BIZONYTALAN INFORMÁCIÓK KEZELÉSE LOGIKAI ADATMODELLEKBEN

Achs Ágnes, *achs@pmmf.hu*
Janus Pannonius Tudományegyetem Pollack Mihály Fõiskolai Kara, Pécs

Abstract

In this lecture we deal with the concept of logical data models and give an extension to handling uncertain information. We define the fuzzy Datalog programs as a set of Horn formulae with degrees and give their meaning by defining the deterministic and nondeterministic semantics.

1. Alapfogalmak

Term-nek nevezünk egy változót, konstansot vagy egy $f(t_1,...,t_n)$ alakú kifejezést, ahol f függvényszimbólum és $t_1,...,t_n$ termek. *Atom* egy $p(\underline{t})$ alakú formula, ahol p egy n változós predikátumszimbólum és $\underline{t}$ egy n elemû termsorozat. *Literál*-nak nevezünk egy atomot (pozitív literál) vagy a negáltját (negatív literál). Literálok diszjunkciója a *klóz*. A változó nélküli termet (atomot, literált, klózt) *alap*termnek (alapatom, alapliterál, alapklóz) nevezzük.

2. Logikai adatmodell

Egy logikai adatmodell - vagy más szóval tudásbázis - tényekbõl és következtetési szabályokból épül fel. A tények bizonyos ismeretekre, melyekbõl a szabályok segítségével következtethetünk újabb ismeretekre. A deduktív adatbázisok elméletében közismertek a Datalog-szerû nyelvek.

A Datalog Horn-klózok, azaz

$$A \leftarrow B_1, ..., B_n$$

alakú szabályok halmaza, ahol A, B_i (i = 1,...,n) pozitív literálok.

Azok a predikátumok, melyekhez tartozó relációkat az adatbázisban tároljuk, az EDB predikátumok, a megfelelõ relációk az EDB relációk. Ezeket szokták tényeknek is nevezni. (Alakja : $A \leftarrow$)

Azokat a predikátumokat (relációkat), melyeket logikai szabályok definiálnak, IDB predikátumoknak (relációknak) nevezzük.

Minden Datalog programhoz készíthetõ egy úgynevezett megelõzési gráf. A gráf csúcsai a programban szereplõ közönséges predikátumok. A p predikátumból a q predikátumba akkor vezet él, ha van olyan szabály, melynek feje q és törzsében szerepel p. Ha a gráf kört tartalmaz, akkor a program rekurzív.

A Datalog szabályok kiértékelésekor két utat követhetünk. Az egyik módszer szerint minden IDB predikátumhoz meghatározunk egy relációt az EDB predikátumokhoz tartozó relációk segítségével. Rekurzív programok esetén egy Datalog egyenletrendszerhez jutunk, melyet iteratív módon oldhatunk meg. Ezt a módszert preferálja [1]. [2] a másik utat részesíti elõnyben, amely szerint egy rákövetkezési operátort definiálhatunk, s ez alapján juthatunk új tényekhez. Mindkét megoldási algoritmus egy fixpontkeresés, és a

legkisebb fixpontot eredményezi. Ez a fixpont egyúttal a Datalog program egyértelmû minimális modellje. (Vagyis azon tények minimális halmaza, melyek kielégítik a program összes szabályát.)

Nem ilyen egyszerû a helyzet, ha negációt is megengedünk, hiszen a negatív infor mációk kezelése nehéz feladat.

A negatív adatok tömege jóval nagyobb, mint a pozitívaké, ezért csak a pozitívakat tároljuk, s a negatívakra következtetünk. Többféle következtetési mód is ismert, de mindegyik kapcsolódik valamilyen módon a zártvilág-feltételezéshez (CWA). A CWA lényege, hogy ha egy tény logikailag nem következik egy klózhalmazból, akkor arra következtetünk, hogy a tény negáltja igaz.

Ha a CWA elvet a közönséges Datalogra alkalmazzuk, akkor is tudunk negatív tényekre következtetni, de ezekbõl a negatív tényekbõl már nem határozhatunk meg újabbakat. Hogy tovább tudjunk belõlük következtetni, meg kell engednünk negatív literálok használatát a szabálytörzsben. Így jutunk a Datalog $\neg$ nyelvhez.

A negáció miatt elõfordulhat, hogy nem kapunk legkisebb fixpontot, hanem több minimálisat. (Ez hasonlóan több minimális modellt is jelent.) Kérdés, melyiket vá lasszuk közülük, azaz milyen szemantikát rendeljünk a Datalog $\neg$ - hoz.

