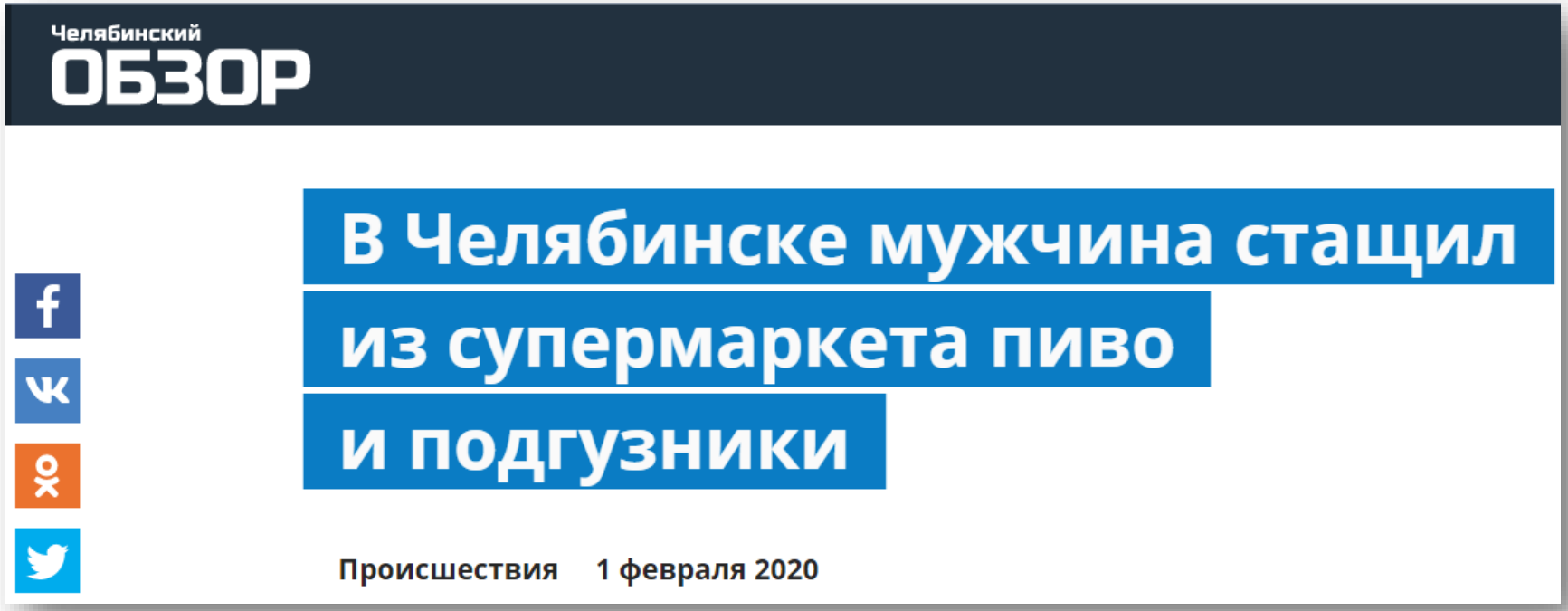
Задача поиска шаблонов в данных

Мыслить по шаблону –
вернейший способ завалить дело.

Дж.Ф. Энрайт



https://obzor174.ru/v-chelyabinske-muzhchina-stashchil-iz-supermarketa-pivo-i-podguzniki

Содержание

- Постановка задачи
  - Частые наборы
  - Ассоциативные правила
- Основные алгоритмы поиска частых наборов
  - Apriori
  - Eclat
  - FP-Growth
- Меры полезности шаблонов:
  - support
  - confidence
  - lift и др.
- **Поиск шаблонов в Больших данных**
  - **Компактное представление частых наборов**
  - **Иерархии наборов**
  - **Фрагментация и сэмплинг для поиска частых наборов**

