Anfragebeantwortung in Probabilistischen Datenbanken und Wissensbasen[1]

İsmail İlkan Ceylan[2]

Abstract: Probabilistische Datenbanken und Wissensbasen werden immer wichtiger in der Wissenschaft und Industrie. Sie werden ständig mit neuen Daten erweitert, angetrieben durch moderne Informationsextraktionssysteme, die Fakten mit Wahrscheinlichkeiten assoziieren. Der Stand der Technik, solche Daten zu speichern und zu verarbeiten, basiert auf probabilistischen Datenbanksystemen, die breit und erfolgreich eingesetzt werden. Jenseits von allen Erfolgsgeschichten fehlt solchen Systemen aber immer noch die grundlegende Maschinerie, um das in ihnen gespeicherte wertvolle Wissen an die Endnutzer weiterzugeben, was ihre potenziellen Anwendungen in der Praxis begrenzt. In ihrer klassischen Form basieren solche Systeme in der Regel auf starken, unrealistischen Einschränkungen, wie der *Welt-* und *Domänenabgeschlossenheit*, der *Tupelunabhängigkeit* und *dem Mangel an Allgemeinwissen*. Diese Einschränkungen führen nicht nur zu unerwünschten Konsequenzen, sondern setzen diese Systeme auch bei wichtigen Aufgaben, wie der Anfragebeantwortung, auf ein schwaches Fundament. In dieser Arbeit erweitern wir probabilistische Datenbanken und Wissensbasen mit realistischeren Datenmodellen und ermöglichen damit bessere Mittel für die Anfragebeantwortung. Aufbauend auf dem langen Bestreben, Logik und Wahrscheinlichkeit zu integrieren, entwickeln wir unterschiedliche Semantiken für probabilistische Datenbanken und Wissensbasen, analysieren ihre algorithmischen Eigenschaften und entwerfen, wann immer möglich, effiziente Anfragebeantwortungsalgorithmen.

1 Einführung

Es besteht ein starkes Interesse daran, große probabilistische Wissensbasen aus Daten auf automatisierte Weise aufzubauen, was zu einer Reihe von Systemen wie DeepDive [Sh15], NELL [Mi15], Reverb [FSE11], Microsoft Probase [Wu12], IBM Watson [Fe12] und Google Knowledge Vault [Do14] geführt hat. Diese Systeme durchsuchen kontinuierlich das Web und extrahieren *strukturierte* Informationen und füllen so ihre Datenbanken mit Millionen von Entitäten und Milliarden von Tupeln auf. Die Forschung auf dem Gebiet der großen Wissensbasen dient als neue Ära für die Integration von Logik und Wahrscheinlichkeit.

Inwieweit können diese Such- und Extraktionssysteme bei realen Anwendungen helfen? Obwohl sie sich noch in einem frühen Entwicklungsstadium befinden, werden Systeme wie DeepDive routinemäßig zum Aufbau von Wissensbasen für Bereiche wie *Paläontologie*, *Geologie*, *medizinische Genetik* und *menschliche Bewegung* eingesetzt; siehe, z.B., [Ku15] und [Pe14]. IBM Watson revolutioniert *Gesundheitssysteme* [Fe13] und viele andere Anwendungsgebiete der *Naturwissenschaften*. Google Knowledge Vault hat mehr als

[1] Originaltitel: Query Answering in Probabilistic Data and Knowledge Bases

[2] University of Oxford, ismail.ceylan@cs.ox.ac.uk (Nominierung bei der Technischen Universität Dresden).



(a) Mozart (b) Beethoven

Abb. 1: Informationsfelder für die Google-Suche (a) Mozart und (b) Beethoven

eine Milliarde Fakten aus dem Web zusammengestellt und wird hauptsächlich zur Verbesserung der Qualität von Suchergebnissen im Web verwendet.

