שאלה 3:

<u>נתחיל מפרגמנטציה אופקית על הרלציה הנתונה:</u>

לפי השאילתות הנתונות, נקבע את קבוצת הפרדיקטים:

$$Pr = \left\{ \begin{array}{ll} p_1 : \ NetWorth > 5 & \forall 1 \leqslant i \leqslant |distinct(DMA)|, p_{3_i} : \ \forall y \in DMA \ s.t \ DMA = y \\ p_2 : \ householdSize > 2 & \forall 1 \leqslant j \leqslant |distinct(zipCode)|, p_{4_j} : \ \forall x \in zipCode \ s.t \ zipCode = x \end{array} \right\}$$

נפעיל את אלגוריתם COM-MIN על הקבוצה

$$Pr = \{p_1, p_2, \ \forall 1 \leqslant i \leqslant |distinct(DMA)| : \ p_{3_i}, \forall 1 \leqslant j \leqslant |distinct(zipCode)| : \ p_{4_j}\}:$$

$$Pr' = \{p_1\} \ ; \ Pr = \{p_2, \ \forall 1 \leqslant i \leqslant |distinct(DMA)| : \ p_{3_i}, \forall 1 \leqslant j \leqslant |distinct(zipCode)| : \ p_{4_j}\}$$

נקבע פרגמנט בהתאם לפרדיקט: $f_1 = \sigma_{p1} MediaData \Longrightarrow F = \{f_1\}$

p_1 רלוונטי לפרגמנטציה ונמשיך לפרדיקט הבא:

$$Pr' = \{p_1, p_2\} \ ; \ Pr = \{ \ \forall 1 \leqslant i \leqslant |distinct(DMA)| : \ p_{3_i}, \forall 1 \leqslant j \leqslant |distinct(zipCode)| : \ p_{4_j}\}$$

נקבע פרגמנט בהתאם לפרדיקט: $f_2 = \sigma_{p2} MediaData \Longrightarrow F = \{f_1, f_2\}$

p_2 רלוונטי לפרגמנטציה ונמשיך לפרדיקטים הבאים:

$$Pr' = \{p_1, p_2, \forall 1 \leqslant i \leqslant |distinct(DMA)| : \ p_{3_i}\} \ ; \ Pr = \{\forall 1 \leqslant j \leqslant |distinct(zipCode)| : \ p_{4_j}\}$$

נקבע פרגמנטים בהתאם לפרדיקטים:

$$\forall 1 \leqslant i \leqslant |distinct(DMA)| : \ f_{3_i} = \sigma_{p_{3_i}} MediaData$$
$$\Longrightarrow F = \{f_1, f_2, \forall 1 \leqslant i \leqslant |distinct(DMA)| : f_{3_i}\}$$

$\forall 1 \leqslant i \leqslant |distinct(DMA)| : \ p_{3_i}$ רלוונטיים לפרגמנטציה ונמשיך לפרדיקטים הבאים:

$$Pr' = \{p_1, p_2, \forall 1 \leqslant i \leqslant |distinct(DMA)| : \ p_{3_i}, \forall 1 \leqslant j \leqslant |distinct(zipCode)| : \ p_{4_j}\} \ ; \ Pr = \{\}$$

נקבע פרגמנטים בהתאם לפרדיקטים:

$$\forall 1 \leqslant j \leqslant |distinct(zipCode)| : \ f_{4_j} = \sigma_{p_{4_j}} MediaData$$
$$\Longrightarrow F = \{f_1, f_2, \forall 1 \leqslant i \leqslant |distinct(DMA)| : \ f_{3_i}, \forall 1 \leqslant j \leqslant |distinct(zipCode)| : \ f_{4_j}\}$$

$\forall 1 \leqslant j \leqslant |distinct(zipCode)| : \ p_{4_j}$ רלוונטיים לפרגמנטציה,

לכן סיימנו את ריצת האלגוריתם COM-MIN וקיבלנו:

$$Pr' = \left\{ \begin{array}{l|l} p_1 : \ NetWorth > 5 & \forall 1 \leqslant i \leqslant |distinct(DMA)|, p_{3_i} : \ \forall y \in DMA \ s.t \ DMA = y \\ \hline p_2 : \ householdSize > 2 & \forall 1 \leqslant j \leqslant |distinct(zipCode)|, p_{4_j} : \ \forall x \in zipCode \ s.t \ zipCode = x \end{array} \right\}$$

נסמן: $m = |distinct(DMA)|$, $n = |distinct(zipCode)|$. נוסיף לפרדיקטים p_1, p_2 את המשלימים שלהם $p_1' = NetWorth \leqslant 5, \ p_2' = householdSize \leqslant 2$ ואז הקבוצה M תהיה מורכבת מכל זוג של שני פרדיקטים, למעט זוגות פרדיקטים שבהם שני הפרדיקטים שייכים לאותו אטריביוט. לכן, גודל הקבוצה M יהיה:

$|M| = 4*2 + 4m + 4n + m*n$.

