XML ÉS FÉLIG-STRUKTURÁLT ADATBÁZISOK FÜGGŐSÉGEI

A doktori értekezés tézisei

Készítette:
Szabó Gyula István
okleveles matematikus

Témavezető:
Dr. Benczúr András egyetemi tanár
Az MTA doktora

INFORMÁCIÓS RENDSZEREK TANSZÉK
EÖTVÖS LORÁND TUDOMÁNYEGYETEM, INFORMATIKAI KAR

INFORMATIKA DOKTORI ISKOLA
A Doktori Iskola vezetője:
Dr. Benczúr András egyetemi tanár
Az MTA doktora

INFORMÁCIÓS RENDSZEREK DOKTORI PROGRAM
Vezető:
Dr. Benczúr András egyetemi tanár
Az MTA doktora

Budapest, 2013.

Bevezetés

Az Internet (World Wide Web) mindennapjaink részévé vált, ide fordulunk információkért, ide helyezzük az információkat, amelyeket a világgal közölni kívánunk. Ennek a szédületes információfolyamnak ki nem mondottan standard formátumává lett az XML. Az XML eredetileg egyedi dokumentumok készítésére szolgáló eszköz volna, a gyakorlatban, mint az online adatközlés egyeduralkodója, adatbázis kezelő eszközzé fejlődött. A relációs adatbázisok Internetes közzétételéhez is az XML formátumot használják.
A relációs adatbázis merev sémaszerkezete nehezen illeszkedik az on-line alkalmazásokban megszokott vizuális felületek követelményeihez. Az on-line alkalmazások elvárásaihoz kapcsolódva alakultak ki az önleíró adatszerkezetek (HTML, félig-strukturált adatbázisok, SGML, XML). A relációs adatbázisban tárolt adatok megjelenítésénél felmerült a relációs függőségek „konverziójaként" keletkező függőségek, azaz a félig-strukturált adatbázisok adatfüggőségei vizsgálatának igénye.

Az Értekezés az XML adatbázisokon értelmezett függőségek (épségi megszorítások) új családját vezeti be, a szakirodalomból ismert eddigi definícióktól lényegesen eltérő új szemlélet alapján. A dolgozat központi új, eredeti eredménye az XML konvencionális, fa-szerkezetű („kétdimenziós") modelljének egydimenziós, reguláris kifejezéssel leírható modellre való szűkítése, függőségeknek ezen a modellen, reguláris nyelv mondataiból felépített adatbázison való vizsgálata. A reguláris nyelvet adatbázisként interpretáló modell számos ismert függőségtípust (funkcionális, tartalmazási, összekapcsolási) megalapozhat, az Értekezés ezek közül a publikációinkban már megjelent, reguláris funkcionális függőségekre (RFD) vonatkozó eredményeket foglalja egységbe.

A funkcionális függőség (functional dependency - FD) valószínűleg a legfontosabb épségi megszorítás bármely adatmodell esetén (ahol egyáltalán értelmezhető, azaz, szinte minden modellben), de mindenképpen a legalaposabban elemzett függőség. A relációs FD definíciója természetesen adódik: egy Y attribútumhalmazon felvett értékek egy másik, X attribútumhalmazon felvett értékektől függenek, azaz, „Y az X függvénye". Másképpen fogalmazva, ha egy reláció két sora megegyezik az X-beli attribútumokon, akkor meg kell, hogy egyezzen az Y-beli attribútumokon is. Az XML funkcionális függőséget (XFD) számos különböző módon definiálták, ám nem született általánosan elfogadott meghatározás. A fő probléma a funkcionális függőség XML környezetben való definiálásakor a „sor" fogalmának hiánya. A relációs séma egy előfordulása azonos szerkezetű sorok (véges) halmaza, így természetesen adó-

dik sor-párok kiválasztása a funkcionális függőség ellenőrzéséhez. Egy XML dokumentum szerkezetileg gyökeres fa, amelyben a belső csomópontok XML elemek (jelölők), a levelek az elemek értékei, ezért az XML világban nincs magától értetődő, természetesen adódó, általánosan elfogadott „sor" fogalom, és ha ki is választjuk az elemek valamely együttesét és „sor"-nak nyilvánítjuk őket, általában nehéz találni egy „jó" összehasonlító eljárást.