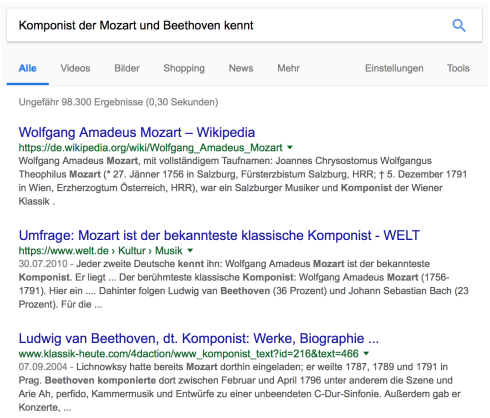
Aus einem größeren Blickwinkel betrachtet, ist die Suche nach dem Aufbau großer Wissensbasen ein neuer Meilenstein für die Forschung im Bereich der künstlichen Intelligenz (KI). Bereiche wie Informationsextraktion, natürliche Sprachverarbeitung (z.B. Beantwortung von Fragen), relationales und tiefgehendes Lernen, Wissensrepräsentation und -schlussfolgerung und Datenbanken ergreifen Initiative für ein gemeinsames Ziel. Inferenzverfahren und die Anfragebeantwortung auf großen probabilistischen Wissensbasen wird allgemein als das Kernstück dieser Bemühungen angesehen. Vielleicht liegt die sichtbarste Anwendung von probabilistischen Wissensbasen in Suchmaschinen. Heutzutage wird die Standardliste relevanter Webseiten oft um eine Tabelle mit strukturierten Daten erweitert, die sich auf die Suchanfrage bezieht. Zum Beispiel zeigt die Suche nach Mozart (Figur 1a) oder Beethoven (Figur 1b) eine Box, die ihre Kinder, Ehepartner, Schüler, Geburtsorte usw. identifiziert, was eindeutig mit der zugrundeliegenden Wissensbasis verknüpft ist.

Neben allen Erfolgsgeschichten fehlt es den probabilistischen Wissensbasen jedoch immer noch an der grundlegenden Maschinerie, um einen Teil des wertvollen Wissens, das sich in ihnen versteckt, dem Endnutzer zu vermitteln [WHS16], was ihre potenziellen Anwendungen in der Praxis stark einschränkt. Zum Beispiel ist die Information, die neben den Suchergebnissen in der Google-Suche angezeigt wird, sehr stark moderiert und zeigt nur Fakten, über die der Suchmaschinenanbieter absolut sicher ist. Andere, unsicherere Informationen sind vor dem Benutzer verborgen.

Darüber hinaus gibt es keine Unterstützung für relationale Anfragen über diesen Datenbanken. Selbst die einfache Anfrage “Gibt es einen Komponisten, der sowohl Mozart als auch Beethoven kennt?” kann von diesen Systemen derzeit nicht beantwortet werden, trotz der Tatsache, dass es eine mögliche Antwort gibt: Haydn, eine Person, die der probabilis-



(a) Haydn



(b) Gibt es ein Komponist der sowohl Mozart als auch Beethoven kennt?

Abb. 2: Unfähigkeit, relationale Anfragebeantwortung auf großen Wissensbasen durchzuführen. Obwohl Mozart (Figur 1a), Beethoven (Figur 1b) und Haydn (Figur 2a) der probabilistischen Wissensbasis bekannt sind, kann die Anfrage, ob es einen Komponisten gibt, der sowohl Mozart als auch Beethoven kennt (Figur 2b), von diesem System nicht beantwortet werden.

tischen Wissensbasis bekannt ist (Figur 2a). Was macht die Auswertung einer solch einfachen Anfrage so schwierig, und warum kann dieses Wissen nicht an den Benutzer weitergegeben werden? Diese Fragen bilden unsere globale Motivation, und die Antworten sind mit tiefen theoretischen Problemen sowie mit technischen Einschränkungen verbunden, die wir als Nächstes skizzieren.

2 Probabilistische Datenbanken

Das grundlegendste Modell ist das der tupelunabhängigen probabilistischen Datenbanken (PDBs) [Su11], die tatsächlich vielen dieser Systeme zugrunde liegen. Wir betrachten ein (endliches) relationales Vokabular σ bestehend aus *endlichen* Mengen **R** von *Prädikaten*, **C** von *Konstanten* und **V** von *Variablen*. Ein *Term* ist eine Konstante oder eine Variable; ein *Atom* hat die Form $\mathsf{P}(s_1,\ldots,s_n)$, wobei P ein n -stelliges Prädikat ist, und $s_1,\ldots,s_n$ Terme sind. Ein *Grundatom* ist ein Atom ohne Variablen.

