SIMLEARN – Ontologiegestützte Integration von Simulationsmodellen, Systemen für maschinelles Lernen und Planungsdaten

Nils Reinosch[1], Alexander Münzberg[2], Daniel Martini[1], Alexander Niehus[1], Liv Seuring[1], Christian Troost[3], Rajiv Kumar Srivastava[4], Thomas Berger[3], Thilo Streck[4] und Ansgar Bernardi[2]

Abstract: Maschinelle Lernverfahren bieten gerade im Agrarbereich mit kaum kontrollierbaren, natürlichen Einflüssen und entsprechender Unsicherheit eine große Chance für betriebliche Entscheidungsunterstützung. Im Projekt SIMLEARN werden die für einen solchen Ansatz benötigten großen Mengen an aufgearbeiteten Trainingsdaten durch in Simulationsmodellen kodifiziertes Wissen mit fortschreitenden Erkenntnissen erlernter Modelle iterativ ergänzt. Durch vorhandene Simulationsmodelle werden umfangreiche synthetische Trainingsdatensätze erzeugt und für das initiale Training eines lernenden Systems verwendet. Das so initiierte lernende System wird durch im landwirtschaftlichen Betrieb erhobene Daten erweitert und an die individuelle betriebliche Situation angepasst. Im Ergebnis soll das trainierte System verbesserte, für den konkreten Betrieb adaptierte Vorhersagen liefern, für die umfangreiche Datenintegration werden dabei Ontologien erprobt. Ontologien bieten hier große Vorteile in der Datenabfrage durch die mehrdimensionale Struktur und logische Verknüpfungen. Für eine bessere Handhabung wird die standardisierte Mappingsprache R2RML verwendet, um die großen Mengen tabellarischer Daten in Ontologien zu überführen. SIMLEARN betrachtet exemplarisch betriebliche Entscheidungen im Getreideanbau auf operativer und taktischer Ebene mit Vorhersagen zu Ertrags-, Einkommens- und Umwelteffekten. Expertenwissen in Form von Faustzahlen und Planungswerten füllt lückenhafte Daten. In dieser Arbeit wird die entwickelte Ontologie vorgestellt.

Keywords: Ontologie, Simulation, maschinelles Lernen, relationale Daten, Datenmanagement, Smart Data

[1] Kuratorium für Technik und Bauwesen in der Landwirtschaft; Team Digitale Technologien, Bartningstraße 49, 64289 Darmstadt, n.reinosch@ktbl.de, https://www.ktbl.de/
[2] Deutsches Forschungszentrum für Künstliche Intelligenz; Smart Data&Knowledge Services Dept., Trippstadter Str. 122, 67663 Kaiserslautern, Germany, www.dfki.de
[3] Universität Hohenheim, Ökonomik der Landnutzung (490d), Wollgrasweg 43, 70599 Stuttgart, landuse-economics.uni-hohenheim.de
[4] Universität Hohenheim, Biogeophysik (310d), Emil-Wolff-Str. 27, 70593 Stuttgart, biogeophysik.uni-hohenheim.de

1 Einleitung

Im 20. Jahrhundert fand die Industrialisierung der Landwirtschaft statt, heute erleben wir ihre Digitalisierung. Daten können automatisiert verarbeitet werden und so eine wichtige Basis für Entscheidungen in der Landwirtschaft bilden. Die richtigen Entscheidungen zum richtigen Zeitpunkt zu treffen ist in der Landwirtschaft essenziell und kann ökonomisch und ökologisch bedeutende Auswirkungen haben. Bislang bedienten sich Landwirte aus einer Vielzahl von Quellen, um an Informationen bezüglich Wetter, Bodenanalysen, Monitoring, Marktpreise, etc. zu gelangen, kombinierten diese mit eigenen Erfahrungswerten und zogen Schlüsse zum Anbauverfahren. In einem solch komplexen System wie der Landwirtschaft mit kaum kontrollierbaren, natürlichen Einflüssen und entsprechenden Unsicherheiten, wird es zunehmend schwerer, sich an immer schnellere Veränderungen anzupassen und die besten Entscheidungen aus ökonomischer und ökologischer Sicht zu treffen. Maschinelle Lernverfahren bietet in einem solchen komplexen System große Chancen für die betriebliche Entscheidungsunterstützung.

