Merjenje podobnosti (označenih) podatkovnih tabel

Measuring similarity of (annotated) data tables

Melita Hajdinjak

Univerza v Ljubljani, Fakulteta za elektrotehniko / Tržaška cesta 25, 1000 Ljubljana

E-Mail: melita.hajdinjak@fe.uni-lj.si

* Avtor za korespondenco; Tel.: +386-1-4768-385; Fax: + 386-1-4768-316

Povzetek: V članku predlagamo mero podobnosti klasičnih relacij oziroma podatkovnih tabel. Dobimo jo kot posplošitev Egli-Milnerjeve urejenosti in Hausdorffove metrike. Ta mera omogoča, da primerjamo različne podatkovne tabele. Mero podobnosti klasičnih relacij poskušamo razširiti na *D*-relacije, imenovane tudi relacije s podobnostmi, ki posplošujejo veliko skupino označenih relacij. V splošni obliki takšne mere zdaj nastopa še funkcija, ki meri podobnost označb. Izpostavimo lastnosti, ki naj jim ta funkcija zadošča, in poiščemo ustrezno obliko funkcije v primeru nekaterih posebnih označevalnih domen.

Ključne besede: relacijska algebra; mera podobnosti; označene relacije; De Morganov okvir; Egli-Milnerjeva urejenost; Hausdorffova razdalja

Abstract: We propose a measure of similarity for classical relations or data tables. It is obtained as a generalization of the Egli-Milner ordering and the Hausdorff metric. This measure allows us to compare different data tables. We attempt to extend the measure from classical relations to *D*-relations, also called relations with similarities, which generalize a large group of annotated relations. The general form of such a measure now contains a function for comparing annotations. We expose some properties of the annotation-comparing function and find suitable candidates in the case of some special annotation domains.

Keywords: relational algebra; similarity measure; annotated relation; De Morgan frame; Egli-Milner ordering; Hausdorff distance

1. Uvod

Vračanje informacij, ki niso natančni odgovori na zastavljeno vprašanje, so pa relevantni oziroma blizu tistemu, kar je bilo vprašano, je splošna lastnost človeške komunikacije, znana kot sodelujoče odgovarjanje [1]. Ker poizvedovanja po netočnih, približnih informacijah klasična relacijska algebra ni sposobna modelirati, so bile predlagane številne posplošitve in razširitve tega najbolj priljubljenega podatkovnega modela. Članek temelji na označenih relacijah, imenovanih *D*-relacije, ki jih dobimo tako, da domene atributov razširimo z merami podobnosti [2].

1.1. Klasične relacije

Naj bodo τ_i množice, *relacijska shema*

$$U = \{a_1{:}\tau_1,\ldots,a_n{:}\tau_n\}$$

pa preslikava, ki atributu a_i priredi domeno τ_i, torej $U(a_i) = \tau_i$ za $i = 1,\ldots,n$. Naj bo *U-terica* $t = \{a_1{:}v_1,\ldots,a_n{:}v_n\}$ preslikava, ki atributu a_i priredi element $v_i \in \tau_i$, torej $t(a_i) = v_i$ za $i = 1,\ldots,n$. Množico vseh *U*-teric imenujmo $U - \mathrm{Tup}$.

Označena relacija nad U imenujemo preslikavo oblike

$$A{:}U - \mathrm{Tup} \to K,$$

ki vsaki *U*-terici priredi nek element izbrane domene *K*. Pišemo $A \in Rel(U)$. Klasična relacija, ki jo običajno pojmujemo kot množico *U*-teric in jo predstavimo s tabelo, je poseben primer označene relacije. Dobimo jo, če

izberemo $K = \{0,1\}$ in definiramo $A(t) = 1$, če $t \in A$, ter $A(t) = 0$ sicer.

