetriken angezeigt. Dies erlaubt dem Nutzer unpassende Tabellen aus der Menge der zu verarbeitenden Tabellen (im Arbeitsspeicher) zu entfernen. Weitere Schritte werden lediglich auf der Teildatenmenge der übrigen Daten durchgeführt.

In der Anwendungsoberfläche werden Zusatzinformationen, wie Verteilungen statistischer Qualitätsmetriken betrachteter Tabellen angezeigt. Verteilungen werden dabei als Histogramme dargestellt und bieten eine einfache Möglichkeit einen schnellen Überblick über gefundene Daten zu erhalten. Einige der Metriken sind:

- *Coverage*: Coverage ist die Anzahl an Übereinstimmung gefundener Sucheinträge geteilt durch die Anzahl der verfügbaren Einträge der Suchtabelle.
- *Ratio*: Ratio ist das Verhältnis aus Übereinstimmungen gefundener Sucheinträge zu der Anzahl gefundener Tabellen.

Zur weiteren Unterstützung der Entscheidungsfindung des Nutzers wurden zwei weitere explorative Visualisierungen implementiert.

Die erste Visualisierungsform stellt eine interaktive Dokumenten-Karte basierend auf der Technologie selbstorganisierender Karten dar, welche gefundene Datensätze in Form von Clustern darstellt. Jede Tabelle wird dabei als Punkt auf einer Karte angezeigt. Mit Hilfe dieser Darstellungstechnik können Zusammenhänge benachbarter Strukturen erkannt werden. Es stehen unterschiedliche Abstandsmaße (P-Matrix, U-Matrix, sowie U*-Matrix) zur Wahl. Zudem kann auf verschiedene Farbkodierungen für Darstellung der Nachbarschaft zurückgegriffen werden. Eine Möglichkeit zur genaueren Evaluation von Teilbereichen ist durch eine Zoom-Funktionalität gegeben. Als farbgebende Dimension der Datenpunkte können drei Optionen aus der Übersichtstabelle gewählt werden. Diese Optionen sind:

- Anzahl der Schema-Übereinstimmungen für jede Tabelle,
- Anzahl der Instanz-Übereinstimmungen für jede Tabelle und
- der Name der Tabelle.

Jeder Datenpunkt ist zudem direkt mit der repräsentierten Tabelle verknüpft, sodass ein Klick des Nutzers auf einen Datenpunkt ein Anzeigen des entsprechenden Datensatzes veranlasst. Dies stellt einen Drill-Down-Mechanismus dar. Diese innovative Kartendarstellung erfüllt somit die Vision, die im Aufgabenpaket **T2.1** beschrieben wird und die sich auf Clustering Methoden zur visuellen Exploration großer Dataset-Collections anhand von Kennzahlen bezieht. Diese Art der Verwendung von SOMs als Basis der entwickelten innovativen Visualisierungstechnik bringt den Vorteil mit sich, dass die Visualisierung nicht angepasst werden muss, wenn die Übersichtstabelle, die die Basis der Visualisierung bildet, durch weitere Daten wie Coverage oder Ratio ergänzt wird. Neue Daten werden somit automatisch für das Clustering und die Visualisierung verwendet. Dies ist möglich, da SOMs intern eine Dimensionsreduktion durchführen und somit ohne Probleme die Dimensionalität des zugrundeliegenden Datensatzes erhöht werden kann. Solch komplexe Aufarbeitung wäre bei einer rein textuellen Darstellung nicht so leicht möglich.

Als zweite Form der Darstellung, welche auch auf der Übersichtstabelle basiert, werden Punkte in einem drei-achsigen orthogonalen System dargestellt. Die Punkte sind dabei durch die jeweiligen Tabellennamen annotiert. Die Dimensionalitäten der Achsen sind gegeben

durch: die Anzahl der Schema-Übereinstimmungen (x-Achse) und die Anzahl an Instanz-Übereinstimmungen, welche es schon erlauben die Pareto-Grenze zu identifizieren, während der Name (z-Achse) als Herausstellungsmerkmal die Übersicht erhöht. Die Pareto-Grenze stellt einen Schwellwert für den besten Kompromiss dar, der erreicht wird, wenn Tabellen genutzt werden, die den höchsten Wert auf der jeweiligen Achse besitzen aber gleichzeitig nicht die Werte der anderen Achse dominieren.

Im Gesamten stellt die Data-Search-Erweiterung für RapidMiner somit eine geführte Nutzer-erfahrung für den kompletten Daten-Entdeckungs-Prozess bereit. Gleichzeitig werden die aufwendigen Schritte der Übersetzung und Zusammenführung durch die Translate- und Fuse-Operatoren übernommen. Die folgende Abbildung zeigt ein 3D Streudiagramm. Datenpunkte sind mit Tabellennamen annotiert und der Anzahl an Instanz-Übereinstimmungen entsprechend eingefärbt. Tabellennamen die rot, orange oder gelb eingefärbt sind zeigen Kandidaten, die im Vergleich zu anderen (in blau dargestellten) Tabellen signifikanter sind. Die Kreise wurden nachträglich zur Abbildung hinzugefügt, um interessante Regionen hervorzuheben.

Das dritte Jahr

Ein neuer Operator „Google Table Search" wurde der „Data Search" Erweiterung hinzugefügt. Dieser bietet eine Keyword-basierte Suche zur Exploration von im Netz zur Verfügung stehender Daten Tabellen. Die Tabellen-Suche unterstützt das übergeordnete Ziel der Organisation von privaten und unternehmerischen Repositories. Der neue Operator bietet ein einfaches Interface um den „Google Table Search" Service[1] innerhalb von RapidMiner zu nutzen.

Das Ergebnis listet mehrere öffentliche Webseiten, die über Daten Tabellen verfügen. Diese können direkt dem neuen „Read HTML Tables" Operator der „Web Table Extraction" Erweiterung (**T3.2**, **T4.1**) übergeben werden um eine Massen Extraktion der Daten Tabellen von diesen Webseiten durchzuführen.

Die gefundenen Tabellen können verwendet werden um neue Daten Repositories zu erstellen oder bestehende Repositories zu erweitern. Dafür kann die neue „Data Table Upload" Eigenschaft verwendet werden (**T2.2**). Mit Hilfe dieser neuen Funktionalität erschließt die „Data Search" Erweiterung das Netz als eine mögliche Datenquelle.

T2.2 Design und Implementierung von Operationen für Repräsentation und manuelle Verfeinerung von Datenintegrations-Workflows

Die Entwicklung begann planmäßig im August 2015. Diese Aufgabe umfasst Arbeiten über alle drei Projektjahre hinweg.

Das erste Jahr

Um eine Vision für das gesamte Projekt zu schaffen, wurde von RapidMiner ein Design für interaktive Datenintegrationsumgebung entwickelt. Dieser vereint alle Komponenten, die in

[1] Google table search service. Web-Link: https://research.google.com/tables

diesem Projekt entwickelt werden sollen in einer Form, die es auch für einen unerfahrenen Nutzer möglich macht Datensuche und –integration zu betreiben.

Zu diesem Zeitpunkt boten die Operatoren noch keine Möglichkeiten die Ergebnisse weiter zu verfeinern oder durch interaktive Grafiken verrauschte Daten zu entfernen.

Das zweite Jahr

Wie auch das Ergebnis des Data-Search-Operators, können die Daten, welche vom „Translate“ Operator bereitgestellt werden, manuell angepasst werden. Dieser Operator automatisiert den Datenintegrationsprozess, während der Nutzer trotzdem über eine graphische Schnittstelle Eingriffsmöglichkeiten hat. So kann der Nutzer ungewollte Tabellen herausfiltern und diese Selektion in den weiteren Integrationsschritt einfließen lassen, der von dem „Fuse“ Operator gehandhabt wird.

In zweiten Projektjahr, wurden zudem Qualitätsmaße zur Anwendungsoberfläche, die nach dem Übersetzungsschritt angezeigt wird, hinzugefügt. Diese Maße beinhalten Querabstandsähnlichkeit (Cross-Distance Similarity) und Unähnlichkeit für jeden nicht leeren Wert (auf Zellenebene) zu allen anderen (Zell-)Werten aller Zielattribute der Übersetzungstabellen. Auf diese Weise kann eine Art Vertrauenswert für eine ganze Tabelle oder ihren Beitrag auf Zellenebene abgeleitet werden. Es muss erwähnt werden, dass es selbst für einen wohldefinierten Datensatz sehr schwer ist etwas wie einen Vertrauenswert zu definieren. Da basierend auf der Grundannahme einer unsicheren Datenlage, welche inhärent bei der Suche nach neuen Daten gegeben ist, Vertrauen basierend auf dem Nachschlagen eines Wortes in einem bekannten Wörterbuch, nicht möglich ist. Die neuen Maße sind:

- Levenstein Mistrust:

  Der Wert „Levenstein Mistrust“ beschreibt ein fehlendes Vertrauen in einen Wert begründet durch den Abstand eines Zellenwertes zu allen weiteren, nicht leeren Einträgen derselben Spalte bestimmt durch das Levenstein Abstandsmaß.

- Jaro Winkler Trust:

  Der Wert „Jaro Winkler Trust“ beschreibt ein Vertrauen in einen Wert begründet durch den Abstand eines Zellenwertes zu allen weiteren, nicht leeren Einträgen derselben Spalte bestimmt über das Jaro-Winkler-Ähnlichkeitsmaß.

- Fuzzy Trust:

  Der Wert „Fuzzy Trust“ beschreibt ein Vertrauen in einen Wert begründet durch den Abstand eines Zellenwertes zu allen weiteren, nicht leeren Einträgen derselben Spalte bestimmt über die Fuzzy-Ähnlichkeit.

- Missing Values:

  Der Wert „Missing Values“ beschreibt die Anzahl leerer Einträge in Bezug auf die gesuchte Spalte geteilt durch die Anzahl an Einträgen der übersetzten Tabelle.

Zusätzlich werden beim Übersetzungsschritt fehlende Einträge gehandhabt. Die aufgeführten Qualitätsmaße helfen zusammen mit der Aufbereitung des Datensatzes zu Übersetzungs-tabellen bei der anschließenden Zusammenführung.

Zur Bekanntmachung der Arbeitsergebnisse und zur Hervorhebung der Praktikabilität der „Data Search for Data Mining“ Erweiterung, wurde ein Blogeintrag in der RapidMiner Community veröffentlicht. Beiträge in der RapidMiner-Community sind auch für nicht registrierte Nutzer sichtbar. Der Blog-Beitrag ist unter folgendem Link erreichbar:

https://community.rapidminer.com/discussion/38231/the-data-search-for-data-mining-extension-release

Das dritte Jahr

Im drittem Projektjahr nutzte die Aufgabe **T2.2** die neuesten Ergebnisse aus den Aufgaben **T1.1**, **T1.3** und **T1.4**, welchen Anwendern ermöglichen die Suchmaschineninstanz durch neue Eigenschaften der „Data Search Back-End API“[2] präziser zu kontrollieren.

Die neuen Operatoren “Unconstrained Search” und „Correlation-Based Search“ wurden der “Data Search” Erweiterung hinzugefügt. Diese sind die Client-Seite der neuen Suchmethoden, welche im dritten Jahr dem Back-End hinzugefügt wurden (**T1.4**).

„Unconstrained Search“ Operator

Dieser Operator ruft die “Unconstrained Search” Methode im Back-End Server auf. Er sendet die Eingabetabelle an den Server, zusammen mit den notwendigen Parametern, und erhält als Antwort die erweiterte Tabelle. Das Schema-Matching, Instance-Matching und die Data Fusion werden vom Back-End Server durchgeführt. Dabei werden die Daten von einem vom Nutzer spezifizierten Repository genutzt. Obwohl die Suche und Fusion vom Back-End durchgeführt werden, ermöglichen die Parameter des Operators eine größtmögliche Konfigurierbarkeit.

„Correlation-Based Search“ Operator

Dieser Operator ruft die Correlation Search im Back-End -Server auf. Die Correlation Search funktioniert ähnlich zur Unconstrained Search außer, dass die hinzugefügten Attribute eine hohe Korrelation zum ausgewählten in der Eingangstabelle haben. Der Nutzer muss außerdem ein Attribut der Eingabetabelle als Korrelationsattribut auswählen und einen Mindestwert für die Korrelation angeben.

„Enrich Data Table by Data Fusion” Operator

Ein weiterer neuer Operator „Enrich Data Table by Data Fusion”, wurde ebenfalls entwickelt. Dieser liefert eine, vom Back-End unabhängige, Methode für die „Unconstrained“ und „Correlation-Based” Suche. Mit diesem Operator hat RapidMiner einen eigenen Data Enrichment Algorithmus implementiert, welcher auf den Konzepten zu Unconstrained und Correlation Search der Universität Mannheim beruht (**T1.4**), und eine alternative Anwendung der Methoden ermöglicht. Als Besonderheit ist der Operator vollkommen unabhängig von einem Back-End Server und kann in Fällen eingesetzt werden, in denen die Installation und

[2] DataSearch Back-End API, Web-Link: http://web.informatik.uni-mannheim.de/ds4dm/API-definition.html

die Verwaltung eines Back-End Servers zu aufwändig ist, oder die Daten aus Gründen der Vertraulichkeit nicht an andere Stelle gespeichert werden dürfen. Der Operator liefert ähnliche Ergebnisse wie die Unconstrained oder Correlation-Based Suche, verarbeitet hierfür aber einen lokalen Korpus von Tabellen. Er findet relevante Tabellen durch Schema- und Instanz-Übereinstimmung. Er findet so relevante Tabellen und erweitert die Eingabetabelle um neue Attribute abhängig von der gewählten Dichte und Korrelation zu einem ausgewählten Attribut.