Maschinelle Lernverfahren können in komplexen Situationen mit vielen Variablen Muster und Abhängigkeiten erkennen und mit den erlernten Modellen Klassifikationen, Vorhersagen oder Entscheidungshilfen liefern, um die Gesamtwirtschaftlichkeit sowie umweltbewusstes Handeln zu verbessern. In der Praxis sind die für solche Ansätze notwendigen großen Mengen an aufgearbeiteten Trainingsdaten aber oft nicht verfügbar. Um hier Abhilfe zu schaffen, wird im Rahmen dieses Projektes ein innovatives Vorgehen entwickelt und erprobt: Vorhandene Daten und in Simulationsmodellen kodifiziertes Wissen werden mit fortschreitenden Erkenntnissen erlernter Modelle iterativ kombiniert. Durch vorhandene Simulationsmodelle werden umfangreiche synthetische Trainingsdatensätze erzeugt und für das initiale Training eines lernenden Systems verwendet. Exemplarisch betrachtet das Projekt SIMLEARN betriebliche Entscheidungen im Getreideanbau auf operativer und taktischer Ebene mit Vorhersagen zu Ertrags-, Einkommens- und Umwelteffekten.

Im Projekt werden ontologiegestützte Ansätze für die Datenintegration erprobt, die eine systemübergreifende Vereinheitlichung, mit eindeutigen Bezeichnern für Datenfelder und deren Beschreibung nach Datentypen, Wertebereichen und Zusammenhängen, mit anderen Datenfeldern ermöglichen. Abfragen und Inferenz auf Basis der erstellten Ontologie demonstrieren die automatische Transformation, Ableitung und Plausibilisierung von Daten. Die vorliegende Ontologie beschreibt Ein- und Ausgabeparameter einer Teilmenge der Daten. Sie annotiert Daten mit Informationen zur Verwendung im Prozess des maschinellen Lernens und strukturiert Parameter und Objekte in Eigenschafts- und Klassenhierarchien. Inferenz-Regeln zur Klassifizierung von Bodenarten und Mapping-Definitionen auf Basis des R2RML-Vokabulars zeigen die Leistungsfähigkeit. Hier soll erprobt werden, ob Ontologien eine sinnvolle Ergänzung sind, um Daten für den Einsatz in maschinellen Lernverfahren zu vervollständigen und zu vereinheitlichen sowie zusätzliches Wissen zu extrahieren und entsprechend aufzubereiten.

2 Material und Methoden

2.1 Ontologien und Vokabularien zur Datenintegration

Eine Ontologie ist eine formale Beschreibung von existierenden Konzepten und Beziehungen in einer Fachdomäne, d.h. eine Konzeptualisierung eines Wissensbereiches [Gr08]. Im Kontext des Semantic Web werden Ontologien meist als Wissensgraphen auf Basis des Resource Description Framework (RDF) und der Web Ontology Language (OWL) abgebildet [Hi12]. Solche Graphen bestehen aus sogenannten Tripeln, die Aussagen der Form Subjekt-Prädikat-Objekt darstellen. Dabei werden „Dinge“ als Entitäten (Subjekt und Objekt) miteinander über Prädikate in Relation gesetzt. Auf diese Weise können Texte und Werte einzelnen Klassen und Instanzen zugeordnet werden (z. B. Name des Arbeitsgerätes und seine Arbeitsbreite), aber auch die Klassen und Instanzen untereinander in Relationen gesetzt werden, um fachlogische Zusammenhänge darzustellen. Darüber hinaus bieten Ontologien eine mehrdimensionale Datenstruktur gegenüber den zwei Dimensionen von Tabellen [Pr12].

In der Regel werden Ontologien auf Basis von RDF aus verschiedenen Vokabularien zusammengesetzt [Hi12]. Neben dem RDF-Basisvokabular und RDFS (RDF Schema [BG14]) werden dabei in diesem Projekt insbesondere XML Schema Datentypen [Pe12]), SKOS (Simple Knowledge Organisation System, [MB09]), QUDT (Quantities, Units, Dimensions, and Types Ontology [Qu11]) und OWL (Web Ontology Language, [Hi22]) verwendet.

