HUA NAM SON[1], GUBÁN MIKLÓS[2]

A bevásárlókosár modelljének alkalmazása a fluidumcsomópontok osztályozására[3]

Absztrakt

A cikkben megmutatjuk a bevásárlókosár-modellezést, illetve alkalmazását a szolgáltatási folyamatok vizsgálatában. A gyakori vásárlói kosarak felfedezése, a termékek közötti asszociációs kapcsolatok feltárása a kutatók számára egy vonzó téma. Korábbi vizsgálatokban a vevő vásárlásait, a tranzakciókat elemezték termékszinten. A közelmúltban néhány kutatás a termék mennyisége alapján vizsgálja a vevő vásárlásait, tranzakcióit: nem a vevők által választott termékek vagy az ügyfelek által választott szolgáltatások alapján, hanem a vevők által választott termék mennyisége, illetve az ügyfelek által választott szolgáltatások mennyisége alapján vizsgálja a termékek összességében lévő struktúrát, a szolgáltatási folyamatok szervezését. Jelen kutatásunkban az utóbbi szempontból írjuk le a bevásárlókosár-modellezést. A modell formális leírása után megmutatjuk, hogy a vevő vásárlásait, az ügyfél részére végzett szolgáltatásokat ebben a megközelítésben lehet megvizsgálni. Megmutatjuk, hogy a módszert alkalmazni lehet a vevők, az ügyfelek elemzésére, osztályozására is. A megközelítésnek az az előnye, hogy jobban láthatjuk a tranzakciók közötti természetes kapcsolatokat, amelyeket a modell formális leírása segítségével valóban egy részben rendezés, háló jellegű struktúrával lehet jellemezni. Ebben az általánosabb modellben hálóelméleti eszközökkel újravizsgálunk néhány ismert kérdést. A gyakori bevásárlókosarak és az asszociációs szabályok explicit reprezentációja, illetve az azok felfedezésére alkalmas algoritmusok ismertetése után bevezetjük a vásárlóklasszifikáció és a legközelebbi szomszéd módszerének újabb fogalmát. Megmutatunk egy módszert, amely szerint valamely esetben gyorsabban lehet meghatározni a termékkészlet, illetve a szolgáltatási folyamatban lévő csomópontok közötti d-szomszédságot.

Kulcsszavak: vásárlói kosár, gyakori termékek, asszociációs szabály, osztályozás

Bevezetés

Az adatbányászati módszerek hasznos eszköznek bizonyulnak az üzleti tevékenységek elemzésében, a folyamatok modellezésében, illetve más kutatási területeken. A bevásárlókosár-mo-

1 Főiskolai docens, BGF PSzK, Budapest; e-mail: Hua.NamSon@pszfb.bgf.hu.
2 Főiskolai tanár, BGF PSzK, Budapest; e-mail: Guban.Miklos@gkz.bgf.hu.
3 A cikk a LOST in Services kutatási projekt keretein belül készült, mely az EMMI-26130-2/2013/TUDPOL támogatásából valósult meg.

dellezés területén a kutatók nagy erőfeszítéseket tettek, hogy felfedezzék a vevői vásárlásban elrejtett információkat. A vásárlói kosarak (MB) és az asszociációs szabályok felfedezése igen fontos feladat különböző alkalmazási területeken, például a kiskereskedelmi szektor döntéshozatali, illetve stratégiameghatározási folyamataiban (Agrawal – Srikant 1994: 487–499). A megvásárolt termékek vagy az ügyfelek által választott szolgáltatások alapján megvizsgálhatjuk a vásárlók vagy az ügyfelek közösségének a jellemzőit. A feladat fontosságát könnyen megláthatjuk a kereskedelem ügyfélmenedzsment-, marketing- és egyéb folyamataiban. Más aspektusban a megvásárolt termékek vagy az ügyfelek által választott szolgáltatások alapján elemezhetjük a termékekben, illetve a szolgáltatásokban rejlő kapcsolatokat. Az ebben az irányban végzett elemzések eredményei fontos információt biztosítanak ahhoz, hogy a vállalat vezetői döntést hozhassanak. A megvásárolt termékek vagy az ügyfelek által választott szolgáltatások mennyiségi elemzése azonban eltér a korábbi kutatásokban használt módszerektől: az elemzés nem a megvásárolt tételek szintjén, hanem a megvásárolt tételek mennyisége alapján történik. A következő példával illusztráljuk az eltérést a két módszer között: Egy üzletben az üzlet vezetője a vásárlás adatai alapján belátja, hogy a vásárlók 70%-a vajat vásárolt, és a korábbi kutatásokban használt módszerekkel megtudhatja, hogy a vaj az egyik gyakran vásárolt (gyakori) termék. Azonban az üzlet vezetője a mennyiségi elemzés szerint megjegyezheti, hogy a (0,5 kg vaj, 1 kg liszt) az egyik gyakrabban vásárolt árukészlet, a (10 kg vaj, 0,5 kg liszt) pedig nem. Ennek alapján a korábbi kutatásokban használt módszerekkel feltárt vaj és liszt közötti asszociációs kapcsolat helyett pontosabban tudja, hogy a 0,5 kg vaj és az 1 kg liszt között fennáll az asszociációs kapcsolat, a 10 kg vaj és a 0,5 kg liszt között azonban nem.