Mivel az XML elemeket a sémanyelvek reguláris kifejezésekkel írják le, kézenfekvő, hogy a függőségeket egy reguláris nyelven (a nyelv mondatain) értelmezzük. Ezzel a megközelítéssel nyilván az XFD-nél (vagy más, XML függőségnél) lényegesen szűkebb fogalomhoz jutunk, hiszen így az XML hierarchikusan tagolt fa-szerkezete helyett egyetlen elem ugyan összetett, de zárt világára korlátozódunk. Mindenesetre, ezen az úton az egyes függőségek egyszerű, de elég általános definícióit sikerül megadni, amely definíciók az adatmodellek egy tág osztályán érvényesek: egyetlen feltétel van, az „adatsorok" legyenek egy reguláris nyelv mondatai, azaz, generálja azokat egy reguláris grammatika vagy ezzel ekvivalens módon, feleljenek meg egy reguláris kifejezésnek.

A reguláris nyelveken értelmezett funkcionális függőség (RFD) modelljét az XML teljes leírására alkalmas kiterjesztett környezetfüggetlen nyelvű modellben is elhelyeztük, összetett értékű adatok bevezetésével kerülve el a fa-szerkezetű modell felhasználását. Az RFD standard modellje a teljes általánosságot megszorító feltételeket (megkötés) használ, azért, hogy az üres adatok, végtelen ábécék, rendezetlen halmazok összehasonlítása, számlálók kezelése ne váljon szükségessé. Ezekkel a megkötésekkel a logikai implikáció kezelése is egyszerűbbé vált. Külön témakörként foglalkozik az Értekezés a megkötések feloldásával. Sikerült megmutatni, hogy a megkötések elhagyásával is érvényesek, bár természetesen szűkített formában, az eredeti modell eredményei, a tetszőleges számú diszjunkciót tartalmazó reguláris kifejezésre épített RFD például, az XFD modelltől eltérően, végesen axiomatizálható, ha az üres jelsorozatot érvényes jelsorozatnak tekintjük.

1. Reguláris nyelv duális nyelve és a kiterjesztett reláció

1.1 Definíció (Duális nyelv)
Legyen L reguláris nyelv amelyet az $M(L)$ véges automata fogad el. Az L nyelv duális nyelvének ábécéjét az $M(L)$ automata állapotai alkotják (jelölése $\Phi(L)$), a $D(L)$ duális nyelv mondatai az állapotszimbólumok sorozatai: az $M(L)$ automata gráfjának az L nyelv mondatait elfogadó bejárásai (START-

tól END-ig). Ha $t \in L$ *és* $\mathrm{w} \in D(L)$ *a t-t elfogadó bejárás, akkor azt mondjuk, hogy t típusa* w*, azaz* $\mathrm{w} = type(t)$*.*

A $D(L)$ duális nyelv mindegyik mondata egy sortípust határoz meg; ezen típusok összessége a „kiterjesztett reláció" sémája. A kiterjesztett reláció egy I előfordulása egy rögzített típushoz (duális mondathoz) tartozó kiterjesztett sorok egy halmaza.