Die ausführliche Beschreibung des Algorithmus findet sich im Hilfetext des Operators. Der Operator ermöglicht dem Nutzer die Suche über mehrere Parameter zu steuern. Die Auswirkung verschiedener Einstellung kann aufgrund der lokalen Ausführung direkt beobachtet werden. Sinnvolle Einstellung sind als Voreinstellung bereits ausgewählt.

Als Eingabe erwartet der Operator eine Datentabelle und eine Sammlung von weiteren Tabellen, aus welcher die relevanten Teilmengen für die Datenfusionierung ausgewählt werden.

„Create Correspondences“ Operator

Ein weiterer, neuer Operator „Create Correspondences”, wurde ebenfalls entwickelt. Dieser Operator implementiert eine Variante der „Keyword-based Search“ und der „Correspondence-based Search“ Algorithmen der Backen-API. Der Vorteil dieses Operators ist, dass der Nutzer in der Lage ist die „Correspondence Search“ vollkommen unabhängig von einem bestehenden Back-End Server durchzuführen. Dieser Operator kann somit den “Data Search“ Operator in der typischen “Search-Join“ Prozesskette, bestehend aus „Data Search“, „Translate“ und „Advance Fuse“ Operatoren, ersetzen. Der neue „Create Correspondence“ Operator unterstützt ebenfalls die manuellen, grafischen Auswahlmethoden, wie der zuvor verwendete „Data Search“ Operator.

Der “Create Correspondence” Operator führt einen Schema- und Instance-Vergleich durch und erlaubt es dem Nutzer durch iterative Vergleiche eine passende Parameterkombination zu finden, um die Ergebnisse für die resultierende Tabelle zu verfeinern. Ähnlich zu dem „Enrich Table by Data Fusion” Operator, führt der „Create Correspondences” Operator ebenfalls während der Schema- und Instanzvergleiche, interne Konfliktauflösungen vor. Dieser Operator ist ein wichtiger Schritt die „Data Search“ Erweiterung unabhängiger von einem bestehenden Back-End zu machen, da es dem Nutzer ermöglicht wird auf bereits bestehende Datensammlungen zurückzugreifen. Insbesondere auch, wenn die Daten sensible Informationen enthalten, welche nicht auf einem anderen Server gespeichert werden sollen oder dürfen.

„Create Repository“ Operator

Dieser Operator ermöglicht es ein neues Repository in der Suchmaschineninstanz zu erzeugen. Das Benutzerinterface macht den Operator einfach in der Benutzung, selbst bei verschiedenen Back-End Instanzen.

„Data Table Upload“ Operator

In einigen Fällen reicht es aus eine einzelne Datentabellen zu einem Repository hoch zu laden. Für solche Fälle ist der „Data Table Upload“ Operator entwickelt worden. Man kann ihn nutzen um eine Tabelle in einem spezifischen Repository zu aktualisieren.

„Data Tables Upload“ Operator

In vielen Fällen ist es notwendig mehrere Daten Tabellen zum Back-End Repository hoch zu laden. Dieses Mehrfach-Hochladen („Bulk-upload“) ist als eine notwendige Eigenschaft für die Verwaltung von privaten wie auch unternehmerischen Repositories identifiziert worden. Zwar können solche „Bulk“ Daten nicht mit Datenmengen aus dem Bereich des „Big Data“ verglichen werden, allerdings können sie trotzdem verhältnismäßig groß werden, im Bereich von mehreren Hundert Megabytes bis zu einigen Gigabytes. Daher muss die Integrität des Hochladens speziell sichergestellt werden.

Zur Bekanntmachung der Arbeitsergebnisse und zur Hervorhebung der Praktikabilität der „Data Search for Data Mining“ Erweiterung, wurde ein Blogbeitrag. Der Blog-Beitrag ist unter folgendem Link erreichbar:

https://community.rapidminer.com/discussion/43306/the-web-as-a-new-data-source-for-rapidminer

T2.3 Design und Implementierung von Konfliktauflösungs-Operationen

Die Entwicklung begann planmäßig im Oktober 2015. Diese Aufgabe umfasst Arbeiten in den Projektjahren 1 und 3.

Das erste Jahr

„Fuse“ Operator

Eine erste Version für die Konfliktauflösung wurde mit dem „Fuse“ Operator umgesetzt. Dies half dabei den ersten Prototypen für die „Search-Join“ Implementierung im ersten Projektjahr.

Der „Fuse“ Operator führt die Datenintegration durch. Die Datenintegration ist der Schritt, bei dem auf Zellenebene Werte für das gesuchte Attribut zum richtigen Eintrag der Suchtabelle zugeordnet werden.

Dabei erhält der Operator eine Sammlung von Übersetzungstabellen, die Suchtabelle (mit dem Zielschema) und eine Instanz-Korrespondenz. Mit diesen Daten werden im Operator alle möglichen Werte identifiziert, die einen Wert des gesuchten Attributes darstellen. Da es mehrere Kandidaten pro Zellwert geben kann, wird eine Zusammenführungspolitik angewandt. Bei der Anwendung einer solchen Politik werden Konflikte aufgelöst, um die Entscheidung für einen einzelnen Zellwert herbeizuführen. Als Ergebnis wird eine neue Tabelle erzeugt, die der Suchtabelle angereichert um das gesuchte Attribut, entspricht.

Das dritte Jahr

In Aufgabe **T2.3** wurde ein neuer Operator zur Konfliktbehebung bei der Search-Join Methode implementiert. Dieser enthält die Correspondence Search als Strategie (ein Hauptergebnis des zweiten Projektjahres). Die Auflösung von Konflikten ist auch ein wichtiger Bestandteil der Arbeiten an Aufgabe **T2.2** (siehe die Beschreibung des „Enrich Table by Data Fusion“ und „Create Correspondences“ Operatoren).

„Advanced Fuse“ Operator

Der „Advanced Fuse” Operator wurde entwickelt um einige Nachteile des „Fuse“ Operators aus dem ersten Projektjahr auszugleichen. Der „Fuse“ Operator bietet lediglich eine einfache Methode, welche wie eine „First Fit“ Heuristik arbeitet. Er wählt den ersten verfügbaren Wert aus mehreren Kandidaten aus um ein erweitertes Attribut zu befüllen. Die „Advanced Fuse” Methode identifiziert die am vertrauenswürdigsten Werte durch die Berechnung verschiedener Hilfsfunktionen. Die Parameter der Methode sind Gewichte für die, auf den Metadaten basierenden, Statistiken, welche vom Benutzer angepasst werden können.

Einige dieser Statistiken basieren auf der Tabelle, wie zum Beispiel „coverage weight“, „ratio weight“ und „non missing values weight“. Andere Statistiken beruhen auf einzelnen Beispielen, wie das „Jaro Winkler similarity weight”, „Lenvenshtein dissimilarity“ und „fuzzy similarity weight“. Diese drei Statistiken stellen den Durchschnitt der relativen Abstände zwischen jedem Wertepaar dar.

Im Allgemeinen sollten die Gewichtsparameter so gesetzt werden, dass der ähnlichste Wert aus der Kandidatenmenge ausgewählt wird. Gleichzeitig kann dieser Ansatz auch genutzt werden um verschiedenartige Werte zu bevorzugen. Diese Konfigurierbarkeit wird erst durch die Verwendung der Bewertung durch die Hilfsfunktionen ermöglicht.

„Data Search Verbindungsmanager“ Frontend

Zur Verbesserung der Integration zwischen dem Frontend und dem Back-End, wurde ein Verbindungsmanager im Frontend implementiert. Dieser hilft graphisch die Verbindungseinstellung zu einem Back-End Server zu verwalten.

T2.4 Projekte mit Pilotanwendern

Diese Aufgabe umfasst Arbeiten im zweiten Projektjahr.

Das zweite Jahr

Zur Umsetzung dieser Aufgabe trat RapidMiner an zwei professionelle IT-Beraterfirmen heran, die mehrere Jahre Erfahrung im Bereich Data Mining und dessen Anwendung auf kommerzielle Probleme haben. Diese Firmen sind i) Lindon Ventures[3] und ii) Genzer Consulting[4]. Die Kontaktpersonen haben zwar Erfahrung mit RapidMiner, sind jedoch externe Personen, die unabhängig von RapidMiner agieren. Um maximalen Nutzen aus der Studie zu ziehen, wurden die Probanden gebeten, alle im Rahmen des Projektes entwickelten Operatoren zu testen. Dies umfasst die „Data Search for Data Mining“ Erweiterung aus **AP2**, und somit durch die Protokollabfragen der Operatoren auch das Suchmaschinen Back-End, die „Web Table Extraction“-Erweiterung für RapidMiner von **AP3**, die „PDF Table Extraction“-

[3] Lindon Ventures, USA: Web-Link: www.lindonventures.com
[4] Genzer Consulting, USA: Web-Link: www.genzerconsulting.com

Erweiterung für RapidMiner aus **AP3**, sowie die „Spreadsheet Table Extraction“-Erweiterung für RapidMiner aus **AP4**. Die Berater erhielten eine Vorlage die:

i) Die Erweiterungen vorstellt,

ii) Anleitet eigene Tests durchzuführen und

iii) Hilft strukturierte Bewertungen zur Nutzung und der Operatordokumentation zu geben.

Ziele:

Hauptziel der Befragung war die Erhebung von Informationen zur Nutzbarkeit der Erweiterungen. Dazu gehört die Abfrage verschiedener Teilaspekte, nämlich der Funktionalität, der Anwendung, der Akzeptanz, der Dokumentation und der gegebenen Hilfestellungen/Anleitungen. Im Folgenden wird auf diese Aspekte eingegangen und die bisherige Rückmeldung zusammengefasst. Weiterhin wird eine erste Strategie zur Einarbeitung der gewonnenen Rückmeldungen für zukünftige Änderungen aufgezeigt.

Rückmeldung zur Funktionalität:

Besonders die Datensucherweiterung rief ein hohes Interesse bei den Probanden hervor. Ihr Interesse beruht sowohl auf der Neuheit des Konzeptes der kontextbasierten Suche tabellarischer Daten, als auch der Praktikabilität der Integration. Sie bewerteten die Idee als hervorragend, mächtig und attestieren ein enormes Potential. Bedingt durch die angeregte Begeisterung waren die Erwartungen was die Quantität und Genauigkeit der interessanten Suchergebnisse angeht sehr hochgesteckt.

Die Rückmeldung zur Funktionalität half dabei, folgende Arbeiten für die Arbeitspakete AP1 und AP2 besser zu planen. So zeigt sich, dass das inkrementelle Hinzufügen von Korpora und die damit verbundene erhöhte Genauigkeit der Suchergebnisse von besonderer Wichtigkeit ist. Beispielsweise bringt die Einführung von gewissen Goldstandard Datensätzen die Möglichkeit mit sich Methoden zunächst stärker zu verfeinern, da eine bekannte Testumgebung vorliegt. Weiterhin ist es für die Probanden von größtem Interesse kommerzielle (private) Daten automatisiert zum Back-End hinzufügen zu können. Bisher bedarf dieser Schritt noch Eingriffe seitens des Nutzers. Dementsprechend ist die Entwicklung weiterer Werkzeuge zum Handling des Back-Ends ein entscheidender Schritt, um die Adoption durch Endnutzer zu begünstigen. Dabei kann zudem Bedenken bezüglich der Handhabung unternehmenskritischer Daten zuvorgekommen werden.

Die „Web Table Extraction“-Erweiterung für RapidMiner war auch Teil der Studie. Diese Erweiterung wurde von beiden Probanden verwendet um online verfügbare Daten (von Wikipedia und Google Docs) zu extrahieren.

AP 3 Zugriff auf Web-Daten

Dieses Arbeitspaket wurde planmäßig im Oktober 2015 begonnen und im April 2018 fertiggestellt. Die Meilensteine **M3.1**, **M3.2**, **M3.3**, **M3.4**, **M3.5**, **M3.6** und **M3.7** wurden jeweils planmäßig erreicht.

Anmerkung:

RapidMiner beteiligte sich an den Aufgaben **T3.2** und den Meilensteinen **M3.3**, **M3.4** und **M3.5.** Arbeitspaket **AP3** wird von der Universität Mannheim geleitet, weswegen weitere Details über die Arbeit im Bericht der Universität Mannheim zu finden sind.

