TECHNISCHE UNIVERSITÄT DRESDEN

Fakultät Informatik, Institut für Künstliche Intelligenz, Professur Computational Logic

THEORETISCHE INFORMATIK UND LOGIK

22. Vorlesung: Datalog

Hannes Straß

TU Dresden, 4. Juli 2022

Rückblick: Formeln als Anfragen

linien:

Linie	Typ
85	Bus
3	Tram
F1	Fähre
...	...

haltestellen:

SID	Name	Rollstuhl
17	Hauptbahnhof	true
42	Helmholtzstr.	true
57	Stadtgutstr.	true
123	Gustav-Freytag-Str.	false
...	...	...

verbindung:

Von	Zu	Linie
57	42	85
17	789	3
...	...	...

Die einfache Arität der Prädikatenlogik wird durch ein Schema mit Namen (und oft auch Datentypen) ersetzt:

- linien[Linie:string, Typ:string]
- haltestellen[SID:int, Halt:string, Rollstuhl:bool]
- verbindung[Von:int, Zu:int, Linie:string]

Relationale Algebra: Parameter (Spalten) durch Namen adressiert
Prädikatenlogik: Parameter durch Reihenfolge adressiert

Die Anfrage $\exists z_{\text{Linie}}.(\text{verbindung}(x_{\text{Von}}, x_{\text{Zu}}, z_{\text{Linie}}) \wedge \text{linien}(z_{\text{Linie}}, x_{\text{Typ}}))$ entspricht einer (natürlichen) Join-Operation ($\wedge$) mit anschließender Projektion ($\exists$):
$\pi_{\text{Von,Zu,Typ}}(\text{verbindung} \bowtie \text{linien})$.

Rekursive Anfragen

Rückblick: Mit Prädikatenlogik können nur lokale Eigenschaften getestet werden.

Nichtlokale Eigenschaften wie die Erreichbarkeit in Graphen sind praktisch relevant (speziell in Graphdatenbanken).

Wie kann man solche Anfragen logisch ausdrücken?

Idee: Um beliebig weit zu schauen muss man Rekursion einführen.

Beispiel: Eine Haltestelle ist von Helmholtzstr. aus erreichbar, wenn
(1) sie die Haltestelle Helmholtzstr. ist, oder
(2) sie neben einer Haltestelle liegt, die von Helmholtzstr. aus erreichbar ist.

Rekursion in Logik

Beispiel: Eine Haltestelle ist von Helmholtzstr. aus erreichbar, wenn
(1) sie die Haltestelle Helmholtzstr. ist, oder
(2) sie neben einer Haltestelle liegt, die von Helmholtzstr. aus erreichbar ist.

Wie kann man das in Logik ausdrücken?

Beispiel: Das Prädikat Erreichbar enthält alle von Helmholtzstr. erreichbaren Haltestellen, wenn die folgenden Formeln erfüllt sind:
(1) $\text{Erreichbar}(\texttt{helmholtzstr})$
(2) $\forall x, y, z.\big((\text{Erreichbar}(x) \wedge \text{verbindung}(x, y, z)) \rightarrow \text{Erreichbar}(y)\big)$

Model Checking?

Beispiel: Das Prädikat Erreichbar enthält alle von Helmholtzstr. erreichbaren Haltestellen, wenn die folgenden Formeln erfüllt sind:

(1) Erreichbar(helmholtzstr)

(2) $\forall x, y, z.\big((\text{Erreichbar}(x) \wedge \text{verbindung}(x, y, z)) \rightarrow \text{Erreichbar}(y)\big)$

Sei Q die Menge der Formeln (1) und (2).

- Die Modelle von Q sind alle Interpretationen, in denen Erreichbar mindestens die von Helmholtzstr. aus erreichbaren Haltestellen enthält.
- Eine gegebene Datenbankinstanz, betrachtet als Interpretation, ist normalerweise kein Modell von Q, sofern es nicht schon eine entsprechende Tabelle für Erreichbar gibt.

$\rightsquigarrow$ Model Checking führt hier nicht zum gewünschten Ergebnis.

