ASOCIĀCIJU LIKUMU PIELIETOŠANAS IESPĒJAS STATISTISKO DATU ANALĪZĒ

APPLICATION POSSIBILITIES OF ASSOCIATION RULES IN STATISTICAL DATA ANALYSIS

Pēteris GRABUSTS

Dr. sc. ing., asoc. prof.,Rēzeknes Augstskola
Rēzekne, Latvija
E–pasts: peter@ru.lv

Abstract. *This paper studies one of intelligent data processing methods: using association rules for data analysis. The method of association rule obtaining what was initially developed to analyse consumer's basket has turned to be a good tool for other tasks too. The method helps search and find regularities of the form $X \Rightarrow Y$ in different kinds of data. Nowadays this method is widely applied in the tasks of large scale database processing and analysing. As a result, methods of association rule construction occupy their place among the basic methods of intelligent data processing. The paper consists of two parts: theoretical and experimental. The theoretical part examines the mathematical aspects of association rule construction in detail and describes basic concepts and algorithm application possibilities. The experimental part presents implementation results and analysis of experiments. Conclusions have been drawn concerning the efficiency of association rules' application in search of regularities. Even though the association rules mining method is among the fundamental data processing methods, in Latvia this method is not widely used, therefore, the article under consideration reveals the potential possibilities of the association rule mining in the analysis of statistical data.*

***Keywords*:** *Apriori algorithm, association rules, confidence, Data Mining, support.*

Ievads

Statistisko datu analīzes jomā aizvien biežāk tiek pielietotas tā saucamās intelektuālās datu analīzes metodes, jo ir uzkrājies liels daudzums datu dažādās tautsaimniecības nozarēs un tradicionālās klasiskās statistikas metodes daudzos gadījumos nespēj piedāvāt risinājumus. Kā piemēru var minēt vienu no problēmām, ar kuru sastopas lielveikalu menedžeri: ja pircējs nopērk konkrētu preci, tad X% gadījumos viņš nopērk arī citu preci, kas ir pastarpināta pirmajai precei. Sākotnēji šis uzdevums tika izmantots lielveikalos tipisku iepirkumu šablonu atrašanai, tāpēc to dažkārt sauc par iepirkumu groza analīzi. Likumsakarības, pēc kurām varētu spriest par šādu notikumu saistību, nosauca par asociācijām. Asociācijas jeb asociāciju likumi ļauj atrast likumsakarības starp vairākiem saistītiem notikumiem. Šādu likumu pamatā ir apgalvojums: ja ir izpildījies notikums A, tad ar iespēju X% būs spēkā arī notikums B.

Asociāciju likumu iegūšanas pamatā ir 1993.g. izstrādātie teorētiskie pieņēmumi par šādu likumu esamību (1.). 1994.g. tika publicēts efektīvs algoritms asociāciju likumu iegūšanai (2.). Šie pētījumi stimulēja daudzu līdzīgu algoritmu izstrādāšanu, kas ļāva analizēt, piemēram, liela apjoma pirkumu operācijas un vispārināt šo uzdevumu par vienu no intelektuālās datu analīzes pamatmetodēm (5.,9.). Asociāciju likumus var izmantot ne tikai pircēju groza analīzei, bet var pielietot jebkuru datu analīzei – rūpīgi analizējot iegūtās likumsakarības.

Asociāciju likumu iegūšanas metodes pielietošana neaprobežojas tikai ar uzskaitītajiem pielietojumiem. Tās tiek plaši izmantotas arī lielās datu bāzēs, ko arī parāda pieaugošais izstrādāto asociāciju likumu iegūšanas metožu skaits (3.,4.,7.,8.).

Lai arī pamatoti asociāciju likumu ieguves metode ir starp galvenajām intelektuālās datu apstrādes pamatmetodēm, Latvijā šāda metode netiek plaši izmantota, tāpēc autora nolūks ir parādīt asociāciju likumu ieguves potenciālās iespējas tautsaimniecības statistisko datu analīzē.

Darba mērķis ir parādīt asociāciju likumu iegūšanas algoritmu darbības iespējas statistisko datu analīzē, noskaidrot iegūto likumu skaita atkarību no sākumvērtībām, izvērtēt tādu riska faktoru, kā strauju likumu skaita pieaugumu pie noteiktām parametru sākumvērtībām.

Mērķa realizācijas nolūkā tika veikta eksperimentu sērija, lai noskaidrotu iegūto likumu skaita atkarību no atbalsta sākumvērtībām.

Par eksperimentālajiem datiem kalpoja Latvijas Centrālās statistikas pārvaldes dati par respondentu apsekošanu (11.). Eksperimentam izvēlētās datu izlases bija saistītas ar iedzīvotāju mājsaimniecību ekonomisko pašnovērtējumu, iedzīvotāju migrācijas procesu pētīšanu un iedzīvotāju darba apstākļu apsekošanas jautājumiem.

Galvenās pētījuma metodes dotajā darbā ir aprakstošā metode, matemātiskā modelēšana un statistiskā analīze. Pētījuma laika periods ir 2008.–2013. g.

Asociāciju analīzi raksturojošie lielumi

Visiem asociāciju likumiem veidā IF(X) THEN(Y) (Ja...Tad) ir divi raksturlielumi (2.):

- Ticamība – gadījumu daļa, kad likums izpildās, starp visiem tā pielietošanas gadījumiem (gadījumu Y daļa starp X gadījumiem).
- Atbalsts – gadījumu daļa, kad likums izpildās, starp visiem gadījumiem, kad izpildās Y (gadījumu X daļa starp gadījumiem Y).