ולכן הפרגמנטים הסופיים הם:

$$MediaData_1 = \sigma_{(NetWorth>5)\wedge(householdSize>2)} MediaData$$
$$MediaData_2 = \sigma_{(NetWorth>5)\wedge(householdSize\leqslant 2)} MediaData$$
$$MediaData_3 = \sigma_{(NetWorth\leqslant 5)\wedge(householdSize>2)} MediaData$$
$$MediaData_4 = \sigma_{(NetWorth\leqslant 5)\wedge(householdSize\leqslant 2)} MediaData$$
$$\forall 1 \leqslant i \leqslant m:\ MediaData_{1i} = \sigma_{(NetWorth>5)\wedge(y_i\in DMA\ s.t\ DMA=y_i)} MediaData$$
$$\forall 1 \leqslant i \leqslant m:\ MediaData_{2i} = \sigma_{(NetWorth\leqslant 5)\wedge(y_i\in DMA\ s.t\ DMA=y_i)} MediaData$$
$$\forall 1 \leqslant i \leqslant m:\ MediaData_{3i} = \sigma_{(householdSize>2)\wedge(y_i\in DMA\ s.t\ DMA=y_i)} MediaData$$
$$\forall 1 \leqslant i \leqslant m:\ MediaData_{4i} = \sigma_{(householdSize\leqslant 2)\wedge(y_i\in DMA\ s.t\ DMA=y_i)} MediaData$$
$$\forall 1 \leqslant j \leqslant n:\ MediaData_{5j} = \sigma_{(NetWorth>5)\wedge(x_j\in zipCode\ s.t\ zipCode=x_j)} MediaData$$
$$\forall 1 \leqslant j \leqslant n:\ MediaData_{6j} = \sigma_{(NetWorth\leqslant 5)\wedge(x_j\in zipCode\ s.t\ zipCode=x_j)} MediaData$$
$$\forall 1 \leqslant j \leqslant n:\ MediaData_{7j} = \sigma_{(householdSize>2)\wedge(x_j\in zipCode\ s.t\ zipCode=x_j)} MediaData$$
$$\forall 1 \leqslant j \leqslant n:\ MediaData_{8j} = \sigma_{(householdSize\leqslant 2)\wedge(x_j\in zipCode\ s.t\ zipCode=x_j)} MediaData$$
$$\forall 1 \leqslant j \leqslant nm:\ MediaData_{9j} = \sigma_{(y_i\in DMA\ s.t\ DMA=y_i)\wedge(x_j\in zipCode\ s.t\ zipCode=x_j)} MediaData$$

<u>נבצע פרגמנטציה אנכית על הרלציה המקורית:</u>

על מנת לחשב את מטריצת ה-attribute affinity נעקוב אחרי השלבים באלגוריתם שלמדנו בהרצאה:

- לפי הנתון, נבנה את מטריצת הגישה לאתרים:

$$\left(\begin{array}{c|cccc} & S1 & S2 & S3 & S4 \\ q1 & 60 & 60 & 60 & 60 \\ q2 & 20 & 20 & 20 & 20 \\ q3 & 20 & 20 & 20 & 20 \end{array}\right)$$

- נמצא את מטריצת ה-use: לפי הגדרה, $use(q_i, A_j) = \begin{cases} 1 & , A_j\ is\ referenced\ by\ q_i \\ 0 & , o.w \end{cases}$.

לכן המטריצה תיראה כך:

* לא הכנסנו את כל השדות- פירוט בהמשך

$$use = \left(\begin{array}{c|ccccc} & NetWorth & NumOfAdults & householdSize & DMA & zipCode \\ q1 & 1 & 1 & 1 & 0 & 0 \\ q2 & 1 & 0 & 0 & 1 & 0 \\ q3 & 1 & 0 & 0 & 1 & 1 \end{array}\right)$$

נחשב את ערך ה-Attribute Affinity של כל שני שדות ונסדר במטריצת ה-Attribute Affinity:

כל ערך במטריצת ה-affinity יחושב לפי הנוסחה הבאה, כאשר A_i, A_j הם שני אטריביוטים ברלציה שלנו:

$$aff(A_i, A_j) = \sum_{k|use(q_k,A_i)=1 \wedge use(q_k,A_j)=1} \sum_{\forall S_l} ref_l(q_k) acc_l(q_k)$$

לדוגמה עבור השדה DMA החישוב יבוצע כך:

$$aff(DMA, NetWorth) = 4*1*20 + 4*1*20 = 160$$
$$aff(DMA, NumOfAdults) = 0$$
$$aff(DMA, householdSize) = 0$$
$$aff(DMA, DMA) = 4*1*20 + 4*1*20 = 160$$
$$aff(DMA, zipCode) = 4*1*20 = 80$$

ובאופן דומה עבור כל שדה, למעט השדות HHID, deviceID, GreenLiving.

* השדות HHID, deviceID הם מפתחות ראשיים ברלציה MediaData ולכן חייבים להימצא בכל אחד מהפרגמנטים, כלומר נוסיף אותם ידנית לכל פרגמנט שיווצר.

נשים לב כי בשדות GreenLiving ו-deviceID לא נעשה שימוש באף אחת מהשאילתות ולכן לא נתחשב בהם בחלוקה.

המטריצה תיראה כך:

$$aff = \left(\begin{array}{c|ccccc} & NetWorth & NumOfAdults & householdSize & DMA & zipCode \\ NetWorth & 400 & 240 & 240 & 160 & 80 \\ NumOfAdults & 240 & 240 & 240 & 0 & 0 \\ householdSize & 240 & 240 & 240 & 0 & 0 \\ DMA & 160 & 0 & 0 & 160 & 80 \\ zipCode & 80 & 0 & 0 & 80 & 80 \end{array} \right)$$

נסמן את האטריביוטים הכתובים בשורה במטריצה aff בהתאמה משמאל לימין ב-A_i עבור $1 \leqslant i \leqslant 5$.