1.2. D-relacije

Naj bodo $L_i = (L_i, \vee_i, \wedge_i, 0_i, 1_i, \neg_i)$ De Morganovi okvirji, tj. polne mreže z negacijo $\neg_i : L_i \to L_i$, v katerih so končni infimumi distributivni glede na supremume. Naj bo *De Morganova shema*

$$D = \{a_1 : L_1, \ldots, a_n : L_n\}$$

preslikava, ki atributu a_i priredi De Morganov okvir L_i, torej $D(a_i) = L_i$ za $i = 1, \ldots, n$. Naj bo $D(U)$ -terica $s = \{a_1 : l_1, \ldots, a_n : l_n\}$ preslikava, ki atributu a_i priredi element $l_i \in L_i$. Množico vseh $D(U)$ -teric imenujmo $D(U) - \mathrm{Tup}$. Predpostavimo, da je na vsaki množici τ_i definirana refleksivna *mera podobnosti*

$$\rho i : \tau_i \times \tau_i \to L_i,$$

$\rho i(t(a_i), t(a_i)) = 1_i$, kjer je 1_i največji element De Morganovega okvirja L_i. Primeri takšnih refleksivnih mer podobnosti so:

- *enakost* $=_i : \tau_i \times \tau_i \to \{0,1\}$, ki vrne 1_i, če sta elementa enaka, in 0_i, če sta različna,
- *mehka enakost* $\approx_i : \tau_i \times \tau_i \to [0,1]$, ki vrne mehko stopnjo enakosti obeh elementov,
- *metrika* $d_i : \tau_i \times \tau_i \to [0, d_{max}]$, ki vrne razdaljo obeh elementov v metričnem prostoru.

Označeno relacijo

$$A : U - \mathrm{Tup} \to D(U) - \mathrm{Tup},$$

ki *U*-terico *t* preslika v $D(U)$ -terico $A(t) = \{a_1 : l_1, \ldots, a_n : l_n\}$, dobljeno s pomočjo mer podobnosti ρ_i, imenujemo *relacija s podobnostmi* ali *D-relacija*. Pišemo $A \in Rel(D(U))$.

Hajdinjak in Bierman [2] sta pokazala, da lahko na tako označenih relacijah definiramo vse operacije klasične relacijske algebre (unijo, projekcijo, izbiro, spoj, razliko). Na začetku poizvedovanja sta *U*-terico, ki ni v podatkovni tabeli, označila z najmanjšim elementom, $0 = \{a_1 : 0_1, \ldots, a_n : 0_n\}$, *U*-terico, ki je v podatkovni tabeli, pa z največjim elementom, $1 = \{a_1 : 1_1, \ldots, a_n : 1_n\}$. Relacijsko algebro na *D*-relacijah sta poimenovala *relacijska algebra s podobnostmi*. Taka algebra pravilno modelira glavne primere označenih relacij, kot so klasične relacije, c-tabele, tabele dogodkov in mehke relacije [3].

2. Podobnost relacij

Problem, ki ga obravnava članek, izvira s področja sodelujočega odgovarjanja. Odgovore na isto poizvedbo, ki so lahko samo približni, bi radi primerjali med sabo in tudi določili njihovo relevantnost.

2.1. Podobnost klasičnih relacij

Naj bo na vsaki množici τ_i iz $U = \{a_1 : \tau_1, \ldots, a_n : \tau_n\}$ definirana *mera podobnosti* $\rho_i : \tau_i \times \tau_i \to L_i$. Iščemo funkcijo $\rho_U : Rel(U) \times Rel(U) \to D(U) - \mathrm{Tup}$, ki bo paru klasičnih relacij ali podatkovnih tabel $A, B \in Rel(U)$ priredila njuno podobnost. Ker ima shema *U* v splošnem več različnih atributov, lahko delamo več različnih primerjav, zato smo predpostavili, da bo kodomena iskane mere ρ_U množica $D(U) - \mathrm{Tup}$. Primerjavo vrednosti atributov *U*-teric v relacijah *A* in *B* omogočajo mere ρ_i.

