FUTÓ PÉTER

A cluster analízis egy új modellje és algoritmusa

1. A cluster analízis gráf és hipergráf modelljei

A cluster analízis alapfeladata az, hogy objektumok és jellemzőik bonyolult rendszerének struktúráját feltárja; az objektumokat – előzetes ismeretek, tapasztalatok nélkül – kizárólag a jellemzőikből adódó kapcsolataik alapján természetes csoportokba ún. clusterekbe sorolja úgy, hogy az egymáshoz hasonló objektumok azonos clusterekbe, az egymáshoz kevésbé hasonló objektumok különböző clusterekbe kerüljenek.

A clusteranalízis feladatkörébe tartozik a jellemzők csoportosítása is, az általuk jellemzett objektumokból adódó kapcsolatok felhasználásával. A bonyolult rendszerek struktúrájának feltárása nagy jelentőségű és gyakori feladat az orvostudományban, biológiában, a közgazdaságtanban, a mérnöki tudományokban, az információ-tudományban és még sok más területen.

Tehát a cluster analízis modelljeinek, eljárásainak felhasználási területe igen széles. Ezt tükrözi terminológiájának heterogenitása is. A biológusok, orvosok numerikus taxonomiáról, a mérnökök tanító nélküli tanuló algoritmusokról, a statisztikusok, információtudományi szakemberek cluster analízisről, az operációkutatók partícionálási feladatról beszélnek, de valamennyien ugyanazon probléma megoldására, egy bonyolult rendszer struktúrájának feltárására törekszenek. A különböző felhasználási területeken alkalmazott modellek és eljárások összefonódásával a hetvenes évek elejére a cluster analízisnek két fő ága alakult ki: a matematikai-statisztikai és az információtudományi cluster elemzés.

A matematikai-statisztikai clusteranalízis a vizsgált rendszer valamennyi objektumát egy adott rendezett jellemző halmaz felhasználásával írja le. Tehát adott az objektumoknak a $D = \{d_1, \ldots d_m\}$ halmaza, és az objektumok mindegyikén megfigyelhető jellemzők $C = (C_1, \ldots C_p)$ vektora.

A jellemzők mérési skálájuk alapján négy fő csoportba sorolhatók.

Jelölje a C jellemzőnek a d_j, ill. d_k objektumra jellemző értékét $C(j)$, ill. $C(k)$.

Ha C nominális skálájú, akkor csak azt tudjuk, hogy $C(j) = C(k)$, vagy $C(j) \neq C(k)$ (pl. szem színe, születési hely).

Ha C ordinális skálájú, akkor azt tudjuk, hogy $C(j) = C(k)$ vagy $C(j) > C(k)$ vagy $C(j) < C(k)$ (pl. indiai kasztrendszerben elfoglalt hely).

Ha C intervallum skálájú, akkor ismerjük $C(j) - C(k)$ értékét (pl. hőmérséklet °C-ban).

Ha C hányados skálájú, akkor ismerjük $C(j)/C(k)$ értékét (pl. hőmérséklet °K-ban).

Hányados, ill. intervallum skálájú változók esetén nyilván nem mellékes a mértékegység megválasztása sem.

A matematikai-statisztikai cluster elemzés alapvető problémája a különböző skálájú jellemzők értékeinek felhasználásával az objektum-párokra hasonlósági mérőszám konstruálása. A probléma igen bonyolult, és sikeres megoldása a rendszer-struktúra feltárásának szükséges feltétele. Nem véletlen, hogy a cluster analízissel foglalkozó cikkek nagy hányada a hasonlósági mérőszám meghatározására szolgáló heurisztikus eljárások konstruálásával és összehasonlításával foglalkozik, szinte háttérbe szorítva a cluster kereső algoritmusok kidolgozását is (Anderberg [2]). Az információtudományi cluster elemzésnél adott az objektumoknak (általában dokumentumoknak) a $D = \{d_1, \ldots d_m\}$ halmaza, és az objektumok jellemzésére (indexelésére, leírására) használt tárgyszavaknak a $K = \{k_1, \ldots k_n\}$ halmaza.

Az objektumok jellemzésére a K halmaznak egy-egy általában nem rögzített elemszárú részhalmazát használják.

Az objektumok közötti hasonlóság és a hasonlósági mérőszám az objektumokat jellemző tárgyszó részhalmazok megegyező és egymástól különböző elemei számának valamilyen függvényén alapul.

A hasonlósági mérőszám konstruálása itt is igen bonyolult és döntő jelentőségű probléma. Ezt tükrözi a gyakorlatban használt mérőszámok széles skálája is.

A cluster elemzés mindkét ágában használt hasonlósági mérőszámok $[s(d_i, d_j)]$ két objektum között értelmezettek és a legtöbb esetben a következő tulajdonságokkal rendelkeznek:

1. $0 \leq s(d_i, d_j) \leq 1 \qquad (i, j = 1 \ldots m)$
2. $s(d_i, d_i) = 1 \qquad (i = 1 \ldots m)$
3. $s(d_i, d_j) = s(d_j, d_i) \qquad (i, j = 1 \ldots m)$

Ennek alapján megszerkeszthető a cluster analízis súlyozott gráf modellje, amelyen értelmezhető a gyakorlatban használt valamennyi cluster kereső algoritmus.

1. Modell

Adott a $D = \{d_1, \ldots d_m\}$ objektum halmaz, és valamennyi objektumpár esetén a hasonlósági mérőszámuk: $s(d_i, d_j)$ $(i, j = 1 \ldots m)$. Legyen $X = \{x_1, \ldots X_m\}$ egy gráf pontjainak a halmaza. Modelezze az x_i pont a d_i objektumot. Ha $d_i \neq d_j$ és $s(d_i, d_j) > 0$, akkor a gráf x_j és x_i pontja között irányítatlan élt (E_k) húzunk, amelyhez súlyként hozzárendeljük a $v(E_k) = s(d_i, d_j)$ értéket. Legyen $\varepsilon = \{E_1, \ldots E_m\}$. Tehát a $G = (X; \varepsilon)$ gráf és az élein értelmezett $v(E_j) > 0$ $(j = 1 \ldots n)$ súlyfüggvény modellezi az objektumokat és az objektumpárok hasonlósági mérőszámait.

A cluster kereső eljárásokat két nagy csoportba sorolhatjuk:

- hierarchikus eljárások,
- nem hierarchikus eljárások.

A hierarchikus eljárások a $G = (X; \varepsilon)$ gráf pontjai közül olyan részhalmazokat jelölnek ki, amelyek vagy egymást tartalmazzák, vagy diszjunktak. Minden esetben a kijelölt részhalmazok között lesznek a gráf pontjai is. Ezeket

a részhalmazokat a tartalmazási reláció felhasználásával hierarchiába rendezhetjük. (Innen adódik az eljárások neve is.) A hierarchikus eljárásokat két további csoportba sorolhatjuk:

– agglomeratív jellegű eljárások,
– nem agglomeratív jellegű eljárások.

Az agglomeratív jellegű eljárások a gráf pontjainak fokozatos összekapcsolásával alkotják a clustereket (pl. legközelebbi szomszéd módszere, centroid módszer, Ward-módszer) (*Sibson* [27], *Gower* [15], *Ward* [31]).

A nem agglomeratív jellegű eljárások a gráf fokozatos szétdarabolásával határozzák meg a clustereket. Ezek hatékony particionálási eljárások hiánya miatt egyenlőre nem nagyon elterjedtek, pl. Edwards és Cavalli-Sforza (*Anderberg* [2]).

A nem hierarchikus eljárások a $G = (X; \varepsilon)$ gráf pontjai közül olyan részhalmazokat (clustereket) jelölnek ki, amelyek diszjunktak (tehát a tartalmazás nem megengedett). Általában nem megkötés az, hogy a gráf valamennyi pontja legyen eleme legalább egy részhalmaznak.

A nem hierarchikus eljárásokat további két csoportba oszthatjuk:

– nem strukturális jellegű kritérium alapján osztályozó eljárások,
– strukturális kritérium alapján osztályozó eljárások.

A nem strukturális jellegű kritérium alapján osztályozó eljárásoknál általában előre adott a kívánt cluster szám, és valamilyen célfüggvény (pl. csoportokon belüli szórásnégyzetek összege), mint vezérfonal felhasználásával keresik a gráf pontjainak jobb elosztását (*MacQueen* [24], *Forgy* [10]).

A strukturális kritérium alapján működő eljárások nagy hányadánál súlyozott élű $G = (X; \varepsilon)$ gráfból küszöbszintek (pl. $t = 0{,}1; 0{,}3$) bevezetésével olyan $G = (X; \varepsilon_t)$ gráfokat állítanak elő, amelynél az x_i és x_j pontok között akkor húzódik él (amelyhez már nem rendelnek súlyt), ha $s(d_i, d_j) > t$. A $G = (X; \varepsilon_t)$ gráf(ok) komponenseit (*Auguston, Minker* [3]), vagy maximális teljes részgráfjait (*Osteen* [26]), vagy az ezekből képzett struktúrákat tekintik clustereknek. A strukturális kritériumok alapján működő eljárások közé sorolhatók a gráfelmélet particionálási módszerei is (*Lawler* [22]).

A cluster kereső eljárásokkal szemben a következő észrevételek tehetők:

– Több elem hasonlóságát csak elempárok hasonlóságával képesek kifejezni.
– Az algoritmusok futtatása előtt semmit, vagy csak igen keveset tudnak mondani az eredményként adódó clusterek tulajdonságairól. (Nincs explicit cluster definíció.)
– Az eljárások között nincs igazán hatékony, bizonyíthatóan az optimumhoz konvergáló algoritmus.

Jelenleg is széles körű nemzetközi kutatómunka folyik, amelynek célja a gyakorlatban jól használható egzakt cluster definíció megszerkesztése, valamint hatékony és konvergens cluster kereső algoritmusok kidolgozása. Ez a dolgozat a fenti kutatómunkát szeretné előbbre vinni a cluster elemzés hiper-

gráf modelljeinek megszerkesztésével, strukturális kritériumon alapuló cluster definíció, és olyan polinommal fedhető lépésszámú, konvergens, hierarchikus cluster eljárás kidolgozásával, amelynek alkalmazásával elkerülhető a hasonlósági mátrix megalkotása.

A szerzőt a hipergráf modellek megalkotására az információtudományi cluster elemzés gráf modelljének és eljárásainak kritikája sarkallta, míg a hipergráf kvázi-komponense fogalmának megalkotásához – ami az új cluster definíció és eljárás alapköve – *Lawler* [22] cikke adta az ötletet, aki *Luccio* és *Sami* [23] eredményeinek felhasználásával észrevette, hogy vannak a hipergráfoknak olyan ponthalmazai, amelyek a minimális két részre vágás során nem vágódnak el. Lawler erre a felismerésre alapozva hatékony heurisztikus eljárásokat dolgozott ki hipergráfok több részre vágására, de nem foglalkozott mélyebben az el nem vágódó ponthalmazok tulajdonságaival.