נרצה להפעיל את אלגוריתם Bond Energy על מנת למצוא את מטריצת ה-CA. לצורך כך, נבחר בחירה אקראית ששני השדות ההתחלתיים שלנו יהיו $A_1 = NetWorth$ ו-$A_2 = NumOfAdults$, וכעת על מנת להוסיף את האטריביוט $A_3 = householdSize$ לחלוקה נבדוק את קשרי ה-bond בין כל שני אטריביוטים עבור כל אחת מהאפשרויות למיקום האטריביוט השלישי.

הוספת A_3 משמאל ל-A_1:

A_3, A_1, A_2 :

$$bond(A_0, A_3) = 0$$
$$\begin{aligned} bond(A_3, A_1) = {} & aff(A_1, A_3)*aff(A_1, A_1) + aff(A_2, A_3)*aff(A_2, A_1) + \\ & aff(A_3, A_3)*aff(A_3, A_1) + aff(A_4, A_3)*aff(A_4, A_1) + \\ & aff(A_5, A_3)*aff(A_5, A_1) \\ & = 240*400 + 240*240 + 240*240 + 0*160 + 0*80 = 211,200 \end{aligned}$$
$$bond(A_0, A_1) = 0$$
$$cont(A_0, A_3, A_1) = 2bond(A_0, A_3) + 2bond(A_3, A_1) - 2bond(A_0, A_1) = 422,400$$

הוספת A_3 בין A_1 ל- A_2:

$\mathbf{A_1, A_3, A_2}$:

$bond(A_1, A_3) = 211,200$

$$\begin{aligned} bond(A_3, A_2) = {} & aff(A_1, A_3)*aff(A_1, A_2) + aff(A_2, A_3)*aff(A_2, A_2) + \\ & aff(A_3, A_3)*aff(A_3, A_2) + aff(A_4, A_3)*aff(A_4, A_2) + \\ & aff(A_5, A_3)*aff(A_5, A_2) \\ & = 240*240 + 240*240 + 240*240 + 0*0 + 0*0 = 172,800 \end{aligned}$$

$bond(A_1, A_2) = 0$

$cont(A_1, A_3, A_2) = 2bond(A_1, A_3) + 2bond(A_3, A_2) - 2bond(A_1, A_2) = 768,000$

הוספת A_3 מימין ל- A_2:

$\mathbf{A_1, A_2, A_3}$:

$bond(A_2, A_3) = 172,800$

$bond(A_3, A_4') = 0$

$bond(A_2, A_4') = 0$

$cont(A_2, A_3, A_4') = 2bond(A_2, A_3) + 2bond(A_3, A_4') - 2bond(A_2, A_4') = 345,600$

לכן נבחר להוסיף את $\mathbf{A_3}$ בין $\mathbf{A_1}$ ל- $\mathbf{A_2}$. והמצב הנוכחי הוא: $\mathbf{A_1, A_3, A_2}$

נבדוק לאן להוסיף את האטריביוט $A_4 = DMA$:

הוספת A_4 משמאל ל-A_1:

$\mathbf{A_4, A_1, A_3, A_2}$:

$bond(A_0, A_4) = 0$

$$\begin{aligned} bond(A_4, A_1) = {} & aff(A_1, A_4)*aff(A_1, A_1) + aff(A_2, A_4)*aff(A_2, A_1) + \\ & aff(A_3, A_4)*aff(A_3, A_1) + aff(A_4, A_4)*aff(A_4, A_1) + \\ & aff(A_5, A_4)*aff(A_5, A_1) \\ & = 160*400 + 0*240 + 0*240 + 160*160 + 80*80 = 96,000 \end{aligned}$$

$bond(A_0, A_1) = 0$

$cont(A_0, A_4, A_1) = 2bond(A_0, A_4) + 2bond(A_4, A_1) - 2bond(A_0, A_1) = 192,000$

הוספת A_4 בין A_1 ל- A_3:

$\mathbf{A_1, A_4, A_3, A_2}$:

$bond(A_1, A_4) = 96,000$

$$\begin{aligned} bond(A_3, A_4) = {} & aff(A_1, A_3)*aff(A_1, A_4) + aff(A_2, A_3)*aff(A_2, A_4) + \\ & aff(A_3, A_3)*aff(A_3, A_4) + aff(A_4, A_3)*aff(A_4, A_4) + \\ & aff(A_5, A_3)*aff(A_5, A_4) \\ & = 240*160 + 240*0 + 240*0 + 0*160 + 0*80 = 38,400 \end{aligned}$$

$bond(A_1, A_3) = 211,200$

$cont(A_1, A_4, A_3) = 2bond(A_1, A_4) + 2bond(A_3, A_4) - 2bond(A_1, A_3) = 153,600$

הוספת A_4 בין A_3 ל- A_2:

A_1, A_3, A_4, A_2 :

$bond(A_3, A_4) = 38,400$

$$\begin{aligned} bond(A_4, A_2) = {} & aff(A_1, A_4)*aff(A_1, A_2) + aff(A_2, A_4)*aff(A_2, A_2) + \\ & aff(A_3, A_4)*aff(A_3, A_2) + aff(A_4, A_4)*aff(A_4, A_2) + \\ & aff(A_5, A_4)*aff(A_5, A_2) \\ & = 160*240 + 0*240 + 0*240 + 160*0 + 80*0 = 38,400 \end{aligned}$$