Pieņem, ka I={i_1,i_2, ..., i_m] ir literāļu kopa, ko sauc par vienumiem. Apakškopu X $\subseteq$ I sauc par vienumu kopu. k–tā vienumu kopa ir kopa, kas satur k vienumus. Pieņem, ka datu kopa D={T_1, T_2, ..., T_n} ir transakciju kopa, kur katra transakcija T_i ir vienumu kopa. Katra transakcija saistās jeb tiek asociēta ar unikālu identifikatoru, sauktu par TID. Transakcija T satur vienumu kopu X, ja ir spēkā X $\subseteq$ T.

Asociāciju likums formāli ir implikācija formā X $\Rightarrow$ Y, kur X $\subset$ I, Y $\subset$ I un X $\cap$ Y=0. X tiek dēvēts par likuma antecedenta daļu un Y – par konsekventa daļu. Katrai vienumu kopai ir zināms statistisks nozīmības mērs, ko sauc par atbalstu. Likumam X $\Rightarrow$ Y ir atbalsts s transakciju kopā D, ja s% no transakcijām D kopā satur X $\cup$ Y:

$$s(X) = \frac{|\{T \in D | X \subseteq T\}|}{|D|} \quad (1)$$

Saka, ka likums X $\Rightarrow$ Y ir spēkā transakciju kopā D ar ticamību c, ja c% no D transakcijām, kas satur X – satur arī Y:

$$c(X,Y) = \frac{s(X \cup Y)}{s(X)} \quad (2)$$

Ticamība nosaka likuma „stiprumu”. Ticamības robežvērtība c_{min} tiek izmantota, lai izslēgtu likumus, kas nav pietiekoši stipri. Attiecīgi atbalsta robežvērtība s_{min} izslēdz visus likumus, kuriem transakcijas satur antecedenta un konsekventa daļas ar nepietiekošu apjomu.

Atbalsta robežvērtība ir definēta caur visām vienumu kopām. Attiecībā uz asociāciju likumiem tā procentuāli raksturo transakciju skaitu, kas satur visus vienumus, kuri parādās likumā. Ticamības robežvērtība raksturo minimālo varbūtību, ka konsekventa daļa ir patiesa tad, ja antecedenta daļa ir patiesa. Ticamības vērtība tuvu pie 100% raksturo ļoti stiprus likumus.

Asociāciju likumu atrašanas problēmu var raksturot sekojoši. Dots: vienumu kopa I, vienumu kopu datu bāze D, atbalsta robežvērtība s_{min}, ticamības robežvērtība c_{min}. Ir jāatrod visus asociāciju likumus veidā X $\Rightarrow$ Y. Tādējādi, vajag atrast visus asociāciju likumus X $\Rightarrow$ Y kopā D ar atbalsta vērtību s(X$\cup$Y)$\geq s_{min}$ un ticamību c(X,Y) $\geq c_{min}$.

Ilustrējam izklāstīto ar piemēru. 1. tab. dota transakciju datu bāze. Tā ir kopa I, kas sastāv no produktiem A,B,C un D.

1. tabula

Transakciju TID datu piemērs *(adaptēts no10.)*

TID	Pirkumu grozs
1	{A,C}
2	{B}
3	{A,B,C,D}
4	{B,D}
5	{A,B,D}

TID	Pirkumu grozs
6	{A,B}
7	{A,D}
8	{B,C,D}
9	{C,D}
10	{A,B,D}

Katra tabulas rinda satur transakcijas identifikatoru TID, kas raksturo pircēja veikto operācijas numuru un iegādāto produktu kopu. Vienumu kopas {A,B} atbalsts ir 0,4. {A,B,D} atbalsta vērtība ir 0,3. Sekojoši, likuma {A,D} $\Rightarrow$ {B} ticamība ir 0,75, jo pēc formulas (2) iegūst:

$$c(A, D \Longrightarrow B) = \frac{s(A \cup B \cup D)}{s(B)} = \frac{0{,}3}{0{,}4} = 0{,}75$$

Ja atbalsta robežvērtība s_{min} ir mazāka vai vienāda ar 0,3 un ticamības robežvērtība c_{min} ir mazāka vai vienāda ar 0,75, tad šis likums tiek uzlūkots kā pieņemams asociāciju likums. Var teikt, ka ir iegūts asociāciju likums „Ja pircējs nopērk produktus A un D, tad ir iespējams, ka 75% gadījumos viņš nopirks arī produktu B". Protams, šādi apgalvojumi ir pieņemami tikai lielām transakciju datu bāzēm. Atbalsta un ticamības vērtības vēl negarantē likuma pielietojamību pircēja uzvedības modeļa darbībā. Tās var tikai palīdzēt lēmumu pieņemšanas procesā.

Vispārīgā gadījumā visu asociāciju likumu iegūšanas process tiek reducēts uz diviem uzdevumiem: 1) atrast visas vienumu kopas, kurām transakciju atbalsta vērtības pārsniedz atbalsta robežvērtību. Vienumu kopa ar minimālo atbalstu tiek dēvēta par lielo vienumu kopu, pārējās par mazajām vienumu kopām; 2) izmantojot lielo vienumu kopu, atrast attiecīgos likumus.

Asociāciju likumu iegūšanas algoritms

Apriori algoritms un kandidātkopu ģenerēšanas algoritms pseido – C notācijā (2.tab. un 3.tab.) ir viens no biežāk citētajiem algoritmiem vienumu kopu atrašanā (1.). Lai arī pastāv virkne atšķirīgu algoritmu, šis algoritms kalpo par pamatu līdzīga tipa konstrukciju izstrādāšanai.