Als syntaktisches Zielformat wird die Terse RDF Triple Language (TURTLE) genutzt [Pr12]. Da teilweise recht umfangreiche Klassenhierarchien aufzubauen waren, wurde ein spezieller Ansatz gewählt: Die initiale Sammlung abzubildender Entitäten und ihrer Eigenschaften erfolgte zunächst in MS-Excel. Anschließend wurden die dabei entstandenen Tabellenblätter über die standardisierte Mapping-Sprache R2RML [Da12] und zugehöriger Prozessierungswerkzeuge aus CSV-Dateien in die Teilontologien umgewandelt. Dieser Vorgang bietet den Vorteil, dass Änderungen sehr einfach zu handhaben sind und dass Domänenexperten ohne großes Wissen über Ontologien, unter Anleitung des Ontologieentwicklers, mit bekannten Werkzeugen wie MS-Excel an Ontologien arbeiten können [Da12]. Zur Übertragung der tabellarischen Daten nach RDF werden Regeln in R2RMLformuliert und mit einer R2RML Engine (hier R2RML-F [CH01]) ausgeführt. So können aus einer CSV-Datei oder Datenbank mittels SQL-Abfrage Daten aus dem Tabellenformat in eine Ontologie umgewandelt werden. Auf diese Art können sehr große Datenmengen mit einem vergleichsweise überschaubaren R2RML Mapping in eine Ontologie übersetzt werden. Die Inhalte von Tabellenspalten werden dabei als Subjekt und Objekt mit festgelegten Prädikaten verknüpft [Da12]. Für die Abfrage der Ontologie wird ein SPARQL-Endpoint (SPARQL Protocol and RDF Query Language [HS13]) verwendet. Dabei handelt es sich um eine graphenbasierte Abfragesprache für RDF-Modelle.

3 Ergebnisse

Experimentell wird für das maschinelle Lernverfahren eine Ontologie zur Bereitstellung, gezielten Selektion, Vereinheitlichung und Zusammenführung der großen Mengen an Daten verwendet. Dabei werden verschiedene Teilbereiche abgebildet: Einerseits bezieht sich die Ontologie auf Ein- und Ausgabedaten der genutzten Simulationsmodelle, die wie oben beschrieben beispielsweise Boden- und Pflanzenparameter sowie durchgeführte Maßnahmen beinhalten. Die KTBL-Datenbank bietet weitreichende Standard- und Planungswerte, mit denen lückenhafte Datensätze vervollständigt werden können; über eingefügte Property-Hierarchien können verwandte oder ähnliche Eigenschaften unmittelbar identifiziert werden. Außerdem werden Daten vom jeweiligen landwirtschaftlichen Betrieb, auf den sich das System automatisiert anpassen lassen soll, integriert. Die Ontologie ist in mehrere Teilontologien unterteilt: Maschinen, Pflanzen, Farmaktivitäten (z. B. Maßnahmen in der Kultur), Stoffe (z. B. Betriebsmittel wie Dünger, Saatgut, Kraftstoff etc.), Böden und Realfarmdaten, wobei erstere derzeit noch am umfangreichsten ist.

Die Teilontologien Maschinen, Pflanzen, Farmaktivitäten bestehen aus einer von Fachexperten erstellten Klassenhierarchie auf bis zu fünf Ebenen mit multipler Vererbung, an den die Instanzen des entsprechenden Themengebietes angehängt wurden. In der Teilontologie Farmaktivitäten findet sich dabei die Verknüpfung zu den Pflanzen, Maschinen und Stoffen. So kann mittels SPARQL eine Abfrage gestaltet werden, die für eine festgelegte Pflanze, Maschine, Farmaktivität oder Stoff angibt, welche übrigen Maschinen, Pflanzen, Farmaktivitäten und Stoffe damit kompatibel sind.