$bond(A_3, A_2) = 172,800$

$cont(A_3, A_4, A_2) = 2bond(A_3, A_4) + 2bond(A_4, A_2) - 2bond(A_3, A_2) = -192,000$

הוספת A_4 מימין ל- A_2:

A_1, A_3, A_2, A_4 :

$bond(A_2, A_4) = 38,400$

$bond(A_4, A_5') = 0$

$bond(A_2, A_5') = 0$

$cont(A_2, A_4, A_5') = 2bond(A_2, A_4) + 2bond(A_4, A_5') - 2bond(A_2, A_5') = 76,800$

לכן נבחר להוסיף את A_4 משמאל ל-A_1. והמצב הנוכחי הוא: A_4, A_1, A_3, A_2

נבדוק לאן להוסיף את האטריביוט $A_5 = zipCode$:

הוספת A_5 משמאל ל-A_4:

A_5, A_4, A_1, A_3, A_2 :

$bond(A_0, A_5) = 0$

$$\begin{aligned} bond(A_5, A_4) = {} & aff(A_1, A_5)*aff(A_1, A_4) + aff(A_2, A_5)*aff(A_2, A_4) + \\ & aff(A_3, A_5)*aff(A_3, A_4) + aff(A_4, A_5)*aff(A_4, A_4) + \\ & aff(A_5, A_5)*aff(A_5, A_4) \\ & = 80*160 + 0*0 + 0*0 + 80*160 + 80*80 = 32,000 \end{aligned}$$

$bond(A_0, A_4) = 0$

$cont(A_0, A_5, A_4) = 2bond(A_0, A_5) + 2bond(A_5, A_1) - 2bond(A_0, A_1) = 64,000$

הוספת A_5 בין A_4 ל- A_1:

A_4, A_5, A_1, A_3, A_2 :

$bond(A_4, A_5) = 32,000$

$$\begin{aligned} bond(A_5, A_1) = {} & aff(A_1, A_5)*aff(A_1, A_1) + aff(A_2, A_5)*aff(A_2, A_1) + \\ & aff(A_3, A_5)*aff(A_3, A_1) + aff(A_4, A_5)*aff(A_4, A_1) + \\ & aff(A_5, A_5)*aff(A_5, A_1) \\ & = 80*400 + 0*240 + 0*240 + 80*160 + 80*80 = 51,200 \end{aligned}$$

$bond(A_4, A_1) = 96,000$

$cont(A_4, A_5, A_1) = 2bond(A_4, A_5) + 2bond(A_5, A_1) - 2bond(A_4, A_1) = -25,600$

הוספת A_5 בין A_1 ל- A_3:

A_4, A_1, A_5, A_3, A_2 :

$bond(A_1, A_5) = 51,200$

$$bond(A_5, A_3) = aff(A_1, A_5)*aff(A_1, A_3) + aff(A_2, A_5)*aff(A_2, A_3) + aff(A_3, A_5)*aff(A_3, A_3) + aff(A_4, A_5)*aff(A_4, A_3) + aff(A_5, A_5)*aff(A_5, A_3) = 80*240 + 0*240 + 0*240 + 80*0 + 80*0 = 19,200$$

$$bond(A_1, A_3) = 211,200$$

$$cont(A_1, A_5, A_3) = 2bond(A_1, A_5) + 2bond(A_5, A_3) - 2bond(A_1, A_3) = -281,600$$

הוספת A_5 בין ל- A_2:

A_4, A_1, A_3, A_5, A_2 :

$$bond(A_3, A_5) = 19,200$$

$$bond(A_5, A_2) = aff(A_1, A_5)*aff(A_1, A_2) + aff(A_2, A_5)*aff(A_2, A_2) + aff(A_3, A_5)*aff(A_3, A_2) + aff(A_5, A_5)*aff(A_4, A_2) + aff(A_5, A_5)*aff(A_5, A_2) = 80*240 + 0*240 + 0*240 + 80*0 + 80*0 = 19,200$$

$$bond(A_3, A_2) = 172,800$$

$$cont(A_3, A_4, A_2) = 2bond(A_3, A_5) + 2bond(A_5, A_2) - 2bond(A_3, A_2) = -268,800$$

הוספת A_5 מימין ל- A_2:

A_4, A_1, A_3, A_2, A_5 :

$$bond(A_2, A_5) = 19,200$$

$$bond(A_5, A_6') = 0$$

$$bond(A_2, A_6') = 0$$

$$cont(A_2, A_5, A_6') = 2bond(A_2, A_5) + 2bond(A_5, A_6') - 2bond(A_2, A_6') = 38,400$$

לכן נבחר להוסיף את A_5 משמאל ל-A_4. והמצב הנוכחי הוא: A_5, A_4, A_1, A_3, A_2.