Az információtudományi cluster elemzésben két objektumot akkor tartanak hasonlónak, ha a jellemzésükre (indexelésükre) használt deszkriptorok (tárgyszavak) közül legalább egy közös.

Ez a bináris reláció nyilván szimmetrikus, reflexív, de nem tranzitív, tehát tolerancia reláció (*Srejder* [29]). Ez a reláció nyilván egy olyan több elemű relációból származik, amelynél minden egyes deszkriptor kapcsolatot létesít azon (nem feltétlenül kettő) objektumok között, amelyek jellemzésére az adott deszkriptort felhasználták.

Ezzel teljesen analóg módon értelmezhető egy több elemű reláció a deszkriptorok között is úgy, hogy minden egyes objektum kapcsolatot létesít a jellemzésére használt deszkriptorok között.

A matematikai-statisztikai cluster elemzésben az objektumok közötti hasonlóság fogalmát általában az objektumokat jellemző vektorok között definiált valamilyen távolság fogalmából vezetik le. Tehát a hasonlósági reláció itt is bináris, mégpedig szimmetrikus, reflexív és általában nem tranzitív (tolerancia) reláció. Ez a tolerancia reláció is nyilván egy olyan több elemű relációból származik, ahol az objektumok közötti hasonlóság alapja az, hogy egy vagy több jellemzőjük értéke nagyon közeli, vagy megegyezik.

Ezen alapul az a gondolat, hogy az objektumok jellemzésére szolgáló mátrix cellái értékeinek felhasználásával itt is deszkriptorokat alakítsunk ki. Például deszkriptor lehet az, hogy egy adott jellemző értéke egy adott intervallumba esik. A feladat természetétől függően deszkriptorokat definiálhatunk úgy is, hogy több jellemző értékét szorítjuk határok közé.

A deszkriptorok definiálásánál nem kikötés az, hogy segítségükkel az objektumoknak egy osztályozását hozzuk létre; tehát megengedhetjük azt is, hogy például a „súly 1" deszkriptor a 10 kp és a 15 kp közötti súlyú objektumok jellemzésére szolgáljon, míg a „súly-hossz" deszkriptor a 14 kp és a 16 kp közötti súlyú és a 1 m és a 2 m közötti hosszúságú objektumok jellemzésére szolgáljon.

Természetesen csak olyan deszkriptorokat definiálunk, amelyek legalább egy objektum jellemzésére szolgálnak, és a definiált deszkriptorok között található minden egyes objektumhoz legalább egy, amely az illető objektum jellemzésére szolgál.

Bár jelentősége nem olyan nagy mint az információtudományi cluster elemzésben, de itt is definiálható a deszkriptorok között egy több elemű reláció

úgy, hogy minden egyes objektum kapcsolatot létesít a jellemzésére használt deszkriptorok között.

A fentiek alapján mind az információtudományi, mind a matematikai-statisztikai cluster elemzésben jól használható a következő két hipergráf modell.

2. *Modell*

Adott a $D = \{d_1, \ldots d_m\}$ objektum halmaz. Jelölje a d_j objektum jellemzésére használt deszkriptorok halmazát $E_j (j = 1, \ldots m)$. Mivel az objektumok jellemzésére legalább egy, de véges sok deszkriptort használnak, ezért

$$1 \leq | E_j | \leq K \quad (j = 1, \ldots m). \tag{1}$$

Az objektum halmaz objektumainak jellemzésére szolgáló deszkriptorok halmazát jelölje: X

$$X = \bigcup_{j=1}^{m} E_j \tag{2}$$

Ha ε jelöli az objektumok jellemzésére szolgáló deszkriptor halmazok osztályát: $\varepsilon = \{E_1, \ldots E_m\}$, akkor $H = (X; \varepsilon)$ hipergráf, ugyanis (vö. 1. Definíció).

(I) $X = \{x_1, \ldots x_n\}$ véges halmaz (1), (2) miatt,
(II) $E_j \neq \emptyset (j = 1, \ldots m)$ (1) miatt,
(III) $\bigcup_{j=1}^{m} E_j =$ (2) miatt.

A $H = (X, \varepsilon)$ hipergráf pontjai tehát deszkriptorok, élei pedig az egy-egy objektum jellemzésére szolgáló deszkriptor halmazok.

3. *Modell*

A $H = (X; \varepsilon)$ hipergráf duálisa a $H^* = (E; X_1, \ldots X_n)$ hipergráf, amelynek pontjai $(e_1, \ldots e_m)$ a $H = (X; \varepsilon)$ hipergráf élet $(E_1, \ldots E_m)$ reprezentálják, élei pedig $(X_1, \ldots X_n)$ a $H = (X; \varepsilon)$ hipergráf pontjainak felelnek meg a következő értelemben:

$$X_i = \{e_j \mid j \leq m, \; x_i \in E_j\}. \tag{3}$$

$H^* = (E; \mathfrak{X})$ valóban hipergráf, ugyanis

(I′) $E = \{e_1, \ldots e_m\}$ véges halmaz (I) miatt,
(II′) $X_i \neq \emptyset (i = 1, \ldots n)$ (III) és (3) miatt.
(III′) $\bigcup_{i=1}^{m} X_i = E$ (II) és (3) miatt.

A $H^* = (E; \mathfrak{X})$ hipergráf pontjai objektumok, élei pedig az egyes deszkriptorok által meghatározott olyan objektum halmazok, amelyek objektumai jellemzésére az adott deszkriptort felhasználtuk.

A modellezés mindkét modell esetén az éleken értelmezett pozitív súlyfüggvény $v(E_j) > 0$ $(j = 1 \ldots m)$, illetve $u(X_i) > 0$ $(i = 1, \ldots n)$ bevezetésével finomítható. Ha finomításra nincs szükség, vagy nem lehetséges, akkor is bevezetünk egy súlyfüggvényt, mégpedig úgy, hogy minden egyes élhez az 1 súlyt rendeljük hozzá.

A dolgozat 2. része azt a strukturális kritériumon alapuló cluster definíció ismerteti, amely a cluster elemzés mindhárom fent említett modelljére sikerrel alkalmazható.

2. Matematikai alapok. A hipergráf kvázi-komponensének fogalmán alapuló új cluster definíció

A dolgozatban szereplő – hasonlósági mérőszámot nem használó – cluster eljárás a csoportosítandó objektumok és a jellemzésükre használt deszkriptorok kapcsolatait feltáró hipergráf modelleken, valamint a cluster definícióként alkalmazott kvázi-komponens fogalmán alapul.

A dolgozatnak ez a része tartalmazza azokat a matematikai alapokat, amelyek a kvázi-komponens definíciójához és a cluster elemzés szempontjából fontos tulajdonságainak leírásához szükségesek. A kvázi-komponens definícióját követik azok a megjegyzések, lemmák, tételek, amelyek alátámasztják a kvázi-komponens fogalom alkalmazhatóságát a cluster elemzésben. A dolgozatban csak azokat a tételeket bizonyítjuk, amelyek részletes bizonyítása a *Futó* [13] dolgozatban nem szerepel.

2.1. Definíció: Adott az $X = \{x_1, \ldots, x_n\}$ véges halmaz és $\varepsilon = \{E_1, \ldots, E_m\}$ az X halmaz részhalmazainak osztálya.

A $H = (X;\varepsilon)$ pár *hipergráf*, ha

$$E_j \neq \emptyset \qquad (j = 1, \ldots m), \tag{1}$$

$$\bigcup_{j=1}^{m} E_j = X. \tag{2}$$

Az X halmaz elemeit pontoknak, az ε halmaz elemeit éleknek nevezzük.

Ha $X = \emptyset$, akkor a hipergráfot üresnek nevezzük.

Értelmezzük a hipergráf ponthalmazai és élhalmazai között az $\mathcal{E}$: $2^X \to 2^\varepsilon$ és az $\mathcal{H}$: $2^\varepsilon \to 2^X$ leképezéseket:

2.2. Definíció: Tetszőleges S ponthalmaz $(S \subseteq X)$ esetén

$$\mathcal{E}(S) = \{E_j | E_j \in \varepsilon,\ \exists x_i \in S\colon\ x_i \in E_j\}.$$

2.3. Definíció: Tetszőleges $\mathcal{F}$ élhalmaz $(\mathcal{F} \subseteq \varepsilon)$ esetén

$$\mathcal{H}(\mathcal{F}) = \{x_i | x_i \in X,\ \exists E_j \in \mathcal{F}\colon\ x_i \in E_j\}.$$

2.1. Megjegyzés: Egyszerűen belátható, hogy tetszőleges S ponthalmaz $(S \subseteq X)$ és $\mathcal{F}$ élhalmaz $(\mathcal{F} \subseteq \varepsilon)$ választása esetén $S \subseteq \mathcal{H}\big(\mathcal{E}(S)\big)$ és $\mathcal{F} \subseteq \mathcal{E}\big(\mathcal{H}(\mathcal{F})\big)$, tehát az $\mathcal{E}$: $2^X \to 2^\varepsilon$ és az $\mathcal{H}$: $2^\varepsilon \to 2^X$ leképezések egymásnak nem inverzei.

Új fogalmak bevezetését, tételek egyszerűbb bizonyítását és a kvázi-komponensek meghatározására szolgáló algoritmus gyorsítását teszi lehetővé az $\mathscr{E}'\colon 2^X \otimes 2^X \to 2^\varepsilon$ leképezés, amely az $\mathscr{E}\colon 2^X \to 2^\varepsilon$ leképezés általánosítása.

2.4. Definíció: Tetszőleges S és T ponthalmazok $(S \subseteq X)$, $(T \subseteq X)$ esetén

$$\mathscr{E}'(S|T) = \{E_j | E_j \in \varepsilon;\ \exists x_i \in S\colon x_i \in E_j,\ E_j \subseteq T\}.$$

2.2. Megjegyzés: Egyszerű számolással bizonyíthatók az $\mathscr{E}'\colon 2^X \otimes 2^X \to 2^\varepsilon$ leképezés következő tulajdonságai, amelyeket a továbbiakban gyakran felhasználunk:

Ha $S \subseteq X$ és $T = X$, akkor $\mathscr{E}'(S|T) = \mathscr{E}(S)$.
Ha $T \subseteq S \subseteq X$, akkor $\mathscr{E}'(S|T) = \mathscr{E}'(T|T)$.
Ha $S \subset S' \subseteq X$ és $T \subseteq X$, akkor $\mathscr{E}'(S|T) \subseteq \mathscr{E}'(S'|T)$.
Ha $S \subseteq X$ és $T \subset T' \subseteq X$, akkor $\mathscr{E}'(S|T) \subseteq \mathscr{E}'(S|T')$.

A gráfelméleti szakirodalomban széleskörűen használt az élhalmaz által generált rész-hipergráf és a ponthalmaz által generált alhipergráf fogalma.

2.5. Definíció: A $H = (X; \varepsilon)$ hipergráfnak az $\mathscr{F}$ élhalmaz $(\mathscr{F} \subseteq \varepsilon)$ által *generált rész-hipergráfja:* $H = \big(\mathscr{H}(\mathscr{F}); \mathscr{F}\big)$.