Datenbanken als logische Theorien

Wir nehmen daher eine andere Perspektive ein:

- Datenbankinstanzen $\mathcal{I}$ werden als endliche Menge von Fakten $\mathcal{F}_\mathcal{I}$ dargestellt:

$$\mathcal{F}_\mathcal{I} = \left\{ p(c_1, \ldots, c_n) \;\middle|\; c_1, \ldots, c_n \in \mathbf{C} \text{ und } \langle c_1^\mathcal{I}, \ldots, c_n^\mathcal{I} \rangle \in p^\mathcal{I} \right\}$$

  (Wie zuvor nimmt man vereinfachend oft an, dass $c^\mathcal{I} = c$ gilt.)
- Rekursive Anfragen können wie im Beispiel als prädikatenlogische Formelmenge Q dargestellt werden.
- Ein Fakt $p(c_1, \ldots, c_n)$ ist genau dann eine Antwort auf die Anfrage Q auf der Datenbank $\mathcal{I}$, wenn gilt: $\mathcal{F}_\mathcal{I} \cup Q \models p(c_1, \ldots, c_n)$.

Das heißt: Wir fassen nun Anfragebeantwortung als prädikatenlogisches Schließen auf (nicht Model Checking).

Beispiel

Beispiel: Für die Anfrage

$$Q = \Big\{ \text{Erreichbar}(\texttt{helmholtzstr}),\ \forall x, y, z.\big((\text{Erreichbar}(x) \wedge \text{verbindung}(x, y, z)) \rightarrow \text{Erreichbar}(y)\big) \Big\}$$

und die Datenbankinstanz $\mathcal{I}$ mit

$$\begin{aligned}\mathcal{F}_\mathcal{I} = \{\ & \text{verbindung}(\texttt{helmholtzstr}, \texttt{stadtgutstr}, 85)\\ & \text{verbindung}(\texttt{stadtgutstr}, \texttt{räcknitzhöhe}, 85)\\ & \text{verbindung}(\texttt{räcknitzhöhe}, \texttt{zellescher\_weg}, 11)\\ & \text{verbindung}(\texttt{schillerplatz}, \texttt{körnerplatz}, 61)\ \}\end{aligned}$$

folgen genau die Antworten Erreichbar(helmholtzstr), Erreichbar(stadtgutstr), Erreichbar(räcknitzhöhe) und Erreichbar(zellescher_weg).

Sind rekursive Anfragen praktikabel?

Wir wissen: Prädikatenlogisches Schließen ist unentscheidbar.

Ist die Beantwortung rekursiver Anfragen dann überhaupt möglich?

Ja, wenn wir uns auf bestimmte Formen von Anfragen beschränken.

Eine Datalog-Regel ist eine Formel der Form

$$\forall x_1, \ldots, x_\ell.\big((B_1 \wedge \ldots \wedge B_n) \rightarrow H\big)$$

wobei $B_1, \ldots, B_n$ und H prädikatenlogische Atome sind und $x_1, \ldots, x_\ell$ eine Liste aller in den Atomen vorkommenden Variablen ist.

- H heißt Kopf und $B_1 \wedge \ldots \wedge B_n$ Rumpf der Regel. Wir verlangen, dass jede Variable im Kopf auch im Rumpf vorkommt.
- Ein Fakt ist eine variablenfreie Regel mit $n = 0$.
- Ein Datalog-Programm P ist eine Menge von Datalog-Regeln (einschl. Fakten).

Es ist üblich, eine Regel als $H \leftarrow B_1 \wedge \ldots \wedge B_n$ und einen Fakt schlicht als H zu schreiben.

(Die Allquantoren werden weggelassen; das „Umdrehen“ des Implikationspfeils hat historische Gründe.)