1.2 Definíció (Kiterjesztett reláció)
Legyen L reguláris nyelv amelyet az $M(L)$ *automata fogad el, legyen* $D(L)$ *a társított duális nyelv.*
$R(L)=\{t|t \in L, type(t) \in D(L)\}$ *kiterjesztett reláció L felett. Az R kiterjesztett reláció sémája a társított duális nyelv, azaz* $schema(R) = D(L)$*. Legyen* $\mathrm{w} \in D(L)$ *egy duális mondat, akkor az* $I_{\mathrm{w}}(L) = \{t|t \in L, type(t) = \mathrm{w}\}$ *(vagy egyszerűen* I_{w} *ha L, vagy I ha* w *is egyértelmű a szövegkörnyezetben) sorhalmaz* $R(L)$*-nek egy előfordulása.*
A továbbiakban jelöljük $\mathrm{w}=schema(I)$*-vel az I előfordulás sémáját (sorainak azonos* w *típusát), és* $\mathcal{I}(L) = \{I_{\mathrm{w}_1}, I_{\mathrm{w}_2}, \ldots; \mathrm{w}_i \in D(L)\}$*-vel az előfordulások halmazát L felett.* $\mathcal{I}(L)$ *akkor és csak akkor véges ha az L nyelv nem-rekurzív. Ha* $t \in I \in \mathcal{I}(L)$*, akkor* $type(t) = schema(I) \in schema(R)$*.*

2. A kiterjesztett reláció attribútumai

A relációs modellben a függőségek (tartalmazási, többértékű, funkcionális, összekapcsolási) szintaktikus leírásához relációs attribútumok halmazait használjuk. A reguláris nyelvre alapozott modellben is így fogunk eljárni, ehhez először meg kell határoznunk az attribútumok fogalmát, valamint azt is, miként tudunk egy séma attribútumaiból a függőségek számára alkalmas részhalmazokat kiválasztani.

2.1 Definíció (Reguláris attribútum)
Legyen L reguláris nyelv amelyet az $M(L)$ *véges automata fogad el. Legyen* $D(L)$ *az L duális nyelve és legyen* $\mathrm{w} \in D(L)$ *egy nem üres duális mondat (*$\mathrm{w} \neq \epsilon$*). Legyen továbbá* $\mathrm{w}=v_1\, v_2\, \ldots\, v_n; v_i \in \Phi(L), 1 \leq i \leq n$*. A* $v_i, 1 \leq i \leq n$ *(nem feltétlenül különböző, de sorrendjük szerint megkülönböztetett) szimbólumokat a* w *duális mondathoz, mint sortípushoz tartozó reguláris mondatok (sorok) attribútumainak nevezzük. Ha* $t \in I_{\mathrm{w}}$ *(*$type(t) = schema(I_{\mathrm{w}}) = \mathrm{w}$*) és* $t = t_1\, t_2\, \ldots\, t_n$*, akkor* $t[v_i] = t_i$ *a t sor* v_i *attribútumon felvett értéke, vagy, a* v_i *attribútum a t sorból a* t_i *értéket vetíti ki.*

Ha az L nyelv nem rekurzív, azaz ha az $M(L)$ automata gráfja körmentes (a generáló reguláris grammatika / kifejezés nem-rekurzív), akkor a társított

$D(L)$ duális nyelv véges, mivel az $M(L)$ gráfnak véges sok bejárása van (azonos jelölést - $M(L)$ - használunk a véges automatára és gráf reprezentációjára). Ezen duális mondatok mindegyikén kijelölhetünk a függőségeket specifikáló attribútumhalmazokat.

Ha az $M(L)$ automata gráfja tartalmaz kört (a generáló reguláris grammatika / kifejezés rekurzív), akkor a társított $D(L)$ duális nyelv végtelen. Használhatjuk a pumpálási eljárást arra, hogy kiválasszunk részsorozatokat a duális mondatokból és függőségeket értelmezzünk ezeken a részsorozatokon. Kiválaszthatunk utakat a nem-pumpált részen, azután vehetünk részsorozatokat a pumpált részen és együtt pumpálhatjuk a kiválasztott részeket. Természetesen, a bejárt csúcsokat mindig egy teljes bejárásból kell kiválasztanunk ($\mathcal{START}$-tól $\mathcal{END}$-ig) a bejárás sorrendjében, és az ismétlődések is a bejárás sorrendjében szerepelnek, mivel a (bejárás által generált) duális mondat a definiálandó függőség értelmezési tartománya.

Megjegyezzük, hogy a függőség szintaktikus definícióját a $D(L)$ duális nyelven adjuk meg; a függőség szemantikáját (a kielégítettség ellenőrzését) az L nyelv mondatain értelmezzük.

