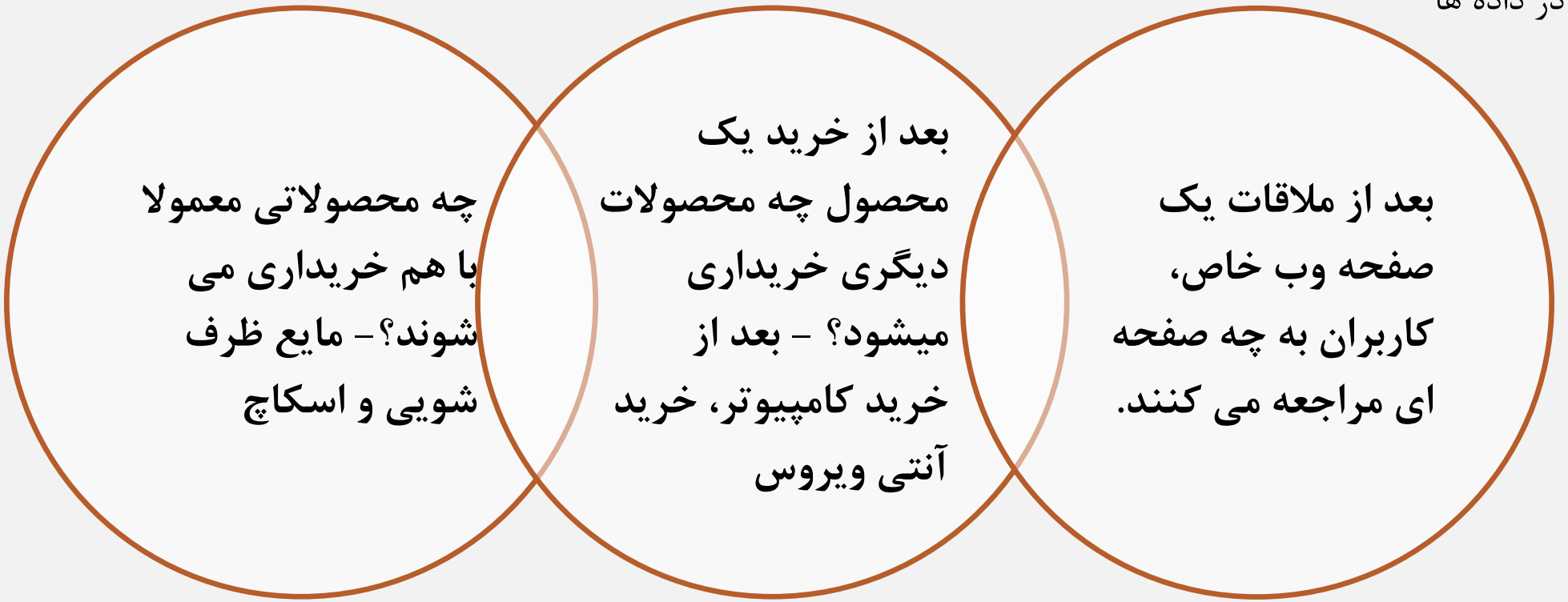
کاوش الگوهای پر تکرار، همبستگی و وابستگی: مفاهیم اولیه و روشها