Необходимость компактного представления $\mathcal{L}$

- Пример

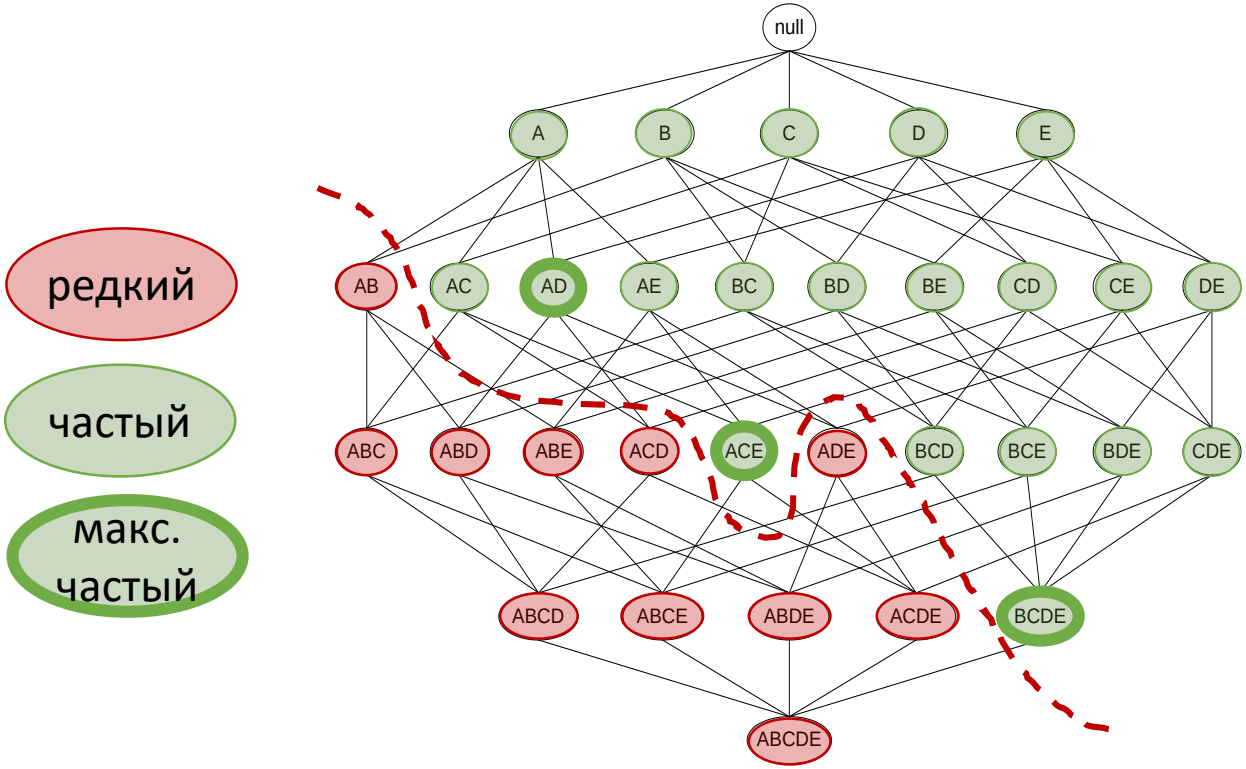
TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

- $|\mathcal{L}| = 3 \cdot \sum_{k=1}^{10} C_{10}^{k} = 3 \cdot \sum_{k=1}^{10} \frac{10!}{(10-k)! \cdot k!} = 3069$
- $X \subseteq Y \wedge \sup(X) = \sup(Y) \Longrightarrow X$ избыточен