2. tabula

Apriori algoritms *(adaptēts no 1.)*

Ievadā:	Transakciju datu bāze D un atbalsta robežvērtība s_{min}.
Izejā:	Kopa L, kas satur visas D vienumu kopas.
(1)	L_1 = {lielā 1-vienumu kopa}
(2)	**for** (k=2; $L_{k-1} \neq 0$; k++) **do begin**
(3)	C_k = AprioriGen(L_{k-1}); // Jauns kandidāts
(4)	**forall** transakcijām t ∈ D **do begin**
(5)	C_t = apakškopa(C_k, t); // Kandidāti iekš t
(6)	**forall** kandidāti c ∈ C_t **do**
(7)	c.count++;
(8)	**end**
(9)	L_k={c ∈ C_k | c.count ≥ minsup}
(10)	**end**
(11)	Atbilde=$\cup_k L_k$;

3. tabula

AprioriGen funkcija *(adaptēts no 1.)*

(1)	**function** AprioriGen(L_{k-1});
(2)	// Pievienošana
(3)	**insert into** C_k
(4)	**select** p.item$_1$, p.item$_2$,..., p.item$_{k-1}$, q.item$_{k-1}$
(5)	**from** L_{k-1} p, L_{k-1} q
(6)	**where** p.item$_1$=q.item$_1$,..., p.item$_{k-2}$=q.item$_{k-2}$, p.item$_{k-1}$<q.item$_{k-1}$;
(7)	// Saīsināšana
(8)	**forall** itemkopai c ∈ C_k **do**
(9)	**forall** (k-1)-apakškopai s no c **do**
(10)	**if** (s ∉ L_{k-1}) **then**
(11)	**delete** c no C_k;

Iteratīvi tiek meklēta lielā vienumu kopa. Pirmajā iterācijā par lielo vienumu kopu tiek ņemta visa datubāze D. Nākamajās iterācijās ar funkcijas AprioriGen palīdzību tiek ģenerētas kandidātkopas.

Literatūrā nav pārāk daudz priekšlikumu, kādā veidā iegūt likumus no atrastajām lielajām vienumu kopām, jo vienumu kopu ģenerācijas process tiek uzskatīts par ļoti būtisku, kam arī atvēlēta nozīmīgāka loma asociāciju likumu ieguves procesā. Par pamatu šāda tipa algoritmiem ir ņemti pieņēmumi par algoritma eksistenci asociāciju likumu iegūšanai (2.).

Pieņem, ka dota k–vienumu kopa (k ≥ 2) un apakškopa l, kur 0 ⊂ l ⊂ f. Tad likums l ⇒ f\l ir asociāciju likums, ja izpildās nosacījums:

$$\frac{s(f)}{s(l)} \geq c_{min} \tag{3}$$

kur s(f) – f atbalsta vērtība, s(l) – l atbalsta vērtība un c_{min} – minimālā ticamības vērtība.

Ja likumam $l \Rightarrow f \backslash l$ nav minimālās ticamības vērtības, tad neizpildās arī likums $l' \Rightarrow f \backslash l'$, kam $0 \subset l' \subset l$. Tādējādi, tā vietā, lai izskatītu visas f apakškopas likumu ģenerēšanai, var lietot funkciju, kas ģenerē likumus ar (k–1) – apakškopu l' no k–tās vienumu kopas l kā likuma antecedentu, ja izpildās nosacījums (3). Algoritma un likumu ģenerācijas funkcijas detalizētākus aprakstus var atrast (2.). Šis algoritms tiek uzskatīts par pamata algoritmu likumu ģenerēšanai no iepriekš iegūtajām vienumu kopām, taču tas ir samērā lēns, tāpēc ir izstrādāta vesela virkne algoritmu, kas ļauj ātrāk atrast asociāciju likumus (10.).

Asociāciju likumu iegūšana no statistiskajiem datiem

Pētījuma daļā tika veikta eksperimentu sērija, lai noskaidrotu iegūto likumu skaita atkarību no atbalsta sākumvērtībām. Par eksperimentālajiem datiem kalpoja Latvijas Centrālās statistikas pārvaldes dati (11.) ar respondentu apsekošanas atbilžu variantiem. Eksperimentiem izvēlētās datu izlases bija veltītas iedzīvotāju mājsaimniecību ekonomiskajam pašnovērtējumam, iedzīvotāju migrācijas procesu pētīšanai un iedzīvotāju darba apstākļu apsekošanas jautājumiem.

Pētījumu daļas pirmajā etapā tika sagatavoti dati eksperimentiem. Sākotnēji visi dati bija SPSS formātā un šajā etapā tika fiksēti visi nepieciešamie atribūti un izstrādāta kodifikatoru sistēma, lai datiem varētu pielietot asociāciju likumu algoritmu. Otrajā etapā ar programmatūras palīdzību (izstrādāta Matlab vidē) tika realizēts asociāciju likumu iegūšanas algoritms un pie sākotnējiem nosacījumiem tika iegūti datus raksturojošie likumi un veikta to analīze.