⇐ מטריצת ה-Clustered Affinity תיראה כך:

$$CA = \left(\begin{array}{c|ccccc} & A_5 & A_4 & A_1 & A_3 & A_2 \\ A_5 & 80 & 80 & 80 & 0 & 0 \\ A_4 & 80 & 160 & 160 & 0 & 0 \\ A_1 & 80 & 160 & 400 & 240 & 240 \\ A_3 & 0 & 0 & 240 & 240 & 240 \\ A_2 & 0 & 0 & 240 & 240 & 240 \end{array}\right)$$

נרצה לפצל את המטריצה לקבוצת TA וקבוצת BA, לצורך כך נחפש את החלוקה שתמקסם את הביטוי $z = CTQ*CBQ - COQ^2$, המחושב לפי הנוסחאות הבאות:

$$CQ = \sum_{q_i \in Q} \sum_{\forall S_j} ref_j(q_i) acc_j(q_i)$$

$$CTQ = \sum_{q_i \in TQ} \sum_{\forall S_j} ref_j(q_i) acc_j(q_i)$$

$$CBQ = \sum_{q_i \in BQ} \sum_{\forall S_j} ref_j(q_i) acc_j(q_i)$$

$$COQ = \sum_{q_i \in OQ} \sum_{\forall S_j} ref_j(q_i) acc_j(q_i)$$

כאשר:

קבוצת כל האטריביוטים שהשאילתה ניגשת אליהם - $AQ(q_i) = \{A_j | use(q_i, A_j) = 1\}$

קבוצת כל השאילתות שמקיימות שכל האטריביוטים שהן ניגשות אליהם נמצאים בתחום של TA -

$TQ(q_i) = \{q_i | AQ(q_i) \subseteq TA\}$

קבוצת כל השאילתות שמקיימות שכל האטריביוטים שהן ניגשות אליהם נמצאים בתחום של BA -

$BQ(q_i) = \{q_i | AQ(q_i) \subseteq BA\}$

$OQ = Q - \{TQ \cup BQ\}$

לצורך מציאת החלוקה שתמקסם את הביטוי, קודם כל ננסה לשים את האטריביוט הראשון בפרגמנט אחד, ואת שאר האטריביוטים בפרגמנט השני.

פרמוטציה נוכחית A_5, A_4, A_1, A_3, A_2:

- פיצול נוכחי: $A_5 \mid A_4, A_1, A_3, A_2$

אף אחת מהשאילתות לא מקיימת שכל האטריביוטים שהיא ניגשת אליהם נמצאים ב-TA (שמכילה רק את כניסה 1,1 במטריצה) ולכן TQ יהיה ריק. שאילתות 1 ו-2 מקיימות שכל האטריביוטים שלהן נמצאים ב-BQ (שמכילה את 4 השורות התחתונות וארבע העמודות השמאליות של המטריצה) ולכן BQ יכיל את שאילתות 1,2. שאילתה מס' 3 מכילה גם את האטריביוט ZipCode שנמצא ב-TA וגם את DMA ו-NetWorth שנמצאים ב-BA ולכן שאילתה 3 תהיה בקבוצה OQ. *הנימוקים בשלבים הבאים באלגוריתם יהיו דומים לנימוקים אלו.

$TQ = \{\},\ BQ = \{q_1, q_2\},\ OQ = \{q_3\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-TA, כלומר אין שאילתות בקבוצה TQ, ולכן CTQ=0.

$CBQ = 60*4 + 20*4 = 320\ ,\ \ COQ = 20*4 = 80\ \ \Rightarrow\ \ z = 0*320 - 80^2 = -6400$

- פיצול נוכחי: $A_5, A_4 \mid A_1, A_3, A_2$

$TQ = \{\},\ BQ = \{q_1\},\ OQ = \{q_2, q_3\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-TA, כלומר אין שאילתות בקבוצה TQ, ולכן CTQ=0.

$CBQ = 60*4 = 240\ ,\ \ COQ = 20*4 + 20*4 = 160\ \ \Rightarrow\ \ z = 0*240 - 160^2 = -25,600$

- פיצול נוכחי: $A_5, A_4, A_1 \mid A_3, A_2$

$TQ = \{q_3\},\ BQ = \{\},\ OQ = \{q_1, q_2\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-BA, כלומר אין שאילתות בקבוצה BQ, ולכן CBQ=0.

$CTQ = 20*4 = 80\ ,\ \ COQ = 60*4 + 20*4 = 320\ \ \Rightarrow\ \ z = 80*0 - 320^2 = -102,400$

- פיצול נוכחי: $A_5, A_4, A_1, A_3 | A_2$

$TQ = \{q_2, q_3\},\ BQ = \{\},\ OQ = \{q_1\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-BA, כלומר אין שאילתות בקבוצה BQ, ולכן CBQ=0.

$CTQ = 20*4 + 20*4 = 160\ ,\ \ COQ = 60*4 = 240\ \ \Rightarrow\ \ z = 160*0 - 240^2 = -57,600$

$\Leftarrow$ קיבלנו שעבור הפרמוטציה $(A_5, A_4, A_1, A_3, A_2)$ החלוקה הטובה ביותר היא $(A_5 \mid A_4, A_1, A_3, A_2)$

פרמוטציה נוכחית A_4, A_1, A_3, A_2, A_5:

- פיצול נוכחי: $A_4 \mid A_1, A_3, A_2, A_5$

$TQ = \{\},\ BQ = \{q_1\},\ OQ = \{q_2, q_3\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-TA, כלומר אין שאילתות בקבוצה TQ, ולכן CTQ=0.