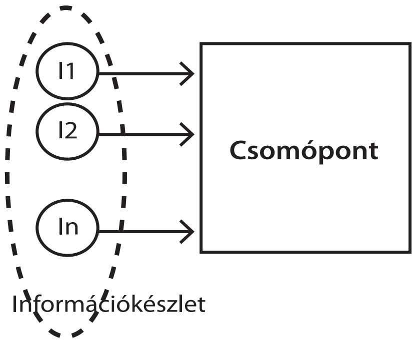
A részletesebb, mennyiségi vizsgálat lehetővé teszi, hogy:

1. Egységesen tudjuk vizsgálni a termékkészletek struktúráját, valamint a vásárlók (akik az általuk megvásárolt termékekkel azonosíthatók) közötti kapcsolatokat. Nem különböztetjük meg a vásárlókat az általuk megvásárolt termékkészlettől. Következésképpen a termékkészletekre vonatkozó eredmények természetesen fennállnak a vásárlók elemzésére.
2. A termékkészletek között adódó struktúra alapján vizsgálhatjuk a termékkészletek (vásárlócsoportok) jellemzőit, illetve a köztük lévő kapcsolatokat. A termékkészletek (vásárlócsoportok) vizsgálatait ezen alapstruktúra figyelembevételével kell végezni.

Ezen túlmenően a mennyiségi módszer még több előnyt hoz magával.

A bevásárlókosár-modellezést kiterjeszthetjük a fluidumfolyamatok vizsgálatára. Egy fluidumfolyamat az egymással kapcsolódó információfeldolgozó csomópontokból áll. A folyamatban működő csomópontokat egyféle „vásárlók”-nak tekinthetjük, amelyek a működésük során különböző információkészleteket igényelnek. Feltételezzük, hogy a csomópontok által igényelt információkészlet típusai végesek, mint a termékek egy üzletben. A fent említett modellezés szerint a csomópontok által igényelt információkészletetek alapján a csomópontok közötti kapcsolatokat, így a folyamat struktúráját lehet vizsgálni, illetve a folyamat csomópontjait lehet osztályozni.

1. ábra: *A fluidumfolyamat információfeldolgozó csomópontja, mint egy „vásárló"*



Az alábbi fejezetekben megmutatjuk a bevásárlókosár modelljét, amely a termékkészletek vagy szolgáltatások mennyiségi elemzésére alkalmas. Emlékeztetőül megmutatjuk a korábbi kutatásokban alkotott modellt (Demetrovics et al. 2011: 170–173). Megismertetjük a gyakori termékekre, illetve a termékek közötti asszociációs kapcsolatokra vonatkozó elemzést. A modellezés formalizmusa alapján megvizsgáljuk az osztályozási problémát. A kapott eredmények közvetlenül felhasználhatók a szolgáltatási folyamatokat igénybe vevők szegmentálására.

A bevásárlókosár-modell

Adott $P = \{p_1, p_2, \ldots, p_n\}$ áruk véges halmazára vásárlói kosárnak tekintsük az a = $(\alpha[1], \alpha[2], \ldots, \alpha[n])$ sorozatot, ahol $\alpha[i] \in \aleph$ a P_i áru mennyisége az α kosárban. A vásárlói kosarak összességét Ω-val jelöljük. Valójában egy $\alpha = (\alpha[1], \alpha[2], \ldots, \alpha[n])$ sorozat alatt egy termékkészletet vagy fluid folyamatok esetén egy információkészletet érthetünk. Egy vásárló, vagy fluid folyamatok esetén egy csomópont a vásárlói kosárral, illetve az információkészlettel azonosítható. Ebben az értelemben az alábbiakban megmutatott vásárlói kosarakra, termékkészletekre vonatkozó eredmények értelmesek a vásárlókra, csomópontokra vonatkozóan is.

Adott $\alpha, \beta \in \Omega$-ra, ahol $\alpha = (\alpha[1], \alpha[2], \ldots, \alpha[n])$, $\beta = (\beta[1], \beta[2], \ldots, \beta[n])$ írjuk $\alpha \leq \beta$ ha minden $i = 1, 2, \ldots, n$-re teljesül $\alpha[i] \leq \beta[i]$. A $\langle \Omega, \leq \rangle$ pár a $\leq$ természetes részberendezéssel rendelkező háló. Adott $A \subseteq \Omega$ halmazra jelöljük:

$$U(A) = \{\alpha \in \Omega \mid \forall \beta \in A : \beta \leq \alpha\} \tag{1}$$

és

$$L(A) = \{\alpha \in \Omega \mid \forall \beta \in A : \alpha \leq \beta\}. \tag{2}$$

Jelöljük továbbá

$$sup(A) = \{\alpha \in U(A) \mid \nexists \beta \in U(A) : \beta < \alpha\} \tag{3}$$