M3.3 Erste Version der Extraktionsbibliothek. Die hiermit erstellten Daten sind im Demonstrator geladen

Die erste Version der Extraktionsbibliothek wurde planmäßig im April 2016 zusammen mit Universität Mannheim fertiggestellt.

Der Schwerpunkt lag dabei darauf, die Möglichkeit zu untersuchen Tabellen aus HTML Dokumenten zu verarbeiten. Studien von Google zeigen, dass eine große Menge von Datentabellen auf Webseiten existieren, aber weniger als 1,1% der erfassten 14 Milliarden Tabellen als relationale Tabellen vorliegen. Die erste Priorität war, eine Methode zu entwickeln um diese relationalen Tabellen, bei denen Spalten Attribute repräsentieren und die Zeilen die konkreten Werte zu jedem Attribut repräsentieren, auszulesen.

M3.4 Zweite Version der Extraktionsbibliothek

Die zweite Version der Extraktionsbibliothek enthält Arbeiten zur Extraktion von tabellarischen Daten aus HTML Dokumenten. Diese können sowohl online als auch offline vorliegen. Mit der Veröffentlichung der „Web Table Extraction"-Erweiterung für RapidMiner wird der Meilenstein erfüllt. Details sind der Beschreibung von **AP3** (**T3.2**) zu entnehmen.

M3.5 Finale Version der Extraktionsbibliothek

Der finalen Version der Extraktionsbibliothek wurde um die Funktionalität zur Extraktion von tabellarischen Daten aus PDF Dokumenten erweitert. Damit können Daten sowohl aus online als auch offline verfügbaren PDF Dokumenten extrahiert werden. Mit der Veröffentlichung der „PDF Table Extraction"-Erweiterung für RapidMiner wird der Meilenstein erfüllt. Details sind der Beschreibung von **AP3** (**T3.2**) zu entnehmen.

T3.2 Weiterentwicklung von Extraktionsmechanismen für die als relevant erkannten Datenformate

Dieser Task behandelt die Entwicklung von Extraktionsmechanismen von Datenformaten, die eine hohe Relevanz in Bezug auf das Internet aufweisen.

Die Entwicklung begann planmäßig im Januar 2016. Diese Aufgabe umfasst Arbeiten im ersten und zweiten Projektjahr. Darüber hinaus arbeitete RapidMiner auch im dritten Jahr weiter an dieser Aufgabe, um weitere Verbesserungen zu entwickeln.

Das erste Jahr

Im ersten Jahr wurde der Java Quellcode der Universität Mannheim von RapidMiner auf Wikipedia Daten getestet. Diese Arbeit beinhaltet ein Klassifikationsmodell, welche von der Universität Mannheim auf einer großen HTML-Datensammlung trainiert wurde. Für den Quelltext einer Webseite kann das Modell eine Vorhersage treffen, ob der HTML-tag <table> eine relationale Tabelle beschreibt oder nicht. Der ursprüngliche Quellcode der Universität Mannheim konnte HTML Dokumente von einem lokalen Verzeichnis lesen.

Das zweite Jahr

Eines der Ziele des Tasks **T3.2** war die Entwicklung fortschrittlicher Methoden zur Extraktion tabellarischer Daten aus HTML-Dokumenten und deren Konvertierung in Formate, die durch RapidMiner verarbeitet werden können.

Ähnlich liegen viele online verfügbaren Daten in (herunterladbaren) PDF (Portable Document Format) Dokumenten vor. Das PDF-Format wird häufig als universelles Datenformat bezeichnet, da eine Vielzahl proprietärer und offener Formate in das Format konvertiert werden können. Zudem erlaubt es die Beeinflussung der Darstellungsqualität bei der Erzeugung. Aus diesem Grund ist das Format sehr populär geworden und wird meist als Format der Wahl für die Veröffentlichung wissenschaftlicher Ergebnisse verwendet. Um Zugriff auf Daten in PDF-Format zu erhalten, wurde im Rahmen des Tasks **T3.2** eine RapidMiner-Erweiterung zur Extraktion tabellarischer Daten aus PDF-Dokumenten entwickelt.

Auf Grund dieser Aspekte wurde in den Meilensteinen **M3.4** und **M3.5** der Fokus auf das HTML- und das PDF-Format gelegt, da die Gewinnung von Daten aus den genannten Formaten von großem kommerziellem und wissenschaftlichem Interesse ist. Heutige Extraktions-verfahren nutzen bisweilen dokumentspezifische Datengewinnungstechniken. Unser produktorientierter Zugang erlaubte die Entwicklung einer allgemeingültigeren Methodik der Datenextraktion. Durch die Nutzung von Algorithmen, können die Nachteile von regulären Ausdrücken umgangen werden, da die Abstraktionsstufe weiter erhöht werden kann.

Die entwickelten Algorithmen verwenden Methoden zur automatischen Auffindung und Extraktion relevanter Daten und nehmen somit dem Nutzer den Aufwand der Generierung Regulärer Ausdrücke ab. Das so designte Verfahren schwächt auch die enge Bindung an vordefinierte Dokumentstrukturen ab. Dadurch können beispielsweise Dokumente mit ähnlichen Strukturen in einer recht allgemeinen Art genutzt werden. Dieser grobkörnige Ansatz basiert jedoch auch auf gewissen Regeln und hat damit auch Limitierungen. Nichtsdestotrotz zeigt sich, auch beim Test mit Probanden (**M2.3**), eine Überlegenheit der neu entwickelten Extraktionsverfahren über bisher bestehende. Folglich kann objektiv von einer Verbesserung bestehender Techniken im Rahmen der Meilensteine **M3.4** und **M3.5** gesprochen werden. Als direkt anwendbare Ergebnisse dieser Arbeit, entstanden zwei RapidMiner-Erweiterungen („Web Table Extraction“ und „PDF Table Extraction“), die bereits im RapidMiner Marketplace veröffentlicht wurden.

„Web Table Extraction“-Erweiterung für RapidMiner

Die „Web Table Extraction“-Erweiterung für RapidMiner ist ein Ergebnis der Zusammenarbeit zwischen der Universität Mannheim und RapidMiner. Die Erweiterung wurde am 22.03.2017 veröffentlicht und am 23.03.2017, 28.03.2017, 29.03.2017, 24.07.2017, 26.07.2017, 19.10.2017, 30.01.2018 und zuletzt am 06.02.2018 um einige Verbesserungen erweitert und in RapidMiner Marketplace aktualisiert.

„Read HTML Table” Operator

Die Erweiterung stellt einen Operator namens „Read HTML Table“ zur Extraktion tabellarischer Daten aus lokal oder online verfügbaren HTML-Dokumenten bereit, also für die Extraktion von tabellarischen Daten aus Web-Seiten oder Intranet-Seiten. Dabei findet ein von der Universität Mannheim entwickelter und bereits in vorausgegangenen Projekten getesteter Algorithmus Anwendung. Diese Projekte umfassen Web Data Commons[5], sowie DWTC[6](Dresden Web Table Corpus), welche wiederum auf Daten aus dem Common Crawl Projekt[7], dem größten öffentlich verfügbaren Web-Korpus, zurückgreifen. RapidMiner erweiterte den Mechanismus zum Klassifizieren von HTML-Tabellen so, dass Webseiten direkt angegeben werden können, ohne dass sie lokal vorliegen müssen.

Der Algorithmus funktioniert wie folgt: In einem ersten Schritt wird ein HTML-Dokument in den Arbeitsspeicher geladen. Anschließend wird eine Gruppe von Vorverarbeitungsschritten durchgeführt, um Tabellen definierende Elemente zu identifizieren. Danach wird ein Klassifikationsmodel angewandt, mit dem Ziel zwischen Tabellen für das Layout und Daten enthaltenden (relationalen) Tabellen zu unterscheiden. Dieses Klassifikationsmodel wurde, im Rahmen der bereits genannten Projekte, auf einer Vielzahl von aufwendig aufbereiteten HTML-Datentabellen trainiert. Nach der Einordnung wird der Inhalt der relationalen Tabellen als eigenes Objekt hinterlegt. Dazu gehört der Tabellenkopf, die Datenzeilen, sowie benötigte Metadaten, wie der Tabellenname oder der Name der Internetseite, von der die Daten extrahiert wurden. Im Anschluss an all diese Schritte wird das interne Datenmodel in ein RapidMiner-ExampleSet umgewandelt, welches der tabellarischen Datenstruktur innerhalb RapidMiners entspricht.

Zur Bekanntmachung der Arbeitsergebnisse und zur Hervorhebung der Praktikabilität einer Extraktion tabellarischer Daten von Web-Seiten, wurde ein Blog-Beitrag mit dem Titel „The Web Table Extraction Operator“ in der RapidMiner-Community veröffentlicht. Dieser Beitrag zeigt, wie eine solche Extraktion im Kontext größerer Data-Mining-Szenarien oder einer kommerziellen Anwendung verwendet werden kann. Beiträge in der RapidMiner-Community sind auch für nicht registrierte Nutzer sichtbar. Der Blog-Beitrag ist unter folgendem Link erreichbar:

https://community.rapidminer.com/discussion/37353/the-web-table-extraction-operator

[5] Web Data Commons: Web-Link: http://webdatacommons.org
[6] Dresden Web Table Corpus: Web-Link: https://wwwdb.inf.tu-dresden.de/misc/dwtc
[7] Common Crawl Project: Web-Link: http://commoncrawl.org

„PDF Table Extraction“-Erweiterung für RapidMiner

Die „PDF Table Extraction“-Erweiterung für RapidMiner besteht aus dem „Read PDF Table“-Operator. Die Erweiterung wurde am 23.03.2017 veröffentlicht und am 29.03.2017, 24.07.2017, 26.07.2017, 30.01.2018 und zuletzt am 09.02.2018 um einige Verbesserungen erweitert und in RapidMiner Marketplace aktualisiert.

„Read PDF Table“ Operator

Mit dem „Read PDF Table“ Operator können PDF-Dokumente sowohl von einem lokalen als auch von einem Online-Pfad eingelesen werden. Dabei werden Tabellen in PDF-Dokumenten erkannt, extrahiert und in RapidMiner-Datentabellen („ExampleSets“) umgewandelt. Die automatische Erkennung der Tabellen wird mit dem Nurminen-Algorithmus[8] durchgeführt. Dabei werden sowohl einzelne als auch mehrere Regionen auf einer Seite des betrachteten Dokuments identifiziert, die tabellarische Daten enthalten können. Anschließend werden, je nach Bedarf, der Basis-Algorithmus und/oder die Tabellen-Extraktions-Algorithmen des Tabula-Java-Frameworks[9], basierend auf den zuvor bestimmten Koordinaten der Tabelle(n) innerhalb des Dokumentes, angewandt. Wie gefordert extrahiert der Operator lediglich Daten aus Tabellen und ignoriert dabei umliegende Dokumentinhalte. Dabei ist es möglich mehrere Tabellen in einem Schritt zu extrahieren, unabhängig davon, ob sich mehrere Tabellen auf einer Seite befinden, oder mehrere Seiten Tabellen enthalten.

Die gewonnenen Daten werden in RapidMiner-Datentabellen („ExampleSets“) umgewandelt und als Sammlung („Collection“) solcher Datentabellen in RapidMiner bereitgestellt. Die Anzahl der ExampleSets entspricht dabei der im Dokument gefundenen Tabellen. Da die Erkennungs-Algorithmen eine Kalibration (Anpassung) an das Dokument vornehmen und Dokumente unterschiedlichste Inhalte enthalten, kann es vorkommen, dass einzelne Tabellen als mehrere Tabellen erkannt werden. Jedoch enthält in diesem Fall oft eine der erzeugten Tabellen die gewünschten Daten.

Für die „PDF Table Extraction“ Erweiterung wurde auch ein Blog Beitrag geschrieben, der die Bequemlichkeit der Nutzung dieser Lösung zur effizienten Datengewinnung aus PDF Dokumenten zeigt. Im Beitrag wird gezeigt, wie der Operator, und damit der PDF-Datentabellen-Extraktionsvorgang, nahtlos in einen Data-Mining-Prozess integriert werden kann. Der Beitrag ist unter folgendem Link zu erreichen:

https://community.rapidminer.com/discussion/37490/pdf-table-extraction-extension-released

Das dritte Jahr

„Read HTML Tables“ Operator

Die „Web Table Extraction“ Erweiterung wurde aktualisiert um die gleichzeitige Extraktion von mehreren HTML Daten Tabellen („Bulk Extraction“) zu ermöglichen. Das ist nun mit Hilfe des neuen Operators „Read HTML Tables“ möglich. Zusätzlich wurde die Hilfsdokumentation verbessert und zwei Tutorial Prozesses zur Bulk Extraction wurden hinzugefügt. Dieser Operator extrahiert HTML Daten Tabellen aus einer Liste von Website URLs oder Datei

[8] Anssi Nurminen: Master Thesis: Web-Link: http://dspace.cc.tut.fi/dpub/bitstream/handle/123456789/21520/Nurminen.pdf
[9] Tabula Java Framework: Web-Link: https://github.com/tabulapdf/tabula-java

Pfaden zu HTML Dokumenten. Diese Liste kann als ExampleSet dem Operator übergeben werden.