$CBQ = 60*4 = 240\ ,\ \ COQ = 20*4 + 20*4 = 160 \ \Rightarrow\ z = 0*240 - 160^2 = -25,600$

- פיצול נוכחי: $A_4, A_1 \mid A_3, A_2, A_5$

$TQ = \{\},\ BQ = \{\},\ OQ = \{q_1, q_2, q_3\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-TA או ב-BA, כלומר אין שאילתות בקבוצות TQ ו-BQ, ולכן CTQ=0 , CBQ=0.

$COQ = 60*4 + 20*4 + 20*4 = 400 \ \Rightarrow\ z = 0*0 - 400^2 = -160,000$

- פיצול נוכחי: $A_4\ , A_1, A_3 | A_2, A_5$

$TQ = \{q_2\},\ BQ = \{\},\ OQ = \{q_1, q_3\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-BA, כלומר אין שאילתות בקבוצה BQ, ולכן CBQ=0.

$CTQ = 20*4 = 80\ ,\ \ COQ = 60*4 + 20*4 = 320 \ \Rightarrow\ z = 0*80 - 320^2 = -102,400$

- פיצול נוכחי: $A_4\ , A_1, A_3, A_2 | A_5$

$TQ = \{q_1, q_2\},\ BQ = \{\},\ OQ = \{q_3\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-BA, כלומר אין שאילתות בקבוצה BQ, ולכן CBQ=0.

$CTQ = 60*4 + 20*4 = 320\ ,\ \ COQ = 20*4 = 80 \ \Rightarrow\ z = 0*320 - 80^2 = -6,400$

$\Leftarrow$ קיבלנו שעבור הפרמוטציה $(A_4, A_1, A_3, A_2, A_5)$ החלוקה הטובה ביותר היא $(A_4\ , A_1, A_3, A_2 | A_5)$

פרמוטציה נוכחית A_1, A_3, A_2, A_5, A_4:

- פיצול נוכחי: $A_1 \mid A_3, A_2, A_5, A_4$

$TQ = \{\},\ BQ = \{\},\ OQ = \{q_1, q_2, q_3\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-TA או ב-BA, כלומר אין שאילתות בקבוצות TQ ו-BQ, ולכן CTQ=0 , CBQ=0.

$COQ = 60*4 + 20*4 + 20*4 = 400 \ \Rightarrow\ z = 0*0 - 400^2 = -160,000$

- פיצול נוכחי: $A_1, A_3 \mid A_2, A_5, A_4$

$TQ = \{\},\ BQ = \{\},\ OQ = \{q_1, q_2, q_3\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-TA או ב-BA, כלומר אין שאילתות בקבוצות TQ ו-BQ, ולכן CTQ=0 , CBQ=0.

$COQ = 60*4 + 20*4 + 20*4 = 400 \ \Rightarrow\ z = 0*0 - 400^2 = -160,000$

- פיצול נוכחי: $A_1, A_3, A_2 \mid A_5, A_4$

$TQ = \{q_1\},\ BQ = \{\},\ OQ = \{q_2, q_3\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-BA, כלומר אין שאילתות בקבוצה BQ, ולכן CBQ=0.

$CTQ = 60*4 = 240\ ,\ \ COQ = 20*4 + 20*4 = 160 \ \Rightarrow\ z = 0*240 - 160^2 = -25,600$

- פיצול נוכחי: $A_1, A_3, A_2, A_5 \mid A_4$

$TQ = \{q_1\},\ BQ = \{\},\ OQ = \{q_2, q_3\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-BA, כלומר אין שאילתות בקבוצה BQ, ולכן CBQ=0.

$$CTQ = 60*4 = 240 \ , \ COQ = 20*4 + 20*4 = 160 \ \Rightarrow \ z = 0*240 - 160^2 = -25,600$$

⇐ קיבלנו שעבור הפרמוטציה $(A_1, A_3, A_2, A_5, A_4)$ החלוקה הטובה ביותר היא $(A_1, A_3, A_2, A_5 \mid A_4)$

(מבין שתי החלוקות האחרונות- עם הערך הכי גבוה, נבחר באופן אקראי)

פרמוטציה נוכחית A_3, A_2, A_5, A_4, A_1:

- פיצול נוכחי: $A_3 \mid A_2, A_5, A_4, A_1$

$$TQ = \{\}, \ BQ = \{q_2, q_3\}, \ OQ = \{q_1\}$$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-TA, כלומר אין שאילתות בקבוצה TQ, ולכן CTQ=0.

$$CBQ = 20*4 + 20*4 = 160 \ , \ COQ = 60*4 = 240 \ \Rightarrow \ z = 0*160 - 240^2 = -57,600$$

- פיצול נוכחי: $A_3, A_2 \mid A_5, A_4, A_1$

$$TQ = \{\}, \ BQ = \{q_2, q_3\}, \ OQ = \{q_1\}$$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-TA, כלומר אין שאילתות בקבוצה TQ, ולכן CTQ=0.

$$CBQ = 20*4 + 20*4 = 160 \ , \ COQ = 60*4 = 240 \ \Rightarrow \ z = 0*160 - 240^2 = -57,600$$

- פיצול נוכחי: $A_3, A_2, A_5 \mid A_4, A_1$

$$TQ = \{\}, \ BQ = \{q_2\}, \ OQ = \{q_1, q_3\}$$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-TA, כלומר אין שאילתות בקבוצה TQ, ולכן CTQ=0.

$$CBQ = 20*4 = 80 \ , \ COQ = 60*4 + 20*4 = 320 \ \Rightarrow \ z = 0*80 - 320^2 = -102,400$$

- פיצול נוכחי: $A_3, A_2, A_5, A_4 \mid A_1$

$$TQ = \{\}, \ BQ = \{\}, \ OQ = \{q_1, q_2, q_3\}$$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-TA או ב-BA, כלומר אין שאילתות בקבוצות TQ ו-BQ, ולכן CBQ=0 , CTQ=0.

$$COQ = 20*4 + 20*4 + 60*4 = 400 \ \Rightarrow \ z = 0*0 - 400^2 = -160,000$$

⇐ קיבלנו שעבור הפרמוטציה $(A_3, A_2, A_5, A_4, A_1)$ החלוקה הטובה ביותר היא $(A_3 \mid A_2, A_5, A_4, A_1)$.