- Če je ρ_i enakost, ki slika v $L_i = \{0,1\}$, lahko za mero podobnosti relacij oziroma podatkovnih tabel (glede na atribut a_i) izberemo funkcijo, ki slika enaki množici v *1* in različni v *0*.
- Če je ρ_i refleksivna relacija, ki slika v $L_i = \{0,1\}$, lahko za mero podobnosti relacij (glede na atribut a_i) izberemo *Egli-Milnerjevo urejenost*, v kateri sta dve množici v relaciji, če je vsak element prve množice v relaciji z nekim elementom druge množice in za vsak element druge množice obstaja nek element v prvi množici, ki je z njim v relaciji [4,5]. Pomembnost te urejenosti v podatkovnih modelih so prepoznali že Buneman, Jung in Ohori [6].
- Če je ρi funkcija razdalje oziroma metrika, ki slika v $L_i = [0, d_{max}]$, lahko za mero podobnosti relacij (glede na atribut a_i) izberemo *Hausdorffovo metriko*, kjer sta dve množici blizu ena drugi, če je vsak element ene množice blizu nekemu elementu druge množice. Hausdorffova razdalja je najdaljša pot, ki jo moramo prepotovati od izbrane točke v eni množici do druge množice [4,7]. Ta razdalja je zelo široko uporabna (npr. v geometriji fraktalov, numerični matematiki in pri razpoznavanju vzorcev).

Zgoraj omenjene mere podobnosti posplošuje mera, ki sta jo na množicah že predlagala Hajdinjak in Bauer [4]. To mero sedaj prenesemo na klasične relacije.

Definicija 1. [Mera podobnosti na klasičnih relacijah] Naj bosta $A,B \in Rel(U)$, kjer je $U = \{a_1{:}\tau_1,\ldots,a_n{:}\tau_n\}$. Naj bo $\rho_i{:}\tau_i \times \tau_i \to L_i$ mera podobnosti na množici τ_i oziroma atributu a_i, kjer je $L_i = (L_i, \vee_i, \wedge_i, 0_i, 1_i, \neg_i)$ De Morganov okvir in $D = \{a_1{:}L_1,\ldots,a_n{:}L_n\}$ De Morganova shema. Na klasičnih relacijah nad shemo U definiramo naslednjo mero podobnosti:

$$\rho_U{:}Rel(U)\times Rel(U) \to D(U) - \mathrm{Tup},$$

$$\rho_U(A,B)(a_i) = \left(\wedge_{i,t\in A} \vee_{i,u\in B} \rho_i(t(a_i),u(a_i))\right) \wedge_i \left(\wedge_{i,u\in B} \vee_{i,t\in A} \rho_i(t(a_i),u(a_i))\right).$$

Tudi neskončni infimumi ($\wedge_i$) in supremumi ($\vee_i$) obstajajo, saj je L_i De Morganov okvir, ki je polna mreža.

Izrek 1. [Lastnosti mere ρ_U] Naj bo $D = \{a_1{:}L_1,\ldots,a_n{:}L_n\}$ De Morganova shema z $L_i = (L_i, \vee_i, \wedge_i, 0_i, 1_i, \neg_i)$, naj bo $\rho_U{:}Rel(U)\times Rel(U) \to D(U) - \mathrm{Tup}$ mera podobnosti na klasičnih relacijah nad U, naj bosta $A,B \in Rel(U)$ *in* $B \neq \{\}$. Potem velja:

$$\rho_U(A,A) = \{a_1{:}1_1,\ldots,a_n{:}1_n\} = \mathbf{1},$$

$$\rho_U(\{\},B) = \rho_U(B,\{\}) = \{a_1{:}0_1,\ldots,a_n{:}0_n\} = \mathbf{0}.$$

Vidimo, da ima mera podobnosti ρ_U nekatere pomembne lastnosti. Vsaka relacija je najbolj podobna (**1**) sama sebi, kar velja tudi za prazno relacijo, in prazna relacija je popolnoma različna (**0**) od vsake neprazne relacije. Ti dve lastnosti sta še posebej pomembni, ko med poizvedovanjem pričakujemo vsaj sodelujoče odgovore, ki so lahko le približek dejansko iskanih informacij. Opazimo tudi, da iz $\rho_U(A,B) = \mathbf{1}$ ne sledi nujno $A = B$.

