

Mis on relatsioon?

... ja kuidas ta on seotud SQL'is kasutatavate tabelitega.

Käesolev tekst on loodud neile, kel tekib raskusi aine Andmebaasid MTAT.03.264 neljanda õppenädala loengus esitatud relatsioonilise andmebaaside teooria põhimõistega „relatsioon“. Siit leiate nii pikema selgituse, näited kui ka seosed SQL andmetabelitega.

Relatsiooni definitsioon

Kui me vaatame oma andmebaasi graafilist mudelit, siis sisaldab see infot olemitüüpide, nende tunnuste ja olemitüüpide omavaheliste seoste ning nende seoste tüüpide kohta (1:1, 1:n, n:m seosed). Alustame seekord arutlust tunnustest. Igal mudelis esineval tunnusel on mingi hulk võimalikke väärtusi. Näiteks TÜ tudengite matriklinumbrid on tüüpiliselt tähega A või B algavad numbrijadad, kus on reeglina 5 numbrit, näiteks „A12345“. Tähistame seda kõikvõimalike matriklinumbrite hulka D.

Kuigi matriklinumbrite puhul on hulga D kõikvõimalike väärtuste hulk üsna suur (ülikoolis õpib ca 18000 üliõpilast, kõigil on erinevad matriklinumbrid ja kui me võtame arvesse ka neid, kes on juba lõpetanud, kasvab võimalike matriklinumbrite arv veel suuremaks), siis on ta siiski lõplik. Mõne tunnuse väärtuste hulk võib olla ka lõpmatu – näiteks, kui Maaülikooli uurijad hoiavad andmebaasis vastsündinud vasikate andmeid ja ühe tunnusena on kasutusel vasika sünnikaal, on see tavaline reaalarv, küll mingites mitte väga suurtes piirides muutuv, aga ikkagi reaalarv ja erinevate väärtuste arv põhimõtteliselt lõpmatu. Mõne tunnuse võimalike väärtuste arv on väga väike: näiteks isiku tunnus „sugu“ võib omandada ainult 2 erinevat väärtust, eksamil saadud hinne - kuus erinevat väärtust A, B, C, D, E või F.

Olgu meil mingi hulk selliseid tunnuste võimalike väärtuste hulki, tähistame neid D_1, D_2, \dots, D_n . Oluline on, et kõik hulgad sisaldaks ainult lihtväärtusi, s.t. et mitte üheski hulgas D_i ei oleks struktuurseid väärtusi – hulki, korteeže jne. Defineerime nüüd nende hulkade baasil mõiste relatsioon.

Def. Realatsioon. Relatsiooniks nimetatakse nende lihtväärtushulkade otsekorrutise alamhulka:

$$R \subseteq D_1 \times D_2 \times \dots \times D_n \quad (*)$$

Otsekorrutise näide

Näide: Olgu D_1 matriklinumbrite hulk, D_2 tudengi sugu, D_3 elu esimese ülikoolieksami tulemus. Näite lihtsuse huvides olgu hulgas D_1 ainult 3 väärtust: A12345, A54321 ja B99999. Hulgas D_2 on 2 väärtust: M ja N. Hulgas D_3 on kuus väärtust: A, B, C, D, E, F.

Siis see otsekorrutis $D_1 \times D_2 \times D_3$ on hulk, mis sisaldab kõikvõimalikke kolmikuid, mille esimene element pärineb hulgast D_1 , teine – hulgast D_2 ja kolmas – hulgast D_3 . Antud näite puhul on selliseid kõikvõimalikke kolmikuid 36 tükki. On selge, et reaalsete andmete peale mõeldes on otsekorrutise elementide hulk üüratult suur. Õnneks pole seda otsekorrutist realselt mitte kunagi vaja leida 😊.

Esitame need otsekorrutisena saadud kolmikud tabelina:

D1	D2	D3
A12345	M	A
A12345	M	B
A12345	M	C
A12345	M	D
A12345	M	E
A12345	M	F
A12345	N	A
A12345	N	B
A12345	N	C
A12345	N	D
A12345	N	E
A12345	N	F
A54321	M	A
A54321	M	B
A54321	M	C
A54321	M	D
A54321	M	E
A54321	M	F
A54321	N	A
A54321	N	B
A54321	N	C
A54321	N	D
A54321	N	E
A54321	N	F
B99999	M	A
B99999	M	B
B99999	M	C
B99999	M	D
B99999	M	E
B99999	M	F
B99999	N	A
B99999	N	B
B99999	N	C
B99999	N	D
B99999	N	E
B99999	N	F

Vahemärkus: Kui seda tabelit on tüütu lugeda, siis veel tüütum oli teda kirja panna, aga loodan, et ta selgitab, mida tähendab kolme hulga otsekorrutis.