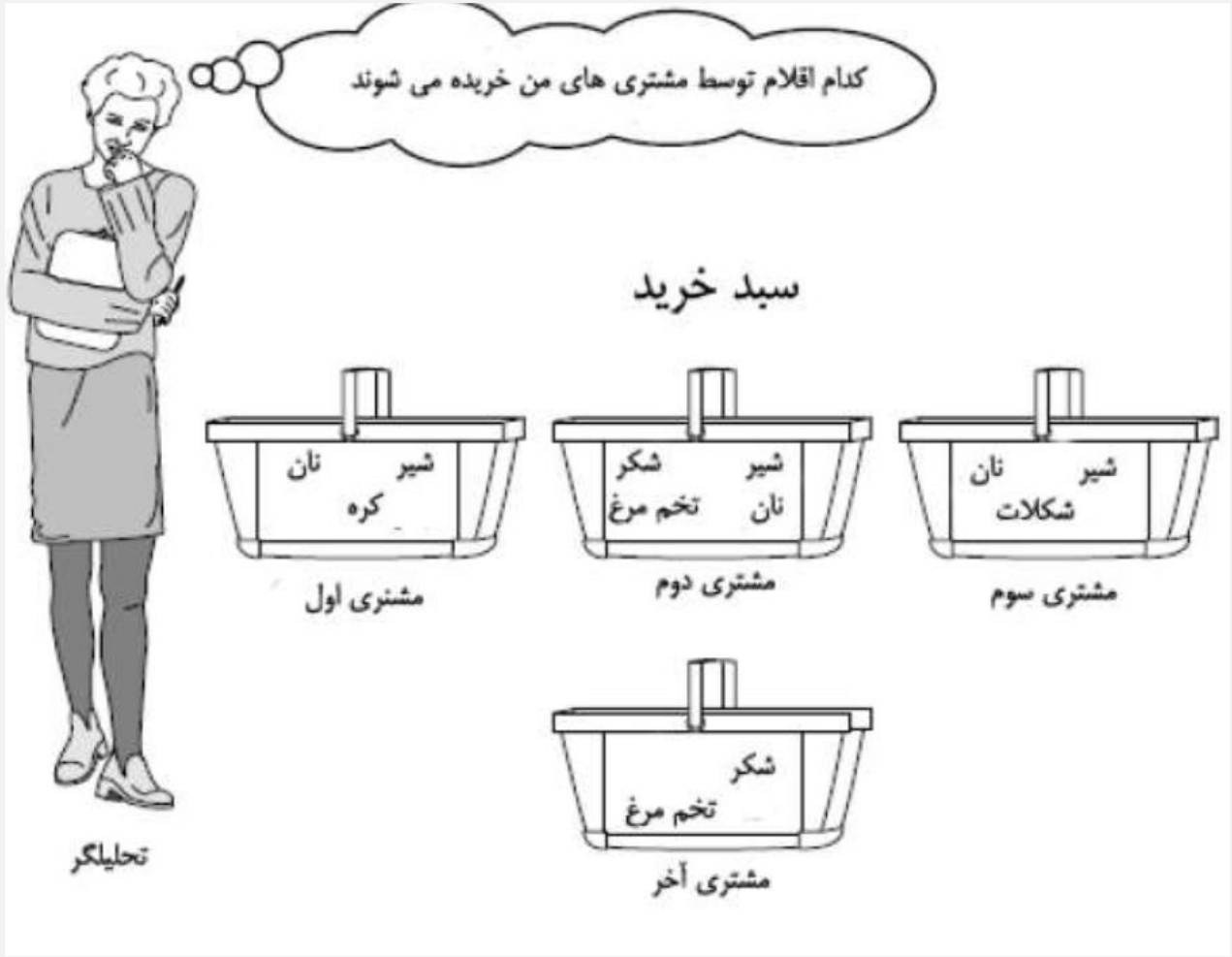
الگوی پر تکرار: یک الگو (مجموعه ای از آیتم ها، زیر دنباله ها، ساختارها و ...) که مکررا در مجموعه داده ها تکرار می شوند.