és

$$inf(A) = \{\alpha \in L(A) \mid \nexists \beta \in L(A) : \alpha < \beta\}. \tag{4}$$

Megjegyezhetjük, hogy $sup(A)$ és $inf(A)$ az Ω egyes elemei, nevezetesen $sup(A) = u \in \Omega$, ahol $u[i] = max\{\alpha[i] \mid \alpha \in A\}$ és $inf(A) = v \in \Omega$, ahol $v[i] = min\{\alpha[i] \mid \alpha \in A\}$.
Adott $A \subseteq \Omega$ és $\alpha \in \Omega$-ra jelöljük

$$supp_A(\alpha) = \frac{|\{\beta \in A \mid \alpha \leq \beta\}|}{|A|}. \tag{5}$$

A $supp_A(\alpha)$ a hányada azoknak a vásárlói kosaraknak az egész A-ra, amelyek meghaladják az adott, a minta-bevásárlókosárként ismert α küszöböt. A $supp_A(\alpha)$ az α kosár A-ban való támogatottságát jelenti. A vásárlói kosár támogatottsága egy statisztikai mutató, és természetesen a nagyobb támogatottsággal rendelkező vásárlói kosár fontosabb, és felkelti az üzleti menedzserek, valamint a kutatók figyelmét.

Megjegyezhetjük, hogy egy p_i áru a vizsgálatunkban $U(\alpha_i)$-vel azonosítható, ahol $\alpha_i = (\alpha[1], \alpha[2], \ldots, \alpha[n])$, $\alpha[k] = 0$, ha $k \neq i$ és $\alpha[i] = 1$. Ne keverjük össze p_i-t α_i-vel.

Adott $\alpha, \beta \in \Omega$-ra, ahol $\alpha = (\alpha[1], \alpha[2], \ldots, \alpha[n])$ és $\beta = (\beta[1], \beta[2], \ldots, \beta[n])$ írjuk $\gamma = \alpha \cup \beta$, ha $\gamma[i] = max\{\alpha[i], \beta[i]\}$ minden $i = 1, 2, \ldots, n$-re. A β kosár α kosárhoz való *asszociációs szabályát* $\alpha \to \beta$-val jelöljük. Az $\alpha \to \beta$ asszociációs szabály A MB halmazára való bizalmassága alatt az alábbi hányadot értjük:

$$conf_A(\alpha \to \beta) = \frac{supp_A(\alpha \cup \beta)}{supp_A(\alpha)}. \tag{6}$$

Mint Agrawal és Srikant (1994: 487–499) megjegyezték, az MB halmazának támogatottsága egyféle statisztikai mutató, az asszociációs szabály bizalmassága a szabály egyik erősségmutatója.

Gyakori bevásárlókosarak

A gyakori bevásárlókosarak feltárása mindig vonzó téma a kutatók számára (Pasquier et al. 1999: 398–416). Ebben a cikkben a korábbi kutatásban használt mószerrel elemezzük a gyakori bevásárlókosarakat a termék mennyisége alapján.

Adott $A \subseteq \Omega$, $\alpha \in \Omega$-ra és $0 \le \varepsilon \le 1$-re mondjuk, hogy α egy ε-gyakori MB, ha $supp_A(\alpha) \ge \varepsilon$. Az ε-gyakori MB összességét Φ_A^{ε}-vel jelöljük.

A priori elv: Adott $A \subseteq \Omega$, $\alpha, \beta \in \Omega$-ra és $0 \le \varepsilon \le 1$-re, ha $\alpha \le \beta$ és β egy ε-gyakori kosár, akkor α is ε-gyakori kosár.

Példa 1: Tekintsük a $P = \{a, b, c\}$ áruhalmazt és az $A = \{\alpha, \beta, \gamma, \delta\}$ tranzakciók halmazát, ahol $\alpha = (2, 1, 0)$, $\beta = (1, 1, 1)$, $\gamma = (1, 0, 1)$, $\delta = (2, 2, 0)$. A $\sigma = (1, 1, 0)$, $\eta = (1, 2, 0)$ kosarakra $supp_A(\sigma) = \frac{3}{4}$ és $supp_A(\eta) = \frac{1}{4}$. Adott $\varepsilon = \frac{1}{2}$ küszöbre az A ε-gyakori MB-k az alábbiak:

$$\Phi_A^{\frac{1}{2}} = \{(2,1,0),(1,0,1),(1,1,0),(2,0,0),(0,0,1),(0,1,0),(1,0,0),(0,0,0)\}.$$

Jelöljük a

$$\Phi_{A,k} = \{\alpha \in \Omega \mid \exists \alpha_1, \alpha_2, \ldots, \alpha_k \in A : \alpha \le \{\alpha_1, \alpha_2, \ldots, a_k\}\}.$$

Belátható, hogy ha $k \le l$, akkor $\Phi_{A,k} \supseteq \Phi_{A,l}$ és $\Phi_A^{\varepsilon} = \Phi_{A,k}$, ahol $k = \lceil \varepsilon |A| \rceil$ a legkisebb egész szám, amelyik nem kisebb, mint $\varepsilon |A|$.