Mājsaimniecību ekonomiskais pašnovērtējums

Respondentiem bija uzdoti sekojoši jautājumi:

- Ņemot vērā Jūsu mājsaimniecības kopējo ekonomisko situāciju, lūdzu, pasakiet, kurš no izteikumiem vislabāk raksturo Jūsu situāciju (doti 7 atbilžu varianti);
- Ja Jūsu mājsaimniecībai pēkšņi ievajadzētos 120 latus, vai Jūs varētu sagādāt šādu naudas summu nedēļas laikā? (doti 6 atbilžu varianti);
- Vai Jūsu mājsaimniecības ekonomiskā situācija patlaban ir labāka, tāda pati vai sliktāka nekā pirms 5 gadiem? (doti 5 atbilžu varianti);

- Vai Jūsu mājsaimniecības ekonomiskā situācija pēc 5 gadiem būs labāka, paliks tāda pati vai būs sliktāka, salīdzinot ar pašreizējo situāciju? (doti 5 atbilžu varianti).

Pirmajā eksperimenta daļā tika pieņemts, ka ticamības robežvērtība c_{min}=75 un atbalsta robežvērtība s_{min}=75. Pie šīm sākumvērtībām (c_{min}=75 un s_{min}=75) tika iegūti 6 likumi. Katram likumam tika izskaitļota atbalsta vērtība un ticamības vērtība. Iegūtie likumi ir sekojoši:

1) **25=>34** Atbalsts=247 un ticamība=82;
2) **32 42 =>23** Atbalsts=205 un ticamība=75;
3) **25 43 =>34** Atbalsts=168 un ticamība=83;
4) **25 53 =>34** Atbalsts=106 un ticamība=84;
5) **25 59 =>34** Atbalsts=75 un ticamība=87;
6) **25 43 53 =>34** Atbalsts=78 un ticamība=85.

Var konstatēt (ņemot talkā atbilžu kodu atšifrējumus), ka

Pirmais likums nosaka: JA „*Mūsu saimniecība ir nabadzīga*" **TAD** „*Mēs nevarētu pēkšņi sagādāt 120 latus nedēļas laikā*";

Otrais likums nosaka: JA „*Mums būtu nepieciešama citu palīdzība sagādāt 120 latus nedēļas laikā"* **UN** „*Mūsu mājsaimniecības ekonomiskā situācija ir tādi pati, kā 5 gadus atpakaļ*" **TAD** „ *Mēs neesam ne bagāti, ne nabadzīgi*";

Trešais likums nosaka: JA „*Mēs esam nabadzīgi*" **UN** „*Mūsu mājsaimniecības ekonomiskā situācija ir sliktāka, kā 5 gadus atpakaļ*" **TAD** „*Nebūtu iespējams sagādāt 120 latus nedēļas laikā*";

Ceturtais likums nosaka: JA „*Mēs esam nabadzīgi*" **UN** „*Mūsu mājsaimniecības ekonomiskā situācija pēc 5 gadiem būs sliktāka nekā tagad*" **TAD** „*Nebūtu iespējams sagādāt 120 latus nedēļas laikā*";

Piektais likums: kļūdains, jo anketā nav tāda koda 59. Visticamāk, tā ir anketas aizpildītāja kļūda;

Sestais likums nosaka: JA „*Mēs esam nabadzīgi*" **UN** „*Mūsu mājsaimniecības ekonomiskā situācija ir sliktāka, kā 5 gadus atpakaļ*" **UN** „*Mūsu mājsaimniecības ekonomiskā situācija pēc 5 gadiem būs sliktāka nekā tagad*" **TAD** „*Nebūtu iespējams sagādāt 120 latus nedēļas laikā*".

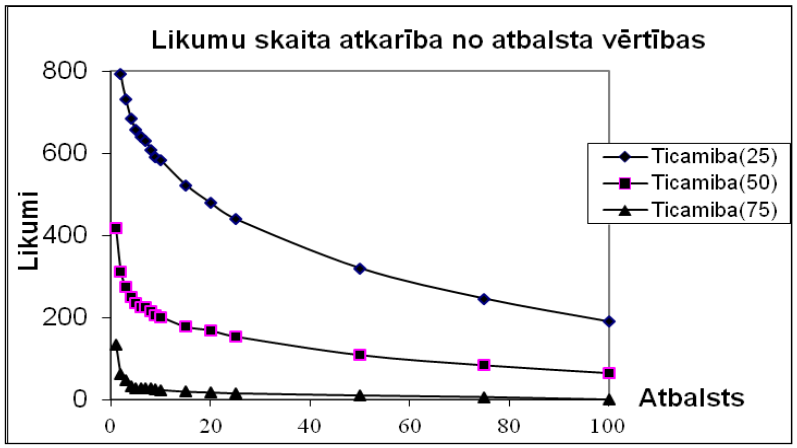
Otrajā eksperimenta daļā tika iegūtas likumu skaita vērtības pie dažādām atbalsta vērtībām un fiksētām ticamības robežvērtībām. Rezultāti uzrādīti 4. tabulā.

4. tabula

Likumu skaita atkarība no uzdotās atbalsta vērtības
(autora aprēķinu rezultāts)

Atbalsts	1	2	3	4	5	6	7	8	9	10	15	20	25	50	75	100
Ticamība(25)	957	792	731	684	656	639	629	608	591	582	522	478	439	319	245	189
Ticamība(50)	417	310	273	249	235	226	225	215	205	201	177	168	153	108	83	64
Ticamība(75)	133	62	48	32	28	28	28	26	24	23	19	17	15	10	6	n/a

Grafika veidā atbilstība parādīta 1. attēlā.