Eine Datenbank $\mathcal{D}$ ist eine endliche Menge von *Grundatomen*. Eine *probabilistische Datenbank* $\mathcal{P}$ ist eine endliche Menge von *(probabilistischen) Atomen* der Form $\langle t : p\rangle$, wobei t ein Grundatom ist und $p \in [0,1]$, und immer wenn $\langle t : p\rangle, \langle t : q\rangle \in \mathcal{P}$, dann muss $(p = q)$ gelten. Eine PDB $\mathcal{P}$ ordnet jedem Atom t die Wahrscheinlichkeit p zu, wenn $\langle t : p\rangle \in \mathcal{P}$, und andernfalls die Wahrscheinlichkeit 0 zu. Unter der *Tupelunabhängigkeitsannahme* induziert eine solche Wahrscheinlichkeitszuordnung P die folgende *eindeutige gemeinsame Wahrscheinlichkeitsverteilung* über klassischen Datenbanken $\mathcal{D}$:

$$\mathrm{P}_{\mathcal{P}}(\mathcal{D}) := \prod_{t\in\mathcal{D}} \mathrm{P}_{\mathcal{P}}(t) \cdot \prod_{t\notin\mathcal{D}} (1-\mathrm{P}_{\mathcal{P}}(t)).$$

Aus Gründen der Effizienz fehlt probabilistischen Datenbanken typischerweise ein geeigneter Umgang mit Unvollständigkeit in der Praxis. Insbesondere kann jedes der oben genannten Systeme nur einen Teil der realen Welt modellieren, und diese Beschreibung ist zwangsläufig unvollständig. Wenn es jedoch um die Anfragebeantwortung geht, verwenden die meisten dieser Systeme starke und in den meisten Fällen unrealistische Vollständigkeitsannahmen, die ihre Anwendbarkeit einschränken. Wir werden nun einen Blick auf diesen Annahmen werfen, die inhärent mit der Semantik von probabilistischen Datenbanken verknüpft sind.

Aus modelltheoretischer Sicht basieren probabilistische Datenbanken auf Annahmen i) wie der *Weltabgeschlossenheit (WA)*, ii) der *Tupelunabhängigkeit (TU)*, iii) dem *Mangel an Allgemeinwissen (MA)* und iv) der *Domänenabgeschlossenheit (DA)*. Hier bedeutet WA, dass alle Fakten, die nicht in der probabilistischen Datenbank erscheinen, die Wahrscheinlichkeit 0 haben. Dies kann auch als eine probabilistische Variante der klassischen WA [Re78] gesehen werden. Die TU besagt, dass jedes Tupel in der probabilistischen Datenbank als unabhängige Bernoulli-Zufallsvariable interpretiert wird. MA bedeutet, dass es nicht möglich ist, explizites Domänenwissen zu kodieren. Und die DA impliziert, dass die betrachtete Domäne auf eine endliche Menge bekannter Konstanten festgelegt ist. Alle diese Annahmen kommen von klassischen Datenbanken, während die TU eine zusätzliche Annahme auf dem Wahrscheinlichkeitsraum ist. Wir konzentrieren uns zuerst auf die WA und illustrieren ihre Konsequenzen an einem einfachen Beispiel.

Beispiel 1. Wir betrachten eine tupelunabhängige PDB, die die Wahrscheinlichkeit 0.5 mehreren selbsterklärenden Fakten zuordnet:

$$\langle \mathsf{Komponist}(\mathsf{haydn}) : 0.5 \rangle, \langle \mathsf{LehrerVon}(\mathsf{haydn}, \mathsf{beethoven}) : 0.5 \rangle,$$
$$\langle \mathsf{Kennt}(\mathsf{haydn}, \mathsf{beethoven}) : 0.5 \rangle, \langle \mathsf{FreundVon}(\mathsf{haydn}, \mathsf{mozart}) : 0.5 \rangle.$$

Unter der GWA haben alle fehlenden Fakten die Wahrscheinlichkeit 0, dass heißt, sie sind falsch. Folglich erhalten die folgenden beiden Anfragen die Wahrscheinlichkeit **0**:

$$Q_1 = \exists x (\mathsf{LehrerVon}(x, \mathsf{beethoven}) \wedge \mathsf{GeborenIn}(x, \mathsf{österreich})),$$
$$Q_2 = \exists x (\mathsf{Person}(x) \wedge \neg \mathsf{Person}(x)).$$

Insbesondere wird angenommen, dass $\mathsf{GeborenIn}(\mathsf{haydn}, \mathsf{österreich})$ die Wahrscheinlichkeit 0 hat (d.h. falsch ist); jedoch kann diese Annahme selbst falsch sein. Tatsächlich kann $\mathsf{GeborenIn}(\mathsf{haydn}, \mathsf{österreich})$ sogar die Wahrscheinlichkeit 1 haben (d.h. kann wahr sein), was dazu führen würde, dass Q_1 die Wahrscheinlichkeit **0.5** hat.