Die Boden-Ontologie enthält 31 Klassendefinitionen, die zu unterscheidende Bodenarten gemäß der bodenkundlichen Kartieranleitung, anhand der prozentualen Anteile der Kornfraktionen, definieren. Dabei werden OWL Restrictions zur Einschränkung der gültigen Wertebereiche genutzt. Anhand allgemeiner, logischer Schlussregeln, die in der OWL-Spezifikation beschriebenen sind, können so Instanzen mit Hilfe eines sogenannten Reasoners oder Reasoning Engines automatisiert klassifiziert werden. Im konkreten Fall bedeutet dies, dass Böden (bzw. Bodenproben) mit bekannten Anteilen der Sandfraktion, Schlufffraktion und Tonfraktion durch den Reasoner automatisch den benannten Bodenarten zugeordnet werden können. Ein Vorteil der Nutzung von OWL und ihrer zugrundeliegenden logischen Grundlage zur Beschreibung von Klassifikationssystemen und -regeln wie des Systems der Bodenarten ist es, dass auf diese Art beschriebene Systeme kombinierbar sind, d.h. weitere Klassen und zugehörige Werteeinschränkungen können flexibel hinzugenommen werden. Eine Erprobung soll daher auch für weitere Betriebsdaten erfolgen. Außerdem sind solchermaßen beschriebene Klassendefinitionen mit Hilfe von SPARQL auch in „umgekehrter“ Richtung abfragbar: d. h. mit einer gegebenen Klassenzuweisung zu z. B. einer Bodenart lässt sich ermitteln, welche Wertebereiche möglich sind. Auf diesem Weg können beispielsweise auch sinnvolle Variationen von Daten erstellt werden.

Einer Instanz sind mehrere Klassifizierungen zugeordnet mit rdf:type; diese Typen entsprechen allen in der Hierarchie übergeordneten Klassen. Man nennt dieses Modellierungsmuster „materialisierte Inferenz“, da die sich durch logische Schlussfolgerung (Inferenz) aus der Transitivität der rdfs:subClassOf-Beziehung ergebende Aussagen explizit mit abgebildet („materialisiert“) werden. Vorteile hat dies bei SPARQL-Abfragen: Es kann an einer beliebigen Stelle der Hierarchie abgebrochen werden. Mit der Abfrage „?Instanz rdf:type ktbl-plant:Weizen“ werden so Instanzen vom Typ Weizen und alle untergeordneten Instanzen als Resultat zurückgegeben.

Die Farmaktivitäten-Ontologie stellt das Bindeglied zwischen Pflanzen-, Maschinen- und Stoff-Ontologie dar und ist über entsprechende Prädikate mit diesen verknüpft. Zum Beispiel für „Winterweizen – Backweizen“ und die Aktivität „Säen von Weizen mit Kreiselegge und Sämaschine“ ist eine der verknüpften Maschinen der „Standardtraktor, Allradantrieb, Lastschaltgetriebe, 40 km/h“. Dieser ist vom rdf:type „Traktoren und Trägerfahrzeuge“ mit der URI „ktbl-mash:MachinenGroup-1726“. Jede Maschine besitzt eine Reihe von Eigenschaften, beispielsweise Leergewicht und zulässiges Gesamtgewicht. Des Weiteren sind noch Eigenschaften wie Maximalgeschwindigkeit, Kaufpreis, Motorleistung, usw. verfügbar. Bei Prädikaten, die physikalische Größen darstellen wie z. B. ktbl:hasEmptyWeight wird die QUDT-Ontologie sowie ein in dem Zusammenhang gängiges Modellierungsmuster zur Abbildung von Werten und Einheiten genutzt: Dabei wird ein zusätzlicher Knoten erzeugt, der typisiert ist als ktbl:EmptyWeight und mit den Properties qudt:value und qudt:unit verknüpft ist. Die zugehörigen Werte geben Zahlenwert bzw. zugehörige Einheit an. Wissen kann hierdurch flexibler modelliert werden, da auch unterschiedliche Einheiten für in diesem Beispiel Leergewicht zugelassen sind. Das Vokabular QUDT bietet eine Vielzahl von bereits definierten Einheiten, die mit hilfreichen Eigenschaften ausgestattet sind.