1. attēls. Likumu skaita atkarība no uzdotām atbalsta vērtībām pie dažādām ticamības robežvērtībām
(autora aprēķinu rezultāts)

No tabulas datiem un grafiskās atbilstības var secināt, ka, jo lielāks uzdotais ticamības līmenis un atbalsta robežvērtība, jo mazāks iegūto asociāciju likumu skaits un līdz ar to likumi ir „stingrāki".

Iedzīvotāju migrācijas procesu pētīšana

Respondentiem bija uzdoti sekojoši jautājumi:

- Kurā valstī Jūs esat dzimis? (doti 11 atbilžu varianti);
- Cik ilgi Jūs dzīvojat šajā apdzīvotajā vietā? (doti 4 atbilžu varianti);
- Kur Jūs dzīvojāt pirms pārcelšanās uz šo apdzīvoto vietu? (doti 3 atbilžu varianti);
- Lūdzu nosauciet apdzīvotās vietas tipu, kurā Jūs dzīvojāt pirms pārcelšanās uz šo vietu? (doti 7 atbilžu varianti);
- Kāds bija iemesls tam, ka Jūs pārcēlāties uz šo apdzīvoto vietu? (doti 6 atbilžu varianti);
- Vai Jūs turpmākajos 3 gados plānojat pārcelties uz citu apdzīvotu vietu? (doti 5 atbilžu varianti).

Pirmajā eksperimenta daļā tika pieņemts, ka ticamības robežvērtība c_{min}=95 un atbalsta robežvērtība s_{min}=95. Pie šīm sākumvērtībām tika iegūti 52 likumi. Katram likumam tika izskaitļota atbalsta vērtība un ticamības vērtība. Zemāk parādīti daži iegūtie likumi ar lielākajām atbalsta vērtībām:

1. **12 74 => 51** Atbalsts= 592 un ticamība= 98;
2. **46 => 83** Atbalsts= 545 un ticamība= 97;
3. **12 74 83 => 51** Atbalsts= 530 un ticamība= 98;
4. **12 46 => 83** Atbalsts= 443 un ticamība= 97;
5. **45 51 65 => 83** Atbalsts= 303 un ticamība= 95.

Ņemot talkā kodu atšifrējumus, var spriest, ka

Pirmais likums nosaka: JA "*Jūs esat dzimis Latvijā*" **UN** "*Pārcēlāties uz šo apdzīvoto vietu ģimenes apstākļu dēļ*" **TAD** "*Pirms pārcelšanās uz šo apdzīvoto vietu Jūs dzīvojāt Latvijā*";

Otrais likums nosaka: JA "*Jūs vienmēr esat dzīvojis šajā apdzīvotajā vietā*" **TAD** "*Turpmākajos 3 gados Jūs neplānojat pārcelties uz citu apdzīvoto vietu*";

Trešais likums nosaka: JA "*Jūs esat dzimis Latvijā*" **UN** "*Pārcēlāties uz šo apdzīvoto vietu ģimenes apstākļu dēļ*" **UN** "*Turpmākajos 3 gados Jūs neplānojat pārcelties uz citu apdzīvoto vietu*" **TAD** "*Pirms pārcelšanās uz šo apdzīvoto vietu Jūs dzīvojāt Latvijā*";

Ceturtais likums nosaka: JA "*Jūs esat dzimis Latvijā*" **UN** "*Jūs vienmēr esat dzīvojis šajā apdzīvotajā vietā*" **TAD** "*Turpmākajos 3 gados Jūs neplānojat pārcelties uz citu apdzīvoto vietu*";

Piektais likums nosaka: JA "*Jūs šajā apdzīvotajā vietā esat nodzīvojis līdz 50 gadiem*" **UN** "*Pirms pārcelšanās uz šo apdzīvoto vietu Jūs dzīvojāt Latvijā*" **UN** "*Pirms tam Jūs dzīvojāt ciemā*" **TAD** "*Turpmākajos 3 gados Jūs neplānojat pārcelties uz citu apdzīvoto vietu*".

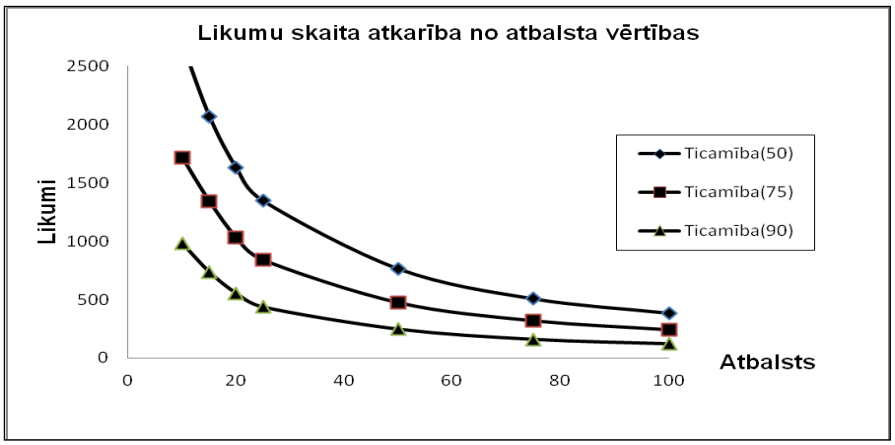
Otrajā eksperimenta daļā tika iegūtas likumu skaita vērtības pie dažādām atbalsta vērtībām un fiksētām ticamības robežvērtībām. Rezultāti uzrādīti 5. tabulā.

5. tabula

Likumu skaita atkarība no uzdotās atbalsta vērtības
(autora aprēķinu rezultāts, adaptēts no autora darba (6.))