Hitro lahko preverimo, da če so mere podobnosti $\rho_i{:}\tau_i \times \tau_i \to L_i$ simetrične, tj. velja $\rho_i(x,y) = \rho_i(y,x)$, je simetrična tudi mera ρ_U, tj. tedaj velja $\rho_U(A,B) = \rho_U(B,A)$.

2.2. Podobnost D-relacij

Ko imamo relacije oziroma podatkovne tabele, ki niso označene le z 0 ali 1, na njihovo podobnost zagotovo vplivajo tudi označbe. Zdaj bomo torej morali primerjati vrednosti atributov in še označbe vrstic v tabelah. Iskana mera podobnosti, imenujmo jo $\rho_{D,U}$, bo razširitev mere ρ_U s funkcijami primerjave označb, $w_i{:}L_i \times L_i \to L_i$.

Definicija 2. [Mera podobnosti na *D*-relacijah] Naj bosta $A,B \in Rel(D(U))$, kjer je $U = \{a_1{:}\tau_1,\ldots,a_n{:}\tau_n\}$ relacijska shema in $D = \{a_1{:}L_1,\ldots,a_n{:}L_n\}$ De Morganova shema z $L_i = (L_i, \vee_i, \wedge_i, 0_i, 1_i, \neg_i)$. Naj bo $\rho_i{:}\tau_i \times \tau_i \to L_i$ mera podobnosti na množici τ_i oziroma atributu a_i. Na *D*-relacijah definiramo naslednjo mero podobnosti:

$$\rho_{D,U}{:}Rel(D(U))\times Rel(D(U)) \to D(U) - \mathrm{Tup},$$

$$\rho_{D,U}(A,B)(a_i) = \wedge_{i,t\in U-\mathrm{Tup}} \vee_{i,u\in U-\mathrm{Tup}} \left(\rho_i(t(a_i),u(a_i)) \wedge_i w_i(A(t)(a_i),B(u)(a_i))\right)$$

$$\wedge_i \wedge_{i,u\in U-\mathrm{Tup}} \vee_{i,t\in U-\mathrm{Tup}} \left(\rho_i(t(a_i),u(a_i)) \wedge_i w_i(A(t)(a_i),B(u)(a_i))\right),$$

kjer je $w_i{:}L_i \times L_i \to L_i$ neka funkcija primerjave označb, povezana z atributom a_i.

Do ustrezne oblike funkcije $w_i{:}L_i \times L_i \to L_i$ lahko pridemo preko posplošitev Egli-Milnerjeve urejenosti in Hausdorffove metrike. Upoštevati pa moramo tudi, da so edina matematična orodja, ki jih imamo na razpolago, da definiramo takšno funkcijo, operacije in konstante, ki jih ponujajo De Morganovi okvirji (to so infimum, supremum in negacija ter najmanjši in največji element).

- Če je $L_i = (\{0,1\}, \vee, \wedge, 0, 1, \neg)$ Boolov De Morganov okvir, kjer je $\neg 0 = 1$ in $\neg 1 = 0$, lahko definiramo $w_i(x,y) = 1$, če $x = y$, in 0 sicer.
- Če je $L_i = ([0,d_{max}], min, max, d_{max}, 0, \neg)$ De Morganov okvir omejenih razdalj, kjer je $\neg x = d_{max} - x$, imamo

več povsem smiselnih možnosti, na primer $w_i(x,y) = \neq x - y \neq$ ali kakšno drugo razdaljo na $[0,d_{max}]$. Če pa želimo zaradi kasnejše posplošitve v definiciji funkcije w_i uporabiti le $min,max,d_{max},0$ in $\neg$, mnoge možnosti odpadejo. Zgornji predlog taksi metrike, na primer, uporablja funkciji razlika in absolutna vrednost, ki sta definirani na intervalu realnih števil $[0,d_{max}]$, v mnogih drugih mrežah pa ne.

- Do podobne ugotovitve kot v prejšnjem primeru pridemo tudi v primeru mehkega De Morganovega okvirja $L_i = ([0,1],max,min,0,1,\neg)$, kjer je $\neg x = 1 - x$.