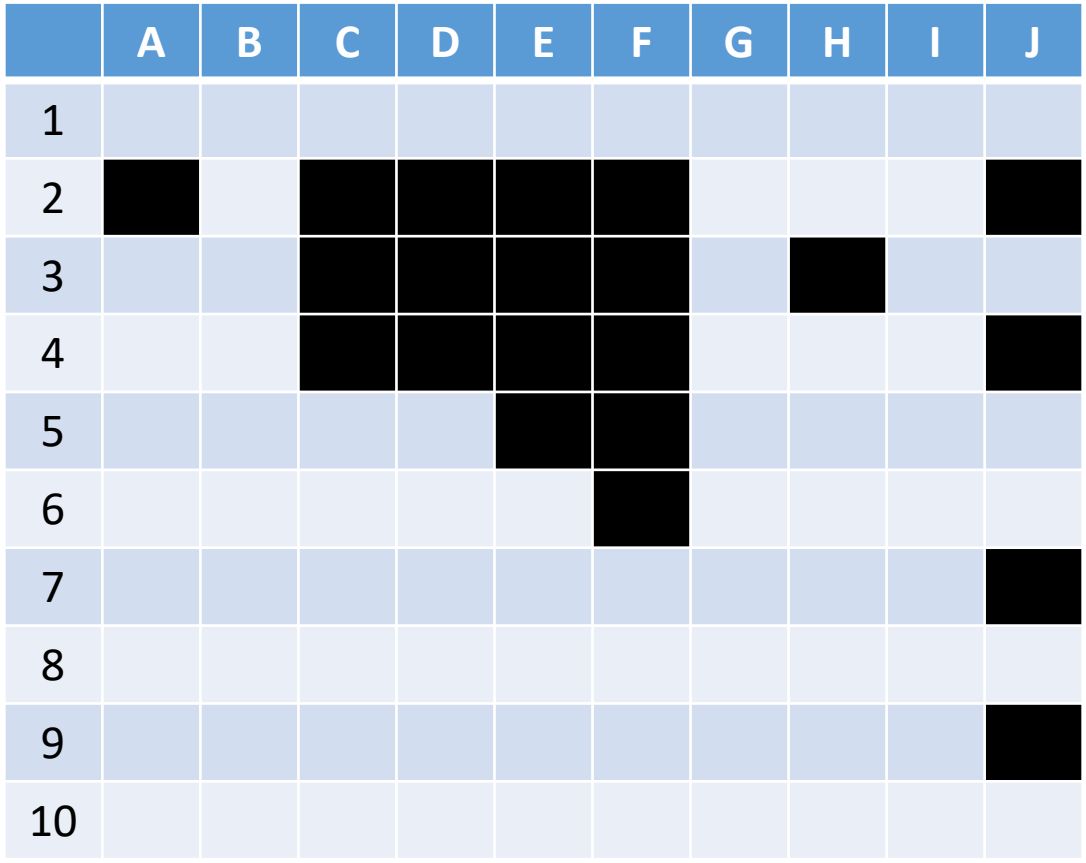
Максимально частый набор

Набор X является *максимально частым*, если он частый и не является непосредственным подмножеством частого набора