Atbalsts	10	15	20	25	50	75	100
Ticamība(50)	2690	2071	1628	1346	765	509	386
Ticamība(75)	1717	1343	1029	839	473	318	241
Ticamība(90)	983	736	553	436	246	159	121

Grafiskā veidā atbilstību var apskatīt 2. attēlā.



2. attēls. Likumu skaita atkarība no atbalsta vērtībām pie dažādām ticamības robežvērtībām
(autora aprēķinu rezultāts, adaptēts no autora darba (6.))

Līdzīgi kā iepriekšējā datu izlasē, no tabulas datiem un grafiskās atbilstības var secināt, ka, jo lielāks uzdotais ticamības līmenis un atbalsta robežvērtība, jo mazāks iegūto asociāciju likumu skaits. Analizējot iegūtos likumus, var secināt, ka aptaujātie iedzīvotāji dzīvo kompaktā nelielā apdzīvotā vietā ar maz izteiktu migrācijas tieksmi.

Iedzīvotāju darba apstākļi

Respondentiem bija uzdoti 8 jautājumi. Anketas jautājumi bija šādi:

- Vai Jūs esat nodarbināts kā ierindas darbinieks vai noteikta līmeņa vadītājs? (doti 7 atbilžu varianti, atbildes kods no 11 līdz 18);
- Kāda veida darba līgums Jums ir ar savu darba devēju? (doti 8 atbilžu varianti, atbildes kods no 21 līdz 28);
- Kāda ir īpašuma forma uzņēmumā, kur Jūs strādājat? (doti 5 atbilžu varianti, atbildes kods no 31 līdz 38);
- Kāds darba režīms vislabāk raksturo Jūsu situāciju? (doti 5 atbilžu varianti, atbildes kods no 41 līdz 48);
- Cik darbinieku ir Jūsu darba vietā? (doti 6 atbilžu varianti, atbildes kods no 51 līdz 47);
- Vai pēdējā gada laikā ir bijusi aizkavēšanās darba algas izmaksā? (doti 3 atbilžu varianti, atbildes kods no 61 līdz 63);
- Vai Jūs uzskatāt, ka nākošo divu gadu laikā Jūsu pašreizējais darbs varētu tikt apdraudēts? (doti 6 atbilžu varianti, atbildes kods no 71 līdz 77);
- Vai Jūsu darba apstākļi, salīdzinot ar stāvokli pirms 5 gadiem, ir...? (doti 7 atbilžu varianti, atbildes kods no 81 līdz 87).

Pirmajā eksperimenta daļā tika analizēti 33 asociāciju likumi, kas tika iegūti pie ticamības robežvērtības c_{min}=90 un atbalsta robežvērtības s_{min}=100. Iegūtie likumi, to atbalsta un ticamības vērtības parādītas 6. tabulā.

6. tabula

Iegūtie likumi un to skaitliskās vērtības *(autora aprēķinu rezultāts)*

N.	Likums	At-balsts	Tica-mība		N.	Likums	At-balsts	Tica-mība
1.	74 =>62	471	91		18.	22 31 74 =>62	156	95
2.	22 74 =>62	373	93		19.	11 31 74 =>62	153	91
3.	41 74 =>62	315	90		20.	22 51 74 =>62	145	94
4.	32 74 =>62	261	91		21.	22 74 83 =>62	136	92
5.	51 74 =>62	199	90		22.	31 41 74 =>62	128	90
6.	31 74 =>62	196	91		23.	41 74 83 =>62	105	91
7.	74 83 =>62	167	91		24.	23 41 62 =>11	104	90
8.	31 43 =>62	111	90		25.	11 22 41 74 =>62	187	93
9.	71 83 =>11	108	92		26.	11 22 32 74 =>62	158	90
10.	52 74 =>62	103	94		27.	22 32 41 74 =>62	138	94
11.	43 74 =>62	103	92		28.	11 22 31 74 =>62	119	96
12.	74 82 =>62	103	90		29.	11 22 74 83 =>62	116	91
13.	11 22 74 =>62	287	92		30.	11 22 51 74 =>62	106	92
14.	22 41 74 =>62	249	94		31.	22 31 41 83 =>62	103	91
15.	22 32 74 =>62	205	92		32.	22 31 41 74 =>62	102	95
16.	32 41 74 =>62	176	91		33.	11 22 32 41 74 =>62	107	92
17.	22 31 83 =>62	158	90					

Ņemot vērā anketas kodu atšifrējumus, piemēra nolūkā var izrakstīt atsevišķus asociāciju likumus:

Pirmais likums nosaka: JA "*Neuzskatu, ka nākošo divu gadu laikā pašreizējais darbs varētu tikt apdraudēts*" **TAD** "*Pēdējā gada laikā nav bijusi aizkavēšanās darba algas izmaksā*";

Otrais likums nosaka: JA "Darba līgums ir pastāvīgs" **UN** "*Neuzskatu, ka nākošo divu gadu laikā pašreizējais darbs varētu tikt apdraudēts*" **TAD** "*Pēdējā gada laikā nav bijusi aizkavēšanās darba algas izmaksā*";

Devītais likums nosaka: JA "*Nākošo divu gadu laikā pašreizējais darbs varētu tikt apdraudēts štatu samazināšanas dēļ*" **UN** "*Darba apstākļi, salīdzinot ar stāvokli pirms 5 gadiem ir tādi paši*"

TAD "*Esmu nodarbināts kā ierindas darbinieks*";