انگیزه: یافتن نظم ذاتی در داده ها



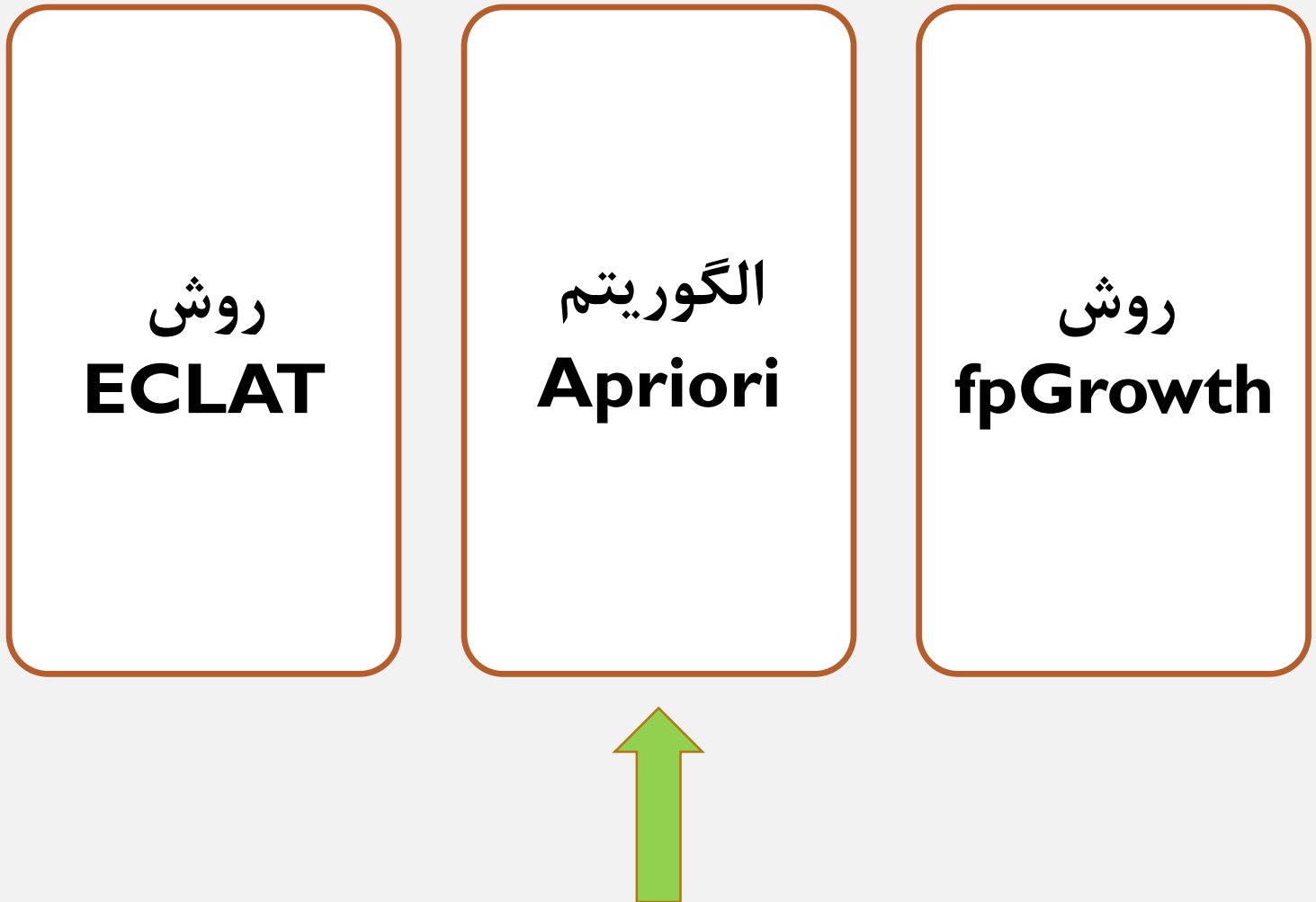
کاربرد ها: تحلیل سبد خرید مشتریان، طراحی کاتالوگ محصولات و تبلیغات، تحلیل فروش و طراحی فروشگاه ها، تحلیل وبلاگ ها(click stream)، تحلیل دنباله DNA ها و ...

آنالیز سبد خرید



- ✓ با پیدا کردن ارتباط زیاد فروش دو محصول می تـوان در چینش محصولات فروشگاه به شکلی عمل کرد که میزان فروش را افزایش دهد.
- ✓ به فروشندگان کمک می کند تا تصمیم بگیرنـد چـه کالاهایی را فروش ویژه اعلام کننـد. در ایـن حالـت، فروش ویژه یک محصول باعث افزایش فروش دیگـر محصولات مرتبط می گردد.

روش های کاوش الگوهای پرتکرار



الگوریتم APRIORI

روشی است مبتنی بر تولید و آزمایش الگوهای کاندید

مفاهیم اولیه

- مجموعه ای از یک یا چند آیتم را itemset میگویند.

- itemset که حاوی K آیتم ($X = \{I_{1+}I_2, \dots, I_K\}$)باشد را K-Itemset گویند.
- فرکانس تکرار یک Itemset را که نشان دهنده تکرار تراکنش هایی است که حاوی Itemset است را Support count آن Item set گویند.
- برای یک Itemset، نسبت تراکنش هایی که حاوی Itemset هستند به کل تراکنش ها را Support گویند.
- Frequent itemset : یک Itemst را پرتکرار گویند اگر مقدار Support آن کمتر از یک حد آستانه مشخص نباشد.