Kétféle szemantikát szoktak értelmezni, az egyik a rétegzés, a másik az inflációs szeman tika.

Rétegzéskor egyfajta rendezést vezetünk be a predikátumok között, s a prog ramokat ebben a sorrendben értékeljük ki, vagyis a deklaratív szabályokhoz procedurális kiértékelés társul. Ha ebben a sorrendben végezzük el a kiértékelést, akkor egy negált predikátumra már csak akkor kerül sor, ha elõbb minden pozitív elõfordulási helyén kiértékeltük.

Az inflációs szemantika esetén nem kívánjuk meg a modell minimalitását, megelég szünk egy inflációs operátor (amely segítségével tényekbõl újabb tényekre következtetünk) legkisebb fixpontjával, s ezt a fixpontot definiáljuk szemantikaként. Ez a szemantika "hatásosabb", mint a rétegzés. Ez azt jelenti, hogy található olyan lekérdezés, amely ki fejezhetõ az inflációs Datalog $\neg$ -ban, de nem fejezhetõ ki a réteg zettben.([2])

A bizonytalan információk kezelésére azonban nem alkalmas a klasszikus Datalog. A további fejezetekben a bizonytalanság kezelésének egy lehetséges módját adjuk meg. A közönséges Datalog szabályokat kiegészítjük egy bizonytalansági szintet jelzõ értékkel és egy implikációs operátorral, amely segítségével következtetni tudunk a szabályfej bizony talansági szintjére. A szabályokhoz és ily módon a predikátumokhoz rendelt bizony talansági szint azt mutatja meg, hogy az illetõ predikátum legalább milyen szinten telje sül. Az így kapott program nyelvet fuzzy Datalognak, vagy röviden fDATALOG-nak nevezzük.

2. A fuzzy elmélet alapfogalmai

A nyelv definiálása elõtt ismerkedjünk meg a fuzzy halmazok elméletének néhány alapvetõ fogalmával!

Legyen D egy halmaz! A D fölötti F *fuzzy halmaz* egy $F : D \to [0,1]$ függvény. Jelölje $\mathcal{F}(D)$ az összes D fölötti fuzzy halmazt. Ekkor $F \in \mathcal{F}(D)$. A fuzzy halmazokon a metszet és unió szokásos értelmezése a következõ:

$$F \cup G\ (d) \stackrel{def}{=} \max(F(d),G(d))$$
$$F \cap G\ (d) \stackrel{def}{=} \min(F(d),G(d))$$

Fuzzy halmazokra vonatkozó rendezési relációt is értelmezhetünk: $F \leq G$ akkor és csak akkor, ha $F(d) \leq G(d)$ $\forall d \in D$. Mivel $\mathcal{F}(D)$ minden részhalmazának létezik legkisebb felsõ, illetve legnagyobb alsó korlátja, ezért ($\mathcal{F}(D), \leq$) teljes háló.

A fuzzy halmazokat gyakran $F= \bigcup_{d\in D} (d,\alpha_d)$ módon jelölik, ahol $(d, \alpha_d) \in D \times [0,1]$.

A következtetések elvégzéséhez szükségünk van az *implikációs operátor* fogalmára. Az implikációs operátorok megválasztásával és tulajdonságaival kapcsolatban sok vizsgálódás folyik. Ezekrõl a vizsgálatokról ad összefoglaló képet [3]. A különbözõ implikációs operátorokat a normák és conormák segítségével értelmezik. Ezek a metszet és az unió mûveletének a fuzzy halmazokra való kiterjesztése. A következõ táblázatban részletezés nélkül összefoglaljuk a leggyakoribb implikációs operátorokat:

jelölés	**név**	**formula**
$I_1(x, y)$	Gödel	1 ha $x \leq y$ y egyébként
$I_2(x, y)$	Lukasiewicz	1 ha $x \leq y$ 1- x + y egyébként
$I_3(x, y)$	Goguen	1 ha $x \leq y$ y/x egyébként
$I_4(x, y)$	Kleene-Dienes	max(1-x, y)
$I_5(x, y)$	Reichenbach	1 - x + xy
$I_6(x, y)$	Zadeh	max(1-x,min(x,y))
$I_7(x, y)$	Gaines-Rescher	1 ha $x \leq y$ 0 egyébként

3. A fuzzy Datalog szintaktikája

1. Definíció:

fDATALOG szabály egy (r ; I ; β) hármas, ahol r egy

$$Q \leftarrow Q_1,...,Q_n \qquad (n\geq 0)$$

alakú formula, Q atom (a szabály feje), $Q_1,...,Q_n$ literálok (a szabály törzse), I egy implikációs operátor és $\beta \in (0,1]$ (a szabály bizonytalansági foka vagy szintje).