Ein dazu verfasster Blog-Beitrag ist unter folgendem Link zu erreichen:

https://community.rapidminer.com/discussion/44089/using-ds4dm-and-web-table-extraction-extensions-for-google-table-and-html-table-extraction

„Read PDF Tables“ Operator

Die „PDF Table Extraction“ Erweiterung wurde ebenso aktualisiert um die Mehrfach-Extraktion (“Bulk Extraction”) von Daten Tabellen durch den neuen “Read PDF Tables” Operator zu ermöglichen. Weiterhin, basierend auf Kunden Feedback, wurden die Parameter „edge detection“ und „extraction algorithm“ als konfigurierbare Parameter hinzugefügt um den Nutzer die Möglichkeit zu geben die Extraktionsqualität zu optimieren.

Dieser Operator extrahiert PDF Datentabelle aus einer Liste von Website URLs oder Datei Pfaden zu PDF Dokumenten. Diese Liste kann als ExampleSet dem Operator übergeben werden. Mehrere Parameter des Operators (z.B. für Anpassung der Extraktionskriterien und Anpassung der Kantendetektionskriterien) können von User konfiguriert werden.

AP 4 Zugriff auf Intranet-Daten

Dieses Arbeitspaket wurde planmäßig im August 2016 begonnen und im April 2018 fertiggestellt. Es besteht aus den Meilensteinen **M4.1**, **M4.2**, **M4.3**, **M4.4**, **M4.5**, **M4.6** und **M4.7**. Diese wurden jeweils planmäßig erreicht.

M4.1 Erhebung über gängige Datenformate und -organisationsformen bei den Partnerunternehmen und weiteren Kunden von RapidMiner

Im Rahmen einer Erhebung zur Nutzung gängiger Datenformate und Datenorganisations-formen (Meilenstein **M4.1**) zeigte sich ein hoher Verbreitungsgrad folgender Datenformate:

- SQL-Datenbanken (werden bereits standardmäßig von RapidMiner unterstützt)
- Microsoft-Excel-Dateien (werden bereits standardmäßig von RapidMiner unterstützt)
- CSV-Dateien (werden bereits standardmäßig von RapidMiner unterstützt)
- Web-Seiten (non-parametric extraction of data tables from HTML-Format wird durch einen im Rahmen des Projektes neu entwickelten Operator von RapidMiner unterstützt)
- PDF-Dokumente (z.B. aus Microsoft Word oder Excel oder Open Office generiert; wird durch einen im Rahmen des Projektes neu entwickelten Operator von RapidMiner unterstützt)
- Google Spreadsheets (Online-Dokumente; werden durch einen im Rahmen des Projektes neu entwickelten Operator von RapidMiner unterstützt)

- Microsoft Office Dokumente (inkl. Online-Dokumente (Service-Name: Microsoft Office 365); Excel Online wird durch einen im Rahmen des Projektes neu entwickelten Operator von RapidMiner unterstützt)

Es ist zu berichten, dass weitere Formate, die im Rahmen der Umfrage berichtet wurden, nun von RapidMiner unterstützt werden. Diese sind durch Anmerkungen in der hier aufgeführten Liste hervorgehoben.

M4.2 Erste Version der Extraktionsbibliothek (gemeinsam mit T3.2)

Zur Erstellung einer ersten Extraktionsbibliothek wurden zusammen mit den Aufgaben aus Task **T3.2** Operatoren zum direkten Zugriff auf Daten aus Dokumenten erarbeitet. Dabei stellte sich heraus, dass moderne Unternehmensstrukturen einen Teil ihrer Daten online oder in Intranet-Infrastrukturen organisieren. Um diese Daten für die Datensuchmaschine greifbar zu machen, wurden weiterhin Zugänge zu diesen Datenquellen evaluiert. So sind Microsoft SharePoint und Informatica als wichtige Datenquellen zu nennen. Dies bildete die Grundlage für weitere Extraktionsmechanismen, die in den Meilensteinen **M4.3** bis **M4.5** erläutert werden.

M4.3 Konzept des Privacy-by-Design-Ansatzes

Eine zentrale Authentifizierungsstelle, sowie ein umfassendes Rechte-Management und die Unterstützung der gängigen Datenformate heben Microsoft Excel Online und Google Spreadsheets als potentielle Kandidaten für eine Privacy-by-Design getriebene Datenintegrationslösung hervor. Die Dienste ermöglichen Unternehmen gleichzeitiges Arbeiten an Dokumenten bei einer granularen Rechteverwaltung. Alle Dokumente sind standardmäßig vor unautorisierten Zugriffen geschützt und eine Freigabe von Dokumenten an weitere Mitarbeiter muss explizit veranlasst werden. Durch diese Konzepte wird ein Zugriff durch nicht autorisierte Dritte von Grund auf gewährleistet.

Im Rahmen des Projektes ist ein Konzept entwickelt worden, das durch direkten Zugriff auf Daten aus Dokumenten, die durch genannte Dienste verwaltet werden, alle Vorteile in ein Privacy-by-Design- Datenintegrationskonzept vereinigt. In Kombination mit dem hohen Verbreitungsgrad dieser Dienste bietet das Konzept Unternehmen eine Privacy-by-Design-Lösung, deren Integration in bestehende Systeme nur einen geringen Aufwand bedeutet.

M4.4 Zweite Version der Extraktionsbibliothek (gemeinsam mit T3.2)

In einer Weiterführung der Extraktionsbibliothek wurden Operatoren zur Extraktion von online verfügbaren tabellarischen Daten aus Google Spreadsheets und Excel Online entwickelt. Dies ermöglicht die Anbindung von Echtzeitdaten direkt an RapidMiner. Diese Operatoren sind Teil der „Spreadsheet Table Extraction“ Erweiterung für RapidMiner und sind detailliert in den Beschreibungen von Aufgaben **T4.1** und **T4.2**.

M4.5 Erster Prototyp der Indexierungkomponente für Unternehmensdaten inklusive Privacy-by-Design, für ausgewählte Datenformate und -organisationsformen

Dieser Meilenstein beinhaltet die Arbeiten zu Aufgabe **T4.2**, welche in der ersten Hälfte von Jahr 3 abgeschlossen wurden. Dies beinhaltet die „Sharepoint Connector“ Erweiterung, welche den Zugang zum Microsoft Cloud-Service Sharepoint erlaubt. Damit können Nutzer Dateien anzeigen und herunterladen, wobei die Zugriffsrechte über die Sharepoint-

Administratoren geregelt werden kann. Unterstützte Dateiformate können in ein RapidMiner ExampleSet eingelesen werden und im Anschluss als JSON Repräsentation zum Back-End hinzugefügt werden.

Die neuen Operatoren in der Repository Management Gruppe (**T4.2**) nutzen dieses Vorgehen um sowohl Sammlungen von ExampleSets ins Back-End zu laden, wo sie direkt indiziert und gespeichert werden. Dieses Feature ist Teil der Data Search Extension, welche am 19. Oktober 2017 veröffentlicht einzelne, als auch wurde und den Abschluss des Meilensteins im Oktober 2017 darstellt.

M4.6 Finale Version der Extraktionsbibliothek (gemeinsam mit T3.2)

Für den Meilenstein **M4.6**, hauptsächlich unter Aufgabe **T4.2**, wurden die Erweiterung („Web Table Extraction“, „PDF Table Extraction“ und „Spreadsheet Table Extraction“) im dritten Jahr weiter verbessert, in dem neue Operatoren hinzugefügt wurden, mehr Parameter zur Optimierung bereitgestellt werden und indem die Dokumentation und Beispielprozesse überarbeitet wurden.

Die überarbeiteten Versionen der „Web Table Extraction“ wurde zuletzt am 06.02.2018 aktualisiert. Die überarbeitete Version der „PDF Table Extraction“ am 09.02.2018 und die „Spreadsheet Table Extraction“ Erweiterung wurde zuletzt am 08.08.2017 aktualisiert. Alle sind frei über den RapidMiner Marketplace verfügbar. Weitere Überarbeitung sind auch über das offizielle Projektende hinaus denkbar, da diese Erweiterung bereits große Beliebtheit in der Community erreicht haben.

M4.7 Finaler Prototyp der Indexierungkomponente für Unternehmensdaten inklusive Privacy-by-Design, für alle als in M4.1 als relevant identifizierten Datenformate und -organisationsformen

In der zweiten Hälfte von drittem Jahr, wurde die Integration von RapidMiner Prozessen in Plattformen Dritter weiterverfolgt. Die Motivation dabei war, die Reichweite aller Arbeiten aus **AP1**, **AP2**, **AP3** und **AP4** zu erhöhen und außerdem auf die Nutzung von RapidMiner als Data Science Plattform aufmerksam zu machen. Als Ergebnis wurde der Zugang zu Unternehmensdaten erweitert durch die Entwicklung eines Adapters zu Informatica.

Für die Entwicklung hierfür ist RapidMiner eine strategische Partnerschaft mit Informatica eingegangen. Informatica ist eine der marktführenden Plattformen für Datenintegration. Die Absicht in DS4DM die Anschlussmöglichkeiten von RapidMiner zu erweitern führte zu der Entwicklung für einen Adapter zu Informaticas Cloud Umgebung. Dieser Adapter erlaubt es RapidMiner Prozesse als Webserviceaus dem Informatica Workflow heraus aufzurufen. Das Design nutzt einen auf Swagger basierende Beschreibung für RESTful Webservices mit JSON oder XML als Datenaustauschformate. Die Beschreibungen können automatisch generiert werden und können Anfragen und Dateneingaben während der Designphase verarbeiten und stellen diese mittels drag&drop dar. Zur Laufzeit werden die Daten von einem Vorgängerknoten (aus Informatica) nahtlos zum RapidMiner Web-Service übergeben werden und die Ausgaben wird ebenso nahtlos zum nächsten Knoten im Informatica Workflow weitergeleitet.

Zusammen mit der SharePoint Connector Erweiterung, welche im zweiten Jahr veröffentlicht wurde, sind die Arbeiten hinsichtlich Anbindungen an Unternehmensdaten nun weitestgehend abgeschlossen.

T4.1: Entwicklung von Adaptern für Unternehmensdatenquellen

Die Arbeit an diesem Teilarbeitspaket wurde plangemäß im August 2016 begonnen. Die Arbeiten zu dieser Aufgabe umfassten die Projektjahre zwei und drei.

Das zweite Jahr

Die Evaluation, welche Adapter für Unternehmensdatenquellen entwickelt werden sollen, bestand aus einer Umfrage unter ausgewählten Kunden und Partnern zur Identifikation der relevantesten Adapter und einer Diskussion mit dem Produktmanagement-Team von RapidMiner, um die Potentiale einzelner Adapter abzuschätzen.

„SharePoint Connector"-Erweiterung für RapidMiner

Zur Lösung der Aufgaben aus Task **T4.1** wurde eine neue RapidMiner-Erweiterung namens „SharePoint Connector" entwickelt. Diese fungiert als Adapter zu Microsoft SharePoint[10]. Microsoft SharePoint wird von vielen Unternehmen als Sammelstelle für Dokumente und Daten aller Art verwendet. Die Erweiterung besteht dabei aus zwei Operatoren „List SharePoint Files" und „Download from SharePoint". Diese Operatoren erfüllen die in Meilenstein **M4.4** geforderten Aufgaben. Die Erweiterung wurde am 27.07.2017 im RapidMiner Marketplace veröffentlicht und wurde am 08.08.2017 und zuletzt am 07.02.2018 aktualisiert.

„List SharePoint Files" Operator

Um eine Übersicht über Unternehmensdaten zu erhalten, die in einer SharePoint-Site-Instanz hinterlegt sind, kann der neu entwickelte „List SharePoint Files" Operator genutzt werden. Über die Nutzung des Operators kann eine RapidMiner-Datentabelle („ExampleSet") erzeugt werden, die Informationen über alle verwalteten Dateien und Ordner auf der aktuellen Site-Verzeichnisebene enthält. Diese Informationen beinhalten Metadaten über die Datei-Erstellung, deren relativen Pfad in der Site, eine Interne ID, sowie einen direkten Download-Link und die Information, ob es sich um einen Ordner handelt oder eine einfache Datei.

Der Operator benötigt ein Token, dieses ist identisch zu dem Token das für den „Read Excel Online" Operator (von „Spreadsheet Table Extraction" Erweiterung) verwendet wird, sowie den Namen der SharePoint Site und die URL der SharePoint Instanz des Unternehmens. Weiterhin kann ein spezifischer Unterordner als Suchdomäne angegeben werden und die Abfrage auf eine gewisse Anzahl an Dateien beschränkt werden.