قوانین وابستگی

- **فرض کنید قانون $A \rightarrow B$ در مجموعه تراکنش های D وجود دار. برای این قانون دو پارامتر تعریف می شود:**

۱- Support : نشان دهنده درصدی از تراکنش های D که حاوی A و B است. این درصد را به صورت احتمال $P(A \cup B)$ بیان می کنیم.

$$\textbf{Support}(A \rightarrow B) = P(A \cup B)$$

۲- Confidence : بیان کننده درصدی از تراکنش های D است که اگر حاوی A باشند، آنگاه B نیز در آن وجود داشته باشد.

$$\textbf{Confidence}\ (A \rightarrow B) = \frac{\textbf{Support}(A \cup B)}{\textbf{Support}(A)} = \frac{\textbf{Support\_count}(A \cup B)}{\textbf{Support\_count}(A)}$$

قوانین وابستگی

قانون A➔B برای دو Itemset به نام های A و B را یک قانون وابستگی گویند اگر:

۱- دارای حداقل Min_Support باشد.

۲- دارای حداقل مقدار Min_Confidence باشد.

مثال

Tid	آیتم های خریداری شده
10	Bread, nuts, butter
20	Bread, Coffee, Butter, Eggs
30	Bread, Butter, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Butter, Eggs, Milk

فرض کنید:

حداقل مقدارsupport برابر ۵۰٪ باشد

حداقل مقدار confidence برابر ۵۰٪ باشد

Frequent Patterns:

Bread: 3, Nuts: 3, Butter: 4, Egg: 4,

(Bread, Butter): 3

Association Rule:

Bread⇒Butter(60%,100%)

Butter⇒Bread(60%,75%)

ویژگی Downward Closure

- یک زیر مجموعه از الگوهای پرتکرار، خود پرتکرار است.

- مثلا: اگر {bread,nuts,butter}پرتکرار باشد آنگاه {bread,butter}نیزپر تکرار است.

الگوریتم Apriori

- نام این الگوریتم به دلیل اینکه از دانش پیشین استفاده می کند A priori انتخاب شده است.
- در این الگوریتم از اطلاعات موجود در تولید K-Itemset برای تولید Itemset-(K+1)استفاده می شود.
- ابتدا برای یافتن 1-itemset های پرتکرار، یکبار مجموعه تراکنش ها اسکن می شوند. نتیجه کاوش L_1 نامیده می شود.
- سپس L_1 برای یافتن L_2 که حاوی 2-itemset های پرتکرار است، مورد استفاده قرار می گیرد.
- این مرحله تا جایی که دیگر هیچ k-itemset پرتکراری پیدا نشود ادامه می یابد.
- برای یافتن هر L_k نیاز به یکبار اسکن مجموعه تراکنش ها می باشد.

چگونگی ایجاد L_k بر اساس L_{k-1}

فرآیند دو مرحله ای برای ایجاد L_k :

گام اتصال (join step): برای ایجاد L_k که معرف مجموعه کاندیداهای k-itemset هستند، L_{k-1} را با خودش اتصال می دهیم. مجموعه بدست آمده را C_k می نامیم.

گام هرس (prune step): بر اساس خاصیت Apriori Property هر عنصر k-itemset که زیر مجموعه هایش در L_{k-1} نباشد، نمی تواند پرتکرار باشد.

مثال ١

Database TDB

tid	items
10	A,C,D
20	B,C,E
30	A,B,C,E
40	B,E

Itemset	Sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	Sup
{A}	2
{B}	3
{C}	3
{E}	3