Izpostavimo lastnosti, ki jih od funkcije w_i pričakujemo:

$$w_i(x,x) = 1_i,$$
$$w_i(0_i,x) = 0_i,$$
$$w_i(1_i,x) = x,$$
$$w_i(x,y) = w_i(y,x).$$

Za mero $\rho_{D,U}$ pa pričakujemo, da ima podobne lastnosti kot mera ρ_U, tj. da za $A,B \in Rel(D(U))$ in $B \neq \{\}$ velja:

$$\rho_{D,U}(A,A) = \{a_1{:}1_1,\ldots,a_n{:}1_n\} = \mathbf{1},$$

$$\rho_{D,U}(\{\},B) = \rho_{D,U}(B,\{\}) = \{a_1{:}0_1,\ldots,a_n{:}0_n\} = \mathbf{0},$$

$\rho_{D,U}(A,B) = \rho_{D,U}(B,A)$, če so mere ρ_i simetrične.

Zavedati se moramo, da nam funkcije, ki zadošča želenim pogojem, v splošnem mogoče ne bo uspelo najti, da mogoče niti ne obstaja. V tem primeru bi lahko pogoje omilili, a pod pogojem, da bi funkcija w_i razlike med označbami še vedno ustrezno kvantitativno vrednotila. Na primer, lahko bi dovolili

$$w_i(x,x) = x \vee_i \neg_i x,$$

kar je v splošnem lahko različno od 1_i, in posledično

$$\rho_{D,U}(A,A) = \wedge_{i,t \in U-\mathrm{Tup}}(A(t)(a_i) \vee_i \neg_i A(t)(a_i)),$$

kar je "blizu" **1**, če so $w_i(x,x)$ "blizu" 1_i.

3. Zaključki

Medtem ko mera podobnosti klasičnih relacij oziroma podatkovnih tabel predvsem omogoča, da ocenjujemo podobnost različnih podatkovnih tabel, lahko mero podobnosti *D*-relacij uporabimo za primerjavo različnih odgovorov na isto poizvedbo, izračun relevantnosti ali zamenljivosti eksaktnega in relaksiranega odgovora, vrednotenje razlik v tabelah iz časovno-odvisne zbirke ali sledenje spremembam v podatkovni tabeli [1,2].

V meri podobnosti *D*-relacij, $\rho_{D,U}$, nastopa funkcija primerjave označb, ki je v meri podobnosti klasičnih relacij, ρ_U, ni. V članku smo predlagali obliko funkcije primerjave označb le za nekatere specifične primere označevalnih domen *D*-relacij (to so Boolov De Morganov okvir, De Morganov okvir omejenih razdalj in mehki De Morganov okvir), splošne oblike pa nismo našli. Izpostavili smo lastnosti, ki naj bi za tako funkcijo veljale, ter se omejili na operacije in konstante, ki v funkcijskem predpisu lahko nastopajo. Na vprašanje splošnega funkcijskega predpisa zaenkrat nismo znali odgovoriti.

Literatura

1. Minker, J. An Overview of Cooperative Answering in Databases. V zborniku konference International Conference on Flexible Query Answering Systems, Roskilde University, Danska, **1998**, 282-285.
2. Hajdinjak, M.; Bierman, G. Extending Relational Algebra with Similarities. Mathematical Structures in Computer Science **2012**, 22 (4), 686-718.
3. Green, J.; Tannen, V. Models for Incomplete and Probabilistic Information. IEEE Data Engineering Bulletin **2006**, 29, 17-24.
4. Hajdinjak, M.; Bauer, A. Similarity Measures for Relational Databases. Informatica **2009**, 33 (2), 135-141.
5. Abramsky, S.; Jung, A. Domain Theory. Claredon Press: Oxford, Velika Britanija, **1994**.
6. Buneman, P.; Jung, P.; Ohori, A. Using Powerdomains to Generalize Relational Databases. Theoretical Computer Science **1991**, 9 (1), 23-55.
7. Rockafellar, R. T.; Wets, R. J.-B. Variational Analysis. Springer Verlag: Berlin, Nemčija, **2005**.