„Download from SharePoint" Operator

Der RapidMiner-Operator „Download from SharePoint" wurde auch als Teil der RapidMiner-Erweiterung „SharePoint Connector" entwickelt. Dieser Operator ermöglicht das Herunterladen von Dateien von einer SharePoint-Site-Instanz auf einen anderen Computer. Für das

[10] Microsoft SharePoint. Web-Link: https://support.office.com/de-de/article/Was-ist-SharePoint-97B915E6-651B-43B2-827D-FB25777F446F

Herunterladen wird ein Token zur Authentifizierung benötigt. Hierbei kann ein Token, wie es bei dem „List SharePoint Files“ Operator gebraucht wurde, verwendet werden. Als Eingabe wird eine Liste an herunterzuladenden Dateien (im RapidMiner-Datentabellen-Format „ExampleSet“) erwartet. Ist diese Liste mit einem Token annotiert, wie es bei der vom Operator „List SharePoint Files“ ausgegebenen Liste der Fall ist, bedarf es keiner erneuten Token-Eingabe. Weiterhin wird ein Zielort als Parameter erwartet. Mit diesen Angaben kann eine Authentifizierung mit zentralen Microsoft Azure Active Directory Servern durchgeführt werden, um im Anschluss über die Microsoft Graph API die Daten anzufragen. Der Operator wandelt die empfangenen Datenströme (Byte-Streams) in Dateien um, und legt diese im Zielverzeichnis ab. Über eine Checkbox hat der Nutzer zudem die Wahl, ein Überschreiben bereits vorhandener Dateien zuzulassen, oder eine Namensänderung zu erzwingen.

Dieser Operator ermöglicht durch das Abfragen von Dateien aus gesicherten Unternehmens-SharePoint-Site-Instanzen eine Einbindung verschiedenster Datentypen in RapidMiner-Prozesse und somit durch das in **T4.2** gezeigte Konzept eine Indexierung seitens des Such-Back-Ends.

Es ist wichtig zu beachten, dass die Kombination aus Operatoren, die in Echtzeit Daten direkt aus Dokumenten auslesen, und Operatoren, die Dateien lokal verfügbar machen, einen soliden Grundbaustein für ein zukunftssicheres Datenintegrationskonzept bieten. Durch die direkte Integrationsmöglichkeit stark frequentierter Datenformate in Echtzeit erschließen sich gänzlich neue Anwendungsmöglichkeiten.

Das dritte Jahr

Unser primärer Ansatz zur Entwicklung von Adaptern für Quellen von Unternehmens-Daten fokussierte bisher auf die Einbindung der Daten in RapidMiner. In der ersten Hälfte von Jahr 3 untersuchten wir den umgekehrten Ansatz, die Bereitstellung von RapidMiner als externe Daten Quelle. Dadurch wird RapidMiner und seine Erweiterungen einer bereits etablierten Unternehmens-Nutzergruppe zur Verfügung gestellt. Dieser Ansatz unterstützt daher das Hauptziel des **AP4**, die Erweiterung von RapidMiners Zugriff auf Intranet Daten. Um diesem Ziel näher zu kommen wurden drei Ansätze zur Integration von RapidMiner mit Dritt-Anbieter Plattformen untersucht. Wir entschieden uns für Informatica als Testfall, da es als ein anerkannter Marktführer[11] innerhalb von Daten Integrationswerkzeugen gesehen wird.

„Informatica Connector für RapidMiner Prozesse”

Es wurden drei verschiedene Ansätze untersucht, wie RapidMiner Prozesse in Informatica Workflows integriert werden können:

i) Integration von RapidMiner Prozessen mit einem definierten Dateiaustausch

ii) ii) Webservices um Datenanfragen und Berechnungsergebnisse auszutauschen

iii) iii) RapidMiner Prozesse mittels Kommandozeile ausführen.

Nach Rücksprache mit den Entwicklern von Informatica, fiel die Entscheidung auf die Webservice basierte Lösung als Anbindung an Informatica Cloud.

[11] Gartner Magic Quadrant for Data Integration Tools, web-link: https://www.informatica.com/data-integration-magic-quadrant.html#fbid=boD9gK4EuAL

Die technischen Aspekte der Anbindung werden im Folgenden kurz erläutert:

1. Die Integration erfolgt über eine Schnittstelle, welche es erlaubt RapidMiner Prozesse (als REST Webservice) aus der Informaticas Cloud Abläufen (Mappings genannt). Damit werden nicht nur weitere Anwendungen und eine nachhaltige Nutzung der Projektarbeit ermöglicht, durch die Ausführung von Datensuche und -extraktion als Teil von Informatica Mappings, es ermöglicht auch das Ausführen von beliebigen RapidMiner Prozessen im Informatica Ökosystem.
2. Die Gestaltung wurde so generisch entworfen, dass jeder Webservice über diese Schnittstelle aufgerufen werden kann, mittels Swagger Spezifikationen[12] um die komplette Signatur des Webservice zu beschreiben. Die Swagger Spezifikationsdatei wird auch Webservice Descriptor genannt. Dieser Descriptor kann, mit Hilfe eines Informatica Hilfsprogramms, automatisch generiert werden. In einigen Sonderfällen sind lediglich kleine Änderungen zu machen.
3. Informatica Mappings werden über den Web-based Workflow Editor erzeugt. Dieser Ansatz ermöglicht es Arbeitsabläufe grafisch zu erstellen. Die Knoten werden in der Informatica Terminologie Transformationen genannt. Daten können aus beliebigen Transformationen gelesen werden und zu jeder Zieltransformation geschrieben werden, sofern ein passender Connector existiert. Daten können auch während mitten in einem Workflow gelesen und gespeichert werden durch sogenannte Mid-stream Transformationen.

Eine erste Version des Connectors wurde am 13.05.2018 auf zwei Produktionsumgebungen von Informatica veröffentlicht und soll Kunden beider Plattformen helfen ihre Vorhersagemodelle in die Informatica Cloud Arbeitsabläufe zu integrieren. Eine Demo wurde von RapidMiner auf der Informatica World[13] in Las Vegas, USA vorgestellt (22-24.05.2018). Eine gemeinsame. Pressemitteilung[14] wurde am 23.05.2018 veröffentlicht.

T4.2 Implementierung Privacy-by-Design-Ansatzes für die Datensuche im Unternehmen

Die Entwicklung begann planmäßig im Februar 2017 begonnen. Die Arbeiten an dieser Aufgabe umfassen die Projektjahre zwei und drei.

Das zweite Jahr

Das Konzept Privacy-by-Design beschreibt ein System, dass bei der Entstehung bereits den Aspekt des Datenschutzes berücksichtigt und eine Integration gewährleistet, die den Datenschutz inhärent werden lässt.

[12] Swagger Specification to describe RESTful webservices, web-link: http://docs.swagger.io/spec.html
[13] Informatica World Event 2018, web-link: https://www.informaticaworld.com
[14] Pressemitteilung über die Partnerschaft zwischen RapidMiner und Informatica, web-link: https://rapidminer.com/news-posts/rapidminer-and-informatica-bring-ai-powered-data-analytics-to-the-enterprise

„Spreadsheet Table Extraction"-Erweiterung für RapidMiner

Eine Analyse des Sachverhalts aus Task **T4.2** führte zu dem Ergebnis, dass Token-basierte Authentifizierungsmethoden der de-facto-Standard sind, um auf sichere Art Daten zu übertragen. Ein sehr bekanntes und erfolgreiches Beispiel ist dabei die Nutzung dieser Methodik zum Zugriff auf Online-Dokumente. Es zeigt sich ein großer Zuspruch sowohl im kommerziellen als auch im privaten Sektor zu dieser neuen Form der Dokumenten-verwaltung.

Dabei wird die Verwaltung der Zugriffsrechte und des Speichers an einen Dritten ausgelagert. Solche Verfahren bringen jedoch neue Fragen technischer und rechtlicher Natur mit sich. Meist wird dies durch rechtliche Absprachen auf der Service-Ebene zwischen Nutzer und Dienstanbieter gehandhabt. Dadurch bietet sich für Plattformanbieter wie RapidMiner die Möglichkeit, Daten aus solchen Diensten für Kunden verfügbar zu machen, ohne dabei eigene Zugriffsverwaltungssysteme implementieren zu müssen. Dies bringt beispielsweise den Vorteil mit sich, dass die jeweiligen Dienstanbieter bedingt durch ihre Spezialisierung eine höhere Expertise einbringen können, von der beide Seiten gleichermaßen profitieren. Bei Anbindung an solche Dienste müssen Plattformanbieter sich den Herausforderungen der Integration des Dienstes und der Erstellung eines praktikablen Interfaces stellen, wodurch dem Endproduktanwender viele Aufgaben abgenommen werden. Im Hinblick auf die genannten Dienste stellen sich Google und Microsoft als führende Anbieter heraus, die ein hohes Maß an Zuverlässigkeit und Funktionsumfang anbieten, um tabellarische Dokumente online programmatisch zu verwalten.

Zur Integration der genannten Online-Dokumente wurde die „Spreadsheet Table Extraction"-Erweiterung für RapidMiner entwickelt. Teil dieser Erweiterung sind die beiden Operatoren „Read Google Spreadsheet" zum Auslesen von Google-Spreadsheet-Daten, sowie „Read Excel Online" zum Auslesen von Excel-Online-Daten. Beide Operatoren authentifizieren sich mit den jeweiligen Diensten über ein Token-basiertes System. Zudem erlaubt die Einbindung dieser Dienste einen Echtzeit-Zugriff auf Unternehmensdaten.

Zur Zusammenführung unserer Ideen und der Erfüllung des Meilensteins **M4.3** wurde die „Spreadsheet Table Extraction"-Erweiterung entwickelt. Dieser Meilenstein deckt dabei sowohl Aufgaben aus Task **T4.1** als auch aus Task **T4.2** ab. Die Erweiterung wurde am 12.05.2017 im RapidMiner Marketplace veröffentlicht und enthielt zum Start den „Read Google Spreadsheet"-Operator. Der „Read Excel Online"-Operator wurde in einem Update am 26.07.2017 hinzugefügt.

Zur Bekanntmachung der Arbeitsergebnisse, wurde noch ein Blog-Beitrag mit dem Titel „The Spreadsheet Table Extraction – Extension Release" in der RapidMiner-Community veröffentlicht. Der Beitrag ist unter folgendem Link zu erreichen:

https://community.rapidminer.com/discussion/38864/the-spreadsheet-table-extraction-extension-release

„Read Google Spreadsheet" Operator

Google Spreadsheets sind eine zunehmend beliebte Datenquelle und ein Format, welches zum Abspeichern und Teilen privater, wie auch öffentlicher Daten Verwendung findet. Es ist Teil

des Google-Docs-Produktportfolios und weist eine hohe Integration mit dem Online-Speicherdienst Google Drive auf. Dieser Speicherdienst stellt eine Cloud-basierte Speicherlösung dar. Der Besitzer eines Google-Spreadsheet-Dokuments kann dieses Dokument leicht mit anderen Teilen und dabei granulare Lese- und Schreibrechte an Einzelpersonen, Gruppen oder die Allgemeinheit erteilen.

Der Nutzer gibt die Web-Adresse (URL) eines Google-Spreadsheets im „Spreadsheet URL"-Parameter an. Der Operator extrahiert dann intern die ID des Dokumentes. Da ein Spreadsheet Dokument mehrere Tabellenblätter enthalten kann, muss ein Tabellenblatt noch über den jeweiligen Namen im Parameter „Sheet Name" angegeben werden. Weiterhin kann der zu extrahierende Zellenbereich in der A1-Filternotation als Teil des Tabellenblattnamens angegeben werden. Der Name und der Zellbereich werden dabei durch ein „!" getrennt angegeben. Weiterhin muss noch der Pfad einer Datei angegeben werden, die das sogenannte Secret, also Nutzerinformationen zur Abfrage eines Tokens zur Authentifizierung, enthält. Dieses Secret liegt in einer JSON Datei vor und kann nach der Freischaltung des Google Accounts für die Verwendung des API-Zugangs, heruntergeladen werden.

Mit diesem Operator kann aus einem Online-Google Spreadsheet-Dokument ein RapidMiner-ExampleSet extrahiert werden.

„Read Excel Online" Operator

Microsoft Excel ist der de-facto-Industriestandard im Bereich Tabellenkalkulation. Im Rahmen der Nutzergruppe von Excel zeigt sich ein zunehmender Trend (Google-Suchtrends[15]) in Richtung Excel Online.

Um wertvolle Unternehmensdaten aus bestehenden Unternehmensstrukturen in Analyse- und Suchprozesse integrieren zu können, wurde ein RapidMiner-Operator zum Auslesen von Excel-Dokumenten entwickelt, der auf online abgelegte Dokumente zugreift. Der Operator „Read Excel Online" ist Teil der RapidMiner-Erweiterung „Spreadsheet Table Extraction". Ähnlich wie beim „Read Google Spreadsheet" Operator, sind folgende Schritte zur Datenabfrage notwendig:

1. Authentifizierung mit den Microsoft Servern.
2. Konfiguration des Operators über die Parameter: Zieldatei, Zieltabelle im Dokument und gegebenenfalls gewünschter Zellbereich.
3. Ausführung des RapidMiner-Prozesses zur Abfrage der Daten.