Beispiele

Das vorige Beispiel war bereits ein Datalog-Programm:

$$\text{Erreichbar}(\texttt{helmholtzstr})$$
$$\text{Erreichbar}(y) \leftarrow \text{Erreichbar}(x) \wedge \text{verbindung}(x, y, z)$$

Ein umfangreicheres Beispiel:

$$\text{vater}(\text{alice}, \text{bob}) \quad \text{mutter}(\text{alice}, \text{carla}) \quad \text{mutter}(\text{evan}, \text{carla}) \quad \text{vater}(\text{carla}, \text{david})$$
$$\text{Elternteil}(x, y) \leftarrow \text{vater}(x, y) \qquad \text{Elternteil}(x, y) \leftarrow \text{mutter}(x, y)$$
$$\text{Vorfahr}(x, y) \leftarrow \text{Elternteil}(x, y)$$
$$\text{Vorfahr}(x, z) \leftarrow \text{Elternteil}(x, y) \wedge \text{Vorfahr}(y, z)$$
$$\text{GleicheGeneration}(x, x) \leftarrow \text{Vorfahr}(x, y)$$
$$\text{GleicheGeneration}(y, y) \leftarrow \text{Vorfahr}(x, y)$$
$$\text{GleicheGeneration}(x, y) \leftarrow \text{Elternteil}(x, v) \wedge \text{Elternteil}(y, w) \wedge \text{GleicheGeneration}(v, w)$$

Auswertung von Datalog

Wie kann man logische Schlüsse aus Datalog ziehen?

Idee: Wende Regeln iterativ auf gegebene Fakten an, um neue Fakten abzuleiten.

- Wir betrachten hier eine Datenbank (Interpretation) als logische Theorie, d.h., Menge von Grundfakten (variablenfreien Atomen).
- Die gegebenen Fakten kann man sich als Teil eines Datalog-Programms vorstellen.

Der Konsequenzoperator T_P für ein Datalog-Programm P bildet endliche Mengen $\mathcal{F}_\mathcal{I}$ von Fakten auf Mengen von Fakten ab:

$$T_P(\mathcal{F}_\mathcal{I}) = \{H\theta \mid H \leftarrow B_1 \wedge \ldots \wedge B_n \in P \text{ und es gibt Substitution } \theta \text{ mit } B_1\theta, \ldots, B_n\theta \in \mathcal{F}_\mathcal{I}\}$$

Beobachtungen:

- Substitutionen θ mit $B_1\theta, \ldots, B_n\theta \in \mathcal{F}_\mathcal{I}$ sind einfach Antworten auf die Datenbankanfrage $B_1 \wedge \ldots \wedge B_n$ über Datenbank $\mathcal{I}$, also praktisch berechenbar.
- θ muss alle Variablen in der angewendeten Regel auf Konstanten (Domänenelemente) aus $\mathcal{F}_\mathcal{I}$ abbilden.

T_P iterativ anwenden

Zur Ermittlung aller Schlüsse muss man T_P iterativ anwenden:

Für ein Datalog-Programm P definieren wir rekursiv:

- $T_P^0 = \emptyset$,
- $T_P^{i+1} = T_P(T_P^i)$ für alle $i \geq 0$.

Beobachtungen:

- $T_P^1 = T_P(\emptyset)$ ist die Menge aller Fakten in P.
- T_P^i enthält nur Fakten, die man bilden kann, indem man einen Regelkopf aus P mit Konstanten aus P instanziiert.
- Es gibt nur endlich viele solcher Atome (über dem Vokabular von P).

T_P erreicht also nach endlich vielen Schritten einen Grenzwert, definiert wie folgt:

$$T_P^\infty = \bigcup_{i \geq 0} T_P^i$$

Beispiel

Für das Programm P mit den Regeln

$$\text{vater}(\text{alice}, \text{bob}) \quad \text{mutter}(\text{alice}, \text{carla}) \quad \text{mutter}(\text{evan}, \text{carla}) \quad \text{vater}(\text{carla}, \text{david})$$
$$\text{Elternteil}(x, y) \leftarrow \text{vater}(x, y) \qquad \text{Elternteil}(x, y) \leftarrow \text{mutter}(x, y)$$
$$\text{GG}(x, x) \leftarrow \text{Elternteil}(x, y)$$
$$\text{GG}(y, y) \leftarrow \text{Elternteil}(x, y)$$
$$\text{GG}(x, y) \leftarrow \text{Elternteil}(x, v) \wedge \text{Elternteil}(y, w) \wedge \text{GG}(v, w)$$