2.6. Definíció: A $H = (X; \varepsilon)$ hipergráfnak az S ponthalmaz $(S \subseteq X)$ által *generált alhipergráfja:* $H = (S; \varepsilon_S)$, ahol

$$\varepsilon_S = \{E_i \cap S | E_i \in \mathscr{E}(S)\}.$$

A generált rész-hipergráf vagy a generált alhipergráf fogalom alkalmazása további munkánkat indokolatlanul elbonyolítaná. Ugyanis a kvázi-komponensek tulajdonságainak leírásához, és a megkeresésükre szolgáló eljáráshoz is olyan rész-hipergráf fogalom szükséges, amelyet ponthalmaz segítségével definiálunk. A rész-hipergráfnak mindazokat és csak azokat az éleket kell tartalmaznia, amelyek a definiáló ponthalmaznak részei.
Ezeket a követelményeket elégíti ki a következő definíció.

2.7. Definíció: A $H = (X; \varepsilon)$ hipergráfnak az S ponthalmaz $(S \subseteq X)$ felhasználásával *kifeszített rész-hipergráfja:*

$$H = \big(\mathscr{H}(\mathscr{E}'(S|S);\ \mathscr{E}'(S|S)\big).$$

Nyilvánvaló, hogy az $(S;\ \mathscr{E}'(S|S)$ pár nem lett volna jó definíció, ugyanis az S halmaznak lehet olyan pontja, amelyet egyetlen él sem tartalmaz. Ezzel szemben $H = \big(\mathscr{H}(\mathscr{E}'(S|S));\ \mathscr{E}'(S|S)\big)$ valóban hipergráf (nincs üres éle és izolált pontja), de ponthalmaza nem feltétlenül egyezik meg S-sel.

2.3. Megjegyzés: Könnyen belátható, hogy $S \subseteq X$ esetén $\mathscr{H}\big(\mathscr{E}'(S|S)\big) \subseteq S$.
A továbbiakban jelöljük a $H = \big(\mathscr{H}(\mathscr{E}'(S|S));\ \mathscr{E}'(S|S)\big)$ hipergráfot röviden H_S-sel.

2.4. Megjegyzés: A H_S kifeszített rész-hipergráf megegyezik az $\mathscr{E}'(S|S)$ élhalmaz által generált rész-hipergráffal.

Mivel a dolgozat következő részeiben csak ponthalmazok által kifeszített rész-hipergráfokkal dolgozunk, ezért ezeket a továbbiakban röviden csak rész hipergráfoknak nevezzük.

2.8. Definíció: A H_S rész-hipergráfban a K ponthalmaz $[K \subseteq \mathscr{H}(\mathscr{E}'(S|S))]$ *hipergráfot kifeszítő* (vagy röviden kifeszítő), ha $\mathscr{H}\big(\mathscr{E}'(K|K)\big) = K$.

Az elnevezést indokolja az, hogy a kifeszítő ponthalmaz megegyezik a felhasználásával kifeszített rész-hipergráf ponthalmazával, azaz $K = \mathscr{H}\big(\mathscr{E}'(K|K)\big)$ miatt $H_K = \big(K;\ \mathscr{E}'(K|K)\big)$.

2.1. Lemma: A H_S rész hipergráfban a K ponthalmaz $\big(K \subseteq \mathscr{H}(\mathscr{E}'(S|S))\big)$, akkor és csak akkor kifeszítő, ha $\exists \mathscr{F}\big(\mathscr{F} \subseteq \mathscr{E}'(S|S)\big)$ élhalmaz, amelyre $K = \mathscr{H}(\mathscr{F})$.

2.5. Megjegyzés: A 2.1. Lemma egyszerű következménye az, hogy egy K ponthalmaz $(K \subseteq X)$ vagy kifeszítő minden olyan H_S rész-hipergráfban, amelyre $K = \mathscr{H}\big(\mathscr{E}'(S|S)\big)$, vagy egyikben sem kifeszítő.

Ez indokolja, hogy a továbbiakban csak kifeszítő ponthalmazról fogunk beszélni (nem tesszük hozzá, hogy a $H = (X;\ \varepsilon)$ hipergráf melyik rész-hipergráfjában), és a K-val jelölt ponthalmazok mindig kifeszítők lesznek.

2.2. Lemma: Ha $S \subseteq X$, akkor $\exists K$ kifeszítő ponthalmaz $\big(K = \mathscr{H}(\mathscr{E}'(S|S))\big)$, amelyre $H_S = H_K$. Az $\mathscr{H}\big(\mathscr{E}'(S|S)\big)$ halmaz az S által tartalmazott maximális kifeszítő ponthalmaz.

2.6. Megjegyzés: A 2.2. Lemma egyszerű következménye, hogy az általánosság megszorítása nélkül feltehetjük bármelyik kifeszített rész-hipergráfról, hogy az egy kifeszítő ponthalmaz felhasználásával keletkezett.

Hasonlóan az $\mathscr{H}\big(\mathscr{E}'(S|S)\big)$ és az S halmaz közötti reláció elemzéséhez (amely a kifeszítő halmaz fogalmának bevezetését eredményezte), $K \supseteq S$ esetén az $\mathscr{H}\big(\mathscr{E}'(S|K)\big)$ és az S halmaz kapcsolatának vizsgálata vezet el a komponens fogalmához. Jelöljük az S halmaz $(S \subseteq X)$ valódi részhalmazainak osztályát $\mathscr{S}_S$-sel:

$$\mathscr{S}_S = \{T|T \neq \emptyset,\ T \subset S\} = 2^S - \{S\} - \{\emptyset\}$$

2.9. Definíció: A H_K rész-hipergráfban a P ponthalmaz $(O \neq P \subseteq K)$ *komponens*, ha

(1) $$\mathscr{H}\big(\mathscr{E}'(P|K)\big) = P,$$

(2) $$T \in \mathscr{S}_P \text{ esetén } \mathscr{H}\big(\mathscr{E}'(T|K)\big) \supset T$$

2.7. Megjegyzés: Ha a H_K rész-hipergráfban a P ponthalmaz komponens, akkor kifeszítő. $[\mathscr{E}'(P|K) = \mathscr{F}$ választással a 2.1. Lemma alkalmazásával adódik.]

2.8. Megjegyzés: A 2.7. Megjegyzés miatt $\mathscr{H}\big(\mathscr{E}'(P|P')\big) = P$. Ezért a 2.2. Lemma alkalmazásával egyszerűen belátható, hogy ha a P halmaz kompo-

nense a H_K rész-hipergráfnak azaz $\mathcal{H}(\mathcal{E}'(P|K)) = P$, akkor $P \subseteq K' \subset K$ esetén $\mathcal{H}(\mathcal{E}'(P|K')) = P$, azaz P komponense a $H_{K'}$ rész-hipergráfnak is.

2.10. Definíció: A H_K rész-hipergráf összefüggő, ha $T \in \mathcal{S}_K$ esetén

$$\mathcal{H}(\mathcal{E}'(T|K)) \supset T.$$

A komponens és az összefüggő rész-hipergráf definíciója már mutatja, hogy hasznos volt a hipergráf ponthalmazai és élhalmazai között értelmezett leképzések bevezetése. Ugyanis a fenti definíciók nyilvánvalóan megegyeznek a szakirodalomban használt definíciókkal, amelyek az út fogalom bevezetése miatt bonyolultabbak.

A kvázi-komponens definíciójához és a kvázi-komponenseket meghatározó eljáráshoz egyaránt szükséges a vágás most következő definíciója.

2.11. Definíció: A H_K rész-hipergráfnak a T *ponthalmaz* ($T \subseteq K$) *által generált vágása:* [jelölése $C_K(T)$].

$$C_K(T) = \mathcal{E}'(T|K) \cap \mathcal{E}'((K - T)|K).$$

Nyilvánvaló, hogy a T és a $K-T$ halmazok által generált vágások megegyeznek, és hogy az üres halmaz és a K által generált vágás mindig az üres halmaz.

2.3. Lemma: $C_K(T) = \mathcal{E}'(T|K) - \mathcal{E}'(T|T)$.

2.4. Lemma: Legyen T ($T \subseteq K$) tetszőleges ponthalmaza a H_K rész-hipergráfnak.
$\mathcal{H}(\mathcal{E}'(T|K)) = T$ akkor és csak akkor, ha $C_K(T) = \emptyset$.

2.5. Lemma: A P ponthalmaz ($P \subseteq K$) akkor és csak akkor komponense a H_K rész-hipergráfnak, ha

(1′) $$C_K(P) = \emptyset,$$

(2′) $$T \in \mathcal{S}_P \text{ esetén } C_K(T) \supset \emptyset.$$

A komponens itt közölt alternatív definíciójának általánosításán alapul a kvázi-komponens definíciója. További vizsgálatainkhoz értelmezzük a hipergráf élein a $v(E_j) > 0$ ($j = 1, \ldots m$) pozitív függvényt. Terjesszük ki a függvény értelmezési tartományát élhalmazokra is. Vezessük be a $w: 2^{\varepsilon} \to R^+$ leképezést, amelynek révén még diszjunkt élhalmazokat is össze tudunk hasonlítani.

2.12. Definíció: Tetszőleges $\mathcal{F}$ élhalmaz ($\mathcal{F} \subseteq \varepsilon$) esetén $w(\mathcal{F}) = \sum_{E_j \in \mathcal{F}} v(E_j)$.

2.9. Megjegyzés: Egyszerű számolással adódnak a $w: 2^{\varepsilon} \to R^+$ függvény következő tulajdonságai, amelyeket a továbbiakban gyakran felhasználunk:

Ha $\mathcal{F} \subseteq \mathcal{G} \subseteq \varepsilon$, akkor $w(\mathcal{F}) < w(\mathcal{G})$.
Ha $\mathcal{F} \subseteq \mathcal{G} \subseteq \varepsilon$, akkor $w(\mathcal{G} - \mathcal{F}) = w(\mathcal{G}) - w(\mathcal{F})$.
Ha $\mathcal{F} \subseteq \varepsilon, \mathcal{G} \subseteq \varepsilon$, akkor $w(\mathcal{F} \cup \mathcal{G}) = w(\mathcal{F}) + w(\mathcal{G}) - w(\mathcal{F} \cap \mathcal{G})$.

Az élhalmazokon értelmezett függvény módot ad a vágások értékének definiálására is.

2.13. Definíció: A H_K rész-hipergráf T ponthalmaza ($T \subseteq K$) által generált vágásának *értéke* [jelölése: $\overline{w}_K(T)$]:

$$\overline{w}_K(T) = w[C_K(T)] = w[\mathcal{E}'(T|K) \cap \mathcal{E}'\big((K - T)|K\big)].$$

A kvázi-komponens definíciója a komponens utolsó definíciójának általánosítása.