El szeretnénk kerülni, hogy a definiálandó függőséget minden egyes (esetleg végtelen sok) duális mondaton meg kelljen adnunk: ehelyett az $M(L)$ gráfon specifikáljuk a függőséget, és ezt a specifikációt alkalmazzuk a duális mondatokra (kiválasztás). Abból a célból, hogy a függőséget specifikálhassuk, kijelölhetünk csúcsokat és kimondhatjuk, hogy egy bejárás során minden áthaladás kiválasztja őket. Kijelölhetünk kezdő és befejező csúcsokat a bejárásból, úgy, hogy ez a két csúcs az útvonal minden záródásánál kerüljön kiválasztásra.

2.2 Definíció (Kijelölés)
Legyen L reguláris nyelv és legyen $M(L)$ a nyelvet elfogadó véges automata gráfja. Az $M(L)$ feletti Y kijelölésnek nevezzük az (Y_1, Y_2) párost, ahol $Y_1 \subseteq \subseteq nodes(M(L))$ és Y_2 az $M(L)$ gráf tranzitív lezártjának egy részgráfja. Y_1-et az $M(L)$ gráf azon csúcsaiból vesszük, amelyek nem tartoznak körhöz, Y_2 pedig olyan csúcsokat tartalmaz amelyek egy bejárás során ismétlődően szerepel(het)nek.

Egy kijelölés egy adott duális mondatból egyértelműen kell, hogy kiválasszon egy részsorozatot. Megadunk két különböző, az alábbiakban részletezett módszert amelyekkel a kijelölés (a módszerre nézve egyértelműen) kiválaszt egy részsorozatot.

Legyen $\mathrm{w} = v_1 v_2 \dots v_n$ duális mondat (w szimbólumsorozat) az $M(L) = (V, E)$ gráf felett, és jelölje $walk\,(\mathrm{w}) = (\mathcal{START}, v_1, e_1, v_2, e_2, \dots, e_{n-1}, v_n, \mathcal{END})$ (rövidítve $walk\,(\mathrm{w}) = (v_1, v_2, \dots, v_n)$) $M(L)$-nek a w-t előállító bejárását.

2.3 Definíció (Szigorú kiválasztás)
Legyen $Y = (Y_1, Y_2)$ egy kijelölés és w *duális mondat $M(L)$ felett. Legyen $walk\,(\mathrm{w}) = (v_1, v_2, \dots, v_n)$. Y_1 szimbólumai bejárásuk szerint kerülnek kiválasztásra (ha a bejárás érinti őket). Minden $e \in E(Y_2)$ él esetén, az él végpontjai az elérés sorrendjében kerülnek kiválasztásra (ha a bejárás érinti őket) a $walk\,(\mathrm{w})$ szerinti bejárás során, mindannyiszor, amikor mindkét végpont sorra kerül. Azaz, ha az e él két végpontja A és B és ha $A = v_i, B = v_j$ valamely $1 \leq i < j \leq n$-re, akkor v_i és v_j akkor és csak akkor választódik ki, ha $v_k \neq A, B; i < k < j$. Y_2 izolált csúcsait minden áthaladásnál (ha van ilyen) kiválasztjuk a $walk\,(\mathrm{w})$ által megadott bejárás során. A kiválasztást Y_2 minden élére és izolált csúcsára egymástól függetlenül elvégezzük. A kiválasztási folyamat végére a* w *duális mondatból kiválasztott szimbólumokból felépül az (esetleg üres)* $\mathrm{w}[Y] = v_{i_1} v_{i_2} \dots v_{i_k}$ *$(1 \leq i_1 < i_2 < \dots < i_k \leq n\,(k \geq 0))$ sorozat.*

Legyen $t \in L$, a megfelelő duális mondat $\mathrm{w} \in D(L)$. A $\mathrm{w}[Y]$ szimbólumsorozatot interpretálhatjuk úgy mint egy (nem szükségképpen különböző) „attribútumokból" álló tömböt (array), amely a t kiterjesztett „sor"-t az értékek (tömb index szerint) megfelelő $t[Y]$ tömbjére vetíti. Ha $\mathrm{w}[Y] = \epsilon$, akkor $t[Y] = \epsilon$ úgyszintén (ϵ jelöli az üres szimbólumsorozatot).