$$\neg\exists Y\ X \subseteq Y \wedge |X| = |Y| - 1 \wedge \sup(Y) \geq minsup$$

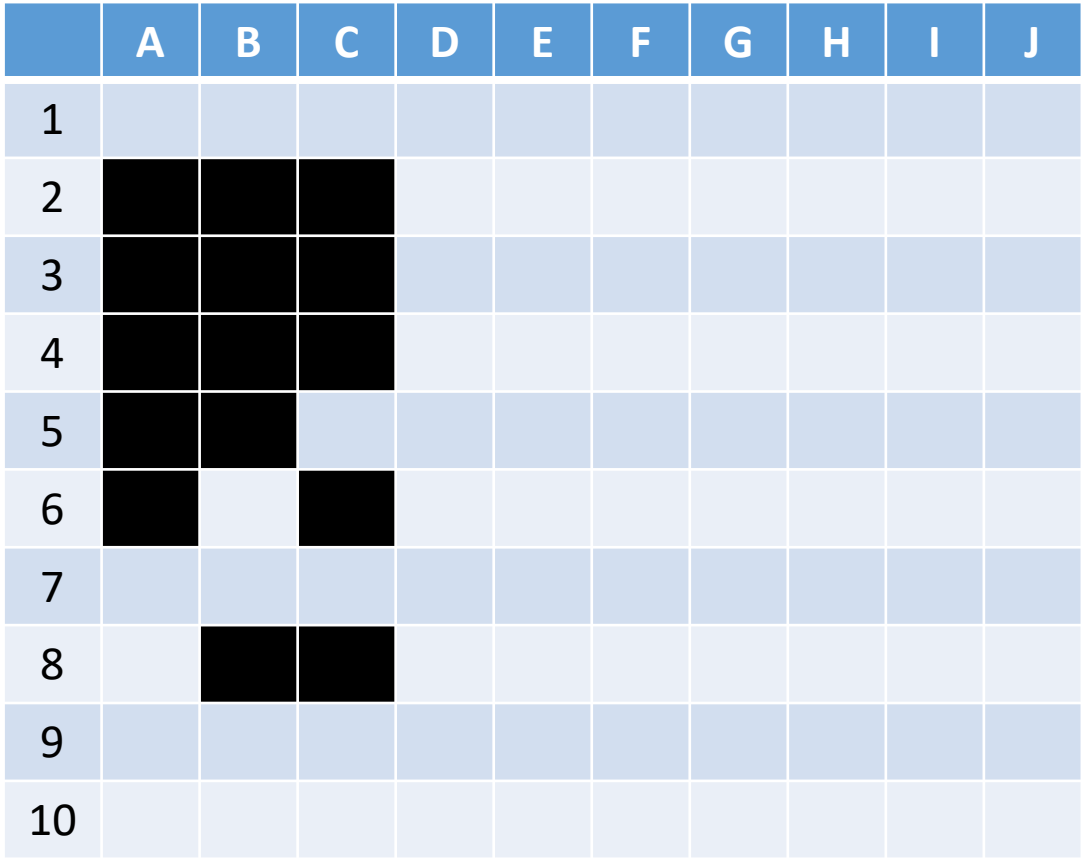


Пример



- $minsup = 5$
  - Частые: $\{F\}$
  - Max частые: $\{F\}$
- $minsup = 4$
  - Частые: $\{E\}$, $\{F\}$, $\{E, F\}$, $\{J\}$
  - Max частые: $\{E, F\}$, $\{J\}$
- $minsup = 3$
  - Частые: $\mathcal{P}(\{C, D, E, F\})$, $\{J\}$
  - Max частые: $\{C, D, E, F\}$, $\{J\}$

Пример



	A	B	C	D	E	F	G	H	I	J
1										
2	■	■	■							
3	■	■	■							
4	■	■	■							
5	■	■								
6	■		■							
7										
8		■	■							
9										
10										

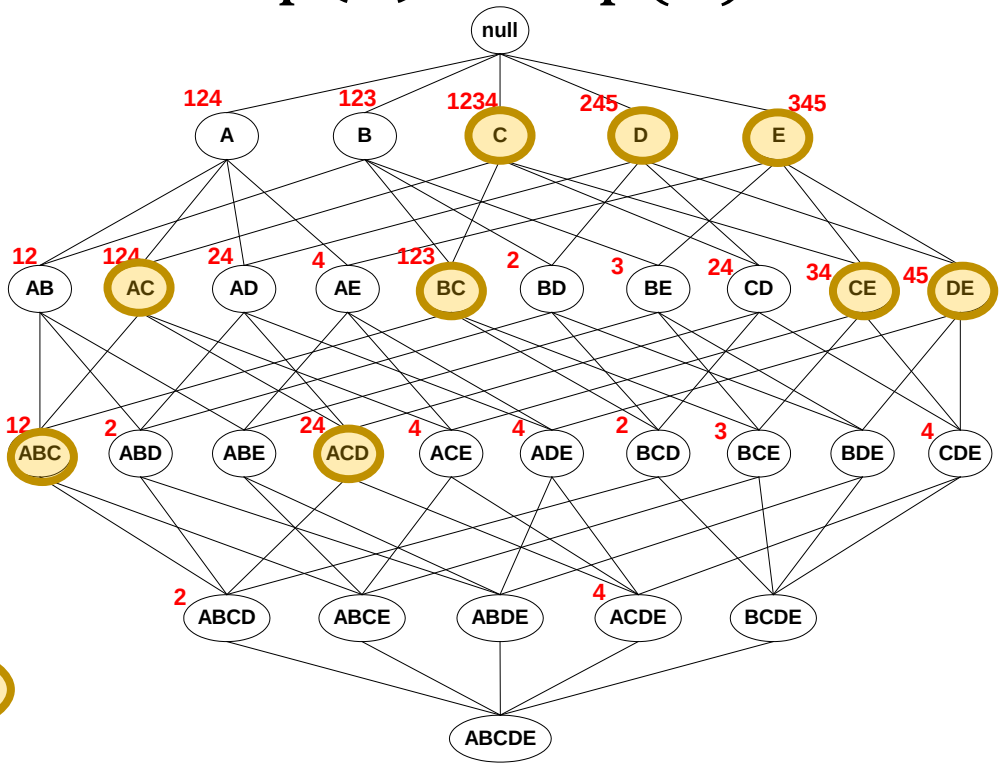
- $minsup = 5$
  - Max частые: $\{A\}, \{B\}, \{C\}$
- $minsup = 4$
  - Max частые: $\{A, B\}, \{A, C\}, \{B, C\}$
- $minsup = 3$
  - Max частые: $\{A, B, C\}$