Tétel 1: Adott $P = \{p_1, p_2, \ldots, p_n\}$ áruhalmazra, $A \subseteq \Omega$ egy kosarak halmazára és egy $0 \le \varepsilon \le 1$ küszöbre az $\alpha \in \Omega$ kosár ε-gyakori akkor, és csak akkor, ha létezik $\alpha_1, \alpha_2, \ldots, \alpha_k \in A$, amire $\alpha \in L(\{\alpha_1, \alpha_2, \ldots, \alpha_k\})$, ahol $k = \lceil \varepsilon |A| \rceil$.

Bizonyítás: Ha létezik $\alpha_1, \alpha_2, \ldots, \alpha_k \in A$, $k = \lceil \varepsilon |A| \rceil$, amire $\alpha \in L(\{\alpha_1, \alpha_2, \ldots, \alpha_k\})$, akkor $\alpha \le \alpha_i$ minden $i = 1, 2, \ldots, k$-ra, azaz $supp_A(\alpha) = \frac{|\{\beta \in A | \alpha \le \beta\}|}{|A|} \ge \frac{k}{|A|} \ge \varepsilon$.

Visszafelé, ha $supp_A \ge \varepsilon$, akkor $|\{\beta \in A \mid \alpha \le \beta\}| \ge \varepsilon . |A|$, azaz létezik $\alpha_1, \alpha_2, \ldots, \alpha_k \in A$, $k = \lceil \varepsilon |A| \rceil$, amire $\alpha \in L(\{\alpha_1, \alpha_2, \ldots, \alpha_k\})$.

A Tétel 1-ből következik:

Algoritmus 1: (Az összes ε-gyakori MB létrehozása adott A tranzakciók halmazára.)

Input: P áruhalmaz, $A \subseteq \Omega$ kosarak halmaza és $0 \leq \varepsilon \leq 1$ küszöb.

Output: Φ_A^ε.

Lépés 1: $\Phi_A^\varepsilon := \emptyset$.
Lépés 2: $k = \lceil \varepsilon |A| \rceil$.
For all $B \subseteq A$, $|B| = k$
$\Phi_A^\varepsilon := \Phi_A^\varepsilon \cup L(B)$
Endfor;
End

Az algoritmus $O\left(\binom{|A|}{k} \cdot (m+1)^n\right)$ futtatási időt igényel, ahol $|P| = n$, $k = \lceil \varepsilon |A| \rceil$ és $m = max\{\alpha[i] \mid \alpha \in A,\ i = 1, 2, \dots n\}$.

Az előző tételből következik:

Tétel 2: (A gyakori M*B* explicit reprezentációja) Adott $P = \{p_1, p_2, \dots, p_n\}$ áruhalmazra, $A \subseteq \Omega$ egy kosarak halmazára és egy $0 \leq \varepsilon \leq 1$ küszöbre létezik $\alpha_1, \alpha_2, \dots, \alpha_s \in \Omega$, ahol $s = \binom{|A|}{\lceil \varepsilon |A| \rceil}$, amire

$$\Phi_A^\varepsilon = \bigcup_{i=1}^{s} L(\alpha_i).$$

Bizonyítás: Legyenek $\{\alpha_1, \alpha_2, \dots, \alpha_s\}$ az $inf\{\beta_1, \beta_2, \dots, \beta_k\}$ halmaza, ahol $k = \lceil \varepsilon |A| \rceil$ és $\beta_i \in A$. A Tétel 2-ből lehet következtetni:

$$\alpha \in \Phi_A^\varepsilon \Leftrightarrow \alpha \leq inf(\{\beta_1, \beta_2, \dots, \beta_k\})$$

valamelyik $\{\beta_1, \beta_2, \dots, \beta_k\} \subseteq A$-ra, ahol $k = \lceil \varepsilon |A| \rceil$. Ez azt jelenti: $\Phi_A^\varepsilon = \bigcup_{i=1}^{s} L(\alpha_i)$.

Belátható, hogy $\alpha_i \leq \alpha_j$ akkor és csak akkor, ha $L(\alpha_i) \subseteq L(\alpha_j)$. Adott A MB halmazára és az ε küszöbre egy $\alpha_1, \alpha_2, \dots, \alpha_s$ MB halmaza, amire teljesül

i. $\Phi_A^\varepsilon = \bigcup_{i=1}^{s} L(\alpha_i)$,
ii. $\forall i, j : 0 \leq i, j \leq s \quad \alpha_i \nleq \alpha_j$ és $\alpha_j \nleq \alpha_i$.