(GG = GleicheGeneration) ergibt sich:

$T_P^0 = \emptyset$

$T_P^1 = \{\text{vater}(\text{alice}, \text{bob}), \text{mutter}(\text{alice}, \text{carla}), \text{mutter}(\text{evan}, \text{carla}), \text{vater}(\text{carla}, \text{david})\}$

$T_P^2 = T_P^1 \cup \{\text{Elternteil}(\text{alice}, \text{bob}), \text{Elternteil}(\text{alice}, \text{carla}), \text{Elternteil}(\text{evan}, \text{carla}), \text{Elternteil}(\text{carla}, \text{david})\}$

$T_P^3 = T_P^2 \cup \{\text{GG}(\text{alice}, \text{alice}), \text{GG}(\text{bob}, \text{bob}), \text{GG}(\text{carla}, \text{carla}), \text{GG}(\text{david}, \text{david}), \text{GG}(\text{evan}, \text{evan})\}$

$T_P^4 = T_P^3 \cup \{\text{GG}(\text{alice}, \text{evan}), \text{GG}(\text{evan}, \text{alice})\}$

$T_P^5 = T_P^4 = T_P^\infty$

Quiz: T_P-Operator

Sei P ein Datalog-Programm und $\mathcal{F}_\mathcal{I}$ eine endliche Menge von Fakten.

- $T_P(\mathcal{F}_\mathcal{I}) = \{H\theta \mid H \leftarrow B_1 \wedge \ldots \wedge B_n \in P$ und es gibt eine Substitution θ mit $B_1\theta, \ldots, B_n\theta \in \mathcal{F}_\mathcal{I}\}$
- $T_P^0 = \emptyset$ und $T_P^{i+1} = T_P(T_P^i)$ für alle $i \geq 0$.
- $T_P^\infty = \bigcup_{i \geq 0} T_P^i$

Quiz: Wir betrachten das folgende Datalog-Programm: ...

Semantische Bedeutung von T_P

Beobachtung 1: Jede Datalog-Regel $H \leftarrow B_1 \wedge \ldots \wedge B_n$ entspricht einer Klausel $H \vee \neg B_1 \vee \ldots \vee \neg B_n$, wobei jeweils alle Variablen allquantifiziert sind.

$\rightsquigarrow$ Datalog-Programme sind syntaktische Varianten skolemisierter Klauseln.
$\rightsquigarrow$ Für die Inferenz von Fakten kann man sich auf Herbrand-Modelle beschränken.

Beobachtung 2: Die Berechnung eines Fakts $H\theta$ durch Anwendung einer solchen Regel entspricht einer (Hyper)-Resolution der Klausel $H \vee \neg B_1 \vee \ldots \vee \neg B_n$ mit den Fakten $B_1\theta, \ldots, B_n\theta$, wobei θ der allgemeinste Unifikator ist.

$\rightsquigarrow$ Abgeleitete Fakten sind logische Konsequenzen (Korrektheit).

Mithilfe dieser Einsichten lässt sich zeigen, dass T_P^∞ zur Berechnung der logischen Schlussfolgerung geeignet ist:

Satz: Für ein Datalog-Programm P ist T_P^∞ das kleinste Herbrand-Modell.
Das heißt: Für einen beliebigen Fakt F gilt

$$F \in T_P^\infty \text{ gdw. } F \text{ in jedem Herbrand-Modell vorkommt gdw. } P \models F.$$

Ableitungsbäume

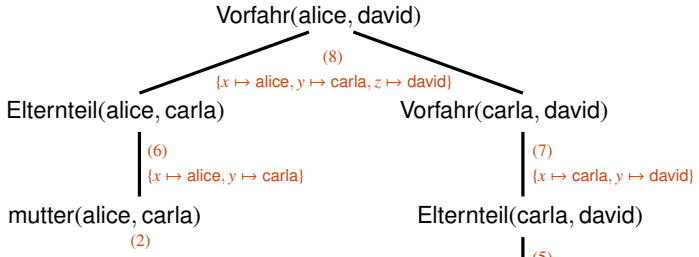
Die Folgerung $P \models F$ lässt sich als endlicher Baum darstellen:

- Jeder Knoten ist ein variablenfreies Atom.
- Jeder Elternknoten entsteht durch Anwendung einer Regel aus P auf seine Kindknoten.
- Jeder Blattknoten ist ein gegebener Fakt aus P.
- Jeder Knoten wird zusätzlich mit der Regel und Substitution θ beschriftet, die zur Ableitung angewendet wurden.