Auf der anderen Seite ist Q_2 unerfüllbar und sollte immer die Wahrscheinlichkeit **0** haben, unabhängig davon, wie unvollständig die PDB ist. Das heißt, die GWA zwingt eine sehr flache Repräsentation, die es sogar unmöglich macht, eine erfüllbare Anfrage von einer unerfüllbaren zu unterscheiden. ◊

Wir können dieses Beispiel natürlich erweitern, um den Effekt der Tupelunabhängigkeitsannahme zu illustrieren.

Beispiel 2. Unter Tupelunabhängigkeit wird der Anfrage

$$Q_3 = \exists x\,(\mathsf{LehrerVon}(x, \mathsf{beethoven}) \wedge \mathsf{Kennt}(x, \mathsf{beethoven}))$$

die Wahrscheinlichkeit $0.5 \cdot 0.5 = \mathbf{0.25}$ zugeordnet. Aber da Haydn ein Lehrer von Beethoven ist, kennt er ihn auch; also sind die beiden Fakten nicht unabhängig. ◊

Diese Beobachtungen werden noch dramatischer, wenn mehrere Einschränkungen dieser Systeme kombiniert auftreten; insbesondere zusammen mit dem Mangel an Allgemeinwissen, der uns zu ontologischen Wissensbasen bringt.

3 Ontologisches Wissen und Probabilistische Wissensbasen

Der Mangel an Allgemeinwissen ist einer der Hauptgründe dafür, dass einige offensichtliche Antworten nicht aus den Wissensbasen abgerufen werden können. Dies zeigt sich in realen Anwendungen: Insbesondere im Kontext der Websuche, bei der die strukturierten Informationsergebnisse eindeutig mit der zugrundeliegenden Wissensbasis verknüpft sind.

Beispiel 3. Eine einfache Anfrage, ob es einen Komponisten gibt, der sowohl Mozart als auch Beethoven kennt,

$$Q_4 := \exists x\,\mathsf{Komponist}(x) \wedge \mathsf{Kennt}(x, \mathsf{beethoven}) \wedge \mathsf{Kennt}(x, \mathsf{mozart}),$$

erhält die Wahrscheinlichkeit 0 und kann daher von diesen Systemen nicht richtig ausgewertet werden. Die Antwort auf diese Frage ist tatsächlich in der Wissensbasis: Es ist bekannt, dass Haydn (i) ein Komponist, (ii) ein Freund von Mozart und (iii) einer der Lehrer von Beethoven ist.

In der Tat erhalten beide Anfragen "Freund von Mozart" und "Lehrer von Beethoven" die richtigen Informationen, die auf Haydn hindeuten, einen der Wissensbasis bekannten Komponisten. Allerdings fehlen explizite Informationen über $\mathsf{Kennt}(\mathsf{haydn}, \mathsf{mozart})$, und daher erhält diese Aussage die Wahrscheinlichkeit 0. ◊

Es ist schwierig, diese einfache Anfrage auszuwerten, weil die aktuellen PDBs kein Allgemeinwissen haben, nämlich, dass zwei befreundete Personen sich kennen, was ontologisch als

$$\forall x, y\,\mathsf{FreundVon}(x, y) \rightarrow \mathsf{Kennt}(x, y),$$

kodiert werden kann. Menschliches Schließen nutzt dieses grundlegende Wissen, um implizite Konsequenzen aus Daten abzuleiten, und diese Art von Wissen ist wesentlich für die Auswertung von Anfragen über großen PDBs in unkontrollierten Umgebungen wie dem Internet. Daher ist die Einbeziehung von Allgemeinwissen sehr wichtig, und dies ist inhärent damit verbunden, die obigen Vollständigkeitsannahmen von PDBs aufzugeben.

Die entscheidende Notwendigkeit, die Unabhängigkeitsannahme zu lockern, wurde bereits in mehreren neueren Ansätzen erkannt, die auf Markov-Logik-Netzen (MLNs) basieren [RD06, GS16]. All diesen Ansätzen ist gemeinsam, dass sie auch die Modellierung von logischem Wissen ermöglichen.