A MB alap ε-gyakori halmazának nevezzük. Könnyen látható, hogy adott A, ε-ra A MB alap

ε-gyakori halmaza egyértelműen meghatározható, amelyet S_A^ε-vel jelölünk. Mivel fontos a Φ_A^ε meghatározása (A-beli ε-gyakori MB halmaza), az S_A^ε MB alap ε-gyakori halmazának a meghatározása érdekes. A fenti tételekből és az S_A^ε meghatározásából közvetlenül következik:

Tétel 3: Adott $P = \{p_1, p_2, \ldots, p_n\}$ áruhalmazra, és adott $0 \leq \varepsilon \leq 1$ küszöbre minden $A \subseteq \Omega$ MB halmazhoz rendelhető egy MB alap ε-gyakori halmaz S_A^ε.

Az egyszerű bizonyítást kihagyjuk. Az alábbi algoritmus egy S_A^ε MB alap ε-gyakori halmazt hozza létre adott $A \subseteq \Omega$ kosarak halmaza és ε küszöb esetén:

Algoritmus 2: (S_A^ε MB alap ε-gyakori halmazának létrehozása)

Input: P áruhalmaz, $A \subseteq \Omega$ kosarak halmaz és $0 \leq \varepsilon \leq 1$ küszöb.

Output: S_A^ε.

Lépés 1: $S_A^\varepsilon := \emptyset$.
Lépés 2: $k = \lceil \varepsilon\, |A| \rceil$.
For $B \subseteq A$, $|B| = k$
For $\alpha \in S_A^\varepsilon$
If $\alpha \leq inf(B)$ or $inf(B) \leq \alpha$ then
$S_A^\varepsilon := S_A^\varepsilon \setminus \{min(\alpha, inf(B))\} \cup \{max(\alpha, inf(B))\}$.
else
$S_A^\varepsilon := S_A^\varepsilon \cup \{inf(B))\}$.
endif
endfor
endfor
end

Belátható, hogy $|S_A^\varepsilon| \leq \binom{|A|}{k}$, ha $|P| = n$, $k = \lceil \varepsilon|A| \rceil$, $m = max\{\alpha[i] \mid i{=}1, 2, \ldots, n;\ \alpha \in A\}$. Ezért az algoritmus $O\left(\binom{|A|}{k}.m.n\right)$ futtatási időt igényel. Megjegyezhetjük, hogy a nagy A esetén az S_A^ε MB alap ε-gyakori halmaza sokkal gyorsabban kiszámítható, mint a Φ_A^ε MB ε-gyakori halmaz.

Példa 2: Tekintsük a Példa 1-et. Adott A tranzakcióhalmazra Algoritmus 2 generálja az alap $\frac{1}{2}$-gyakori MB halmazt $S_A^{\frac{1}{2}} = \{\rho, \theta\}$, ahol $\rho = (2, 1, 0)$, $\theta = (1, 0, 1)$. Ez azt jelenti, hogy A $\frac{1}{2}$-gyakori MB halmaza $\Phi_A^{\frac{1}{2}} = L(\rho) \cup L(\theta)$.

Asszociáció és bizalmasság

A gyakori bevásárlókosarak mellett az asszociációs szabályok feltárása is érdekes téma az adatbányászatban (Ping et al. 2004: 31–47). Az általánosabb modellben egy adott bizalmasságra felfedezhetjük az összes asszociációt. Adott $P = \{p_1, p_2, \ldots, p_n\}$ áruhalmazra, egy $A \subseteq \Omega$ kosarak halmazára és egy $0 \le \varepsilon \le 1$ küszöbre egy $\alpha \to \beta$ asszociációt ε-*bizalmasnak* nevezünk, ha $conf_A(\alpha \to \beta) \ge \varepsilon$. Az összes A-beli ε-bizalmas asszociáció halmazát C_A^{ε}-vel jelöljük. Igaz az alábbi tétel:

Tétel 4: Adott $P = \{p_1, p_2, \ldots, p_n\}$ áruhalmazra, $A \subseteq \Omega$ egy MB halmazára és egy $0 \le \varepsilon \le 1$ küszöbre egy $\alpha \to \beta$ asszociáció ε-bizalmas akkor és csak akkor, ha $\frac{|U(\alpha \cup \beta) \cap A|}{|U(\alpha) \cap A|} \ge \varepsilon$.

Bizonyítás: Megjegyezzük, hogy $supp_A(\alpha \cup \beta) = \frac{|U(\alpha \cup \beta) \cap A|}{|A|}$ és $supp_A(\alpha) = \frac{|U(\alpha) \cap A|}{|A|}$.

A megjegyzéssel a bizonyítás nyilvánvaló.

A keresztmarketing (*cross marketing*), az üzletek elrendezése (*store layout*) stb. területeken felmerülő kérdések egyike az adott bizalmassággal rendelkező asszociációk felfedezése. Az általunk konstruált általánosabb modellben az alábbi tételben valamely értelemben megmutatunk egy explicit reprezentációt a bizalmas asszociációs szabályokra. Pontosabban, megmutatunk egy módszert, amely szerint adott α MB-re és adott bizalmassági küszöbre felfedezhetjük az összes olyan MB β-t, amelyikre $\alpha \to \beta$ bizalmas asszociáció.