פרמוטציה נוכחית A_2, A_5, A_4, A_1, A_3:

- פיצול נוכחי: $A_2 \mid A_5, A_4, A_1, A_3$

$$TQ = \{\}, \ BQ = \{q_2, q_3\}, \ OQ = \{q_1\}$$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-TA, כלומר אין שאילתות בקבוצה TQ, ולכן CTQ=0.

$$CBQ = 20*4 + 20*4 = 160 \ , \ COQ = 60*4 = 240 \ \Rightarrow \ z = 0*160 - 240^2 = -57,600$$

- פיצול נוכחי: $A_2, A_5 \mid A_4, A_1, A_3$

$$TQ = \{\}, \ BQ = \{q_2\}, \ OQ = \{q_1, q_3\}$$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-TA, כלומר אין שאילתות בקבוצה TQ, ולכן CTQ=0.

$$CBQ = 20*4 = 80 \ , \ COQ = 60*4 + 20*4 = 320 \ \Rightarrow \ z = 0*80 - 320^2 = -102,400$$

- פיצול נוכחי: $A_2, A_5, A_4 \mid A_1, A_3$

$TQ = \{\},\ BQ = \{\},\ OQ = \{q_1, q_2, q_3\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-TA או ב-BA, כלומר אין שאילתות בקבוצות TQ ו-BQ, ולכן CTQ=0 , CBQ=0.

$COQ = 20*4 + 20*4 + 60*4 = 400 \quad \Rightarrow \quad z = 0*0 - 400^2 = -160,000$

- פיצול נוכחי: $A_2, A_5, A_4, A_1 \mid A_3$

$TQ = \{q_2, q_3\},\ BQ = \{\},\ OQ = \{q_1\}$

אין שאילתות שמכילות רק אטריביוטים שנמצאים ב-BA, כלומר אין שאילתות בקבוצה BQ, ולכן CBQ=0.

$CTQ = 20*4 + 20*4 = 160 \ ,\ \ COQ = 60*4 = 240 \quad \Rightarrow \quad z = 0*160 - 240^2 = -57,600$

$\Leftarrow$ קיבלנו שעבור הפרמוטציה $(A_2, A_5, A_4, A_1, A_3)$ החלוקה הטובה ביותר היא $(A_2, A_5, A_4, A_1 \mid A_3)$.

לסיכום, הפרמוטציה הטובה ביותר היא $(A_5, A_4, A_1, A_3, A_2)$ והחלוקה הטובה ביותר היא $(A_5 \mid A_4, A_1, A_3, A_2)$, כי עבורה ערך ה-z מקסימלי (-6400). לכן החלוקה האנכית שלנו תיראה כך:
(נוסיף את המפתח הראשי HHID לכל פרגמנט ידנית - אף שאילתה לא עשתה שימוש ב-deviceID,GreenLiving
לכן לא נוסיף אותם לפרגמנטים)

$$schema_1 = (HHID, ZipCode)$$
$$schema_2 = (HHID,\ NetWorth,\ NumOfAdults,\ householdSize,\ DMA)$$

נשתמש בשיטה ההיברידית לחלוקה שלנו - נחיל את החלוקות שמצאנו אחת על השנייה. נשים לב שמאחר שהפרגמנט R_1 מכיל רק את המפתח הראשי ואת ZipCode, נוכל לחלק אותו לפרגמנטים אופקיים רק על סמך תנאים על השדה ZipCode, ולהפך: מאחר ש-R_2 לא מכיל את ZipCode נוכל לחלקו לפרגמנטים אופקיים על סמך כל התנאים שמצאנו למעלה בחלוקה האנכית מלבד תנאים על השדה ZipCode. לבסוף, נקבל את החלוקה ההיברידית הבאה:

$R_1 = \pi_{schema_2}(\sigma_{(NetWorth>5)\wedge(householdSize>2)} MediaData)$
$R_2 = \pi_{schema_2}(\sigma_{(NetWorth>5)\wedge(householdSize\leqslant 2)} MediaData)$
$R_3 = \pi_{schema_2}(\sigma_{(NetWorth\leqslant 5)\wedge(householdSize>2)} MediaData)$
$R_4 = \pi_{schema_2}(\sigma_{(NetWorth\leqslant 5)\wedge(householdSize\leqslant 2)} MediaData)$
$\forall 1 \leqslant i \leqslant m:\ R_{1i} = \pi_{schema_2}(\sigma_{(NetWorth>5)\wedge(y_i \in DMA\ s.t\ DMA=y_i)} MediaData)$
$\forall 1 \leqslant i \leqslant m:\ R_{2i} = \pi_{schema_2}(\sigma_{(NetWorth\leqslant 5)\wedge(y_i \in DMA\ s.t\ DMA=y_i)} MediaData)$
$\forall 1 \leqslant i \leqslant m:\ R_{3i} = \pi_{schema_2}(\sigma_{(householdSize>2)\wedge(y_i \in DMA\ s.t\ DMA=y_i)} MediaData)$
$\forall 1 \leqslant i \leqslant m:\ R_{4i} = \pi_{schema_2}(\sigma_{(householdSize\leqslant 2)\wedge(y_i \in DMA\ s.t\ DMA=y_i)} MediaData)$
$\forall 1 \leqslant j \leqslant n:\ R_{5j} = \pi_{schema_1}(\sigma_{(x_j \in zipCode\ s.t\ zipCode=x_j)} MediaData)$