Anders als bei Ontologiesprachen verwenden MLNs jedoch die Annahme der Domänenabgeschlossenheit, was nicht immer vernünftig ist und zu problematischen (und sogar absurden) Konsequenzen über vielen nicht-trivialen Domänen führt. Wir veranschaulichen dies am folgenden Beispiel.

Beispiel 4. Berücksichtigen Sie das folgende ontologische Wissen:

$$\forall x\, \mathsf{Mensch}(x) \rightarrow \exists y\, \mathsf{Elternteil}(y,x)$$
$$\forall x,y\, \mathsf{Elternteil}(x,y) \rightarrow \mathsf{Vorfahr}(x,y)$$
$$\forall x,y,z\, \mathsf{Vorfahr}(x,y) \wedge \mathsf{Elternteil}(y,z) \rightarrow \mathsf{Vorfahr}(x,z)$$
$$\forall x\, \mathsf{Vorfahr}(x,x) \rightarrow \bot$$

Die erste Einschränkung besagt, dass *jeder einen Elternteil hat* und die anderen definieren die (azyklische) *Vorfahrenbeziehung*.

Wenn das Elternteil einer einzelnen Konstante a in der Datenbank aufgrund der DA explizit nicht erwähnt wird, bedeutet diese Einschränkung in einem MLN, dass a von den bekannten Konstanten ein Elternteil zugewiesen werden muss. Dies ist im Wesentlichen eine völlig zufällige Person in der Datenbank. Unter der Annahme, dass dieses Individuum ein Mensch ist, muss es auch einen Elternteil haben und so weiter. Da die Vorfahrenbeziehung jedoch azyklisch ist, bedeutet dies, dass mindestens eine bekannte Konstante einen Elternteil haben muss, das nicht menschlich ist. Während die Existenz eines solchen Individuums angesichts der Beschränkungen unvermeidlich ist, gibt es keine natürliche Grenze, die man auf die Anzahl der Menschen in der Datenbank setzen kann, bevor dies geschieht—die Domäne aller Menschen ist für alle praktischen Zwecke unbegrenzt.

Die DA, die in PDBs und MLNs verwendet wird, ist eindeutig nicht für unbegrenzte Domänen geeignet und kann nur Approximationen liefern. Zum Beispiel, selbst wenn eine feste Anzahl von Menschen zur Menge der Konstanten hinzugefügt wird, begrenzt dies effektiv die Anzahl der möglichen Generationen von Menschen (d.h. die Tiefe der Elternteil Relation) und die Anzahl der verschiedenen Menschen mit nicht-menschlichem Elternteil (die Anzahl der maximalen Elemente der Elternteil Relation). ◊

MLNs sind im Wesentlichen propositional und können keine unbekannten Individuen ausdrücken, was sie deutlich unterscheidet von vollwertigem prädikatenlogischem Wissen. Insgesamt ist die DA für viele Anwendungen, die von Natur aus über offene Domänen arbeiten, nicht geeignet.

Auf der anderen Seite sind Ontologien prädikatenlogische Theorien, die domänenspezifisches Wissen formalisieren, um dadurch automatisiertes Schließen zu ermöglichen. Ontologiesprachen operieren im Gegensatz zu MLNs über offenen Domänen. Die prominentesten Ontologiesprachen in der Literatur basieren auf Datalog$^{\pm}$ [CGP12, CGL12, CGK13]

und auf Beschreibungslogiken [Ba07]. Das Interpretieren von Datenbanken mit Allgemeinwissen in Form von Ontologien steht in engem Zusammenhang mit dem auf Ontologien basierenden Datenzugriff [Po08], der im Zusammenhang mit klassischen Datenbanken ausführlich untersucht wurde, um eine offene Welt- und Domänenanfragebeantwortung zu ermöglichen. In so einem Szenario wird eine Datenbankanfrage durch eine logische Schnittstelle vermittelt, um implizites Wissen explizit zu machen: das führt zu umfangreicheren Antworten für Anfragen.