Mit diesen drei Schritten werden im Hintergrund komplexe Abfragen und Konzepte abgedeckt. Der Nutzer verifiziert seine Identität gegenüber Microsoft über eine zentrale Nutzerverwaltung (Microsoft Azure Active Directory). Diese Nutzerverwaltung verwendet modernste Techniken, um Unternehmensdaten vor unerlaubten Zugriffen zu schützen und die Authentizität eines Nutzers zu gewährleisten, während der Nutzer lediglich ein Token erhält, welchen er in den RapidMiner-Operator für die Authentifizierung eintragen muss.

[15] Trend in Richtung Excel Online. Web-Link: https://trends.google.com/trends/explore?cat=12&date=today%205-y,today%205-y&geo=,&q=%2Fm%2F0drzplk,%2Fm%2F052zb&hl=de&tz=-120

Die Repository Management Operatoren ermöglichen die „Data Search“ Erweiterung auch unter Real Welt Bedingungen effektiv einzusetzen. Die „Data Table Upload“ und „Data Tables Upload“ Operatoren benötigen ein Repräsentationsmodell um ExampleSets in die Datenstruktur, die von der Back-End API akzeptiert wird, zu konvertieren. In Aufgabe **T4.2** wurde die Modell Konvertierung implementiert, die hauptsächlich ExampleSets oder Collections von ExampleSets zu simplen Daten Transfer Objekten konvertiert. Dieser Beitrag schließt den Meilenstein **M4.5** ab, allerdings sind Aktualisierungen möglich, für den Fall, dass die API erweitert wird.

Wie bereits in der Beschreibung für Meilenstein **M4.6** genannt, unterstützen die Erweiterungen die im Jahr 2 (Aufgabe **T3.2**) entwickelt wurden, die Daten Extraktion von privaten und unternehmerischen Datenspeichern wie zum Beispiel SharePoint. In der ersten Hälfte des dritten Projektjahres, haben wir einen größeren Aufwand in diese Aufgabe (**T4.2**) investiert um die Erweiterungen weiter zu aktualisieren, speziell die „Web Table Extraction“ (letztes Update am 06.02.2018), die „PDF Table Extraction“ (letztes Update am 09.02.2018) und „Spreadsheet Table Extraction“ (letztes Update am 07.02.2018) Erweiterungen.

AP 5 Dissemination und Verwertung

Dieses Arbeitspaket wurde planmäßig im Oktober 2015 begonnen und im Juli 2018 fertiggestellt. Es besteht aus den Meilensteinen **M5.1**, **M5.2** und **M5.3**. Diese wurden jeweils planmäßig erreicht.

M5.1 Projektwebseite mit initialem Inhalt

Für die offizielle Projektwebseite hat RapidMiner eine Cloud-basierte Infrastruktur (Amazon S3 bucket) mit den Domainnamen des Projekts (http://ds4dm.de und http://ds4dm.com) eingerichtet. Der Code der Website wird durch einen statischen Site-Generators (Jekyll) generiert, und auch in Bitbucket Repositories gespeichert.

M5.2 Projektwebseite um Informationen zum Online Demonstrator und zu den Pilotprojekten erweitert

Die Projekt Website wurde regelmäßig aktualisiert, um die Öffentlichkeit immer direkt über Neuigkeiten rund um das DS4DM Projekt zu informieren. Im zweiten Projektjahr wurden dementsprechend mehrere neue Beiträge (Blog Posts) veröffentlicht. Diese Beiträge enthalten zudem Hinweise zu Beiträgen aus der RapidMiner-Community, die die jeweiligen Erweiterungen im Detail vorstellen und sie in einem passenden Szenario kontextual einordnen.

M5.3 Nutzer-Community aufgebaut

Dieser Meilenstein beinhaltet die Aufgaben **T5.1** und **T5.2**, welche sich mit dem Aufbau und der Pflege einer Nutzer-Community, der Pflege der offiziellen Projektwebseite und der Öffentlichkeitsarbeit rund um das Projekt beschäftigen.

Der Aufbau der Community läuft seit dem zweiten Projektjahr und alle veröffentlichen Blogposts sind im RapidMiner Community-Portal veröffentlich, mit über 471.00 registrierten Benutzern und öffentlich über Google durchsuchbar ist. Im dritten Jahr nahm RapidMiner an einigen Veranstaltungen und Konferenzen teil um die Ergebnisse des DS4DM Projektes zu präsentieren. Dies beinhaltete Vorträge, Demonstratoren und Poster-Session. Ein gemeinsamer wissenschaftlicher Artikel wurde gemeinsam mit der Universität Mannheim geschrieben und erfolgreich bei der LWDA 2018 Konferenz eingereicht.

Diese Aktivitäten haben dazu beigetragen die Aufmerksamkeit für das Projekt zu wecken und die Ergebnisse zu teilen. Die Teilnahme an Konferenzen, sowie aktualisierte Versionen der Erweiterungen oder veröffentlichte Artikel wurden regelmäßig über die offizielle Projektseite veröffentlicht.

T5.1 Aufbau und Aktualisierung der Projekt-Website

Die Entwicklung begann planmäßig im August 2015 begonnen. Die Arbeiten an dieser Aufgabe umfassen alle drei Projektjahre.

Um eine möglichst breite Öffentlichkeit über den Fortschritt des DS4DM-Projektes informiert zu halten, wird eine Website auf Deutsch und Englisch gepflegt. Die deutsche Seite ist unter der Domain http://ds4dm.de erreichbar, während die Nutzung der Endung „.com“ (http://ds4dm.com) auf die englisch sprachige Seite verlinkt. Auf der Seite werden Informationen zu anstehenden und abgeschlossenen Aktivitäten bzgl. des Projekts veröffentlicht. Dadurch soll eine breite Öffentlichkeit auf die neusten Ergebnisse aufmerksam gemacht werden. Mit diesen Ergebnissen wurde der Meilenstein **M5.2** erfüllt.
Während des ersten und zweiten DS4DM-Projektjahres wurden 16 Beiträge publiziert.
Im dritten Jahr wurde die Webseite weiter gepflegt und 11 neue Blog- und Nachrichteneinträge über die Projektaktivitäten veröffentlicht.

T5.2 Aufbau und Unterstützung der Nutzer-Community

Das RapidMiner Community Portal[16] (community.rapidminer.com) hat mehr als 471.000 angemeldete Nutzer. Wir verbreiten Informationen über das DS4DM Projekt über dieses Portal. Diese Initiative ist bereits im Jahr 2 gestartet worden. Wir erreichen dies durch die Veröffentlichung von Blog Artikeln über die Nutzung der Erweiterungen, wenn diese veröffentlicht oder aktualisiert werden. Dieser Community Aufbau und die Informationsverbreitung führten zu einem erhöhten Interesse am Projekt und wir konnten die folgenden Ergebnisse bezüglich der Adaption unserer Resultate verzeichnen.

Eine wichtige Eigenschaft des RapidMiner-Community-Portals ist die Verwendung von SEO (Search Engine Optimization)-Techniken. Dies bedeutet, dass Beiträge der RapidMiner-Community zeitnah in Suchergebnissen von Suchmaschinen wie Google auftauchen und somit auch die Ergebnisse des DS4DM Projektes einfacher gefunden werden können, da sie mit gleicher Wichtigkeit wie andere Beiträge der aktiven Community behandelt werden. Während des zweiten DS4DM-Projektjahres begannen die Arbeiten des Community-Aufbaus durch eine erste Veröffentlichung von fünf Beiträgen zum DS4DM-Projekt. Diese Beiträge stellen

[16] RapidMiner-Community-Portal: Web-Link: https://community.rapidminer.com

veröffentlichte RapidMiner-Erweiterungen und mögliche Anwendungsszenarien rund um entwickelte RapidMiner-Erweiterungen dar. So lernt ein Leser etwas über das DS4DM-Projekt und erfährt gleichzeitig, wie die erzielten Ergebnisse für ihn in der Praxis anwendbar sind. Dabei wird beispielsweise die Datenextraktion, -anreicherung und -integration im Kontext eines Data-Mining-Prozesses dargestellt. Beiträge enthalten meist Diagramme und ausführbare Prozessdateien, um den Lernprozess des Nutzers bestmöglich zu unterstützen und eine schnelle Übernahme unserer Ergebnisse in die industrielle Praxis zu gewährleisten.

Veröffentlichungen in RapidMiner-Community-Beiträgen

Links zu den veröffentlichten Beiträgen sind im Folgenden aufgeführt:

- „The 'Data Search for Data Mining' – Extension Release!":

  Link: https://community.rapidminer.com/discussion/38231/the-data-search-for-data-mining-extension-release

- „The Web as a new data source for RapidMiner":

  Link: https://community.rapidminer.com/discussion/43306/the-web-as-a-new-data-source-for-rapidminer

- „The Web Table Extraction Operator":

  Link: https://community.rapidminer.com/discussion/37353/the-web-table-extraction-operator

- „PDF Table Extraction Extension Released!":

  Link: https://community.rapidminer.com/discussion/37490/pdf-table-extraction-extension-released

- „Using DS4DM and Web Table Extraction extensions for Google Table and HTML table extraction":

  Link: https://community.rapidminer.com/discussion/44089/using-ds4dm-and-web-table-extraction-extensions-for-google-table-and-html-table-extraction

- „The Spreadsheet Table Extraction – Extension Release":

  Link: https://community.rapidminer.com/discussion/38864/the-spreadsheet-table-extraction-extension-release

- „Finding the right data is crucial (Einladung der Community Nutzer zur Teilnahme an der Pilot Studie)":

  Link: https://community.rapidminer.com/discussion/39024/finding-the-right-data-is-crucial-help-us-improve-new-data-discovering-techniques

- „Connect to any document within your SharePoint”:

  Link: https://community.rapidminer.com/discussion/40582/connect-to-any-document-within-your-sharepoint

- „Reading Excel files directly from your company's OneDrive”:

  Link: https://community.rapidminer.com/discussion/40587/reading-excel-files-directly-from-your-companies-onedrive

- „Highlights of the Extension Updates”:

  Link: https://community.rapidminer.com/discussion/40589/highlights-of-the-extension-updates

RapidMiner-Operator-Nutzungs-Statistiken

Diese Statistiken bieten Feedback über die Nutzung der entwickelten RapidMiner-Erweiterungen wie zum Beispiel die Anzahl an Downloads bis 23.01.2019 und wie oft einzelne RapidMiner-Operatoren durch Nutzer ausgeführt wurden. Die Statistiken werden im Folgenden gelistet:

Download-Anzahlen der RapidMiner-Erweiterungen:

- Data Search for Data Mining Erweiterung: **4788** Downloads
- Web Table Extraction Erweiterung: **5012** Downloads
- PDF Table Extraction Erweiterung: **3605** Downloads
- Spreadsheet Table Extraction Erweiterung: **3570** Downloads
- SharePoint Connector Erweiterung: **1168** Downloads

Anzahl der Operator-Ausführungen

Die Zahlen repräsentieren die Anzahl der absoluten Ausführungen von Nutzern (exklusive der Projektpartner) seit der ersten Veröffentlichung bis zum 22.01.2019

Die Anzahlen sind wie folgt:

- Read HTML Table Operator: **7,283** Ausführungen
- Read Google Spreadsheet Operator: **5,743** Ausführungen
- Read PDF Table Operator: **4,183** Ausführungen
- Read PDF Tables Operator: **2,029** Ausführungen
- Read HTML Tables Operator: **1,601** Ausführungen
- Google Table Search Operator: **443** Ausführungen
- Data Search Operator: **258** Ausführungen
- Translate Operator: **231** Ausführungen

- List SharePoint Files: **198** Ausführungen
- Fuse Operator: **188** Ausführungen
- Read Excel Online Operator: **128** Ausführungen
- Advanced Fuse Operator: **41** Ausführungen
- Download from SharePoint: **37** Ausführungen
- Enrich Table by Data Fusion Operator: **30** Ausführungen
- Create Repository Operator: **10** Ausführungen
- Create Correspondences Operator: **4** Ausführungen
- Data Table Upload Operator: **2** Ausführung
- Unconstrained Search Operator: **2** Ausführung
- Correlation-Based Search Operator: **1** Ausführung
- Data Tables Upload Operator: **1** Ausführung

Die Zahlen zeigen eine wachsende Nutzung der neuen Operatoren seit dem letzten DS4DM-Zwischenbericht für das dritte Projektjahr.

AP 6 Projektmanagement

Dieses Arbeitspaket wurde planmäßig im August 2015 begonnen und im Juli 2018 fertiggestellt. Es besteht aus den Meilensteinen **M6.1, M6.2**, **M6.3**, **M6.4** und **M6.5**. Diese wurden jeweils planmäßig erreicht.

M6.1 Kick-Off-Workshop in Dortmund

Mit dem Kick-Off Meeting in Dortmund am 25. August startete das neue Projekt „Datensuche für Data Mining (DS4DM)“. Anwesend waren von der Universität Mannheim Prof. Dr. Christian Bizer, Prof. Dr. Heiko Paulheim und noch zwei Kollegen, und von RapidMiner Sabrina Kirstein und Dr. Simon Fischer.