Märkus: tulemus on alati hulk, s.t. kolmikute järjekord ei ole oluline ja korduvaid elemente ei ole. Neid ei saa olla juba tänu otsekorrutise definitsioonile ja sellele, nii D_1 , D_2 kui D_3 olid hulgad, s.t. nendes ei ole korduvaid elemente.

Relatsiooni näide

Nüüd on relatsioonini ainult üks samm. Relatsioon on sellise otsekorrutise mistahes alamhulk. Näiteks, kui tahame, et meie relatsioon R kujutaks kolme tudengi ülikoolis saadud kõige esimesi hindeid, võib selle relatsiooni seisund olla näiteks selline (ka relatsiooni paneme seekord kirja tabelina):

D1	D2	D3
A12345	M	A
A54321	N	C
B99999	M	E

Võime selle kirja panna ka näiteks nii: $\{ \langle A12345, M, A \rangle, \langle A54321, N, C \rangle, \langle B99999, M, E \rangle \}$. Viimane kirja pilt võtab väikeste hulkade puhul vähem ruumi, esimene on üldiselt ülevaatlikum.

Relatsiooni seos andmebaasi tabelitega.

Me kasutame relatsiooni mõistet, et esitada mingi olemitüübi reaalsuses esinevat seisundit. Kui meil on 3 tudengit ja meie olem kajastab nende esimese eksami tulemusi, siis ei saa see tabel koosneda kolmest sellisest reast (kuigi ka need 3 rida on selle suure 36-realise tabeli alamhulk):

$\{ \langle A12345, M, A \rangle, \langle A12345, N, C \rangle, \langle B99999, M, E \rangle \}$

See, et andmebaasi tabeli seisund vastaks tegelikkusele, on juba andmehõive küsimus, millega me käesoleva formaalse aparatuuri ülesehitamise hetkel ei tegele. Eeldame alati, et andmed relatsiooni seisundis vastavad reaalsuse seisundile.

Selleks, et oleks võimalik aru saada, milliste tunnustega meil tegemist on, paneme me neile nimed. Iga nimi on seotud tunnuse võimalike väärtuste hulgaga.

Formaalselt tähistame tunnuste nimesid tähestiku alguse suurtähtedega, vajadusel kasutame indekseid. Konkreetsetes näidetes kasutame mõttekaid sõnu või fraase (nagu näiteks: matrikli number, isiku sugu, hinne). Üldisel kujul on meil siis relatsioon $R(A_1, A_2, \dots, A_n)$, kus tunnuse A_1 võimalikud väärtused on hulgast D_1 , tunnuse A_2 väärtused hulgast D_2 jne.

Kirjutist $R(A_1, A_2, \dots, A_n)$ nimetame relatsiooni kirjelduseks. Tabelit, veergudega A_1, A_2, \dots, A_n ja lahtrites vastavate väärtustega, nimetame relatsiooni seisundiks. Seisund peab vastama reaalsuses vastavate olemieksemplaride andmetele.

Võib juhtuda, et mingeid väärtushulki D kasutatakse relatsiooni kirjeldamisel mitu korda. Näiteks, kui esmakursuslaste edasijõudmise uurimiseks on vaja tabelit, kus matrikli numbri ja isiku soo kõrval oleks kolme esimese eksami hinded, siis on kõigi kolme viimase veeru tunnuse väärtuste hulk eelmistes näidetes vaadatud D_3 . See on täiesti lubatud. Nüüd on aga vaja nimesid pannes hoolikalt jälgida, et nimed relatsiooni kirjelduses oleks erinevad. Näiteks: $R(\text{matrikli number, isiku sugu, esimene hinne, teine hinne, kolmas hinne})$.

Seega SQL tabel andmebaasis esindab relatsiooni seisundit, kus iga rida (kirje) esitab ühte olemitüübi eksemplari. Veerud kujutavad tunnuseid, igas lahtris on sellele reale vastava olemieksemplari vastava

tunnuse väärtus. Iga rea (kirje) antud tunnuse tüüp on relatsiooni kirjelduses kasutatud vastava väärtushulgas D_i esindatud mõni väärtus. Väärtushulgad D_i esindavadki antud tunnuse elementaarse andmetüübiga seotud kõikvõimalike väärtuste hulki.

Loodan, et need enam kui kolm lehekülge selgitasid relatsiooni mõistet. Järgmistes abimaterjalides järgmised mõisted.