2.14. Definíció: A H_K rész-hipergráfban a Q ponthalmaz ($\emptyset \neq Q \subseteq K$) *kvázi-komponens*, ha bármely $T \in \mathcal{S}_Q$ választása esetén $\overline{w}_K(Q) < \overline{w}_K(T)$.

2.10. Megjegyzés: A kvázi-komponens fogalom valóban a komponens fogalmának általánosítása, ugyanis 2.5. Lemma felhasználásával triviális, hogy a H_K rész-hipergráf valamennyi komponense egyben kvázi-komponense is.

2.11. Megjegyzés: A H_K rész hipergráfban a Q ponthalmaz ($Q \subseteq K$) kvázi-komponens, ha $|Q| = 1$, ugyanis ez esetben $\mathcal{S}_Q = \emptyset$.

Az egy elemű kvázi-komponenseket a továbbiakban *triviális* kvázi-komponenseknek nevezzük.

2.6. Lemma: Ha a H_K rész-hipergráfban a Q ponthalmaz nem triviális kvázi-komponens ($|Q| \geq 2$, $Q \subseteq K$), akkor Q kifeszítő.

A 2.6. Lemma egyszerű, de a továbbiak során gyakran felhasznált következményét ismerteti a következő megjegyzés.

2.12. Megjegyzés: Ha a Q ponthalmaz ($Q \subseteq K$) nem triviális kvázi-komponens H_K-ban, akkor legalább egy élt tartalmaz, és $x_i \in Q$ esetén $\exists E_j \in \mathcal{E}'(Q|Q)$, amelyre $x_i \in E_j$ (Ugyanis a Q ponthalmaz kifeszítő.)

A következő tétel a kvázi-komponenseknek a cluster elemzés szempontjából nagyon fontos tulajdonságát fejezi ki. Azt mondja ki, hogy a nem-triviális kvázi-komponens bármely valódi részhalmaza „erősebben kapcsolódik" a kvázi-komponens többi részéhez, mint a kvázi-komponens teljes környezetéhez.

2.7. Tétel: A H_K rész-hipergráfban a Q ponthalmaz ($Q \subseteq K$) akkor és csak akkor nem-triviális kvázi-komponens, ha bármely $T \in \mathcal{S}_Q$ ponthalmaz választása esetén:

$$w[\mathcal{E}'(T|Q)] > w[\mathcal{E}'\big(T|(K-(Q-T))\big)].$$

2.13. Megjegyzés: A 2.7. Tétel tovább nem élesíthető, ugyanis $T = Q$ választása esetén a 2.2. és 2.9. Megjegyzések felhasználásával triviálisan adódik, hogy $w[\mathcal{E}'(Q|Q)] \leq w[\mathcal{E}'(Q|K)]$.

2.14. Megjegyzés. A 2.7. Tétel felhasználásával egyszerű számolással adódik a kvázi-komponensek egyik fontos tulajdonsága: Ha a Q halmaz kvázi-komponense a H_K rész-hipergráfnak, akkor $Q \subseteq K' \subset K$ esetén kvázi-komponense a $H_{K'}$ rész-hipergráfnak is.

Ugyanis a triviális kvázi-komponensekre az állítás nyilvánvaló, a nem-triviálisra pedig a 2.2. és 2.9. Megjegyzést felhasználva a következő adódik:

$$w\,\mathcal{E}'(T|Q)] > w[\mathcal{E}'(T|(K-(Q-T)))] \geq w[\mathcal{E}'(T|(K'-(Q-T)))].$$

2.15. Megjegyzés: Ha Q nem-triviális kvázi-komponense a H_K rész-hipergráfnak, akkor a H_Q rész-hipergráf összefüggő. Ugyanis a 2.6. Lemma alapján Q kifeszítő. 2.14. Megjegyzést $Q = K'$-re alkalmazva $w[\mathcal{E}'(T|Q)] > w[\mathcal{E}'(T|T)]$ adódik tetszőleges $T \in \mathcal{S}_Q$ esetén. Ez pedig a 2.4. Lemma alapján pontosan azt jelenti, hogy H_Q összefüggő.

A most következő tétel alapvető jelentőségű a hipergráf összes kvázi-komponense meghatározására szolgáló hatékony algoritmus konstruálásához.

2.8. Tétel: Legyen $K(|K|) \geq 2$) a $H = (X; \varepsilon)$ hipergráfnak egy kifeszítő ponthalmaza, és S olyan ponthalmaz, amelyre teljesül az, hogy $S \subseteq K$ és $|S| \geq 2$. Legyen T^* az a ponthalmaz ($T^* \in \mathcal{S}_S = \{T|T \neq \emptyset; T \subset S\}$), amelyre teljesül az, hogy bármely $T \in \mathcal{S}_S$ esetén $\overline{w}_K(T^*) \leq \overline{w}_K(T)$.

Legyen $Q(Q \subset S)$ tetszőleges kvázi-komponense a $H = (X; \varepsilon)$ hipergráfnak. Ekkor $Q \subseteq T^*$ vagy $Q \subseteq S - T^*$ teljesül.

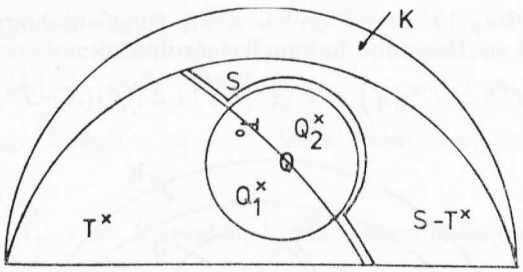
Biz.: Ha Q triviális kvázi-komponens, akkor a tétel állítása triviális.

Tegyük fel, hogy létezik olyan $Q(Q \subset S)$ nem-triviális kvázi-komponense a $H = (X; \varepsilon)$ hipergráfnak, amelyre $Q_1^* = Q \cap T^* \neq \emptyset$ és $Q_2^* = Q \cap (S - T^*) \neq \emptyset$.

Mivel $Q \subset S$, ezért $Q \supseteq T^*$ és $Q \supseteq S - T^*$ egyszerre nem teljesülhet.

1. Eset: Tegyük fel, hogy $Q \supseteq S - T^*$. Legyen $T^0 = T^* \cup Q = T^* \cup Q_2^*$. $T^* \subseteq T^0$ miatt $T^0 \neq \emptyset$ és $Q \supseteq S - T^*$ miatt $T^0 \subset S$, tehát $T^0 \in \mathcal{S}_S$.

Számítsuk ki $\overline{w}_K(T^0)$ értékét $\overline{w}_K(T^*)$ függvényében: $\overline{w}_K(T^0) = w[\mathcal{E}'(T^0|K)] - w[\mathcal{E}'(T^0|T^0)]$.



1. ábra

Az $\mathcal{E}'(T^0|K)$ halmazba tartozó élek attól függően, hogy tartalmaznak-e T^*-beli pontot, két diszjunkt halmazba sorolhatók:

$$\mathcal{E}'((T^* \cup Q_2^*)|K) = \mathcal{E}'(T^*|K) \cup \mathcal{E}'(Q^*|(K - T^*)).$$

Ebből a halmaz egyenlőségből a következő skalár egyenlőség adódik:

$$(1) \qquad w[\mathcal{E}'((T^* \cup Q_2^*)|K)] = w[\mathcal{E}'(T^*|K)] + w[\mathcal{E}'(Q^*|(K-T^*))].$$

Az $\mathcal{E}'(T^0|T^0)$ halmazba tartozó élek attól függően, hogy tartalmaznak-e Q_2^*-beli pontot két diszjunkt halmazba sorolhatók:

$$\mathcal{E}'((T^* \cup Q_2^*)|(T^* \cup Q_2^*)) = \mathcal{E}'(T^*|T^*) \cup \mathcal{E}'(Q_2^*|(T^* \cup Q_2^*)).$$

Ebből a halmaz egyenlőségből a következő skalár egyenlőség adódik.

$$(2) \qquad w[\mathcal{E}'(T^0|T^0)] = w[\mathcal{E}'(T^*|T^*)] + w[\mathcal{E}'(Q_2^*|(T^* \cup Q_2^*))].$$

A (2) egyenlőséget az (1) egyenlőségből kivonva kapjuk, hogy $\overline{w}_K(T^0) = \overline{w}_K(T^*) + w[\mathcal{E}'(Q_2^*|(K-T^*))] - w[\mathcal{E}'(Q_2^*|(T^* \cup Q_2^*))]$.
$\overline{w}_K(T^*) \leq \overline{w}_K(T^0)$ miatt:
$w[\mathcal{E}'(Q_2^*|(K-T^*))] \geq w[\mathcal{E}'(Q_2^*|(T^* \cup Q_2^*))]$
$K-T^* \subseteq K-(Q-Q_2^*)$ és $Q \subseteq T^* \cup Q_2^*$ felhasználásával.

(3) $w[\mathcal{E}'(Q_2^*|(K-(Q-Q_2^*)))] \geq w[\mathcal{E}'(Q_2^*|Q)]$ adódik. Mivel $Q_1^* \neq \emptyset$ és $Q_2^* \neq \emptyset$, tehát $Q_2^* \in \mathcal{S}_Q$.
Az 5. Tétel következménye miatt Q kvázi-komponense a $H_K = (K; \mathcal{E}'(K|K))$ rész-hipergráfnak is. A 2.7. Tétel miatt ekkor Q_2^*-re teljesülnie kell a következő egyenlőtlenségnek:

$$(4) \qquad w[\mathcal{E}'(Q_2^*|(K-(Q-Q_2^*)))] < w[\mathcal{E}'(Q_2^*|Q)].$$

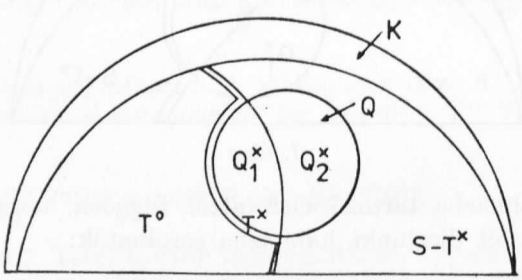
amely ellentmond a (3) egyenlőtlenségnek.