M6.2 Gemeinsame Entwicklungsinfrastruktur aufgebaut

RapidMiner hat in September 2015 eine gemeinsame Infrastruktur eingerichtet. Dazu gehört ein Bitbucket (Git basiert) Projekt, um Code zu teilen und zu verwalten, ein gemeinsamer Google Drive Ordner, eine Mailing-Liste, ein Logo und eine Präsentationsvorlage.

M6.3 Projektworkshop in Dortmund

Der Meilenstein **M6.3** (gemeinsames Meeting der Projektpartner) wurde von August 2016 auf den 7. Oktober 2016 verschoben, da die bisherige Projektmitarbeiterin Sabrina Kirstein das Unternehmen RapidMiner auf eigenen Wunsch hin verlassen hat und ihr Nachfolger Dr. Edwin Yaqub das Team erst seit September 2016 verstärkt. Am 7. Oktober 2016 fand dieses Treffen der Projektpartner bei RapidMiner in Dortmund statt. Anwesend waren von der

Universität Mannheim Dr. Anna Lisa Gentile und Prof. Dr. Christian Bizer (über Skype) und von RapidMiner David Arnu, Dr. Edwin Yaqub, Edin Klapic und Ralf Klinkenberg.

M6.4 Projektworkshop in Mannheim

Die Universität Mannheim veranstaltete ein Arbeitstreffen im Zeitraum vom 28.-29.07.2017. Die Teilnehmer des Treffens waren: Prof. Dr. Chris Bizer, Prof. Dr. Heiko Paulheim und Benedikt Kleppmann (seitens der Universität Mannheim), sowie: Ralf Klinkenberg, Dr. Edwin Yaqub und Philipp Schlunder (seitens der RapidMiner GmbH).

M6.5 Abschlussworkshop in Dortmund

Dieser Meilenstein umfasst den finalen Workshop (Aufgabe **T6.1**), welcher von RapidMiner am 30.05.2018 veranstaltet wurde. Die Kollegen Universität Mannheim nahmen an dem Workshop ebenfalls teil. Gemeinsam wurde der Status des Projektes und der erreichten Meilensteinte für das letzte Projektjahr besprochen. Die Partner waren alle zufrieden mit gelieferten Ergebnissen und sprachen sich für künftige Zusammenarbeiten aus.

T 6.1 Organisation der Zusammenarbeit

Die Entwicklung begann planmäßig im August 2015 mit einem ersten Projekt Kick-off Workshop.
Diese Aufgabe beinhaltet alle Workshops, die während der drei Projektjahre gehalten wurden.

Die Zusammenarbeit zwischen RapidMiner und der Universität Mannheim ist eng und funktioniert sehr gut. Die Arbeiten werden durch wöchentliche Skype-Meetings koordiniert. Beide Seiten tauschen sich regelmäßig in diesen wöchentlichen Skype-Meetings und über E-Mails aus. Außerdem wurden verschiedene Projekttreffen durchgeführt.

- Am 25. August 2015 fand ein erste Projekt Kick-Off Meeting mit allen Partnern statt.
- Am 21. und 22. Januar 2016 wurde ein zweitägiges Projekttreffen aller Projektbeteiligten an der Uni Mannheim durchgeführt.
- Am 7. Oktober 2016 wurde ein Projekttreffen aller Projektbeteiligten bei RapidMiner in Dortmund durchgeführt (Meilenstein **M6.3**).
- Am 28. und 29 Juni 2017 wurde ein Projekttreffen aller Projektbeteiligten an der Universität Mannheim durchgeführt (Meilenstein **M6.4**).
- Am wurde 30. Mai 2018 ein Projekttreffen aller Projektbeteiligten bei RapidMiner in Dortmund durchgeführt (Meilenstein **M6.5**).

T6.2 Aufbau und Unterhaltung gemeinsam genutzter Infrastruktur

Die Entwicklung begann planmäßig im August 2015. Die Arbeiten an dieser Aufgabe umfassen alle drei Projektjahre.

Die geteilte Infrastruktur zur Verwaltung des Projektes wurde bereits zu Beginn des Projektes aufgesetzt. Es besteht aus github[17] und Bitbucket[18] Repositories zur Verwaltung von Quelltext, einem Google Drive Konto zur Verwaltung gemeinsamer Dokumente und Grafiken (wie Projekt Logos, Operator Grafiken, und weiteren), sowie Präsentationsvorlagen. Weiterhin wurden mehrere Poster im Verlauf des zweiten Projektjahres entworfen und auf Konferenzen vorgestellt. Die Universität Mannheim stellt weiterhin technische Informationen zum Back-End auf der Internetseite[19] ihrer Arbeitsgruppe bereit. Abschließend ist zu erwähnen, dass das RapidMiner Community gemeinsam zur Veröffentlichung von Erweiterungen, Blog Beiträgen, Dokumentationen und zum Aufbau einer gemeinsamen Nutzer-Community genutzt wird.

Publikationen und Präsentationen

RapidMiner und die Universität Mannheim beteiligten sich an folgenden Veröffentlichungen und Präsentationen des ersten DS4DM-Demonstrators:

- Edwin Yaqub und Ralf Klinkenberg: DS4DM-Poster und DS4DM-Demonstrator (Live Demo des finalen DS4DM-Demonstrators) bei der 6. BMBF Fachtagung „KMU-innovativ: IKT - Mittelstandskonferenz“ am 19.-20. November 2018 in Berlin.
    - Program Link: https://www.softwaresysteme.pt-dlr.de/de/mittelstandskonferenz-2018.php
    - Link zum Datenintegration Track: https://www.softwaresysteme.pt-dlr.de/de/forschungsvorhaben-kmu.php

- Benedikt Kleppmann, Christian Bizer, Edwin Yaqub, Fabian Temme, Philipp Schlunder, David Arnu and Ralf Klinkenberg: „Density- and Correlation-based Table Extension”. Lernen. Wissen. Daten. Analysen. (LWDA 2018), Mannheim, August 2018.
    - Diese gemeinsame Publikation der Universität Mannheim und RapidMiner wurde auch mit einem Poster passend zu dem Vortag der Konferenz vorgestellt am 22. August 2018.
    - Veröffentlichung Link: http://ceur-ws.org/Vol-2191/paper23.pdf
    - Program Link: https://www.uni-mannheim.de/lwda-2018/program/day-2-thursday-238/#c75092

[17] Github Repositories for DS4DM Back-End, Web-Links: https://github.com/BenediktKleppmann/DS4DM-Backend and https://github.com/AnLiGentile/DS4DM
[18] BitBucket Repositories for the Front-End (Extensions and Website code), Web-Link: https://bitbucket.org/ds4dm
[19] University of Mannheim webpage for DS4DM, Web-Link: http://web.informatik.uni-mannheim.de/ds4dm

- David Arnu: Vortrag „Smart data through automatic data search and extraction". Auf der PAPIs 2018 Konferenz am 06. April, 2018 in London, UK.
    - Link: https://papiseurope2018.sched.com/event/Dope/smart-data-through-automatic-data-search-and-extraction

- Edwin Yaqub, Philipp Schlunder, Ralf Klinkenberg, et al.: „Realizing Smart Data by Automating Tabular Search, Integration, and Extraction Methods", bei der 2. BMBF "Big Data All Hands Meeting and 2nd Smart Data Innovation Conference", Karlsruher Institute of Technology (KIT), am 11.-12. Oktober 2017 am KIT in Karlsruhe.
    - Edwin Yaqub und Philipp Schlunder hielten die Präsentation und Demonstrationen.
    - Program Link: https://indico.scc.kit.edu/event/303/timetable/#20171012
    - Link: https://indico.scc.kit.edu/indico/event/303/contribution/28

- RapidMiner nahm an der „1st Industrial Data Science Conference (IDS 2017)" teil: Die IDS 2017 Konferenz wurde von RapidMiner in Dortmund zusammen mit der Technischen Universität Dortmund organisiert. Die Konferenz fand am 05.09.2017 statt.
    - Von den RapidMiner Angestellten, die am DS4DM Projekt beteiligt sind, haben die folgenden an der Konferenz teilgenommen: Dr. Edwin Yaqub, Philipp Schlunder, David Arnu, Dr. Fabian Temme und Ralf Klinkenberg. Die Konferenz wurde genutzt um Informationen über RapidMiner und seine Forschungsprojekte, inklusive DS4DM, zu verbreiten.
    - Link: http://ids2017.rapidminer.com/about.html

- Petar Petrovski, Christian Bizer: „Extracting Attribute-Value Pairs from Product Specifications on the Web". International Conference on Web Intelligence (WI2017), Leipzig, 23.-26. August, 2017.
    - Veröffentlichung Link: https://pdfs.semanticscholar.org/2b80/2806d964bae95b312cc22e728bd202fd7a5e.pdf

- Edwin Yaqub, Ralf Klinkenberg, et al. DS4DM-Poster „Automated Mechanisms to Discover and Integrate Data from Web-based Tabular Collections", bei der 19. General Online Research Konferenz (GOR 17) am 15.-17. März 2017 in Berlin.

    - Poster Link: https://www.gor.de/gor17/index.php?page=downloadPaper&filename=Yaqub-Automated_Mechanisms_to_Discover_and_Integrate_Data-275.pdf&form_id=275&form_version=final

- Anna Lisa Gentile, Petar Ristoski, Steffen Eckel, Dominique Ritze und Heiko Paulheim (2017): „Entity Matching on Web Tables: A Table Embeddings Approach for Blocking" in „Advances in Database Technology - EDBT 2017: 20th International Conference on Extending Database Technology", Venedig, Italien, 21.-24. März 2017, Proceedings, Seiten 510-513. OpenProceedings, Konstanz, 2017:
    - Veröffentlichung Link: http://www.openproceedings.org/2017/conf/edbt/paper-367.pdf
    - Program Link: http://edbticdt2017.unive.it/?detailed_program

- Anna Lisa Gentile (2016): Vorstellung der Ideen des DS4DM-Projekts auf dem 4. Workshop „LD4IE (Linked Data for Information Extraction)“ auf der 15. International Semantic Web Conference (ISWC 2016), 18. Oktober 2016, Kobe, Japan:
    - Program Link: http://ds4dm.de/2016/10/31/workshop/
    - Program Link: http://web.informatik.uni-mannheim.de/ld4ie2016/LD4IE2016/Overview.html

- Edwin Yaqub und Ralf Klinkenberg: DS4DM-Poster und DS4DM-Demonstrator (Live Demo des ersten DS4DM-Demonstrators) bei der 5. BMBF Fachtagung „KMU innovativ: IKT“ – „Mittelstand: Digital. Innovativ. Vernetzt.“ am 10.-11. Oktober 2016 in Hannover.
    - Program Link: http://www.softwaresysteme.pt-dlr.de/de/fachtagung-2016.php

- Anna Lisa Gentile und Heiko Paulheim: Vortrag „Extending RapidMiner with Data Search and Integration Capabilities“ über das DS4DM-Projekt und erste Projektergebnisse beim „Rhein-Neckar Smart Data Meetup“ am 23. Juni 2016 in Mannheim.

    - Web-Link: http://www.meetup.com/de-DE/Rhein-Neckar-Smart-Data-Meetup/events/231410104/?eventId=231410104

- Anna Lisa Gentile, Sabrina Kirstein, Heiko Paulheim und Christian Bizer (2016): „Extending RapidMiner with Data Search and Integration Capabilities” bei der 13. Extended Semantic Web Conference (ESWC 2016), 29. Mai 2016 – 2. Juni, 2016, Heraklion, Crete, Greece. In Proceedings von Springer, 2016.
    - Die DS4DM Demo mit der ersten Version der RapidMiner Data Search Extension gewann den Best Demonstration Award bei der ESWC 2016:
        - Beschreibung der Demo (Paper): http://ub-madoc.bib.uni-mannheim.de/40718/1/DataSearchDemo.pdf
        - Poster Link: http://dws.informatik.uni-mannheim.de/fileadmin/lehrstuehle/ki/pub/Gentile-etal-RapidminerDataSearch-Poster-ESWC2016.pdf
        - News inkl. Link zum Video: http://ds4dm.de/2016/05/27/escw/
        - News inkl. Link zum Video: http://ds4dm.de/2016/06/07/escw-winner-2016/
        - Video der Demonstration: https://www.youtube.com/watch?v=i2g71G0jdmw
        - ESWC 2016 Best Demo Award: http://dws.informatik.uni-mannheim.de/en/news/singleview/detail/News/rapidminer-data-search-extension-wins-eswc2016-best-demo-award/
        - ESWC 2016: Accepted Posters & Demos.
          Program Link: http://2016.eswc-conferences.org/program/posters-demos.html

- Weitere Ansätze zur Verbreitung der Bekanntheit des Projektes Blog Artikel die im RapidMiner Community Portal veröffentlicht wurden beinhalten eine Referenz auf die offizielle Projektwebseite (http://ds4dm.de und http://ds4dm.com). Neuigkeiten bezüglich der Projektpartner werden auch auf der Webseite veröffentlich. Um eine zusätzliche Zugkraft zu generieren twittern wir auch DS4DM relevante Aktivitäten sowohl über offizielle als auch private Twitter Accounts.