Megjegyezzük, hogy ha ρ, σ, $\rho \le \sigma$, akkor

$$\{\eta \in \Omega \mid \rho \cup \eta \le \sigma\} = L(\sigma).$$

Ebből következik:

Tétel 5: (Bizalmas asszociációs szabályok explicit reprezentációja.) Legyen $P = \{p_1, p_2, \ldots, p_n\}$ egy áruhalmaz, $A \subseteq \Omega$ egy kosarak halmaza és $0 \le \varepsilon \le 1$ egy küszöbérték. Minden $\alpha \in \Omega$ MB-re léteznek $\alpha_1, \alpha_2, \ldots, \alpha_k \in \Omega$, amelyekre:
$\forall \beta \in \Omega : \alpha \to \beta$ ε-bizalmas asszociációs szabály akkor és csak akkor, ha $\beta \in \bigcup_{i=1}^{k} L(\alpha_i)$.

Bizonyítás: Tegyük $s = \lceil \varepsilon |U(\alpha) \cap A| \rceil$. A Tétel 4 szerint $\alpha \to \beta$ egy ε-bizalmas asszociációs szabály akkor és csak akkor, ha $|U(\alpha \cup \beta) \cap A| \ge s$. Legyen α_i az $inf(B)$, ahol $B \subseteq A$, $|B| \ge s$. Nyilvánvaló, hogy $|U(\alpha \cup \beta) \cap A| \ge s$ akkor és csak akkor, ha $\beta \in L(\alpha_i)$. A bizonyítás befejeződik.

A Tétel 5 valamely értelemben megad egy explicit módszert a bizalmas asszociációs szabályok reprezentálására. Az alábbi algoritmus a Tétel 5 egyik közvetlen következménye:

Algoritmus 3: (Generálni az összes $\alpha \to \beta$ ε-bizalmas asszociációs szabályokat adott α-ra és ε-ra.)

Input: P áruhalmaz, $A \subseteq \Omega$ kosarak halmaza, $0 \le \varepsilon \le 1$ küszöb, és α egy kosár.

Output: $\bigcup_{i=1}^{k} L(\alpha_i) = \{\beta \mid \alpha \to \beta \;\; \varepsilon\text{-bizalmas asszociációs szabály}\}$.

Lépés 1: $C := U(\alpha) \cap A = \{\gamma \in A \mid \alpha \le \gamma\}$.
Lépés 2: $s := \lceil \varepsilon |C| \rceil$.
$k := |\{B \subseteq A \,\|B| \ge s, \quad \alpha \le \inf(B)\}|$
For $B \subseteq A$, $|B| \ge s$, $\alpha \le \inf(B)$, calculate $\alpha_i = \inf(B)$, $i = 1, 2, \ldots, k$.
EndFor
Lépés 3:
For $i = 1, 2, \ldots, k$ calculate $L(\alpha_i)$
EndFor
Lépés 4:
Output $\bigcup_{i=1}^{k} L(\alpha_i)$.
End

Példa 3: Tekintsük a Példa 1-et. Adott A MB-k halmazára (lásd Példa 1-et), a $\sigma = (1, 1, 0)$ kosárra és $\varepsilon = \frac{1}{2}$ küszöbre keressük az összes η kosarat, amelyikre $\sigma \to \eta$ asszociációs szabály ε-bizalmas legyen. Kiszámítjuk $U(\sigma) \cap A = \{(2, 1, 0), (1, 1, 1), (2, 2, 0)\}$ és $s := \lceil \varepsilon \mid U(\alpha) \cap A \mid \rceil = 2$. Az Algoritmus 3, Lépés 2 után kaptuk $k = 4$ és $\alpha_1 = (1, 1, 0)$, $\alpha_2 = (2, 1, 0)$. Az összes η kosár, amelyikre az $\sigma \to \eta$ asszociációs szabály $\frac{1}{2}$-bizalmas, a

$$L(\alpha_1) \cup L(\alpha_2) = \{(1, 1, 0), (1, 0, 0), (0, 1, 0), (0, 0, 0), (2, 1, 0), (2, 0, 0)\}$$

halmaz. Ennek eredményeként a $\sigma \to \sigma'$ $(\sigma' \le \sigma)$ formájú triviális asszociációs szabályokon kívül megtaláltuk még a nem triviális asszociációs szabályokat $\sigma \to (2, 1, 0)$ és $\sigma \to (2, 0, 0)$. Ez azt jelenti, hogy az A vásárlókörből az a-t és b-t vásárló ügyfelek több mint 50%-a megvette $2a$-t és $1b$-t, és több mint 50%-a megvette $2a$-t.

Osztályozás

Az osztályozás fontos probléma több területen. A közszolgáltatási, banki szférában az ügyfél-osztályozás kulcsszerepet játszik a vállalatok stratégiájának meghatározásában, az ügyfélkezelés folyamatában (Chicco et al. 2005: 164–172). A korábbi kutatásokban (Qiaohong et al. 2010: 509–520; Thangaraj – Vijayalakshmi 2011: 1–6) az osztályozást, illetve az osztályozási módszer hatékonyságának az értékelését az osztályozandó elemek különböző jellemzői alapján vizsgálták. Azonban megjegyzendő, hogy a Demetrovics–Hua–Guban-cikkben (2011: 170–173) ismertetett modell szerint a termékkészleteket, az ügyfeleket lehet osztályozni magukban a termékkészletekben, illetve az ügyfelek által rendelt termékkészletekben rejtett kapcsolatok alapján.