4 Ein kurzer Blick auf die Ergebnisse

In dieser Dissertation [Ce17] erweitern wir probabilistische Wissensbasen mit realistischeren Datenmodellen und ermöglichen so bessere Antworten auf Anfragen. Wir entwickeln unterschiedliche Semantiken für probabilistische Datenbanken und Wissensbasen, analysieren ihre berechnungstechnischen Eigenschaften, und entwerfen, wann immer möglich, effiziente Abfragebeantwortungsalgorithmen. Um dies zu erreichen, bringt die aktuelle Arbeit einige neuere Paradigmen aus der Datenbanktheorie und Logik für die probabilistische Anfragebeantwortung und den damit verbundenen Inferenzaufgaben zusammen. Die Dissertation ist in vier Teile organisiert, wobei Teil I die Präliminarien einführt und Teil IV den Schlussfolgerungen gewidmet ist. Alle Ergebnisse sind in Teil II (Probabilistische Datenbanken) und Teil III (Logik und Probabilistische Wissensbasen) präsentiert, wie wir zunächst zusammenfassen.

4.1 Resultate für Probabilistische Datenbanken

Wir stellen probabilistische Datenbanken vor, definieren probabilistische Anfragebeantwortung als ein Entscheidungsproblem und untersuchen ihre berechnungstechnische Komplexität. Wir zeigen, dass die Datenkomplexitätsdichotomie des Berechnungsproblems (zwischen Polynomialzeit und #P [DS12]) auf das Entscheidungsproblem unter Turing-Reduktionen (zwischen Polynomialzeit und PP) übertragen werden kann. Über die bekannten Ergebnisse hinaus erhalten wir auch andere Komplexitätsergebnisse.

Anschließend stellen wir *offene probabilistische Datenbanken* vor, die als neues Datenmodell für probabilistische Datenbanken vorgeschlagen werden. Der Hauptunterschied zwischen den probabilistische Datenbanken und ihrer offenen Variante besteht darin, dass letztere die WA nicht anwenden. Wir bieten eine tiefe Diskussion über die semantischen Unterschiede zwischen diesen Modellen und vergleichen sie im Bezug auf die in dieser Arbeit identifizierten Ziele. Neben den semantischen Ergebnissen enthält dieses Kapitel auch eine gründliche Komplexitätsanalyse für eine Vielzahl von Anfragesprachen. Diese Analyse beinhaltet ein Dichotomie-Ergebnis für die Datenkomplexität (zwischen Polynomialzeit und PP) und einen effizienten Algorithmus (einen, der für die in Polynomialzeit berechenbaren Anfragen vollständig ist). Die Hauptergebnisse zu offenen probabilistischen Datenbanken wurden zuvor in [CDV16][3] veröffentlicht, und eine Kurzfassung dieser Arbeit erschien auch als eine eingeladene Publikation in [CDV17].

[3] Ausgezeichnet mit *Marco Cadoli Best Student Paper Prize* bei der KR 2016.

Wir untersuchen auch zwei alternative Inferenzprobleme für probabilistische Datenbanken; nämlich die Suche nach der *wahrscheinlichsten Datenbank* und der *wahrscheinlichsten Hypothese* für eine bestimmte Anfrage, die beide durch die *maximum a posteriori* Berechnungen von Probabilistischen Graphischen Modellen inspiriert sind. Wir argumentieren, dass diese Inferenzprobleme hilfreich sein können, um das volle Potenzial probabilistischer Datenbanken auszuschöpfen. Die meisten dieser Ergebnisse basieren auf der frühen Publikation [CBL17].

4.2 Resultate über Logik und Probabilistische Wissensbasen

Wir erweitern die Ergebnisse in Teil III, um auch Allgemeinwissen in Form von Ontologien einzubeziehen. Bei Ontologien werden Datenbankanfragen zusätzlich mit der Aussagekraft von Ontologien ausgestattet. Nach einer allgemeinen Namenskonvention in diesem Bereich bezeichnen wir solche Anfragen *ontologievermittelte Anfragen*. Wir untersuchen die probabilistische ontologievermittelte Anfragebeantwortung auf probabilistischen Datenbanken und ebenso auf offenen probabilistischen Datenbanken. Die Arbeit in diesen Abschnitten baut auf der früheren Veröffentlichung [BCL17] auf und bezieht sich auch auf [CLP16]. Schließlich betrachten wir die maximum a posteriori Probleme der probabilistische Datenbanken im Zusammenhang mit ontologievermittelten Anfragen, die auf früheren Arbeiten basieren [CBL17]. Alle diese Ergebnisse basieren auf Datalog$^{\pm}$-Ontologien, für die wir auch eine gründliche Komplexitätsanalyse anbieten.