2.4 Definíció (Teljes kiválasztás)
Legyen $Y = (Y_1, Y_2)$ egy kijelölés és w *duális mondat $M(L)$ felett. Legyen $walk\,(\mathrm{w}) = (v_1, v_2, \dots, v_n)$. Y_1 szimbólumai bejárásuk szerint kerülnek kiválasztásra (ha a bejárás érinti őket). Minden $(A, B) \in E(Y_2)$ él esetén, az A végpont minden olyan érintés esetén kiválasztásra kerül, ha ezt követően a B csúcs is sorra kerül a $walk\,(\mathrm{w})$ bejárás során, és a B csúcs minden olyan érintésnél kiválasztásra kerül, ha az A csúcsot követi. Az (A, B) él nem indukálja az A csúcs kiválasztását a bejárás során, ha a B csúcs a bejárás további részében nincs jelen, és hasonlóképpen, a B kiválasztására nem kerül sor, ha a bejárásban őt megelőzően az A csúcs nem szerepelt. A teljes kiválasztás azt jelenti, hogy az elsőként bejárt A csúcs és az utolsóként bejárt B csúcs között valamennyi A és B csúcs kiválasztásra kerül. Y_2 izolált csúcsait minden áthaladásnál (ha van ilyen) kiválasztjuk a $walk\,(\mathrm{w})$ által megadott bejárás során. A kiválasztást Y_2 minden élére és izolált csúcsára egymástól függetlenül elvégezzük. A kiválasztási folyamat végére a* w *duális mondatból kiválasztott szimbólumokból felépül az (esetleg üres)* $\mathrm{w}\{Y\} = v_{i_1} v_{i_2} \dots v_{i_k}$ *$(1 \leq i_1 < i_2 < \dots < i_k \leq n\,(k \geq 0))$.*

Legyen $t \in L$, a megfelelő duális mondat $\mathrm{w} \in D(L)$. A $\mathrm{w}\{Y\}$ szimbólumsorozatot interpretálhatjuk úgy mint egy (nem szükségképpen különböző) „attribútumokból" álló tömböt (array), amely a t kiterjesztett „sor"-t az értékek (tömb index szerint) megfelelő $t\{Y\}$ tömbjére vetíti. Ha $\mathrm{w}\{Y\} = \epsilon$, akkor $t\{Y\} = \epsilon$ úgyszintén (ϵ jelöli az üres szimbólumsorozatot).

3. Reguláris funkcionális függőség

Legyen L reguláris nyelv, legyen $M(L)$ az elfogadó automatája, legyen $D(L)$ a társított duális nyelv. Kiválaszthatunk két részsorozatot mindegyik duális mondaton, mint a függőség bal és jobb oldalát, tekintsük ezt a függőség szintaktikus specifikációjának. Az I előfordulás (az L reguláris nyelv mondatainak halmaza) kielégíti ezt a függőséget, ha nincs két olyan sor I-ben amelyek azonosak a bal oldali részsorozat duális nyelvi szimbólumaihoz illeszkedő szimbólumokban, de legalább egy jobb oldali duális nyelvi szimbólumhoz illeszkedő szimbólumban különböznek.

3.1 Definíció (Reguláris funkcionális függőség)
Legyen L reguláris nyelv és legyen $M(L)$ az L nyelvet elfogadó véges automata gráf reprezentációja. Legyen $X = (X_1, X_2)$ és $Y = (Y_1, Y_2)$ két kijelölés $M(L)$ felett. Az $M(L)$ felett értelmezett funkcionális függőségnek (reguláris FD; RFD) nevezzük az $X \to Y$ kifejezést. Az $I \in \mathcal{I}(L)$ (véges) adatbázis előfordulás kielégíti az $X \to Y$ funkcionális függőséget (jelölve $I \models X \to Y$), ha bármely két $t_1, t_2 \in I$ sorra $t_1[X] = t_2[X]$ csak akkor teljesülhet, ha $t_1[Y] = t_2[Y]$ is igaz. Az $Y = M(L)$ esetet kulcs függőségnek nevezzük.