Az fDATALOG szabály biztonságos, ha
- a fejben elõforduló összes változó szerepel a törzsben is;
- az összes olyan változó, amely negatív literálban szerepel, elõfordul pozitív literálban is.

Egy fDATALOG program biztonságos fDATALOG szabályok véges halmaza.

Az ($A \leftarrow$; I ; β) alakú szabályokat, ahol A alapatom, tényeknek nevezzük.

Egy P program *Herbrand univerzuma* (H_P) a P-ben elõforduló konstansokból és függvény-szimbólumokból képzett összes lehetséges alapatom halmaza. A P *Herbrand bázisa* (B_P) a P-ben elõforduló predikátumszimbólumokból képzett összes lehetséges olyan alapatom halmaza, melynek argumentumai H_P elemei. Egy $(r;I;\beta)$ szabály P-beli *alapelõfordulása* egy olyan szabály, amelyet úgy kapunk r-bõl, hogy az összes r-beli X változót $\Phi(X)$ -szel helyettesítjük, ahol Φ az r-ben elõforduló változók H_P-be való leképezése.

Az $(r;I;\beta)$ szabály összes alapelõfordulásának halmazát (*ground*(r); I ; β)- val jelöljük. A P program alapelõfordulása:

$$ground(P) = \cup_{(r;I;\beta)\in P} (\ ground(r);\ I\ ;\ \beta\ )$$

2. Definíció:

Egy P program interpretációja B_P egy fuzzy részhalmaza:

$$N_P \in \mathcal{F}(B_P),$$

azaz $N_P = \bigcup_{A\in B_P} (A, \alpha_A)$

A konjunkció és negáció szintjét a szokásos módon értelmezzük, azaz tetszõleges $A_1, ..., A_n$ alapatomok esetén :

$$\alpha_{A_1\wedge...\wedge A_n} \stackrel{def}{=} \min(\alpha_{A_1},...,\alpha_{A_n})$$
$$\alpha_{\neg A} \stackrel{def}{=} 1-\alpha_A$$

3. Definíció:

Egy interpretáció a P program modellje, ha minden

$$(A \leftarrow A_1,...,A_n\ ;\ I\ ;\ \beta) \in ground(P)$$

esetén

$$I(\alpha_{A_1\wedge...\wedge A_n}, \alpha_A) \geq \beta\ .$$

Az M legkisebb modell, ha bármely N modell esetén $M \leq N$. M minimális modell, ha nincs olyan $N \neq M$ modell, hogy $N \leq M$ teljesüljön.

Az egyszerûség kedvért $\alpha_{A_1\wedge...\wedge A_n}$ -t $\alpha_{törzs}$ - zsel és α_A -t α_{fej} - jel jelöljük.

Megjegyzés:

Az fDATALOG a közönséges Datalog általánosításának tekinthetõ, hiszen ha minden szabályhoz az I_7 - tel jelölt, ú.n. Gaines-Rescher implikációs operátort és $\beta = 1$ bizonytalansági szintet rendelünk, akkor Datalog szabályok halmazát kapjuk.

4. A fuzzy DATALOG szemantikája

Két rákövetkezési transzformációt értelmezünk, egy determinisztikus és egy nem determinisztikus transzformációt. A determinisztikus transzformáció segítségével párhuzamosan értékeljük ki a program szabályait, míg nem determinisztikus esetben a szabályokat tetszõleges sorrendben, egyenként egymás után véve végezzük a következtetést. A két transzformációnak megfelelõen kétféle szeman tikát rendelhetünk az fDATALOG programokhoz. Ezek a szemantikák pozitív programok esetén megegyez nek, de negációt tartalmazó programoknál különbözõ eredményt adhatnak.