Wir untersuchen auch die sogenannten Bayesschen Ontologiesprachen, die klassische Beschreibungslogiken mit probabilistischer Unsicherheit erweitern. Bayessche Ontologiesprachen wurden in [CP14a] vorgeschlagen und in [CP14b] weiter untersucht; anschliessend, kombiniert in einer Zeitschrift [CP17] als Teil eines Sonderheftes. Es besteht auch eine Proof-of-Concept-Implementierung für Schlussverfolgen in Bayesschen Ontologiesprachen [CMP15]. Unser Fokus ist auf ontologievermittelter Anfragebeantwortung (und den damit verbundenen Problemen), und wir bauen auf früheren Arbeiten auf [CP15b]. Darüber hinaus untersuchen wir auch einen neuartigen Monitoring-Ansatz, der die Macht der Ontologiesprachen mit dynamischen Bayesschen Netzen kombiniert; diese Kombination wurde zuerst in [CP15a] vorgeschlagen. Der resultierende Formalismus wird dann dynamische Bayesschen Ontologiesprachen genannt und erlaubt Projektionen über die zukünftigen Zustände eines Systems.

Literaturverzeichnis

[Ba07] Baader, Franz; Calvanese, Diego; McGuinness, Deborah L; Nardi, Daniele; Patel-Schneider, Peter F, Hrsg. The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, 2nd. Auflage, 2007.

[BCL17] Borgwardt, Stefan; Ceylan, İsmail İlkan; Lukasiewicz, Thomas: Ontology-Mediated Queries for Probabilistic Databases. In: AAAI. 2017.

[CBL17] Ceylan, İsmail İlkan; Borgwardt, Stefan; Lukasiewicz, Thomas: Most Probable Explanations for Probabilistic Database Queries. In: IJCAI. 2017.

[CDV16] Ceylan, İsmail İlkan; Darwiche, Adnan; Van den Broeck, Guy: Open-World Probabilistic Databases. In: KR. 2016.

[CDV17] Ceylan, İsmail İlkan; Darwiche, Adnan; Van den Broeck, Guy: Open-World Probabilistic Databases: An Abridged Report. In: IJCAI. 2017.

[Ce17] Ceylan, İsmail İlkan: Query Answering in Probabilistic Data and Knowledge Bases. Dissertation, TU Dresden, 2017.

[CGK13] Calí, Andrea; Gottlob, Georg; Kifer, Michael: Taming the Infinite Chase: Query Answering under Expressive Relational Constraints. JAIR, 48:115–174, 2013.

[CGL12] Calí, Andrea; Gottlob, Georg; Lukasiewicz, Thomas: A General Datalog-Based Framework for Tractable Query Answering over Ontologies. JWS, 14:57–83, 2012.

[CGP12] Calí, Andrea; Gottlob, Georg; Pieris, Andreas: Towards More Expressive Ontology Languages: The Query Answering Problem. AIJ, 193:87–128, 2012.

[CLP16] Ceylan, İsmail İlkan; Lukasiewicz, Thomas; Peñaloza, Rafael: Complexity Results for Probabilistic Datalog$\pm$. In: ECAI. 2016.

[CMP15] Ceylan, İsmail İlkan; Mendez, Julian; Peñaloza, Rafael: The Bayesian Ontology Reasoner is BORN! In: ORE. 2015.

[CP14a] Ceylan, İsmail İlkan; Peñaloza, Rafael: The Bayesian Description Logic $\mathcal{BEL}$. In: IJCAR. 2014.

[CP14b] Ceylan, İsmail İlkan; Peñaloza, Rafael: Tight Complexity Bounds for Reasoning in the Description Logic BEL. In: JELIA. 2014.

[CP15a] Ceylan, İsmail İlkan; Peñaloza, Rafael: Dynamic Bayesian Ontology Languages. CoRR, abs/1506.08030, 2015.

[CP15b] Ceylan, İsmail İlkan; Peñaloza, Rafael: Probabilistic Query Answering in the Bayesian Description Logic BEL. In: SUM. 2015.

[CP17] Ceylan, İsmail İlkan; Peñaloza, Rafael: The Bayesian Ontology Language $\mathcal{BEL}$. JAR, 58(1):67–95, 2017.

[Do14] Dong, Xin; Gabrilovich, Evgeniy; Heitz, Geremy; Horn, Wilko; Lao, Ni; Murphy, Kevin; Strohmann, Thomas; Sun, Shaohua; Zhang, Wei: Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In: SIGKDD. ACM, 2014.