2. Eset: Tegyük fel, hogy $Q \not\supseteq T^*$. Legyen $T^0 = T^* - Q = T^* - Q_1^*$, azaz $T^* = T^0 \cup Q_1^*$. $T^0 \subseteq T^*$ miatt $T^0 \subset S$, és $Q \not\supseteq T^*$ miatt $T^0 \neq \emptyset$, tehát $T^0 \in \mathcal{S}_S$.
Számítsuk ki $\overline{w}_K(T^*)$ értékét $\overline{w}_K(T^0)$ függvényében:

$$\overline{w}_K(T^*) = w[\mathcal{E}'(T^*|K)] - w[\mathcal{E}'(T^*|T^*)].$$

Az $\mathcal{E}'(T^*|K)$ halmazba tartozó élek attól függően, hogy tartalmaznak-e T^0-beli pontot két diszjunkt halmazba sorolhatók:

$$\mathcal{E}'((T^0 \cup Q_1^*)|K) = \mathcal{E}'(T^0|K) \cup \mathcal{E}'(Q_1^*|(K-T^0)).$$



2. ábra

Ebből a halmaz egyenlőségből a következő skalár egyenlőség adódik:

$$w[\mathcal{E}'(T^*|K)] = w[\mathcal{E}'(T^0|K)] + w[\mathcal{E}'\big(Q_1^*|(K-T^0)\big)]. \tag{5}$$

Az $\mathcal{E}'(T^*|T^*)$ halmazba tartozó élek attól függően, hogy tartalmaznak-e Q_1^*-beli pontot két diszjunkt halmazba sorolhatók:

$$\mathcal{E}'\big((T^0 \cup Q_1^*)|(T^0 \cup Q_1^*)\big) = \mathcal{E}'(T^0|T^0) \cup \mathcal{E}'\big(Q_1^*|(T^0 \cup Q_1^*)\big)$$

Ebből a halmaz egyenlőségből a következő skalár egyenlőség adódik:

$$w[\mathcal{E}'(T^*|T^*)] = w[\mathcal{E}'(T^0|T^0)] + w[\mathcal{E}'\big(Q_1^*|(T^0 \cup Q_1^*)\big)]. \tag{6}$$

A (6) egyenlőséget az (5) egyenlőségből kivonva kapjuk, hogy $\overline{w}_K(T^*) = \overline{w}_K(T^0) + w[\mathcal{E}'\big(Q_1^*|(K-T^0)\big)] - w[\mathcal{E}'\big(Q_1^*|(T^0 \cup Q_1^*)\big)]$.

$\overline{w}_K(T^*) \leq \overline{w}_K(T^0)$ miatt

$w[\mathcal{E}'\big(Q_1^*|(K-T^0)\big)] \leq w[\mathcal{E}'\big(Q_1^*|(T^0 \cup Q_1^*)\big)]$.

$Q \subseteq K-T^0$ és $T^0 \cup Q_1^* \subseteq K-(Q-Q_1^*)$ felhasználásával.

$$w[\mathcal{E}'(Q_1^*|Q)] \leq w[\mathcal{E}'\big(Q_1^*|(K-(Q-Q_1^*))\big)] \text{ adódik.} \tag{7}$$

Mivel $Q_1^* \neq \emptyset$ és $Q_2^* \neq \emptyset$, tehát $Q_1^* \in \mathcal{S}_Q$.

Az 5. Tétel következménye miatt Q kvázi-komponense a $H_K = \big(K; \mathcal{E}'(K|K)\big)$ rész-hipergráfnak is. A 2.7. Tétel miatt ekkor Q_1^*-ra teljesülnie kell a következő egyenlőtlenségnek:

$$w[\mathcal{E}'(Q_1^*|Q)] > w[\mathcal{E}'\big(Q_1^*|(K-(Q-Q_1^*))\big)], \tag{8}$$

amely ellentmond a (7) egyenlőtlenségnek. Q.E.D.

A következő tétel a hipergráf kvázi-komponensei közötti relációra mutat rá. Azt fejezi ki, jogy a kvázi-komponensek vagy tartalmazzák egymást vagy diszjunktak.

2.11. Tétel: Legyenek a $Q(Q \subseteq K)$ és a $Q'(Q' \subseteq K)$ ponthalmazok egymástól különböző kvázi-komponensek a H_K rész-hipergráfban. Ekkor vagy $Q \subset Q'$, $Q' \subset Q$, vagy $Q \cap Q' = \emptyset$ teljesül.

A most következő tétel a hipergráf kvázi-komponenseinek számára ad felső korlátot.

2.12. Tétel: A H_K rész hipergráf kvázi-komponenseinek száma legfeljebb $2|K| - 1$.

A tétel állítása egyébként a fa struktúrájú partíciók jól ismert tulajdonsága.

A következő két lemma már nem a kvázi-komponensek tulajdonságainak feltárására szolgál, hanem a dolgozat 3. részében szereplő algoritmus szerkesztéséhez szükséges.

2.13. Lemma: Legyen adott a $H = (X; \varepsilon)$ hipergráfnak a $K(|K| \geq 2, K \subseteq X)$ kifeszítő ponthalmaza és az S halmaz, amelyre teljesül az, hogy $|S| \geq 2$ és $S \subseteq K$. Legyen $T^* \in \mathfrak{H}_S$ az a ponthalmaz, amelyre teljesül az, hogy $\overline{w}_K(T^*) \leq \overline{w}_K(T)$ tetszőleges $T \in \mathfrak{H}_S$ halmaz választása esetén.

Az S halmaz akkor és csak akkor kvázi-komponens a H_K rész-hipergráfban, ha $\overline{w}_K(S) < \overline{w}_K(T^*)$.

2.14. Lemma: Legyen adott a $H = (X; \varepsilon)$ hipergráfnak a K $(K \subseteq X, |K| \geq 2)$ kifeszítő ponthalmaza és az S ponthalmaz, amelyre teljesül az, hogy $S \subseteq X$ és $|S| \geq 2$. Legyen $T^* \in \mathfrak{H}_S$ az a ponthalmaz, amelyre teljesül az, hogy $\overline{w}_K(T^*) \leq \overline{w}_K(T)$ tetszőleges $T \in \mathfrak{H}_S$ halmaz választása esetén.

Ha $\mathcal{H}(\mathcal{E}'(T^*|T^*)) = \emptyset$, akkor T^* csak triviális kvázi-komponenseket tartalmaz.

Ha $\mathcal{H}(\mathcal{E}'(T^*|T^*)) \neq \emptyset$ és a P ponthalmaz komponense a H_{T^*} rész-hipergráfnak, akkor

$$C_K(T^*) = C_K(\mathcal{H}(\mathcal{E}'(T^*|T^*))) = C_K(P),$$

azaz

$$w_K(T^*) = w_K(\mathcal{H}(\mathcal{E}'(T^*|T^*))) = w_K(P).$$

A kvázi-komponens definíciója és bizonyított tulajdonságai lehetővé teszik a dolgozat első részében említett három modell bármelyikére sikerrel alkalmazható új *cluster definíció* bevezetését:

2.15. Definíció: Az objektumok clusterjei az 1. Modell gráfjának, ill. a 3. Modell hipergráfjának kvázi-komponensei, a deszkriptorok clusterjei a 2. Modell hipergráfjának kvázi-komponensei.

Az így definiált clusterek legfontosabb tulajdonságai a következők:

1. Ha a Q halmaz clustere az R objektum vagy deszkriptor halmaznak, akkor clustere R minden olyan részhalmazának is, amely tartalmazza a Q halmazt. (2.14. Megjegyzés.)
2. Ha a Q halmaz clustere az R halmaznak, akkor bármely $T \in \mathfrak{H}_Q$ részhalmaz erősebben kapcsolódik a Q clusterhez, mint annak környezetéhez (2.7. Tétel).
3. A clusterek vagy diszjunktak vagy tartalmazzák egymást (2.11. Tétel).
4. Az R halmaz clustereinek száma kisebb, mint a rendszer elemeinek számának kétszerese (2.12. Tétel).

A kvázi-komponensek megkeresésére szolgáló eljáráson alapuló új cluster technika

A hipergráf összes kvázi-komponensének meghatározására szolgáló eljárásnak két alapvető rutinja van.

R1: Egy hipergráf komponenseinek meghatározására szolgáló rutin.
R2: Egy hipergráf „minimális két részre vágása" meghatározására szolgáló rutin.

Az R1 rutin feladata a következő:

Adott a $H = (X; \varepsilon)$ hipergráf, és az élein értelmezett $v(E_j) > 0$ $(j = 1, \ldots m)$

függvény. Adott továbbá a $K (K \subseteq X)$ kifeszítő ponthalmaz, és az általa kifeszített $H_K = (K; \mathcal{E}'(K|K))$ rész-hipergráf.

Határozzuk meg a H_K rész-hipergráf összes komponensét, azaz azokat a $P (0 \neq P \subseteq K \subseteq X)$ ponthalmazokat, amelyekre teljesül az, hogy $\overline{w}_K(P) = 0$, és bármely $T \in \mathcal{S}_K = \{T | \emptyset \neq T, T \subset K\}$ esetén $\overline{w}_K(T) > 0$.

A feladat elvégzésére az irodalomban több algoritmus is ismert. Elterjedtek az indexelési technikán alapuló (*Klafszky* [17]) és az ekvivalencia osztályt generáló (*Knuth* [18]) eljárások is.

Az R1 rutin az indexelési technikán alapul és $O(|K| \cdot |\mathcal{E}'(K|K)|)$ lépésben határozza meg a $H_K = \left(K; \mathcal{E}'(K|K)\right)$ hipergráf összes komponensét.

Az R2 rutin feladata a következő:

Adott a $H = (X; \varepsilon)$ hipergráf ($|X| \geq 2$), és az élein értelmezett $v(E_j) > 0$ $(j = 1, \ldots m)$ függvény. Adott S ($|S| \geq 2$ és $S \subseteq X$) ponthalmaz. Határozzuk meg azt a $T^* \in \mathcal{S}_S = \{T | T \neq \emptyset\, T \subset S\}$ ponthalmazt, amelyre teljesül az, hogy $w(T^*) \leq w(T)$ bármely $T \in \mathcal{S}_S$ esetén.

A feladat elvégzése céljából az irodalomban ismertetett eljárások a hipergráf feladatot gráf problémára vezetik vissza, majd Ford – Fulkerson algoritmusával keresik a megoldást (*Lawler* [22]). Ezzel szemben az R2 rutin közvetlenül hipergráfon dolgozik és a maximális folyam problémánál egyszerűbb kereslet-kínálat feladat megoldásán alapul. $O(|X|^3 \cdot |\varepsilon|^2)$ lépésben határozza meg a $H = (X; \varepsilon)$ hipergráf minimális vágását. (*Futó* [13].)

Elemezzük eredeti problémánkat; egy hipergráf összes kvázi-komponensének meghatározására szolgáló eljárás megszerkesztését:

Ha a hipergráf nem összefüggő, akkor kvázi-komponensei a komponensei által kifeszített összefüggő rész-hipergráfoknak is kvázi-komponensei, és az összefüggő rész-hipergráfok kvázi-komponenseinek meghatározásával az eredeti hipergráf valamennyi kvázi-komponensét megkapjuk. (2.14. Megjegyzés, 2.15. Megjegyzés, Futó [13].)

Tehát elegendő olyan algoritmust konstruálni, amely egy összefüggő hipergráf kvázi-komponenseit keresi meg.

Feladat

Adott a $H = (X; \varepsilon)$ *összefüggő* hipergráf és $v(E_j) > 0$ $(i = 1, \ldots m)$ a hipergráf élein értelmezett pozitív függvény. Határozzuk meg a $H = (X; \varepsilon)$ hipergráf összes kvázi-komponensét, azaz azokat a $Q \subseteq X$ ponthalmazokat, amelyekre teljesül az, hogy bármely T esetén $(\emptyset \neq T \subset Q)$, $\overline{w}(Q) < \overline{w}(T)$.