17-tais likums nosaka: JA "*Darba līgums ir pastāvīgs*" **UN** "Darba vieta ir valsts uzņēmums" **UN** "*Darba apstākļi, salīdzinot ar stāvokli pirms 5 gadiem ir tādi paši*" **TAD** "*Pēdējā gada laikā nav bijusi aizkavēšanās darba algas izmaksā*";

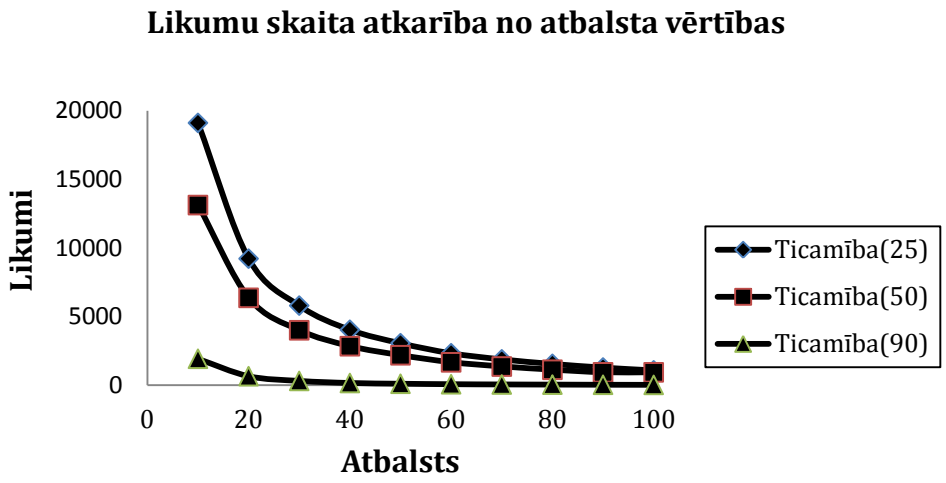
24-tais likums nosaka: JA "*Darba līgums ir pagaidu*" **UN** "Strādāju parastu darba režīmu (starp 6:00 – 18:00)" **UN** "*Pēdējā gada laikā nav bijusi aizkavēšanās darba algas izmaksā*" **TAD** "*Esmu nodarbināts kā ierindas darbinieks*".

Otrajā eksperimenta daļā tika iegūtas likumu skaita vērtības pie dažādām atbalsta vērtībām un fiksētām ticamības robežvērtībām. Rezultāti uzrādīti 7. tab., bet grafika veidā atbilstību var apskatīt 3. att.

7. tabula

Likumu skaita atkarība no uzdotās atbalsta vērtības
(autora aprēķinu rezultāts)

Atbalsts	10	20	30	40	50	60	70	80	90	100
Ticamība (25)	19116	9227	5806	4046	3067	2338	1893	1563	1296	1090
Ticamība (50)	13130	6373	4009	2848	2189	1675	1368	1132	943	943
Ticamība (90)	1934	664	320	168	114	79	62	51	39	33



3.attēls. Likumu skaita atkarība no atbalsta vērtībām pie dažādām ticamības robežvērtībām
(autora aprēķinu rezultāts)

Var pamanīt, ka šai datu izlasei asociāciju likumu konsekventa daļa lielāko tiesu ir 62 (*Pēdējā gada laikā nav bijusi aizkavēšanās darba algas izmaksā*). Tas saistīts ar to, ka anketas sestajā jautājumā bija tikai 3 atbilžu varianti, kas šajā gadījumā jūtami ietekmēja likumu konstruēšanu.

Secinājumi un priekšlikumi

Asociāciju likumu meklēšana ir viens no populārākajiem intelektuālās datu analīzes pielietojumiem. Asociāciju analīze ir lietderīga gadījumos, kad vairāki notikumi ir saistīti savā starpā. Šo metodi

lietderīgi pielietot kā vienu no pirmajiem pētnieciskajiem etapiem, kad zināms (vai būtisks) tikai kāds no datus raksturojošiem lielumiem.

Asociāciju likumu galvenā priekšrocība ir samērā vienkāršu likumu iegūšana. Tādus likumus viegli formulēt un attiecīgi tos var uzreiz izmantot datu analīzē. Cita asociāciju analīzes priekšrocība ir iespēja strādāt ar dažāda garuma ierakstiem, un, visbeidzot, šo metodi ērti lietot datu analīzes sākuma etapā, kad nav skaidra priekšstata par analizējamiem datiem un nav zināms kā sākt risināt konkrēto uzdevumu.

Likumu novērtēšanā izmantojamajām atbalsta un ticamības vērtībām ir būtiska ietekme uz iegūto likumu skaitu. Eksperimentālā ceļā tiek piemeklētas minimālās atbalsta un ticamības sākumvērtības, kā rezultātā tie likumi, kas neatbilst uzdotajām sākumvērtībām, tiek noraidīti un netiek ņemti vērā konkrētā uzdevuma risināšanā.

Ja atbalsta vērtība parāda, cik liels procents transakciju uztur doto likumu, tad ticamības vērtība uzdod varbūtību tam, ka no $X \Rightarrow Y$. Acīmredzot ir spēkā nosacījums: jo lielāka ticamība, jo mazāk likumu tiek iegūts pie atbilstošajām atbalsta vērtībām.