[DS12] Dalvi, Nilesh; Suciu, Dan: The dichotomy of probabilistic inference for unions of conjunctive queries. Journal of ACM, 59(6):1–87, 2012.

[Fe12] Ferrucci, D. A.: Introduction to "This is Watson". IBM J. Res. Dev. 56(3):235–249, 2012.

[Fe13] Ferrucci, David; Levas, Anthony; Bagchi, Sugato; Gondek, David; Mueller, Erik T.: Watson: Beyond jeopardy! AIJ, 199-200:93–105, 2013.

[FSE11] Fader, Anthony; Soderland, Stephen; Etzioni, Oren: Identifying Relations for Open Information Extraction. In: EMNLP. ACL, 2011.

[GS16] Gribkoff, Eric; Suciu, Dan: SlimShot: In-Database Probabilistic Inference for Knowledge Bases. Proceedings of VLDB Endowment, 9(7), 2016.

[Ku15] Ku, Joy P.; Hicks, Jennifer L.; Hastie, Trevor; Leskovec, Jure; Ré, Christopher; Delp, Scott L.: The mobilize center: An NIH big data to knowledge center to advance human movement research and improve mobility. Journal of the American Medical Informatics Association, 22(6):1120–1125, 2015.

[Mi15] Mitchell, T.; Cohen, W.; Hruschka, E.; Talukdar, P.; Betteridge, J.; Carlson, A.; Dalvi, B.; Gardner, M.; Kisiel, B.; Krishnamurthy, J.; Lao, N.; Mazaitis, K.; Mohamed, T.; Nakashole, N.; Platanios, E.; Ritter, A.; Samadi, M.; Settles, B.; Wang, R.; Wijaya, D.; Gupta, A.; Chen, X.; Saparov, A.; Greaves, M.; Welling, J.: Never-Ending Learning. In: AAAI. 2015.

[Pe14] Peters, Shanan E.; Zhang, Ce; Livny, Miron; Ré, Christopher: A Machine Reading System for Assembling Synthetic Paleontological Databases. PLoS ONE, 9(12), 2014.

[Po08] Poggi, Antonella; Lembo, Domenico; Calvanese, Diego; De Giacomo, Giuseppe; Lenzerini, Maurizio; , Riccardo: Linking Data to Ontologies. JDS, 10, 2008.

[RD06] Richardson, Matthew; Domingos, Pedro: Markov Logic Networks. Machine Learning, 62(1):107–136, 2006.

[Re78] Reiter, Raymond: On closed world data bases. Logic and Data Bases, S. 55–76, 1978.

[Sh15] Shin, Jaeho; Wu, Sen; Wang, Feiran; De Sa, Christopher; Zhang, Ce; Ré, Christopher: Incremental Knowledge Base Construction Using DeepDive. Proceedings of VLDB Endowment, 8(11):1310–1321, 2015.

[Su11] Suciu, Dan; Olteanu, Dan; Ré, Christopher; Koch, Christoph: Probabilistic Databases, Jgg. 3. 2011.

[WHS16] Weikum, Gerhard; Hoffart, Johannes; Suchanek, Fabian: Ten Years of Knowledge Harvesting: Lessons and Challenges. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 39(3):41–50, 2016.

[Wu12] Wu, Wentao; Li, Hongsong; Wang, Haixun; Zhu, Kenny Q.: Probase: A Probabilistic Taxonomy for Text Understanding. In: SIGMOD. ACM, S. 481–492, 2012.



İsmail İlkan Ceylan hat Informatik an der Middle East Technical University in der Türkei studiert; danach wechselte er zum International Center for Computational Logic an der TU Dresden, wo er auch seinen Masterabschluss erhalten hat. Er promovierte unter der Leitung von Prof. Franz Baader am Lehrstuhl für Automatentheorie der TU Dresden. Während seiner Promotion absolvierte er einen dreimonatigen Forschungsaufenthalt im Automated Reasoning Lab, unter der Leitung von Prof. Adnan Darwiche an der University of California, Los Angeles. Außerdem hat er mehrere kürzere Forschungsaufenthalte an der University of Oxford gemacht, um mit Prof. Thomas Lukasiewicz zusammenzuarbeiten. Zurzeit arbeitet er als Forscher an der University of Oxford an dem EPSRC Projekt *"Realistic Data Models and Query Compilation for Large-Scale Probabilistic Databases"*, in dem er als Co-Investigator tätig ist.