4. Definíció

A $DT_P : \mathcal{F}(B_P) \rightarrow \mathcal{F}(B_P)$ és $NT_P : \mathcal{F}(B_P) \rightarrow \mathcal{F}(B_P)$ rákövetkezési transz formációkat a következõ módon értelmezzük:

$$DT_P(X)=\{\cup\{(A, \alpha_A)\} \mid (A \leftarrow A_1,...,A_n; I; \beta) \in ground(P), (|A_i|, \alpha_{A_i}) \in X, 1\le i\le n, \alpha_A=\max(0,\min\{\gamma \mid I(\alpha_{törzs}, \gamma) \ge \beta\})\} \cup X$$

és

$$NT_P(X)=\{(A, \alpha_A)\} \cup X$$

ahol $(A \leftarrow A_1,...,A_n; I; \beta) \in ground(P), (|A_i|, \alpha_{A_i}) \in X, 1\le i\le n, \alpha_A=\max(0,\min\{\gamma \mid I(\alpha_{törzs}, \gamma) \ge \beta\})$

Itt $|A_i|$ - val azt a p(**c**) atomot jelöljük, melyre A_i = p(**c**) vagy $A_i = \neg$p(**c**), p egy k argumentumú predikátumszimbólum és **c** k alaptermbõl álló lista.

A vizsgált implikációs operátorok körét le kell szûkítenünk, ugyanis nem minden I esetén létezik a bizonytalansági-szint függvény (bizony talansági-szint függvénynek nevezzük az f (I, α, β) = min ({ γ | I (α, γ) $\ge \beta$ }) függvényt), azaz nem határozható meg tetszõleges $\alpha_{törzs}$, β - hoz a kívánt α_A érték. Könnyen belátható, hogy bármely I $\in$ { I_1, I_2, I_3, I_4, I_5, I_7 } esetén tetszõleges $\alpha_{törzs}$, β értékhez létezik az f (I_i ,$\alpha_{törzs}$, β) (i = 1, 2, 3, 4, 5, 7) érték, míg I = I_6 esetén nem értelmezett minden α, β esetén az f (I_6, α, β) függvény.

A tényhalmazból kiindulva a transzformációk egymás utáni alkal mazásával (vagyis a transzformáció hatványainak meghatározásával) értelmezünk egy halmaz sorozatot. Ezen sorozatok fixpontjaként értelmezzük majd az fDATALOG szeman tikáját.

Belátható a következõ két állítás:

1. Állítás:

A DT_P transzformációnak van fixpontja, vagyis létezik olyan X $\in \mathcal{F}(B_P)$, melyre $DT_P(X) = X$. Az NT_P transzformációnak van fixpontja, vagyis létezik olyan X $\in \mathcal{F}(B_P)$, melyre az X halmaz bármely szabályát is választva $NT_P(X) = X$. Ha P pozitív, akkor a két fixpont megegyezik, és ez a közös érték a legkisebb fixpont.

A fixpontokat lfp(DT_P), illetve lfp(NT_P) -vel jelöljük.

2. Állítás:

lfp(DT_P) és lfp(NT_P) a P modelljei.

Az elõzõ két állítás igazsága teszi lehetõvé, hogy az fDATALOG szemantikáját a következõ módon definiáljuk:

5. Definíció:

lfp(DT_P) az fDATALOG P program determinisztikus szemantikája.
lfp(NT_P) az fDATALOG P program nemdeterminisztikus szemantikája.

Pozitív programok esetén a két szemantika megegyezik, s belátható, hogy ez a fixpont legkisebb modell is.

Belátható az is, hogy az I_1, I_2, I_3, I_4, I_7 operátorok, esetén létezik algoritmus a leg kisebb modell meghatározására, azaz a fixpontszámítási eljárás véges sok lépésben be fejezõdik. Az I_5 operátor alkalmazásakor elõfordulhat, hogy a program nem terminál véges lépésszámon belül.

Példa:

1. r(a) $\leftarrow$; I_1; 0.8

2. $p(x) \leftarrow r(x), \neg q(x); I_1; 0.6$
3. $q(x) \leftarrow r(x); I_1; 0.5$
4. $p(x) \leftarrow q(x); I_1; 0.8$

Ekkor

$$lfp(DT_P) = \{(r(a),0.8); (p(a),0.6); (q(a),0.5)\}$$

Nemdeterminisztikus kiértékeléskor különbözõ megoldásokat kaphatunk, hiszen az függ a kiértékelendõ szabályok sorrendjétõl.

Ha a kiértékelendõ szabályok sorrendje 1.,2.,3.,4. akkor

$$lfp(NT_P) = \{(r(a),0.8); (p(a),0.6); (q(a),0.5)\},$$

míg 1.,3.,2.,4. -es sorrendben

$$lfp(NT_P) = \{(r(a),0.8); (p(a),0.5); (q(a),0.5)\}.$$

Mint a példából is láthatjuk, $lfp(DT_P)$ nem mindig minimális modell, azaz a determinisztikus szemantikát alkalmazva a kapott fixpont nem minden esetben adja a kívánt mini mális modellt. Nem determinisztikus esetben azonban bizonyos feltételek mellett biztosítható a minimalitás. Ez a feltétel a rétegzés. A rétegzés meghatároz egy kiértékelési sorrendet, amelyben a negatív literálokat értékeljük ki elõször, s ily módon mini mális modellt kapunk.