Mit Hilfe von R2RML können Ontologien typischerweise auf der Instanzebene, aber auch auf Ebene der Klassen- und Eigenschaftendefinitionen und der Terminologie aus Tabellen generiert werden. Für das R2RML Mapping wird pro Subjekt-Spalte in der Tabelle eine rr:TriplesMap angelegt und eine rr:subjectMap zugeordnet. Die rr:subjectMap erhält die gewünschte Präfix-URI und in geschweifter Klammer wird die Spalte mit ihrem Namen als Variable angegeben; die Daten dafür werden aus dem rr:logicalTable verwendet, das entweder Ergebnis einer SQL Abfrage sein oder eine CSV-Datei darstellen kann. Jeder rr:subjectMap können beliebig viele rr:predicateObjectMaps zugefügt werden. Die rr:predicateObjectMap legt ein Prädikat fest und eines oder mehrere Objekte, auf die das Prädikat verweist. Das Objekt als rr:objectMap kann dabei einfacher Datentyp (z. B. String, Decimal) sein oder auf mittels eines Internationalized Resource Identifier eine Beziehung zu einem weiteren Knoten etablieren. Für den Knoten selbst wird anschließend eine eigene TriplesMap erstellt.

4 Zusammenfassung und Ausblick

Für die Datenintegrationen wird die standardisierte Mapping-Sprache R2RML verwendet, um aus Datenbanken und CSV-Dateien unmittelbar Ontologien auf Basis des RDF-Basisvokabular, RDFS, XSD, SKOS, QUDT und OWL zu erstellen. Auf diese Art können alle Vorteile relationaler Datenbanken, aber auch einfacher tabellarischer Arbeitsumgebungen wie Tabellenkalkulationen für das Projekt nutzbar gemacht werden und Fachexperten unmittelbar ihr Wissen einbringen. Des Weiteren können über Restriktionen und Reasoning auch automatisierte Zuordnungen zu verschiedenen Klassen erfolgen und so eine noch bessere Handhabung der für maschinelles Lernen notwenigen Daten erreicht werden. Zum derzeitigen Punkt haben sich Ontologien als eine sinnvolle und zukunftsorientierte Ergänzung für die Datenbereitstellung zum maschinellen Lernansatz gezeigt. Die Ontologie bietet vielfältige und genaue Abfragemöglichkeiten und soll in der verbleibenden Projektlaufzeit erweitert werden, um konkrete landwirtschaftliche Betriebe noch detaillierter abzubilden und eine automatisierte Kategorisierung in weiteren Bereichen der realen Farmdaten zu ermöglichen.

Förderhinweis: Das Projekt SIMLEARN wird mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter Förderkennzeichen 01|S19073 gefördert.

Literaturverzeichnis

[BG14] Brickley, G.; Guha, B.: RDF Schema 1.1. W3C Recommendation, 2014, http://www.w3.org/TR/rdf-schema/, Stand: 24.10.2022.

[CH01] Debruyne, C.: R2RML-F: an R2RML Implementation https://github.com/chrdebru/r2rml, Stand: 24.10.2022.

[Da12] Das S., Sundara S., Cyganiak R.: R2RML: RDB to RDF Mapping Language, 2012, https://www.w3.org/TR/r2rml/, Stand: 24.10.2022.

[Gr08] Gruber, T. : Ontology - to appear in the Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2008. https://tomgruber.org/writing/ontology-in-encyclopedia-of-dbs.pdf, Stand: 24.10.2022.

[Hi22] Hitzler, P.; Krötzsch, M.; Parsia, B.; Patel-Schneider, P.; Rudolph S.: OWL 2 Web Ontology Language Primer (Second Edition), 2012, https://www.w3.org/TR/owl2-primer, Stand: 24.10.2022.

[HS13] Harris, S.; Seaborne, A.: SPARQL 1.1 Query Language, 2013, https://www.w3.org/TR/sparql11-query, Stand: 24.10.2022.

[MB09] Miles, A.; Bechhofer, S.: SKOS Simple Knowledge Organization System Reference, 2009, https://www.w3.org/TR/skos-reference, Stand: 24.10.2022.

[Pe12] Peterson, D.; Gao, S.; Malhotra, A.; Sperberg-McQueen, C. M.; Thompson, H. S.: XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes. W3C Recommendation, 2012, http://www.w3.org/TR/xmlschema11-2/, Stand: 24.10.2022.

[Pr12] Prud'hommeaux, E.; Carothers, G.: Turtle: Terse RDF Triple Language, 2012, http://www.w3.org/TR/2012/WD-turtle-20120710/, Stand: 24.10.2022.

[Qu11] QUDT: Quantities, Units, Dimensions and Types, http://qudt.org, DOI: 10.25504/FAIRsharing.d3pqw7, Stand: 24.10.2022.