3.1 Megjegyzés
Azt mondjuk, hogy az L reguláris nyelv gyengén kielégíti a $\sigma = X \to Y$ RFD-t (jelölve $L \overset{w}{\models} \sigma$), ha bármely két $t_1, t_2 \in L$ sor esetén, amelyekre $\mathrm{w} = type(t_1) = type(t_2)$ teljesül, ha a két sor megegyezik X-en, azaz, $t_1[X] = t_2[X]$, akkor $t_1[Y] = t_2[Y]$ is teljesül. Azt mondjuk, hogy az L reguláris nyelv erősen kielégíti a σ RFD-t (jelölve $L \overset{s}{\models} \sigma$), ha bármely két $t_1, t_2 \in L$ sor esetén, amelyekre $\mathrm{w}_i = type(t_i)$, $i = 1,2$ és $\mathrm{w}_1[X] = \mathrm{w}_2[X]$ és $\mathrm{w}_1[Y] = \mathrm{w}_2[Y]$ teljesül, mindannyiszor ha a két sor megegyezik X-en, azaz, $t_1[X] = t_2[X]$, akkor $t_1[Y] = t_2[Y]$ is teljesül. Nyilván, ha $L \overset{s}{\models} \sigma$, akkor $L \overset{w}{\models} \sigma$. A következőkben, ha nem jelezzük másként, a funkcionális függőség kielégítését a 3.1. Definícióban specifikált módon értelmezzük.

4. Reguláris funkcionális függőségek logikai implikációja

4.1 Definíció
Legyen L reguláris nyelv és legyen $D(L)$ a társított duális nyelv és legyen $M(L)$ az L nyelv elfogadó véges automatájának gráf reprezentációja. Legyen X kijelölés $M(L)$ felett és legyen $\mathrm{w} \in M(L)$ duális mondat. Azt mondjuk, hogy w *magában foglalja (subsumes) X-et (jelölve $X \sqsubseteq \mathrm{w}$) ha $nodes(X) \subseteq nodes(\mathrm{w})$. Legyen $\sigma = X \to Y$ egy RFD, azt mondjuk, hogy* w *magában foglalja σ-t (jelölve $\sigma \sqsubseteq \mathrm{w}$) ha $X \sqsubseteq \mathrm{w}$ és $Y \sqsubseteq \mathrm{w}$. Legyen $I \in \mathcal{I}(L)$ egy előfordulás és legyen $\mathrm{w} = schema(I)$. I magában foglalja X-et (jelölva $X \sqsubseteq I$) ha $X \sqsubseteq \mathrm{w}$. Ha Σ RFD-k egy halmaza, akkor $\Sigma \sqsubseteq I$ akkor és csak akkor, ha $\forall \sigma \in \Sigma$-ra teljesül, hogy $\sigma \sqsubseteq I$.*

A következő *Feltételt* fogjuk használni ha az L reguláris nyelv függőségeinek egy Γ halmazáról van szó:

$$\text{Létezik egy } I \in \mathcal{I}(L) \text{ előfordulás úgy, hogy } \Gamma \sqsubseteq I \tag{1}$$

4.1 Megjegyzés
A 3.1. Definíció megadja a $M(L)$ gráfon értelmezett funkcionális függőségek szintaktikus formáját. Az RFD-k szemantikáját az L reguláris nyelv valamely előfordulásán értelmezzük. Ha csak egy RFD-ről van szó, a szemantika fogalma egyértelmű: egy adott előfordulás vagy kielégíti az adott RFD-t vagy nem. Ha viszont az RFD-knek egy Σ halmazáról akarunk valamit megállapítani (például logikai implikációt), meg kell követelnünk az (1) Feltétel teljesülését, azt, hogy az összes Σ-beli függőséget valamely $I \in \mathcal{I}(L)$ (egyazon) előfordulás magában foglalja. Legyen $X \to Y \in \Sigma$, legyen $\mathrm{w} = schema(I)$, akkor (1)-ből következik, hogy $nodes(\mathrm{w}[X]) = nodes(X)$ és $nodes(\mathrm{w}[Y]) = nodes(Y)$. Azaz, az I előfordulás sortípus duális mondata valamennyi szereplő kijelölést teljesen lefedi.