$\rightsquigarrow$ Der Ableitungsbaum stellt die Resolutionsableitung des Fakts an der Wurzel des Baums grafisch dar.

Beobachtung: Für jeden Fakt $F \in T_P^\infty$ gibt es mindestens einen Ableitungsbaum mit Wurzel F.

Beispiel



(1)	vater(alice, bob)
(2)	mutter(alice, carla)
(3)	mutter(evan, carla)
(4)	vater(carla, david)
(5)	Elternteil$(x, y) \leftarrow$ vater(x, y)
(6)	Elternteil$(x, y) \leftarrow$ mutter(x, y)
(7)	Vorfahr$(x, y) \leftarrow$ Elternteil(x, y)
(8)	Vorfahr$(x, z) \leftarrow$ Elternteil$(x, y) \wedge$ Vorfahr(y, z)

Komplexität

Mithilfe von T_P kann man logische Konsequenzen berechnen.
Wie aufwändig ist das?

Worst Case?

- Sei p die Anzahl der Prädikatensmbole, a deren maximale Arität, x die maximale Zahl an Variablen pro Regel und n die Zahl der Konstanten.
- Dann gibt es insgesamt $\leq p \cdot n^a$ variablenfreie Fakten, die abgeleitet werden könnten.
- Für eine Regel gibt es maximal n^x Substitutionen, die bei der Ableitung eine Rolle spielen könnten.

$\rightsquigarrow$ Die Berechnung von T_P^∞ ist in exponentieller Zeit möglich.

Man kann zeigen, dass dies worst-case-optimal ist:

Satz: Das Problem der Schlussfolgerung von Fakten ("$P \models F$?") für Datalog ist ExpTime-vollständig.

Ist Datalog praktisch?

ExpTime ist eine ziemlich hohe Komplexität – ist Datalog praktisch implementierbar?

Ja!

- Die Worst-Case-Komplexität erfordert, dass die Stelligkeit von Prädikaten unbeschränkt wachsen kann.
  $\rightsquigarrow$ In Anwendungen sind sehr große Stelligkeiten jedoch untypisch.
- In Abhängigkeit von der Größe der Datenbank (der Zahl der Fakten) wächst die Laufzeit lediglich polynomiell.
  $\rightsquigarrow$ Gutes Skalierungsverhalten für große Datenmengen.
- Es gibt inzwischen eine Reihe sehr effizienter Implementierungen, z.B.:
  - hochskalierbare speicherbasierte Systeme: z.B. VLog/Rulewerk (VU Amsterdam, TU Dresden), RDFox (Oxford)
  - Systeme basierend auf relationalen Datenbanken: z.B. Llunatic
  - Systeme für komplexere Logiken, die Datalog als Sonderfall unterstützen: z.B. clingo, DLV (Answer Set Programming), E (Theorembeweiser)

Logik höherer Ordnung

Prädikatenlogik ist genau genommen Prädikatenlogik erster Stufe.

Hintergrund:

- Erste Stufe: Quantoren beziehen sich auf Domänenelemente.

  Beispiel: „Jede natürliche Zahl n hat einen Nachfolger $s(n)$."

- Zweite Stufe: Quantoren beziehen sich auf Relationen (über Domänenelementen).

  Beispiel: „Für jede Menge M gilt: Enthält M die Zahl 0 und mit jeder natürlichen Zahl n auch stets deren Nachfolger $s(n)$, so enthält M alle natürlichen Zahlen."