Osztályozás:

Legyen A egy vásárlók halmaza vagy a fluidumfolyamat egy csomóponthalmaza. Az A osztályozása alatt értjük az A részhalmazai egy családját: $C_A = \langle U_1, U_2, \ldots, U_k \rangle$, ahol $U_i \subseteq A$.

A vásárlók közötti távolság és a legközelebbi szomszéd:

Legyen $\alpha = (\alpha[1], \alpha[2], \ldots, \alpha[n])$, $\beta = (\beta[1], \beta[2], \ldots, \beta[n])$ két vásárlói kosár az adott $P = \{p_1, \mathrm{p}_2, \ldots, p_n\}$ áruk halmazából. Az előbbi elemezés alapján α, β két vásárlónak, vagy fluidumfolyamatok esetén két csomópontnak tekinthető. $d(\alpha, \beta)$-val jelöljük a metrikát, azaz a távolságot α és β között. Egy jól ismert metrika a vásárlók, illetve a csomópontok között az euklideszi távolság:

$$d(\alpha, \beta) = \left[\sum_{i=1}^{n} (\alpha[i] - \beta[i])^2\right]^{\frac{1}{2}}.$$

Legyen α és β két vásárló, vagy a fluidumfolyamatok esetén két csomópont egy adott A vásárlók halmazából, illetve a fluidumfolyamat bizonyos csomóponthalmazából. Akkor mondjuk, hogy α egy legközelebbi szomszédja β-nak, ha

i. $d(\alpha, \beta) \leq d$, és
ii. nincs $\gamma \in A$, amelyre $\gamma \neq \beta$ és $d(\gamma, \beta) < d$.

Osztályozás szerinti szomszédság:

Azonban egy adott $C_A = \langle U_1, U_2, \ldots, U_k \rangle$ osztályozásra az A-beli vásárlók, illetve csomópontok közötti legközelebbi szomszédságot másképpen lehet megfogalmazni. Mondjuk, hogy α egy d-szomszédja β-nak az adott $C_A = \langle U_1, U_2, \ldots, U_k \rangle$ osztályozásra, ha van $U_{i_1}, U_{i_2}, \ldots, U_{i_d}$, melyre $\alpha, \beta \in \bigcap_{j=1}^{d} U_{i_j}$. Akkor mondjuk, hogy α egy legközelebbi szomszédja β-nak, ha van olyan d természetes szám, amelyre:

i. α egy d-szomszédja β-nak, és
ii. nincs $\gamma \in A$, amely $\gamma \neq \beta$ és γ *egy* d'-szomszédja β-nak, $d' < d$.

Egy adott $C_A = \langle U_1, U_2, \ldots, U_k \rangle$ osztályozásra jelöljük:

$$C_A^{(d)} = \left\{ \bigcup_{j=1}^{d} U_{i_j} \mid U_{i_j} \in C_A \right\}. \tag{7}$$

Egy A feletti C_A osztályozásra R_{C_A}-val jelöljük a C_A által meghatározott szomszédságot: α áll β-val R_{C_A} kapcsolatban, ha α egy 1-szomszédja β-nak C_A osztályozásra. Mondjuk akkor azt is, hogy R_{C_A} szomszédsági reláció C_A által generálható.

Két $C = \langle U_1, U_2, \ldots, U_k \rangle$, $D = \langle V_1, V_2, \ldots, V_l \rangle$ osztályozásra mondjuk, hogy C dominált D által, jelölve $C \leq D$, ha minden U_i-ra létezik V_j, amire $U_i \subseteq V_j$. Egy C osztályozásra $C^{\max}$-mal jelöljük C-beli maximális halmazok családját. Beláthatjuk, hogy a C és $C^{\max}$ osztályozás egymás által dominált.

Az alábbi lemmákat röviden bebizonyíthatjuk:

Lemma 1. α egy d-szomszédja β-nak a C_A osztályozásra akkor és csak akkor, ha α egy 1-szomszédja β-nak a $C_A^{(d)}$ osztályozásra.

Lemma 2. Ha két C, D osztályozás egymás által dominált, akkor ugyanazt a szomszédsági relációt generálja: $R_C = R_D$.

Lemma 3. Minden C osztályozásra C és $C^{\max}$ ugyanazt a szomszédsági relációt generálja: $R_C = R_{C^{\max}}$.

A fenti lemmákból következtethető:

Tétel 6: Legyen A egy vásárlók halmaza és C_A egy osztályozás A felett. α egy d-szomszédja β-nak a C_A osztályozásra akkor és csak akkor, ha α áll β-val a $R_{C_A^{(d)\max}}$ kapcsolatban.