4.1. Észrevétel. *Ha X egy kijelölés és $X \sqsubseteq I$ (legyen $\mathrm{w} = schema(I)$), akkor $nodes(\mathrm{w}[X]) = nodes(X)$.*

4.2. Észrevétel. *Ha X egy kijelölés és $X \sqsubseteq I$ (legyen $\mathrm{w} = schema(I)$), akkor $nodes(\mathrm{w}[X]) \neq \{\}$.*

4.2 Definíció
Legyen L reguláris nyelv és legyen $M(L)$ az L nyelvet elfogadó véges automata gráf reprezentációja. Legyen Σ RFD-k egy halmaza és legyen $X \to Y$ egy RFD $M(L)$ felett, azt mondjuk, hogy Σ implikálja $X \to Y$-t (jelölve $\Sigma \models X \to Y$)

ha minden olyan(véges) $I \in \mathcal{I}(L)$ adatbázis előfordulás esetén amely kielégíti Σ-t (feltételezve, hogy $(\Sigma \cup X \to Y) \sqsubseteq I$), $I \models X \to Y$ is teljesül.

4.1. Algoritmus. *Algoritmus RFD-k implikációjának ellenőrzésére.*
Input: az $M(L) = (V, E)$ gráf, az RFD-k egy Σ halmaza és a $\sigma : X \to Y$ RFD (ahol $X = (X_1, X_2)$ és $Y = (Y_1, Y_2)$) és mindegyik reguláris funkcionális függőség $M(L)$ felett értendő
Output: igaz, ha $\Sigma \models \sigma$, egyébként hamis

1. Inicializálás
színezzük feketére az $M(L)$ gráfot és kékre a tranzitív lezártját
színezzük zöldre az X csúcsait és éleit mindkét gráfon
színezzük sárgára az Y (nem X-beli) csúcsait és éleit az $M(L)$ gráfon
színezzük pirosra az Y (nem X-beli) csúcsait és éleit a $M(L)$ tranzitív lezártján
2. $FDSET := \Sigma$;
3. $greene := X$
4. ismételjük, amíg van alkalmazható függőség:
ha $W \to Z \in \Sigma$ és $W \subseteq greene$ akkor
i. $FDSET := FDSET - (W \to Z)$;
ii. $greene := greene \cup Z$
iii. színezzük zöldre W csúcsait és éleit mindkét gráfon

5. ha a sárga és piros csúcsok és élek száma mindkét gráfon = 0, akkor output=igaz egyébként output=hamis.

4.1 Lemma
A 4.1. Algoritmus eredménye nem függ a Σ és $X \to Y$ esetén használt (azonos) kiválasztási módszertől.

4.1 Propozíció (Reguláris funkcionális függőségek implikációja)
Legyen L reguláris nyelv és legyen Σ RFD-k egy halmaza és legyen $X \to Y$ egy RFD (valamennyien $M(L)$ felett), akkor $\Sigma \models X \to Y$ akkor és csak akkor, ha a 4.1. Algoritmus az $M(L)$, Σ és $X \to Y$ inputtal igaz eredményt hoz. Az implikáció teljesülése nem függ a Σ és $X \to Y$ esetén alkalmazott (közös) kiválasztási módszertől.

A következő, a relációs adatbázisok világából származó lemma, reguláris funkcionális függőségekre is érvényes.

4.2 Lemma
Legyen Σ RFD-k egy halmaza és $X \to Y$ egy RFD ugyanazon $M(L)$ felett. Akkor $\Sigma \models X \to Y$ akkor és csak akkor, ha $Y \subseteq greene$ amikor a 4.1. Algoritmus véget ér.