Algoritmus

Legyen $\mathcal{K}_j = \{K_j^{(l_j, h_j)}, \ldots, K_{\nu_j}^{(l_{\nu j}, h_{\nu j})}\}$,

az eljárás j-edik lépése előtt azon $K_i^{(l_i, h_i)}$ ponthalmazok rendezett osztálya, melynek elemeiről a j-edik és a további lépések során kell eldönteni, hogy kvázi-komponensek-e.

Legyen $Q_j = \{Q_1, Q_2, \ldots Q_{tj}\}$,

az eljárás j-edik lépése előtt ismert vagy meghatározott kvázi-komponensek halmaza.

0. lépés

$$\mathcal{H}_0 = \{X\},\ \mathcal{Q}_0 = \{\{x_1\}, \ldots \{x_n\}\}.$$

Ha $|X| = 1$, akkor a 2.11. Tétel értelmében $H = (X; \varepsilon)$-nek csak 1 kvázi-komponense van, amelyet azonban már a 0. lépés előtt ismertünk, tehát az eljárás véget ért. Tehát

$$\mathcal{Q}_1 = \mathcal{Q}_0 = \{X\},\ \mathcal{H}_1 = \emptyset.$$

Ha $|X| \geq 2$, akkor mivel a $H = (X; \varepsilon)$ hipergráf összefüggő, ezért $|X| \geq 2$ miatt az X halmaz nem-triviális kvázi-komponens, tehát $X \in \mathcal{Q}_1$, és így $\mathcal{Q}_1 = \{[x_1], \ldots \{x_n\}, X\}$.

Ha $|X| = 2$, akkor az algoritmus véget ér, azaz $\mathcal{H}_1 = \emptyset$, mivel a $H = (X; \varepsilon)$ összes kvázi komponensét meghatároztuk, hiszen a 2.11. Tétel szerint ezek száma legfeljebb 3, és $Q_1 = \{X, \{x_1\}, \{x_2\}\}$.

Ha $|X| > 2$, akkor az R2 rutin segítségével $S_0 = X$ választás mellett meghatározzuk azt a $T_0^* \in \mathcal{S}_{S_0}$ halmazt, amelyre teljesül az, hogy bármely $T \in \mathcal{S}_{S_0}$ halmaz választása esetén $\overline{w}(T_0^*) \leq \overline{w}(T)$.

Jelölje a továbbiakban $\overline{w}(T_0^*)$-ot w_{01} és $\overline{w}(S_0 - T_0^*)$-ot w_{02}. A 2.11. Definíció miatt $w_{01} = w_{02}$, ezért w_{02} is már ismert.

A $H = (X; \varepsilon)$ hipergráf valamennyi X-től különböző kvázi-komponensét vagy T_0^* vagy $X - T_0^*$ tartalmazza (2.8. Tétel).

A nem triviális kvázi-komponenseket (amelyek megkeresése a célunk, ugyanis a triviálisak már a 0. lépés előtt ismertek voltak) a 2.11. Megjegyzés miatt $\mathcal{H}(\mathcal{E}'(T_0^*|T_0^*))$ vagy $\mathcal{H}(\mathcal{E}'((S_0 - T_0^*)|(S_0 - T_0^*)))$ is tartalmazza.

Ha $\mathcal{H}(\mathcal{E}'((S_0 - T_0^*)|(S_0 - T_0^*))) = \emptyset$ és $\mathcal{H}(\mathcal{E}'(T_0^*|T_0^*)) = \emptyset$, akkor T_0^* és $S_0 - T$ csak triviális kvázi-komponenseker tartalmaz, amelyek már a 0. lépés előtt ismertek voltak. Tehát az alogirtmus véget ér. $\mathcal{H}_1 = \emptyset$.

Ha $\mathcal{H}(\mathcal{E}'(T_0^*|T_0^*)) \neq \emptyset$ vagy $\mathcal{H}(\mathcal{E}'((S_0 - T_0^*)|(S_0 - T_0^*))) \neq \emptyset$, akkor a $H = (X; \varepsilon)$ hipergráf nem triviális kvázi-komponensei $H_{T_0^*}$ vagy a $H_{S_0 - T_0^*}$ rész-hipergráfoknak is kvázi-komponensei, sőt ezeket $H_{T_0^*}$ vagy $H_{S_0 - T_0^*}$ komponensei is tartalmazzák (2.14. Megjegyzés).

Határozzuk meg az R1 rutin segítségével $H_{T_0^*}$ és $H_{S_0 - T_0^*}$ komponenseit. (Megjegyzés: az R2 rutin alkalmazása esetén $H_{T_0^*}$ összefüggő lesz.) Jelölje ezeket $K_1^{(0,h_1)}, \ldots K_{v_1}^{(0,h_{v_1})}$.

A felső indexben első helyen álló 0 arra utal, hogy ezek a komponensek a 0. lépésben keletkeztek. A $h_i = 1$ érték esetén a $K_i^{(0,h_i)}$ komponenst T_0^*, a $h_i = 2$ érték esetén a $K_i^{(0,h_i)}$ komponenst az $X - T_0^*$ halmaz tartalmazza.

A 2.14. Lemma következtében $\overline{w}(T_0^*) = \overline{w}(X - T_0^*) = w_{01} = w_{02} = \overline{w}(K_1^{(0,h_1)}) = \ldots = \overline{w}(K_{v_1}^{(0,h_{v_1})})$.

Legyen $\mathcal{H}_1 = \{K_1^{(0,h_1)} \ldots K_{v_1}^{(0,h_{v_1})}\}$.

j. lépés:

($j \geq 1$)

$$\mathcal{H}_j = \{K_j^{(l_j,h_j)} \ldots K_{v_j}^{(l_{vj},h_{vj})}\},$$

$$\mathcal{Q}_j = \{Q_1, Q_2, \ldots Q_{tj}\}.$$

Ha $\mathcal{H}_j = \emptyset$, akkor az eljárás már a $j-1$. lépésben véget ért, ugyanis nincs több olyan ponthalmaz, amely kvázi-komponens lehet.

Ez esetben a Q_j halmaz tartalmazza a $H = (X; \varepsilon)$ hipergráf valamennyi kvázi-komponensét.

Ha $\mathfrak{K}_j \neq \emptyset$, akkor válasszuk ki a $\mathfrak{K}_j$ rendezett halmaz soron következő elemét: $K_j^{(l_j,h_j)}$-t.

Ha $|K_j^{(l_j,h_j)}| = 1$, akkor $K_j^{(l_j,h_j)}$ triviális kvázi-komponens, tehát már a 0. lépés előtt ismert volt. $K_j^{(l_j,h_j)}$ nyilván más kvázi-komponenst már nem tartalmazhat.

Tehát ez esetben: $\mathfrak{K}_{j+1} = \mathfrak{K}_j - \{K_j^{(l_j,h_j)}\}$, $\mathfrak{Q}_{j+1} = \mathfrak{Q}_j$.

A $K_j^{(l_j,h_j)}$ halmaz által generált vágás értékét már keletkezésekor az l_j. lépésben meghatároztuk ($l_j < j$). Ugyanis $h_j = 1$ esetben $K_j^{(l_j,h_j)} \subseteq T_{lj}^*$, $h_j \geq 2$ esetben $K_j^{(l_j,h_j)} \subseteq S_{l_j} - T_{lj}^*$ tartalmazási relációk állnak fenn, és a 2.13. Lemma miatt $\overline{w}(K_j^{(l_j,hj)}) = w_{l_j,h_j}$. A w_{l_j,h_j} értékét már az l_j. lépésben kiszámoltuk.

Ha $|K_j^{(l_j,h_j)}| = 2$, akkor meghatározzuk $K_j^{(l_j,h_j)}$ mindkét elemére (legyenek ezek x_{j1} és x_{j2}) a $\overline{w}(x_{j1})$ és $\overline{w}(x_{j2})$ értékeket. A kvázi-komponens definíciója szerint:

- ha $w_{l_j,h_j} < \overline{w}(x_{j1})$ és $w_{l_j,h_j} < \overline{w}(x_{j2})$, akkor $K_j^{(l_j,h_j)}$ kvázi-komponens, tehát $\mathfrak{Q}_{j+1} = \mathfrak{Q}_j \cup \{K_j^{(l_j,h_j)}\}$ és $\mathfrak{K}_{j+1} = \mathfrak{K}_j - K_j^{(l_j,h_j)}$.
- míg, ha a fenti két egyenlőtlenségnek legalább az egyike nem igaz, akkor $K_j^{(l_j,h_j)}$ nem kvázi-komponens, tehát $\mathfrak{Q}_{j+1} = \mathfrak{Q}_j$ és $\mathfrak{K}_{j+1} = \mathfrak{K}_j - \{K_j^{(l_j,h_j)}\}$.

Ha $K_j^{(l_j,h_j)} > 2$, akkor $S_j = K_j^{(l_j,h_j)}$ választás mellett az R2 rutin segítségével meghatározzuk azt a $T_j^* \in \mathfrak{S}_{sj}$ halmazt, amelyre teljesül az, hogy bármely $T \in \mathfrak{S}_{sj}$ esetén $\overline{w}(T_j^*) \leq \overline{w}(T)$.

Jelöljük a továbbiakban $\overline{w}(T_j^*)$-ot w_{j1}-gyel.

Azt, hogy $S_j = K_j^{(l_j,h_j)}$ kvázi-komponens volt-e, az l_j. lépésben kiszámított w_{l_j,h_j} és a j. lépésben meghatározott w_{j1} összehasonlításával döntjük el.

Ha $w_{l_j,h_j} < w_{j1}$, akkor a $K_j^{(l_j,h_j)}$ halmaz kvázi-komponens és így $\mathfrak{Q}_{j+1} = \mathfrak{Q}_j \cup K_j^{(l_j,h_j)}$.

Ha $w_{l_j,h_j} \geq w_{j1}$, akkor a 2.13. Lemma miatt a $K_j^{(l_j,h_j)}$ halmaz nem kvázi-komponens és így $\mathfrak{Q}_{j+1} = \mathfrak{Q}_j$.

A $H = (X; \varepsilon)$ hipergráf összes – $S_j = K_j^{(l_j,h_j)}$ által valódi részként tartalmazott – kvázi-komponensét vagy T_j^* vagy $S_j - T_j^*$ tartalmazza.

Ezek közül a nem-triviálisakat a 2.14. Megjegyzés miatt

$$\mathfrak{K}\big(\mathfrak{S}'((S_j - T_j^*)|(S_j - T_j^*))\big) \text{ vagy } \mathfrak{K}\big(\mathfrak{S}'(T_j^*|T_j^*)\big)$$

is tartalmazza.

Ha $\mathfrak{K}\big(\mathfrak{S}'((S_j - T_j^*)|(S_j - T_j^*))\big) = \emptyset$ és $\mathfrak{K}\big(\mathfrak{S}'(T_j^*|T_j^*)\big) = \emptyset$, akkor $S_j - T_j^*$ és T_j^* csak triviális kvázi-komponenseket tartalmaz (2.14. Lemma), amelyek már a 0. lépés előtt ismertek voltak.