Замкнутые наборы

Набор X является *замкнутым*, если отсутствует его непосредственное надмножество с той же поддержкой:

$$\neg\exists Y\ X \subseteq Y \wedge |X| = |Y| - 1 \wedge \sup(Y) = \sup(X)$$

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



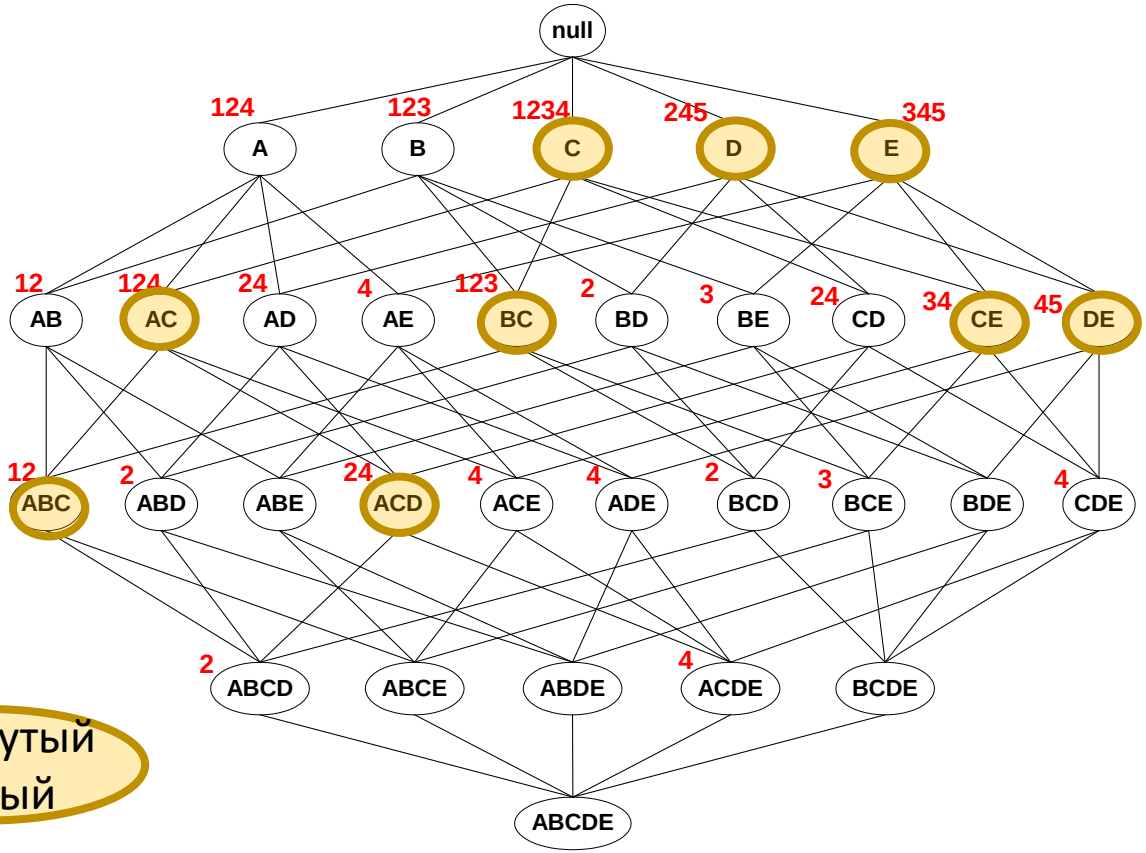
замкнутый

Замкнутые частые наборы

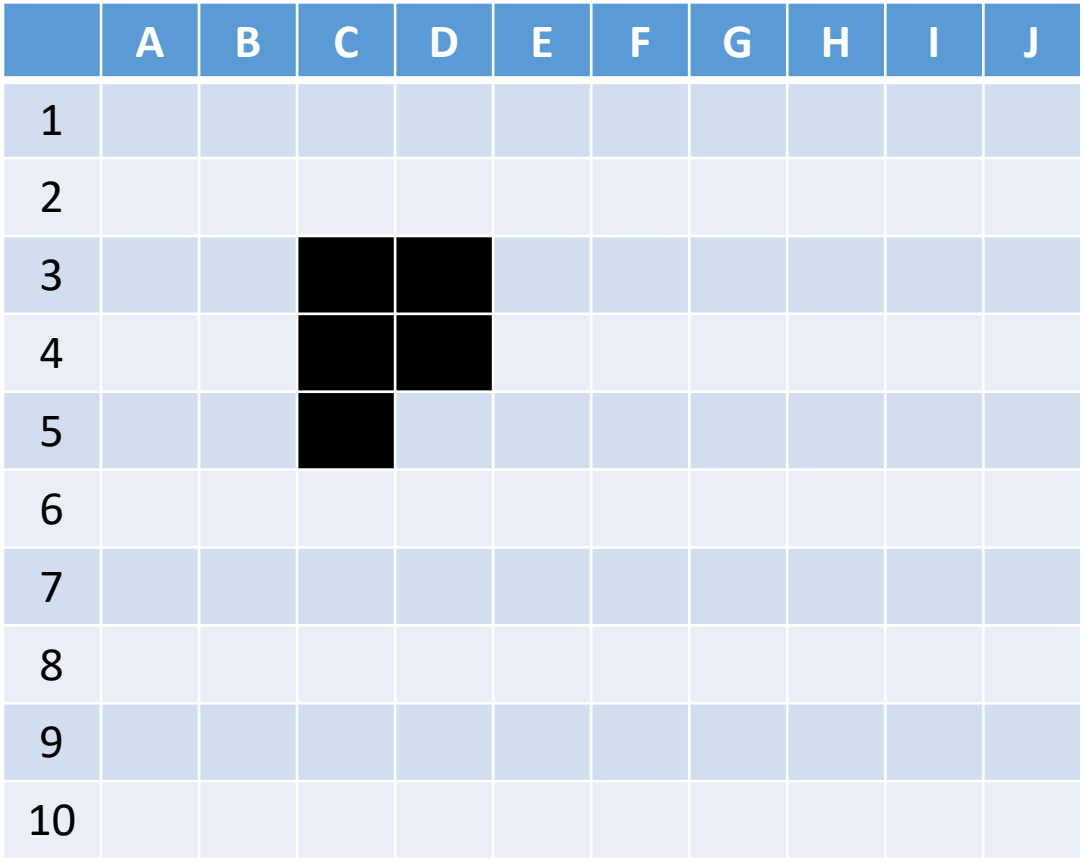
Набор является *замкнутым частым*, если он замкнутый и частый