DS4DM-Resultate in einem weiteren BMBF-geförderten Forschungsprojekt

Eine der im Projekt DS4DM entwickelten RapidMiner-Erweiterung wurde auch in einem weiteren BMBF-geförderten Forschungsprojekt namens „STEPS“ verwendet, was die breite Anwendbarkeit und Übertragbarkeit der in DS4DM entwickelten Methoden zeigt:

- Das STEPS-Projekt (Socio-Technical design and introduction of Cyber-Physical Production Systems in SMEs) entwickelt Strategien und Lösungen, um Cyber-Physical Production Systems (CPPS) und Lösungen in Kleine und Mittlere Unternehmen (KMU) einzuführen.
    - Web-Link: http://www.steps-projekt.de

- Konkret wurde die DS4DM-RapidMiner-Erweiterung „Web Table Extraction“ verwendet, um im STEPS-Projekt eine Bulk Extraktion von Produkt-Katalog-Daten (von Webseiten) für einen Industrie-4.0-Anwendungsfall durchzuführen. Diese Anwendung unserer Arbeit erhöht die Nachhaltigkeit der Arbeit, die im DS4DM-Projekt sowohl von der Universität Mannheim als auch von RapidMiner eingebracht wurde.

Stand des Vorhabens im Vergleich zur ursprünglichen Planung

Bei dem Vorhaben gab es **keine** großen Abweichungen bezüglich der Planung.
Alle Meilensteine wurden plangemäß erreicht.

Relevante F&E-Ergebnisse Dritter

Bei der unter Beachtung von Nr. 6.1 NKBF durchgeführten Informationsrecherche sind während des Projektzeitraums **keine** für die Durchführung des Vorhabens relevanten F&E-Ergebnisse Dritter bekannt geworden.

Jährliche Fortschreibung des Verwertungsplans

Es gab **keine Änderung** des Verwertungsplans. Der Verwertungsplan wurde **ohne Änderungen** wie im Antrag dargelegt ausgeführt.

Literaturliste und Veröffentlichungen

[1] DS4DM offizielle Projekt-Web-Seite (auf deutsch): Web-Link: http://ds4dm.de
(Englische Version: http://ds4dm.com oder http://ds4dm.de/en)

[2] RapidMiner-Marktplatz (*Marketplace*): Web-Link: https://marketplace.rapidminer.com

[3] RapidMiner-Community-Portal: Web-Link: https://community.rapidminer.com

[4] RapidMiner-Erweiterung (*Extension*) "*Data Search for Data Mining*" (DS4DM): Web-Link: https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId=rmx_data_search

[5] RapidMiner-Erweiterung (*Extension*) "*Web Table Extraction*": Web-Link: https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId=rmx_web_table_extraction

[6] RapidMiner-Erweiterung (*Extension*) "*PDF Table Extraction*": Web-Link: https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId=rmx_pdf_table_extraction

[7] RapidMiner-Erweiterung (*Extension*) "*Spreadsheet Table Extraction*": Web-Link: https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId=rmx_spreadsheet_table_extraction

[8] RapidMiner-Erweiterung (*Extension*) "SharePoint Connector": Web-Link: https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId=rmx_sharepoint_connector

[9] Pressemitteilung über die Partnerschaft zwischen RapidMiner und Informatica: web-link: https://rapidminer.com/news-posts/rapidminer-and-informatica-bring-ai-powered-data-analytics-to-the-enterprise

[10] Github-Repositories für das DS4DM-Backend, Web-links: https://github.com/BenediktKleppmann/DS4DM-Backend und https://github.com/AnLiGentile/DS4DM

[11] BitBucket-Repositories für das DS4DM-Front-End: Web-Link: https://bitbucket.org/ds4dm

[12] Universität Mannheim: Web-Seite: http://web.informatik.uni-mannheim.de/ds4dm

[13] STEPS Projekt (Socio-Technical design and introduction of Cyber-Physical Production Systems in SMEs): Web-Link: http://www.steps-projekt.de

[14] Zeyuan Shang, Emanuel Zgraggen, Benedetto Buratti, Ferdinand Kossmann, Philipp Eichmann, Yeounoh Chung, Carsten Binnig, Eli Upfal, Tim Kraska: "Democratizing Data Science through Interactive Curation of ML Pipelines", ACM SIGMOD/PODS International Conference on Management of Data, 30.06-05.07.2019, Amsterdam, The Netherlands. Web-Link: https://sigmod2019.org/

[15] Edwin Yaqub und Ralf Klinkenberg: DS4DM-Poster und DS4DM-Demonstrator (Live Demo des finalen DS4DM-Demonstrators) bei der 6. BMBF Fachtagung „KMU-innovativ: IKT - Mittelstandskonferenz" am 19.-20. November 2018 in Berlin. Program Link: https://www.softwaresysteme.pt-dlr.de/de/mittelstandskonferenz-2018.php

[16] Benedikt Kleppmann, Christian Bizer, Edwin Yaqub, Fabian Temme, Philipp Schlunder, David Arnu, and Ralf Klinkenberg: "Density- and Correlation-based Table Extension", Tagungsbeitrag, Workshop on "Large-scale Data Management and Processing – Applications in Research and Industry" bei der Konferenz "Lernen. Wissen. Daten. Analysen. (LWDA 2018)"

/ “Knowledge Discovery, Data Mining, Machine Learning (KDML 2018)” der Gesellschaft für Informatik (GI e.V.), Mannheim, 22-24.08.2018. In R. Gemulla (Editor): LWDA 2018: Proceedings of the Conference „Lernen, Wissen, Daten, Analysen“, S. 191-194. CEUR Workshop Proceedings, RWTH: Aachen: https://madoc.bib.uni-mannheim.de/46103/ Web-Link: http://web.informatik.uni-mannheim.de/ds4dm/#Unconstrained_Table_Extension

[17] David Arnu: Vortrag „Smart data through automatic data search and extraction”. Auf der PAPIs 2018 Konferenz am 06. April, 2018 in London, UK. Web-Link: https://papiseurope2018.sched.com/event/Dope/smart-data-through-automatic-data-search-and-extraction

[18] Edwin Yaqub, Philipp Schlunder, Ralf Klinkenberg, et al.: „Realizing Smart Data by Automating Tabular Search, Integration, and Extraction Methods”, bei der 2. BMBF “Big Data All Hands Meeting and 2nd Smart Data Innovation Conference”, Karlsruher Institute of Technology (KIT), am 11.-12. Oktober 2017 am KIT in Karlsruhe. Program Link: https://indico.scc.kit.edu/event/303/timetable/#20171012

[19] Teilnahme an der „1st Industrial Data Science Conference (IDS 2017)“ teil: Die IDS 2017 Konferenz wurde von RapidMiner in Dortmund zusammen mit der Technischen Universität Dortmund organisiert. Die Konferenz fand am 05.09.2017 statt. Web-Link: https://www.industrial-data-science.de/talks/2017/01_ralf_klinkenberg__rapidminer__welcome_and_overview/

[20] Petar Petrovski, Christian Bizer: „Extracting Attribute-Value Pairs from Product Specifications on the Web”. International Conference on Web Intelligence (WI2017), Leipzig, 23.-26. August, 2017. Web-Link: https://pdfs.semanticscholar.org/2b80/2806d964bae95b312cc22e728bd202fd7a5e.pdf

[21] Edwin Yaqub, Ralf Klinkenberg, et al. DS4DM-Poster „Automated Mechanisms to Discover and Integrate Data from Web-based Tabular Collections”, bei der 19. General Online Research Konferenz (GOR 17) am 15.-17. März 2017 in Berlin. Web-Link: https://www.gor.de/gor17/index.php?page=downloadPaper&filename=Yaqub-Automated_Mechanisms_to_Discover_and_Integrate_Data-275.pdf&form_id=275&form_version=final

[22] Anna Lisa Gentile, Petar Ristoski, Steffen Eckel, Dominique Ritze und Heiko Paulheim (2017): „Entity Matching on Web Tables: A Table Embeddings Approach for Blocking”. In „Advances in Database Technology - EDBT 2017: 20th International Conference on Extending Database Technology (EDBT)”, Venedig, Italien, 21.-24. März 2017, Proceedings, Seiten 510-513. OpenProceedings, Konstanz, 2017. Web Link: http://www.openproceedings.org/2017/conf/edbt/paper-367.pdf

[23] Anna Lisa Gentile (2016): Vorstellung der Ideen des DS4DM-Projekts auf dem 4. Workshop „LD4IE (Linked Data for Information Extraction)“ auf der 15. International Semantic Web Conference (ISWC 2016), 18. Oktober 2016, Kobe, Japan. Program-Link: http://ds4dm.de/2016/10/31/workshop und http://web.informatik.uni-mannheim.de/ld4ie2016/LD4IE2016/Overview.html

[24] Edwin Yaqub und Ralf Klinkenberg: DS4DM-Poster und DS4DM-Demonstrator (Live Demo des ersten DS4DM-Demonstrators) bei der 5. BMBF Fachtagung „KMU innovativ: IKT“ – „Mittelstand: Digital. Innovativ. Vernetzt.“ am 10.-11. Oktober 2016 in Hannover. Program-Link: http://www.softwaresysteme.pt-dlr.de/de/fachtagung-2016.php

[25] Anna Lisa Gentile und Heiko Paulheim: Vortrag „Extending RapidMiner with Data Search and Integration Capabilities“ über das DS4DM-Projekt und erste Projektergebnisse beim „Rhein-Neckar Smart Data Meetup“ am 23. Juni 2016 in Mannheim. Web-Link: http://www.meetup.com/de-DE/Rhein-Neckar-Smart-Data-Meetup/events/231410104/?eventId=231410104

[26] Anna Lisa Gentile, Sabrina Kirstein, Heiko Paulheim und Christian Bizer (2016): „Extending RapidMiner with Data Search and Integration Capabilities” bei der 13. Extended Semantic Web Conference (ESWC 2016), 29. Mai 2016 – 2. Juni, 2016, Heraklion, Crete, Greece. In Proceedings von Springer, 2016. Beschreibung der Demo (Paper): http://ub-madoc.bib.uni-mannheim.de/40718/1/DataSearchDemo.pdf Video der Demonstration: https://www.youtube.com/watch?v=i2g71G0jdmw

[27] Blog Artikel: „The ‘Data Search for Data Mining’ – Extension Release!”, Mai 2017, Web-Link: https://community.rapidminer.com/discussion/38231/the-data-search-for-data-mining-extension-release

[28] Blog Artikel: „The Web as a new data source for RapidMiner”, November 2017, Web-Link: https://community.rapidminer.com/discussion/43306/the-web-as-a-new-data-source-for-rapidminer

[29] Blog Artikel: „The Web Table Extraction Operator”, März 2017, Web-Link: https://community.rapidminer.com/discussion/37353/the-web-table-extraction-operator

[30] Blog Artikel: „PDF Table Extraction Extension Released!”, April 2017, Web-Link: https://community.rapidminer.com/discussion/37490/pdf-table-extraction-extension-released

[31] Blog Artikel: „Using DS4DM and Web Table Extraction extensions for Google Table and HTML table extraction”, November 2017, Web-Link: https://community.rapidminer.com/discussion/44089/using-ds4dm-and-web-table-extraction-extensions-for-google-table-and-html-table-extraction

[32] Blog Artikel: „The Spreadsheet Table Extraction – Extension Release”, Mai 2017, Web-Link: https://community.rapidminer.com/discussion/38864/the-spreadsheet-table-extraction-extension-release

[33] Blog Artikel: „Finding the right data is crucial (Einladung der Community Nutzer zur Teilnahme an der Pilot Studie)“, Mai 2017, Web-Link: https://community.rapidminer.com/discussion/39024/finding-the-right-data-is-crucial-help-us-improve-new-data-discovering-techniques

[34] Blog Artikel: „Connect to any document within your SharePoint”, Juli 2017, Web-Link: https://community.rapidminer.com/discussion/40582/connect-to-any-document-within-your-sharepoint

[35] Blog Artikel: „Reading Excel files directly from your company's OneDrive", Juli 2017, Web-Link: https://community.rapidminer.com/discussion/40587/reading-excel-files-directly-from-your-companies-onedrive

[36] Blog Artikel: „Highlights of the Extension Updates", Juli 2017, Web-Link: https://community.rapidminer.com/discussion/40589/highlights-of-the-extension-updates

Vorausgehende Arbeiten und Veröffentlichungen

[**Lehmberg 2014**] Oliver Lehmberg, Dominique Ritze, Petar Ristoski, Kai Eckert, Heiko Paulheim and Christian Bizer: „Extending tables with data from over a million websites". In: Semantic Web Challenge, 2014: https://ub-madoc.bib.uni-mannheim.de/37371/1/swc2014_submission_11.pdf

[**Suchanek 2011**] Fabian M. Suchanek, Serge Abiteboul and Pierre Senellart: "PARIS: Probabilistic Alignment of Relations, Instances, and Schema". Proceedings of the VLDB Conference, Vol. 5 (3), pp. 157–168, 2011.