A rétegzéshez elõbb a függõségi gráf fogalmát kell megadnunk. Ez egy olyan irányított gráf, melynek csúcsai a P predikátumai. Egy p predikátumból akkor vezet él egy q predi kátumba, ha P-nek van olyan szabálya, melynek törzsében p vagy $\neg p$ szerepel, és amelynek fej-predikátuma q.

Egy program rétegzett, ha az összes olyan esetben, amikor egy szabály fej-predikátuma p, és a törzs negált literálja $\neg q$, a függõségi gráfban nincs út p-bõl q-ba.

Egy P program rétegzése a P predikátumszimbólumainak olyan $P_1, ..., P_n$ részhalmazokra való partíciója, melyre teljesülnek a következõ feltételek:

a/ ha $p \in P_i$, $q \in P_j$ és nincs él q-ból p-be, akkor $i \geq j$.

b/ ha $p \in P_i$, $q \in P_j$ és van olyan p fejpredikátumú szabály, melynek törzse tartalmazza $\neg q$ - t, akkor $i > j$.

A $P_1, ..., P_n$ halmazokat rétegeknek nevezzük.

Legyen P egy rétegzett fDATALOG program $P_1, ..., P_n$ rétegzéssel. Jelölje P_i^* a P_i réteghez tartozó összes P-beli szabály halmazát, azaz az összes olyan szabály halmazát, melynek fej-predikátuma P_i - ben van.

Legyen

$$L_1 = lfp(NT_{P_1^*}),$$

ahol a fixpontszámolási eljárás kiindulópontja a program tényeinek halmaza.

Legyen

$$L_2 = lfp(NT_{P_2^*}),$$

ahol a számolás kiindulópontja az elõbb kapott L_1,

...

$$L_n = lfp (NT_{P_n^*}) ,$$

ahol a kiindulópont L_{n-1}.

Más szóval, elõször kiszámítjuk a P elsõ rétegéhez tartozó L_1 fixpontot. Ezen fixpont meg határozása után léphetünk a következõ rétegekre.

Megjegyzés:

$$lfp (NT_{P_i^*}) = lfp (DT_{P_i^*})$$

Belátható, hogy L_n a P minimális fixpontja és egyúttal minimális modellje.
Ha P rétegzett fDATALOG program, akkor L_n a P minimális fixpontja.

4. Konklúzió

Az elõadásban a Datalog-szerû nyelvek egy lehetséges fuzzy kiterjesztését adtuk meg úgy, hogy a Datalog szabályokat kiegészítettük egy bizonytalansági szintet jelzõ érték kel és egy implikációs operátorral. Bizonyos rákövetkezési transzformációk fixpontjaként definiáltuk az fDATALOG determinisztikus és nem-determinisztikus szemantikáját és beláttuk, hogy ez a fixpont pozitív programok esetén legkisebb, negácót is tartalmazó programok esetén pedig bizonyos feltételek mellett minimális modell.

Az fDATALOG fuzzy adatokra is kiterjeszthetõ. A hasonlósági relációk felhasz nálásával ilyen irányban indul el [4].

Irodalomjegyzék

[1] J.D.Ullman : Principles of database and knowledge-base systems
Computer Science Press, Rockville, 1988.

[2] S.Ceri - G.Gottlob - L.Tanca : Logic Programming and Databases
Springer-Verlag Berlin, 1990

[3] Didier Dubois - Henri Prade: Fuzzy sets in approximate reasoning, Part 1:
Inference with possibility distributions
Fuzzy Sets and Systems 40 (1991) (143-202)

[4] Ágnes Achs - Attila Kiss : Fuzzy extension of Datalog
Acta Cybernetica 12 (1995) (153-166), Szeged

[5]Ágnes Achs - Attila Kiss : Fixpoint query in fuzzy Datalog
megjelenõben : Annales, Bp

[6] Attila Kiss : On the least models of fuzzy Datalog programs
International Conference on Information Processing and Management of Uncertainty in
Knowledge-based system, Mallorca, 1993. (465-471)

[7] Vilém Novák : Fuzzy sets and their applications
Adam Hilger Bristol and Philadelphia, 1989.