שאלה 4

השאילתה המוצגת בשאלה מחזירה רשימה של ממירים, שהתרחש בהם אירוע צפייה בין השעה שמונה בערב לאחד עשרה בערב (לא כולל), ובנוסף עבור כל ממיר כזה מחזירה את אזור המגורים והמיקוד של המשפחה שהממיר בבעלותה.

על מנת לחשב את מטריצת ה-attribute affinity של הרלציה MediaData נעקוב אחרי השלבים באלגוריתם שלמדנו בהרצאה:

- לפי הנתון, נבנה את מטריצת הגישה לאתרים:

$$\left(\begin{array}{c|cccc} & S1 & S2 & S3 & S4 \\ q1 & 40 & 40 & 40 & 40 \\ q2 & 10 & 10 & 10 & 10 \\ q3 & 20 & 20 & 20 & 20 \\ q4 & 30 & 30 & 30 & 30 \end{array}\right)$$

- נמצא את מטריצת ה-use: לפי הגדרה, $use(q_i, A_j) = \begin{cases} 1 & , A_j \text{ is referenced by } q_i \\ 0 & , o.w \end{cases}$. לכן המטריצה תיראה כך:

* השדות HHID, deviceID הם מפתחות ראשיים ברלציה MediaData ולכן חייבים להימצא בכל אחד מהפרגמנטים, כלומר נוסיף אותם ידנית לכל פרגמנט שיווצר. בנוסף נשים לב כי בשדה GreenLiving לא נעשה שימוש באף אחת מהשאילתות ולכן לא נתחשב בו בחלוקה.

$$use = \left(\begin{array}{c|ccccc} & NetWorth & NumOfAdults & householdSize & DMA & zipCode \\ q1 & 1 & 1 & 1 & 0 & 0 \\ q2 & 1 & 0 & 0 & 1 & 0 \\ q3 & 1 & 0 & 0 & 1 & 1 \\ q4 & 0 & 0 & 0 & 1 & 1 \end{array}\right)$$

- נחשב את ערך ה-Attribute Affinity של כל שני שדות ונסדר במטריצת ה-Attribute Affinity:

לדוגמה עבור השדה DMA החישוב יבוצע כך:

$$aff(DMA, NetWorth) = 4*1*10 + 4*1*20 = 120$$
$$aff(DMA, NumOfAdults) = 0$$
$$aff(DMA, householdSize) = 0$$
$$aff(DMA, DMA) = 4*1*10 + 4*1*20 + 4*1*30 = 240$$
$$aff(DMA, zipCode) = 4*1*20 + 4*1*30 = 200$$

ובאופן דומה עבור כל שדה. המטריצה תיראה כך:

$$aff = \left(\begin{array}{c|ccccc} & NetWorth & NumOfAdults & householdSize & DMA & zipCode \\ NetWorth & 280 & 160 & 160 & 120 & 80 \\ NumOfAdults & 160 & 160 & 160 & 0 & 0 \\ householdSize & 160 & 160 & 160 & 0 & 0 \\ DMA & 120 & 0 & 0 & 240 & 200 \\ zipCode & 80 & 0 & 0 & 200 & 200 \end{array}\right)$$

נשים לב שבהשוואה למטריצת ה-affinity משאלה 3, הקשר בין DMA לבין ZipCode התחזק, במיוחד בהשוואה לקשר של כל אחד מהם עם NetWorth: למשל לפני כן הקשר של ZipCode עם DMA היווה $\frac{80}{240} = 33.3\%$ מהקשרים של ZipCode לשאר השדות במטריצה, וכעת מהווה $\frac{200}{480} = 41.67\%$ מהקשרים שלו. בנוסף נציין כי הקשר של DMA ו-NetWorth התחזק בהשוואה לקשר של householdSize ו-NetWorth במטריצה הזו לעומת המטריצה שבשאלה 3 (לפי חישובים דומים לחישובי האחוזים מעלה). אמנם הוא התחזק, אך לדעתנו אינו גובר על עוצמת הקשר בין DMA ל-ZipCode ולא מספיק על מנת לצרף את DMA לפרגמנט של NetWorth, NumOfAdults, householdSize.

לפי סימונים משאלה 3 ובנוסף נסמן: $A_6 = HHID$. נוסיף את A_6 ידנית (לפי הסברים קודמים) לשני הפרגמנטים שאנחנו סבורות שיווצרו מהחלוקה האנכית הנוכחית.

מהנימוקים לעיל, אנחנו סבורות כי החלוקה האנכית הנוכחית תראה כך:

$$R_1 = (A_6, A_1, A_2, A_3),\ R_2 = (A_6, A_4, A_5)$$