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

$minsup = 2$



Пример

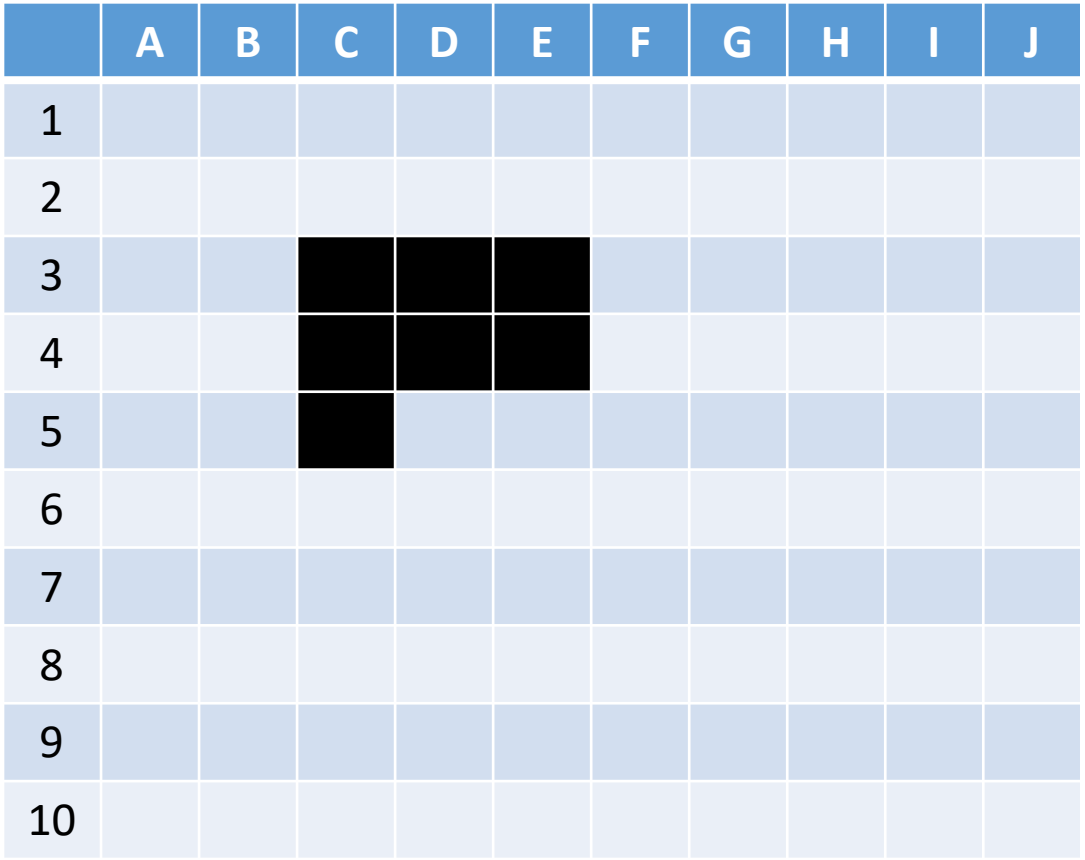


	A	B	C	D	E	F	G	H	I	J
1										
2										
3			■	■						
4			■	■						
5			■							
6										
7										
8										
9										
10										