Logik zweiter Stufe (zweiter Ordnung):

- Ausdrucksstärker: kann z.B. die natürlichen Zahlen exakt charakterisieren.
- Schwieriger: hat kein korrektes und vollständiges Beweisverfahren.

Logik höherer Ordnung: Syntax und Semantik

Syntax: Wie in Prädikatenlogik, aber mit quantifizierten Prädikaten-Variablen.
(Die Stelligkeit einer Prädikaten-Variablen muss jeweils klar festgelegt werden.)

Beispiel: "Für jede Menge M gilt: Enthält M die Zahl 0 und mit jeder natürlichen Zahl n auch stets deren Nachfolger $s(n)$, so enthält M alle natürlichen Zahlen."

$$\forall M.\Big((M(0) \wedge \forall x.(M(x) \rightarrow M(s(x)))) \rightarrow \forall y.M(y)\Big)$$

Wir verwenden hier ein Funktionssymbol s zur Darstellung von Nachfolgern.

Semantik: „Wie zu erwarten." (Gleiche Interpretationen wie in erster Stufe; Interpretation von Prädikaten-Variablen mit Zuweisungen wie bei Objektvariablen in erster Stufe.)

(Intuition: Erste Stufe verhält sich zur zweiten Stufe wie Aussagenlogik zu QBFs.)

Logik höherer Ordnung: logisches Schließen

Offenbar ist Schließen in Logik zweiter Stufe mindestens genauso schwer wie in Logik erster Stufe. Tatsächlich ist es noch deutlich schwerer:

Fakt: Logisches Schließen in Logik höherer Ordnung ist nicht semi-entscheidbar und insbesondere gibt es kein korrektes und vollständiges Ableitungsverfahren für diese Logik.

Logik höherer Ordnung und Datalog

Idee: Die Fakten, die in allen Modellen (eines Datalog-Programms) gefolgert werden können, sind genau diejenigen, die in jeder erfüllenden Interpretation (in Logik zweiter Ordnung) der Datalog-Prädikate gelten.

Beispiel: Das Datalog-Programm

$$\text{Erreichbar}(\texttt{helmholtzstr})$$
$$\text{Erreichbar}(y) \leftarrow \text{Erreichbar}(x) \wedge \text{verbindung}(x, y, z)$$

kann als Formel der Logik zweiter Stufe wie folgt dargestellt werden:

$$\forall \text{Erreichbar}.\Big((\text{Erreichbar}(\texttt{helmholtzstr}) \wedge \forall x, y, z.((\text{Erreichbar}(x) \wedge \text{verbindung}(x, y, z)) \rightarrow \text{Erreichbar}(y))) \rightarrow \text{Erreichbar}(v)\Big)$$

Die Formel hat eine freie Variable v und stellt Erreichbar als Prädikaten-Variable dar. Ein Fakt Erreichbar(a) folgt aus dem ursprünglichen Programm über einer Datenbank $\mathcal{F}_I$ genau dann, wenn $\mathcal{F}_I$ die Formel mit Variablenbelegung $v \mapsto a$ erfüllt.

Model Checking!

Das vorige Beispiel zeigt:

Die Beantwortung von Anfragen in Datalog entspricht einem Auswertungsproblem (Model Checking) für spezielle Formeln zweiter Ordnung über endlichen Interpretationen.

Diese Sicht wird gegenüber der Betrachtung als Logik erster Stufe bevorzugt, weil dadurch abgeleitete Prädikate zu lokalen Variablen des Programms werden, anstatt globaler Teil von Interpretationen (Datenbanken) zu sein.

Zusammenfassung und Ausblick

Datalog erlaubt die Darstellung rekursiver Anfragen in Logik.

Anfragebeantwortung in Datalog:

- = logisches Schließen in Prädikatenlogik;
- = Auswertungsproblem in Logik zweiter Stufe.

(ExpTime-vollständig, aber polynomiell bezüglich der Datenbankgröße.)

Ableitungen in Datalog können berechnet und dargestellt werden:

- mit dem T_P-Operator,
- durch Ableitungsbäume.

Was erwartet uns als nächstes?

- Gödel
- Probeklausur und Prüfung