Itemset
{A,B}
{A,C}
{A,E}
{B,C}
{B,E}
{C,E}

Itemset	Sup
{A,B}	1
{A,C}	2
{A,E}	1
{B,C}	2
{B,E}	3
{C,E}	2

L_2

Itemset	Sup
{A,C}	2
{B,C}	2
{B,E}	3
{C,E}	2

Itemset
{B,C,E}

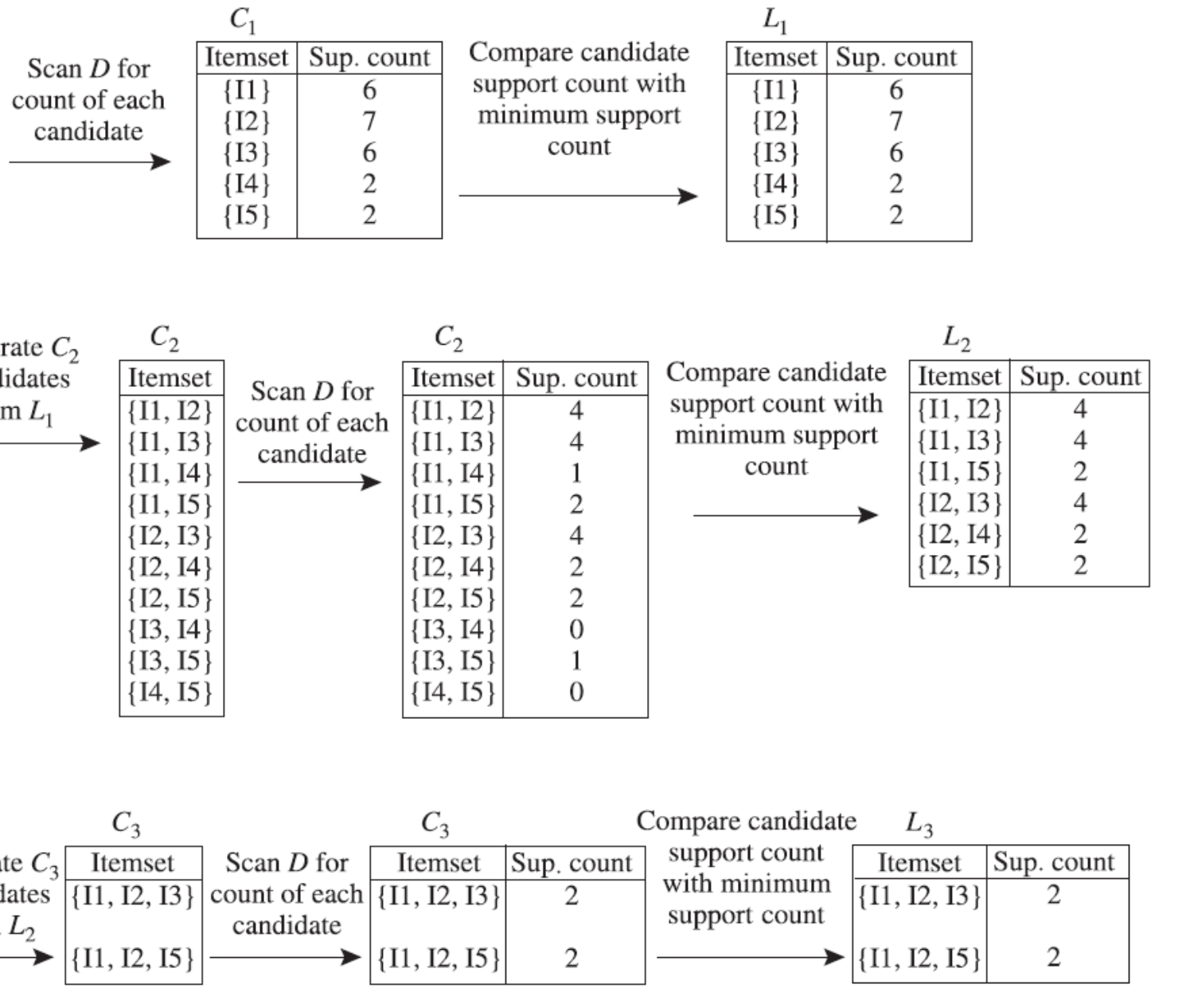
L_3

Itemset	Sup
{B,C,E}	2

Min_Sup = 1

مثال ۲

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3



مثال

- به عنوان مثال برای الگوی پرتکرار {I1,I2,I5} = I قوانین تولید شده عبارتند از:

{I1,I2}➔ I5, confidence= 2/4 = 50%

{I1,I5}➔ I2, confidence= 2/2 = 100%

{I2,I5}➔ I1, confidence= 2/2 = 100%

I1 ➔ {I2,I5} confidence= 2/6 = 33%

I2 ➔ {I1,I5} confidence= 2/7 = 29%

I5 ➔ {I1,I2} confidence= 2/2 = 100%

در صورتی که min_sup = 70% باشد ۲و ۳و ۶ انتخاب می شوند.

منبع