4.2 Megjegyzés
A 4.1. Algoritmus $O(|\Sigma|^2)$-ben fut, ahol $|\Sigma|$ az RFD-k száma Σ-ban. A relációs modellben a logikai implikáció lineáris időben eldönthető, mivel a relációs implikációs probléma az ítéletlogikai HORN-SAT problémával ekvivalens. Egy $X \to Y$ reguláris FD és w *duális mondat esetén, legyen* $\mathrm{w}[X] = \mathrm{w}_1 \ldots \mathrm{w}_n$ *és legyen* $\mathrm{w}[Y] = z$*, akkor a megfelelő $W_1 \ldots W_n$ és Z ítéletlogikai változók száma* w*-től függ, azaz, nem korlátos ha a reguláris nyelv rekurzív.*

5. Reguláris funkcionális függőségek axiomatizálása

Az XML funkcionális függőségekre az általános esetben nem létezik véges axiomatizáció: Arenas és Libkin kimutatták (egy, a k-méretű axiomatizálásra vonatkozó tétel alapján) hogy a logikai implikáció az általuk leírt XML funkcionális függőségi osztály (tree tuples modell) esetében végesen nem axiomatizálható ha a sémaleírás (DTD) tetszőleges számú diszjunkciót tartalmazhat. Úgy látszik azonban , hogy a reguláris funkcionális függőségek esetében van egy helyes és teljes Armstrong-típusú axiómarendszer. Ez azért lehet így, mert az RFD nem igazából XML funkcionális függőség: az RFD mindössze egy „horizontális" dimenzióra korlátozódik, anélkül, hogy az XML fa-struktúráját modellezni tudná. Továbbá, az (1) feltétel kizárja a kezelhetetlen diszjunkciókat.

A reguláris funkcionális függőségre a három Armstrong axiómát a következőképpen adhatjuk meg:
Legyen L reguláris nyelv, legyen $M(L)$ a nyelvet elfogadó automata. Legyenek X, Y, Z kijelölések $M(L)$ felett.

1. Reflexivitás(*RFD-1*): ha $Y \subseteq X$, akkor $X \to Y$
2. Bővítés(*RFD-2*): ha $X \to Y$, akkor $X \cup Z \to Y \cup Z$
3. Tranzitivitás(*RFD-3*): ha $X \to Y$ és $Y \to Z$, akkor $X \to Z$

5.1. Tétel. *Az {RFD-1,RFD-2,RFD-3} axiómarendszer a reguláris funkcionális függőségek implikációjára nézve helyes és teljes.*

Publikációs jegyzék
2013. július 28.

Folyóiratcikk:

[1] Szabó, G. I., Benczúr, A.: Functional Dependencies on Extended Relations Defined by Regular Languages. Annals of Mathematics and Artificial Intelligence, DOI: 10.1007/s10472-013-9352-z

Referált konferenciakiadványok:

[2] Szabó, G. I., Benczúr, A.: Functional Dependencies on Extended Relations Defined by Regular Languages. Proceedings of Foundations of Information and Knowledge Systems, Kiel, Germany, March 2012, Lecture notes in Computer Science, 7153 (eds: T. Lukasiewicz, A. Sali), Springer, p. 385–404

[3] Szabó, G.I., Benczúr, A.: Functional Dependencies on Symbol Strings Generated by Extended Context Free Languages. Proceedings of Advances in Databases and Information Systems, Poznan, Polen, September 2012, Advances in Intelligent Systems and Computing, 186 (eds. T. Morzy, T. Härder, R. Wrembel), Springer, p. 253–264

Referált konferenciakivonatok:

[4] Szabó, G.I., Benczúr, A.: Encapsulated Functional Dependencies for XML Design, Research Conference on Information Technology, 24–25 October, 2011, Pécs, Hungary, C129, ISBN 978-963-7298-46-2

[5] Szabó, G.I.: How Much is XML Involved in DB Publishing? Conference of PhD Students in Computer Science, June 29–July 2, 2010, Szeged Hungary