A Tétel 6 alapján beláthatjuk, hogy az alábbi algoritmus egy C_A osztályozásra a $C_A^{(d)\max}$ osztályozást generálja. A $C_A^{(d)\max}$ osztályozással könnyen ellenőrizhető, hogy két vásárló d-szomszédja-e egymásnak.

Algoritmus:

Input: *A* vásárlók halmaza, C_A egy *A* feletti osztályozás.
Output: $C_A^{(d)\max}$ egy *A* feletti osztályozás.

Lépés 1: C_A alapján ki kell számítani a $C_A^{(d)}$-t:

$$C_A^{(d)} = \left\{ \bigcup_{j=1}^{d} U_{i_j} \mid U_{i_j} \in C_A \right\}.$$

Lépés 2: $C_A^{(d)}$ alapján ki kell számítani a $C_A^{(d)\max}$-ot.

Megjegyzendő, hogy bár az algoritmus 1. lépése általában exponenciális időt igényel, a speciális esetekben, amikor *d*- konstans és jelentősen kisebb a vásárlók számánál, hatékonyan működik az algoritmus.

Konklúzió

A cikk a korábbi kutatás folytatásában elért eredményeket ismertette. Bemutatjuk, hogy a felállított modell és a felhasznált algebrai megközelítés teljesen alkalmas a bevásárlókosár-modellezésre, és alkalmas a fluidumfolyamatok vizsgálatára. A fluidumfolyamatok elemzésében különösen fontos a folyamatos elemzés, amelyet csak a csomópontok jellemzői ismeretével lehet eredményesen végezni. A cikkben említett szempontból a fluidumfolyamatok csomópontjai az információ igényei alapján azonosíthatóak, jellemezhetőek. Eszerint a folyamatok csomópontjait egyféle „vásárlók”-nak lehet tekinteni. Ennek alapján a cikkben ismertetett eredmények szerint tudjuk:

- a csomópontok által igényelt információkészletek struktúráját, mint a termékkészletek struktúráját vizsgálni,
- a gyakori információkészleteket, az információkészletek közötti asszociációs kapcsolatokat megállapítani,
- a csomópontok információjellemzőit megismerni, a csomópontok folyamatban való szerepét, illetve a csomópontok közötti asszociációs kapcsolatokat megállapítani,
- a csomópontok információjellemzőinek ismeretében a fluidumfolyamatok szervezésében felmerülő feladatokat, közülük az optimalizálási feladatokat megoldani. Az opti-

malizálás a csomópontok közötti szomszédsági kapcsolat, a csomópontok osztályozása révén végezhető.

A korábbi kutatásban (Demetrovics et al. 2011: 24–31) ismertetett eredmény valóban a termékkészletekben, az információkészletekben, illetve a csomópontokban rejlő természetes jellemzőket, a köztük lévő természetes rendezést tárta fel. Az eredmény alapján javasoltunk ebben a cikkben egy osztályozási módszert, amellyel a termékkészleteket, az információkészleteket, illetve a csomópontokat csoportosítani lehet a természetes rendezésnek megfelelően.

A korábbi kutatásban és az e cikkben elért eredmények megmutatják, hogy a mennyiségi elemzés valóban egy hasznos módszer a bevásárlókosár-modellezésben, és különösen hatékony eszköz a fluidumfolyamatok elméleti és gyakorlati vizsgálatában.

Hivatkozások

Agrawal, R. – Srikant, R. (1994). *Fast algorithms for mining association rules.* Proceedings of the 20th VLDB Conference, Santiago, Chile, pp. 487–499.

Chicco, G. – Napoli, R. – Piglione, F. – Postolache, P. – Scutariu, M. – Toader, C. (2005). Emergent customer classification, generation, transmission and distribution. *IEE Proceedings, 152*(2), 164–172.

Demetrovics, J. – Hua, N. S. – Guban, A. (2011). *An algebraic approach to market basket model: explicit representation of frequent market baskets and associations rules, CSIT 2011.* Computer science and information technologies. Proceedings of the conference. Yerevan, pp. 170–173.

Demetrovics, J. – Hua, N. S. – Guban, A. (2011). An algebraic representation of frequent market baskets and association rules. *Cybernetics and Information Technologies, 11*(2), 24–31.

Pasquier, N. – Bastide, Y. – Taouil, R. – Lakhal, L. (1999). *Discovering frequent closed itemsets for association rules.* ICDT'99 Proceedings of the 7th International Conference on Database Theory. London: Springer-Verlag, pp. 398–416.

Ping, Y. H. – Yen, L. C. – Chun, C. L. (2004). Algorithms for mining association rules in bag databases. *Information Sciences, 166*(1–4), 31–47.

Zu, Q. – Wu, T. – Wang, H. (2010). A multi-factor customer classification evaluation model. *Computing and Informatics, 29*(4), 509–520.

Thangaraj, M. – Vijayalakshmi, C. R. (2011). A Study on Classification Approaches across Multiple Database Relations. *International Journal of Computer Applications, 12*(12), 1–6.