$minsup = 2$

Itemsets	Support	Closed frequent itemsets
{C}	3	✓
{D}	2	
{C,D}	2	✓

Пример

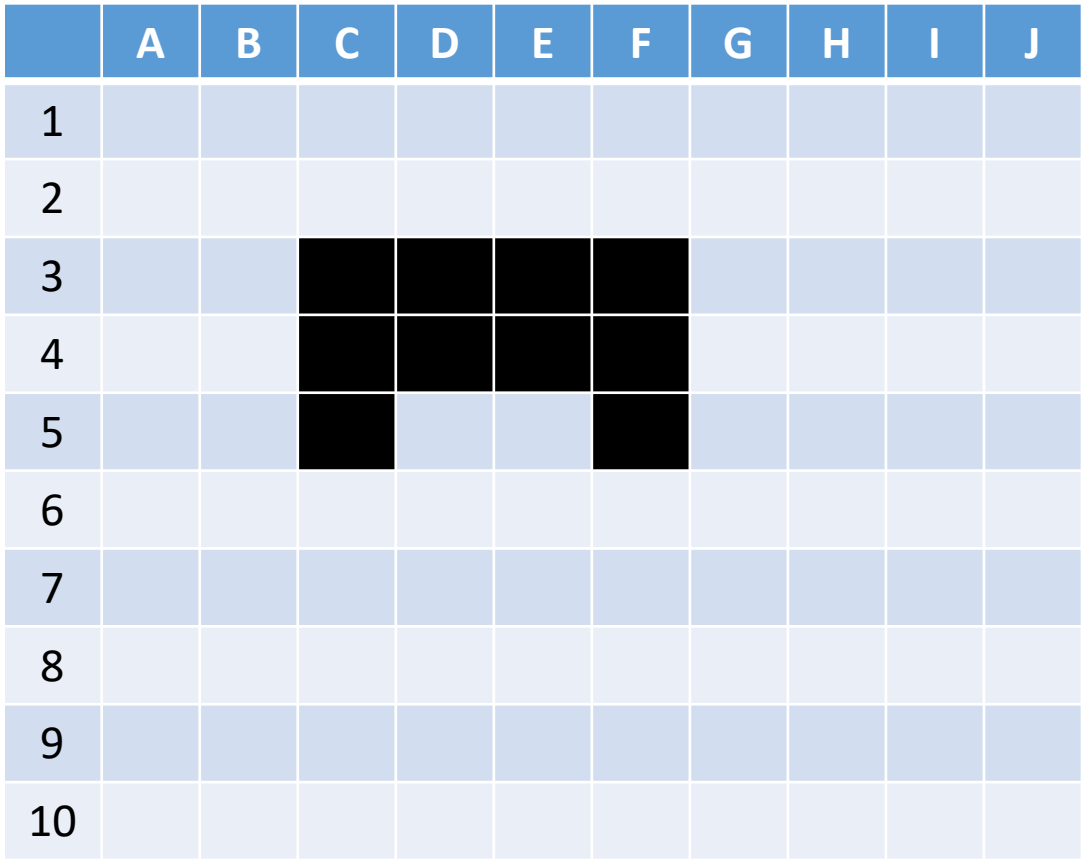


	A	B	C	D	E	F	G	H	I	J
1										
2										
3			■	■	■					
4			■	■	■					
5			■							
6										
7										
8										
9										