Így ez esetben $\mathfrak{K}_{j+1} = \mathfrak{K}_j - \{K_j^{(l_j,h_j)}\}$.

Ha $\mathfrak{K}\big(\mathfrak{S}'((S_l - T_j^*)|(S_j - T_j^*))\big) \neq \emptyset$ vagy $\mathfrak{K}\big(\mathfrak{S}'(T_j^*|T_j^*)\big) \neq \emptyset$, akkor a $H = (X; \varepsilon)$ hipergráfnak az S_j által tartalmazott nem-triviális kvázi-komponensei a $H_{T_j^*}$ vagy a $H_{S_j - T_j^*}$ rész-hipergráfnak is kvázi-komponensei, sőt ezeket $H_{S_j - T_j^*}$ vagy $H_{T_j^*}$ komponensei is tartalmazzák (2.14. Megjegyzés).

Határozzuk meg az R1 rutin segítségével H_{T^*} és $H_{S_j - T_j^*}$ komponenseit. Jelölje $H_{T_j^*}$ komponenseit $K_{v_j+1}^{(j,1)}, \ldots K_{v_j+i_j}^{(j,1)}$.

A 2.14. Lemma miatt $w_{j1} = \overline{w}(K_{v_j+1}^{(j,1)}) = \ldots = \overline{w}(K_{v_j+i_j}^{(j,1)})$.

Jelölje $H_{S_j - T_j^*}$ komponenseit $K_{v_j+i_j+1}^{(j,2)}, \ldots K_{v_j+r_j}^{(j,r_j-i_j+1)})$, $(v_j + r_j = v_{j+1})$ $(r_j - i_j + 1 = h_{v_{j+1}})$. A $H = (X; \varepsilon)$ hipergráfnak a $H_{S_j-T_j^*}$ rész-hipergráf komponensei által generált vágásainak az értékeit meghatározzuk. Jelölje ezeket rendre $w_{j,2}, \ldots w_{j,r_j-i_j+1}$.

Általában $w_{j,1} \neq w_{j,2} \neq \ldots \neq w_{j,r_j-i_j+1}$.

Legyen ez esetben

$$\mathcal{K}_{j+1} = \mathcal{K}_j - \{K_j^{(l_j,h_j)}\} \cup \{K_{v_j+1}^{(j,1)}, \ldots K_{v_{j+1}}^{(j,r_j-i_j+1)}\} = \{K_{j+1}^{(l_{j+1},h_{j+1})}, \ldots K_{v_{j+1}}^{(j,h_{v_{j+1}})}\}.$$

3.1. Tétel: A Feladat megoldására, azaz az összefüggő $H = (X; \varepsilon)$ hipergráf összes kvázi-komponensének meghatározására szolgáló algoritmus legfeljebb $|X| - 1$ számú lépésben véget ér, ha $|X| \geq 2$, és az utolsó $r \cdot (r \leq |X| - 2)$ lépésben kapott Q_{r+1} halmaz tartalmazza a $H = (X; \varepsilon)$ hipergráf összes kvázi-komponensét.

Biz. 1: $|X| = n$ szerinti teljes indukcióval

a) $n = 2$: Ekkor az eljárás már a 0. lépésben véget ért ($\mathcal{K}_1 = \emptyset$), tehát $r = 0 = n - 2 = 0$ teljesül.

b) $n > 2$: Tehát $|X| \geq 3$. Végezzük el az algoritmus 0. lépését.

Ha $\mathcal{H}\big(\mathcal{E}'(T_0^*|T_0^*)\big) = \emptyset$ és $\mathcal{H}\big(\mathcal{E}'(X - T_0^*)|(X - T_0^*)\big) = \emptyset$, akkor az eljárás már a 0. lépésben véget ér, tehát $r = 0 \leq n - 2 \geq 1$.

Ha $\mathcal{H}\big(\mathcal{E}'(T_0^*|T_0^*)\big) \neq \emptyset$ vagy $\mathcal{H}\big(\mathcal{E}'((X - T_0^*)|(X - T_0^*))\big) \neq \emptyset$ akkor jelölje a kapott komponensek $(K_1^{(0,h_1)}, \ldots K_{v_1}^{(0,h_{v_1})})$ elemszámát $n_1, n_2, \ldots n_{v_0}$. Nyilván $n \geq \sum_{i=1}^{v_0} n_i$.

Tudjuk, hogy az X által tartalmazott kvázi-komponensek, a fenti komponensek által kifeszített rész-hipergráfoknak is kvázi-komponensei (2.14. Megjegyzés).

Indukciós feltevésünk értelmében a komponensek által kifeszített hipergráfok kvázi-komponenseit legfeljebb $n_1 - 1, \ldots n_v - 1$ számú lépésben határozhatjuk meg.

Tehát az algoritmus lépésszáma legfeljebb $\sum_{i=1}^{v_0} (n_i - 1) + 1$, ugyanis a 0. lépést már elvégeztük.

Ha $\sum_{i=1}^{v_0} n_i = n$, akkor $v_0 \geq 2$, mivel ez esetben $\mathcal{H}\big(\mathcal{E}'(T_0^*|T_0^*)\big) \neq \emptyset$ és $\mathcal{H}\big(\mathcal{E}'((X - T_0^*)|(X - T_0^*))\big) \neq \emptyset$ kellett legyen.

Ha $v_0 = 1$, akkor $\sum_{i=1}^{1} n_i = n_1 < n$, mivel ez esetben vagy $\mathcal{H}\big(\mathcal{E}'(T_0^*|T_0^*)\big) = \emptyset$, vagy $\mathcal{H}\big(\mathcal{E}'((X - T_0^*)|(X - T_0^*))\big) = \emptyset$ teljesült. Tehát

$$\sum_{i=1}^{v_0} (n_i - 1) + 1 = \sum_{i=1}^{v_0} n_i - v_0 + 1 \geq 2 + 1 = n - 1.$$

Tehát az algoritmus lépésszáma legfeljebb $n - 1$.

Biz. 2: Legyen a Q $(\emptyset \neq Q \subseteq X)$ halmaz a $H = (X; \varepsilon)$ hipergráfnak tetszőleges kvázi-komponense.

Ha Q triviális kvázi-komponens, akkor már $\mathcal{Q}_0$ is tartalmazta és $\mathcal{Q}_j \subseteq \mathcal{Q}_{j+1}$ $(j = 0, \ldots r)$ miatt $Q \in \mathcal{Q}_{r+1}$ is teljesül.

Ha Q nem-triviális kvázi-komponens és $Q = X$ akkor már az algoritmus 0. lépésében ismert, azaz $X \in \mathcal{Q}_1$ és így $\mathcal{Q}_j \subseteq \mathcal{Q}_{j+1}$, $(j = 0, \ldots r)$ miatt $X \subset \mathcal{Q}_{r+1}$.

Ha Q nem-triviális kvázi-komponens, és $Q \neq X$, akkor jelölje $K_j^{(l_j,h_j)}$ a $\bigcup_{j=0}^{r} \mathcal{H}_j$ elemei közül azt a legszűkebb halmazt, amely Q-t tartalmazza.

Ilyen legszűkebb halmaz biztos van, mert $X \subset \mathcal{H}_0$ miatt $X \subset \bigcup_{j=0}^{r} \mathcal{H}_j$.

Jelölje $\mathcal{H}_j$ azt a rendezett halmazt, amelynek $K_j^{(l_j,h_j)}$ az első eleme.

Ha $Q \subset K_j^{(l_j,h_j)}$, akkor a 2.14. Megjegyzés és a 2.8. Tétel miatt az eljárás j. lépésében keletkező komponensek valamelyike tartalmazza Q-t. Ez viszont ellentmond annak, hogy $K_j^{(l_j,h_j)}$ volt az a legszűkebb eleme $\bigcup_{j=0}^{r} \mathcal{H}_j$-nek, amely Q-t tartalmazta.

Tehát $Q = K_j^{(l_j,h_j)}$. Ez pedig azt jelenti, hogy a j. lépésben $K_j^{(l_j,h_j)} \in \mathcal{Q}_{j+1}$ lesz. De $\mathcal{Q}_j \geq \mathcal{Q}_{j+1}$ $(j = 0, \ldots r)$ miatt $Q \in \mathcal{Q}_{r+1}$. Q.E.D.

Mivel az R1 rutin lépésszáma $O(|X| \cdot |\varepsilon|)$ és az R2 rutin lépésszáma $O(|X|^3 \cdot |\varepsilon|^2)$, ezért a 2.12. Tétel alapján állíthatjuk, hogy a hipergráf összes kvázi-komponensének meghatározására szolgáló eljárás lépésszáma $O(|X|^4 \cdot |\varepsilon|^2)$.

A kvázi-komponensek meghatározására szolgáló eljárás a 2. részben bevezetett cluster definíció alapján egy új *hierarchikus cluster technika* lesz, amelynek legfontosabb jellemzői a következők:

1. Az R (objektum vagy deszkriptor) halmaz minden egyes eleme legalább egy clusternek is eleme (2.11. Megjegyzés).
2. Az eljárás konvergens (3.1. Tétel).
3. Az eljárás lépésszáma felülről becsülhető az objektumok és a deszkriptorok számának polinom alakú függvényével (3.1. Tétel).

A cluster elemzés hipergráf és gráf modelljeinek felhasználásával jelenleg folyamatban van az új cluster technika programozása FORTRAN nyelven R20 és TPA/i számítógépekre.

4. Az új cluster modellek és technika alkalmazási lehetőségei

A dolgozat első részében bevezett gráf és hipergráf modellek alkalmasak bonyolult rendszerek szerkezetének leírására. A hipergráf modellek alkalmazása különösen az információtudományi cluster analízis területén nagy jelentőségű. A probléma teljesen kézenfekvő modelljéül szolgálnak, és lehetővé teszik a hasonlósági mérőszámok definiálásának, kiszámolásának elkerülését.

A matematikai-statisztikai cluster elemzés problémáinak modellezésére mind a gráf, mind a hipergráf modellek alkalmasak. Itt azonban már nem kerülhető el vagy a hasonlósági mérőszámok, vagy a deszkriptorok definiálása, amely óhatatlanul a modell torzulására vezet.

A hipergráf vagy gráf kvázi-komponensének fogalmán alapuló cluster definíció és eljárás mindhárom modellre alkalmazható, tehát a klasszikus gráf

modell esetén is egy új lehetőséget nyújt a rendszer struktúrájának feltárására. Igazi jelentősége azonban a hipergráf modellek felhasználásánál látható. Lehetővé teszi olyan objektum rendszerek szerkezetének felderítését is, amelynél az objektumok jellemzésére tárgyszavakat és számadatokat is használnak. A cluster analízissel foglalkozó szakemberek által jól ismert tény, hogy a vizsgálandó rendszerek nagy hányada ilyen tulajdonságú.