Var izdarīt slēdzienu, ka atsevišķu uzdevumu klasēm asociāciju likumu ieguves mehānisms ir ļoti lietderīgs, taču svarīgi ir apzināties, ka iegūtie asociāciju likumi prasa rūpīgu analīzi, lai to izmantošana būtu efektīva. Reizē ar to jāsecina, ka asociāciju likumu analīzē ir arī savas vājās vietas, kuru izpēte varētu būt vērtīgs pielietojumu lauks: asociāciju likumu programmrealizācijas izpilde prasa ievērojamu laika patēriņu; analizējamajiem datiem jābūt pēc iespējas viendabīgiem, kā tas ir anketu datos.

Likumu iegūšanas procesā raksturīga problēma ir tā, ka asociāciju likumu metode tikai dod iespēju iegūt noteiktu likumu skaitu, tālāko analīzi atstājot ekspertu vai analītiķu ziņā.

Izmantotā literatūra un avoti

1. AGRAWAL, R., IMIELINSKI, T., SWAMI, A. *Mining Association Rules Between Sets of Items in Large Databases*. Proc. Conf. on Management of Data, ACM Press, 1993, p. 207-216.
2. AGRAWAL, R., SRIKANT, R. *Fast Algorithms for Mining Association Rules*. Proceedings of the 20th International Conference on Very Large Databases, 1994, p. 487-499.
3. CHEN, M-S., HAN, J., YU, P.S. *Data Mining: An Overview from a Database Perspective*. IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, 1996, p. 866-883.
4. CHEUNG, D.W. et al. *Efficient Mining of Association Rules in Distributed Databases*. IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, 1996, p. 911-922.
5. FAYYAD, U.M., PIATETSKY-SHAPIRO, G., SMYTH, P. *From Data Mining to Knowledge Discovery: An Overview*. In FAYYAD, U.M., PIATETSKY-SHAPIRO, G.,

SMYTH, P. and UTHURUSAMY, R. (editors), Advances in Knowledge Discovery and Data Mining, Chapter 1, AAAI/MIT Press, Menlo Park, California, USA, 1996.
6. GRABUSTS, P. *Using Association Rules to Extract Regularities from Data*. Proc. 6th International Baltic Conference on Data Bases and Information Systems, Riga, 2004, p. 117-126.
7. HOUTSMA, M., SWAMI, A. *Set-Oriented Mining for Association Rules in Relational Databases*. Proceedings of the 11th IEEE International Conference on Data Engineering, Taipei, Taiwan, 1995, p. 25-34.
8. KLEMETTINEN, M. et al. *Finding Interesting Rules from Large Sets of Discovered Association Rules*. 3rd International Conference on Information and Knowledge Management (CIKM), 1994, p. 401-407.
9. NEWQUIST, H.P. *Data Mining: The AI Metamorphosis*. Database Programming and Design, № 9 (Data Mining Special Edition Supplement), 1996.
10. DUNHAM, M.H. et al. (2001). *A Survey of Association Rules*. [atsauce 2013.g. 16.dec.]. Pieejas veids: Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.1602
11. *Statistikas datubāzes* [tiešsaiste]. Centrālās statistikas pārvaldes publikācija [atsauce 2014.g. 7.feb.]. Pieejas veids: http://www.csb.gov.lv/dati/statistikas-datubazes-28270.html

Summary

The search of association rules is one of the most popular intellectual data analysis applications. Analysis of associations is useful in cases where several events are linked to each other. It is useful to use this method at one of the early research stages, when only one of the data characterizing parameters is known (or essential).
The main advantage of association rules is generation of relatively simple rules. The rules in IF–THEN format are easy to understand and interpret. Such rules are easy to formulate and thus they can be directly used in data analysis. Another advantage of association rules analysis is the opportunity to work with variable–length records and, finally, this method is suitable for initial stage of data analysis when there is no clear understanding about the data being analyzed and it is not known how to approach the particular task.
Association rules method works best in situations where the various parameters in the data appear in relatively equal number of cases. Otherwise, the rules will link only frequently repeated parameters and it will not be possible to learn anything new about rarely met parameters, that is, time will be non–effectively spent for processing non-important rules.
Support and confidence values used in the rules evaluation process have a significant impact on the resulting number of rules. Experimentally, minimal support and confidence thresholds have been selected, as a result those rules that do not correspond to defined starting values, are discarded and not used in solving a particular task.
If the support value shows how large percentage of transactions maintains the given rule, then the confidence value instructs that the $X \Rightarrow Y$. Apparently, the condition is in force: the bigger confidence is, the „better" the rule is, nevertheless the confidence value does not allow to evaluate the effectiveness of the rule.
It can be concluded that for certain forms of tasks the association rules extraction mechanism is very useful:

- the aim of the search task for association rules is the determination of frequently occurring object models in data samples;
- the results of the task are expressed in the form of association rules, where the condition and the concluding parts contain such samples;
- the characteristic values of association rules are support and confidence.

The important point is that the resulting association rules require careful analysis in order to use them effectively. At the same time, it shall be concluded that the association rules analysis also has its weaknesses, the study of which could be a valuable application field:

- software implementation of association rules is time-consuming;
- the data under analysis should be possibly homogeneous;
- incorrect or unusual data also participate in rule formation.

The series of experiments were performed in the research part of the paper with a purpose to determine the dependence of the obtained number of rules on the support's initial values. The Latvian Central Statistical Bureau data about the respondents' answer variants served as a basis for the research. The data selected in the research related to the household's economic self-evaluation, the migration process of inhabitants and the evaluation of the working conditions of the population. The software was written in the Matlab environment.

In the experimental part of the study the association rules' mining from the statistical data was carried out, the number of definite rules dependence on the support and confidence values was determined. As a result, the empirical correlation of rules dependence on confidence has been obtained. The obtained rules are logical and reflect the real situation.