Például az orvostudomány területén a csopotosítási probléma megoldása kulcs szerepet játszik a differenciál diagnosztikában, ahol a betegségeket a tünetek és a vizsgálatok eredményei alapján kell csoportosítani, és az analitikus epidemiológiában, ahol a betegségek csoportosítását külső és belső környezeti hatások alapján kell elvégezni. Nyilvánvaló, hogy a tüneteket vagy környezeti hatásokat leíró tárgyszavak kódolása, majd a hasonlósági mérőszámok kidolgozása igen erős torzulásokat eredményez. Hasonló a helyzet a közgazdasági és információtudományi problémák nagy hányadánál is.

A következőkben részletesebben bemutatjuk az új cluster modellek és technika alkalmazási lehetőségeit a kutatásirányítás területén. A választást indokolja egyrészt az, hogy a kutatásirányítás az utóbbi években a figyelem középpontjába került, másrészt, de nem utolsósorban az, hogy a szerző ezen a területen dolgozik, és az itt felmerült problémák sarkallták a cluster analízis mélyebb megismerésére és új módszer kidolgozására. Az Építéstudományi Intézetben dolgozó szűkebb kollektíva közel egy évtizedes kutató és fejlesztő munkájának eredménye – a tudományelmélet, információtudomány és az operációkutatás módszereinek és eredményeinek együttes felhasználásán alapuló LOGEL (*log*ikai *el*járások, vagy angolul: *log*ical mod*el*) tematikai kutatásirányítási módszer, amelynek lényege vázlatosan a következőkben foglalható össze.

1. A koordinált indexelés alapelvének felhasználásával *tárgyszavak* (deszkriptorok) halmazaira képzi le a vizsgált kutatási programok, témák, témajavaslatok tartalmát, módszerét, célját (esetleg ráfordításait, várható eredményeit is). Ezek a tárgyszavak lehetőség szerint egy ellenőrzött, szinonimáktól (különböző szóképek, azonos jelentés) és homonimáktól (azonos szókép, több jelentés) mentes, állandóan bővülő – általában hierarchikus szerkezetű – tárgyszórendszer, az ún. tezaurusz elemei. (A tezaurusz építése sok esetben a kutatási témák tárgyszavazásával párhuzamosan folyik.)
2. A kutatási témák, programok, vagy tudományos tételek, hipotézisek, és az ezeket a különböző szempontok szerint leíró tárgyszavak, valamint mindezek rendszerei közötti kapcsolatokat a LOGEL elmélet logikai modeljeinek felhasználásával írja le, amelyek irányított vagy irányítatlan gráfok, illetve hipergráfok.
3. A különböző kutatásirányítási problémák megoldását a logikai modellek bizonyos részhalmazainak különböző szempontú és mélységű *elemzésével* teszi lehetővé. Az elemzés révén, amely gráfelméleti és matematikai-statisztikai módszerekkel történik, az irányítandó kutatás rendszerről átfogó kép nyerhető.

   A *gráfelméleti* elemzés célja a logikai modellek elemei kapcsolatainak feltárása, míg az elemek eloszlásfüggvényeinek vizsgálata a LOGEL módszer fontos részét képező *deszkriptorstatisztika* feladata.

4. A kutatásirányítási *akciók modellezését* a logikai modellként szolgáló gráfok strukturális változtatásával (pl. bővítésével, szűkítésével, vágásával) segíti elő.

   A gyakorlatban elérni kívánt megoldásoknak a gráfokon strukturális optimumkritériumokat feleltetünk meg. Ezek tényleges eltérését *operációkutatási* algoritmusok teszik lehetővé.
5. A logikai modellek szerkesztését, elemzését, az operációkutatási algoritmusok futtatását az irányítandó kutatási programok vagy témák nagyobb száma esetén a LOGEL módszer számítógépes programrendszere teszi lehetővé, amelynek programjai FORTAN nyelven eddig CDC-3300, SIEMENS 4004/45 és TPA/i számítógépekre dolgozták ki. A fentiekből nyilvánvalóan adódik, hogy a cluster elemzés hipergráf modelljei és az új cluster technika eredményesen alkalmazható a kutatásirányítás területén.

A kutatás helyzetelemzése során lehetővé teszi mind a kutatási témarendszer, mind az indexelésükre felhasznált tárgyszavak clusterjeinek meghatározásával a rendszerek tematikai gócpontjainak és periferikus területeinek meghatározását, ill. extrapolációs módszer felhasználásával a tematikai centrumok elhelyezkedésének előrejelzését.

A kutatások koordinálása során a kutatási célprogramok, irányprogramok kidolgozásánál nyilván ezek magvai a helyzetelemzésnél meghatározott clusterek lesznek.

Az új cluster technika másodlagos alkalmazása az hogy az R2 minimális vágási rutin jól felhasználható kutatási témák optimális csoportosítására, célprogramok, irányprogramok konkrét kialakítására is.

A LOGEL módszeren alapul jelenleg az Építéstudományi Intézet, az ÉVM és az országos számítástechnikai kutatási célprogram kutatásirányítási rendszere. Az új cluster technika számítógépes programjainak elkészülte után 1978 elejétől megkezdjük annak folyamatos alkalmazását mindhárom területen. A számítástechnikai és alkalmazási tapasztalatokról külön cikkben számolunk be 1978 végén.

(Beérkezett: 1977. július 5-én.)

IRODALOMJEGYZÉK

1. Adamson, G. W. – Boreham, J.: The Use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles; Inform. Stor. Retr., Vol. 10. (253 – 260), 1974.
2. Anderberg, M. R.: Cluster Analysis for Applications; Academic Press, 1973.
3. Augustson, J. G. – Minker, J.: An Analysis of Some Graph Theoretical Cluster Techniques; Journal of A.C.M., Vol. 17. (571 – 588), 1970.
4. Balas, E. – Padberg, M.: On the Set Covering Problem: II. An Algorithm for Set Partitioning; Op. Res., No. 23. (74 – 90), 1975.
5. Benedikt, V. – Kelemen, K. – Pintér, Zs. – Vári, P.-né: Cluster analízis és lényegkiemelő eljárás-rendszer terve; SZÁMKI, 1976.
6. Berge, C.: Graphs and Hypergraphs; North Holland/American Elsevier, 1973.
7. Boulton, P. M. – Wallace, C. S.: An Information Measure for Single Link Classification; Comp. Journ., Vol. 18. (236 – 238), 1975.
8. Diday, E. – Schroeder, A.: A New Approach in Mixed Distributions Detection; IRIA, Rapport de Recherche, No. 52, 1974.
9. Edmonds, J. – Karp, R.: Theoretical Improvements in Algorithmic Efficiency for Network Flow Problems; Journal of A.C.M., Vol. 19. (248 – 264), 1972.
10. Forgy, E. W.: Classification so as to Relate to Outside Variables; Final Report,

Conf. Cluster Analysis of Multivariate Data (13.01 – 13.12), Cath. Univ. America, 1966.
11. Fritz, J.: Tanuló algoritmusok alkalmazása az alakfelismerésben; MTA MKI, 1975.
12. Futó, P.: Computer Aided Management of Industrial Research – the Method LOGEL; 12. Progr. Op. Res. (353 – 371), North Holland, 1976.
13. Futó, P.: Hipergráf elméleten alapuló új cluster definíció és technika; Alk. Mat. Lapok (Sajtó alatt).
14. Futóné Szántó, Zs.: Számítástechnika az egészségügyért. ESZTIK, Budapest, 1976.
15. Gower, J. C.: A Comparison of Some Methods of Cluster Analysis; Biometrika, Vol. 23 (623 – 637), 1967.
16. Johnson, S. C.: Hierarchical Clustering Schemes; Psychometr., Vol. 32. (241 – 254), 1976.
17. Klafszky, E.: Hálózati folyamok; Bólyai J. Mat. Társ., 1969.
18. Knuth, D. E.: The Art of Computer Programming; Vol. 1. Fundamental Algorithms, Addison – Wesley, 1968.
19. Kovács, L. B.: A diszkrét programozás kombinatorikus módszerei; Bólyai J. Mat. Társ., 1969.
20. Kunszt, Gy.: A tudományos kutatás logikai modellezése és tematikai irányítása; Akadémiai Kiadó, 1975.
21. Lawler, E. L.: Cutsets and Partitions of Hypergraphs; Networks, Vol. 3. (275 – 286), 1973.
22. Lawler, E. L.: Algorithms, Graphs and Complexity; Networks, Vol. 5. (89 – 92), 1975.
23. Luccio, F. – Sami, M.: On the Decomposition of Networks into Minimally Interconnected Networks; IEEE Trans. Circuit Theory, CT 16. (184 – 188), 1969.
24. MacQueen, J. B.: Some Methods for Classification and Analysis of Multivariate Observations; Proc. Symp. Math. Stat. and Prob., Vol. 1. (281 – 297), 1967.
25. Mulligan, G. B. – Corneil, P. G.: Corrections to Bierstone's Algorithm for Generating Clique; Journal of A.C.M., Vol. 19. (244 – 247), 1972.
26. Osteen, R. E.: Clique Detection Algorithms Based on Line Addition and Line Removal; SIAM Journal Appl. Math., Vol. 26. (126 – 135), 1974.
27. Sibson, R.: SLINK – An Optimally Efficient Algorithm for the Single-link Cluster Method; Comp. Journ., Vol. 16. (30 – 34), 1973.
28. Sparck-Jones, K.: Automatic Indexing "74"; Comp. Lab., Univ. of Cambridge, 1974.
29. Srejder, Ju. A.: Egyenlőség, hasonlóság, rendezés; Gondolat, 1975.
30. Tanimoto, T. T.: An Elementary Mathematical Theory of Classification and Prediction; IBM, 1958.
31. Ward, J. H.: Hierarchical Grouping to Optimize an Objective Funtion; Journ. Amer. Statist. Assoc., Vol. 58. (236 – 244), 1963.
32. Wishart, D.: An Algorithm for Hierarchical Classifications; Biometr., Vol. 22. (165 – 170), 1969.

A NEW MODEL AND ALGORITHM OF CLUSTER ANALYSIS

The first part of the paper presents the hypergraph model of cluster analysis, which enables us to eliminate the clumsy procedure of constructing the similarity matrix. The presentation of the new, hypergraph-based definition and its characteristics (part two) is followed by the detailed description of a new non-agglomerative, hierarchic cluster algorithm (part three). The fourth part of the paper deals with the application possibilities of the new cluster technique.

НОВАЯ МОДЕЛЬ КЛАСТЕРНОГО АНАЛИЗА И ЕЕ АЛГОРИТМ

В первой части работы рассматривается кластерный анализ на модели гиперграфа, посредством использования которой можно обойти несколько сложный метод разработки матрицы подобия. После описания понятия кластера, базирующегося на модели гиперграфа и его свойств (вторая часть) следует детальное изложение нового и неагломеративного по своему характеру иерархического кластерного алгоритма (третья часть). В четвертой части работы рассматриваются возможности применения новой кластерной техники.