10										

$minsup = 2$

Itemsets	Support (counts)	Closed frequent itemsets
{C}	**3**	✓
{D}	2	
{E}	2	
{C,D}	2	
{C,E}	2	
{D,E}	2	
{C,D,E}	**2**	✓

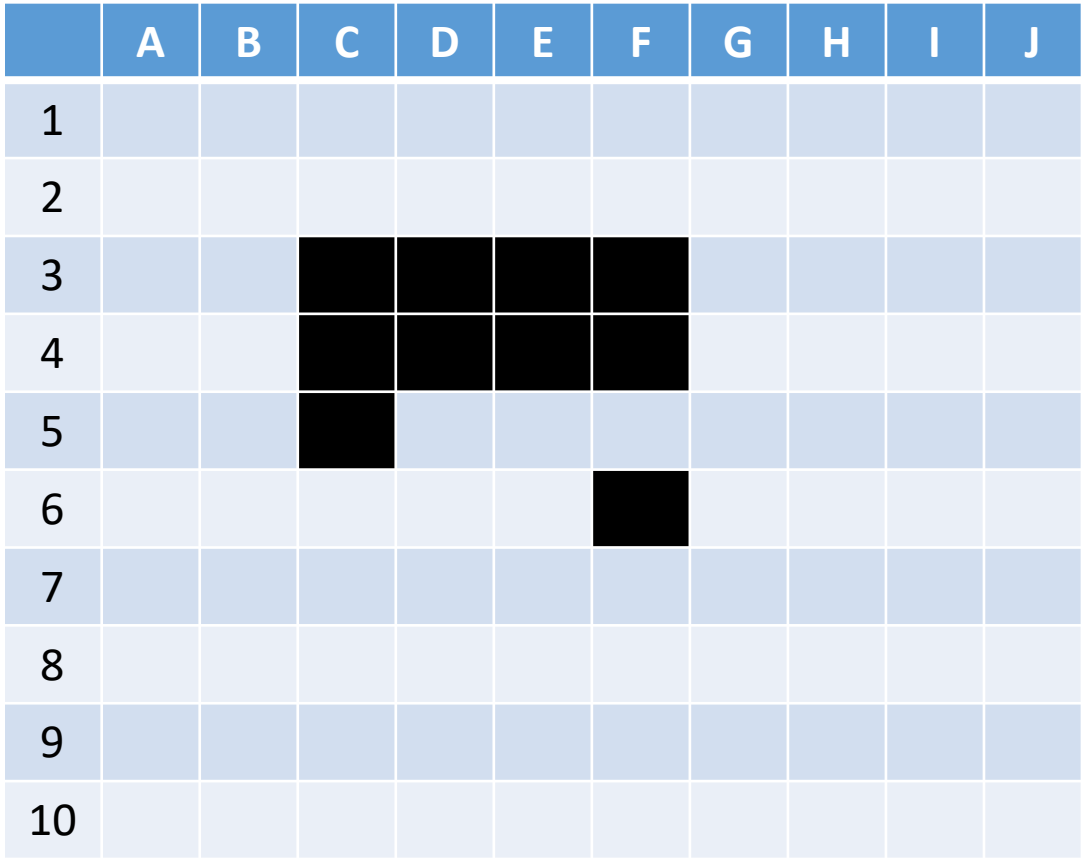
Пример



	A	B	C	D	E	F	G	H	I	J
1										
2										
3			■	■	■	■				
4			■	■	■	■				
5			■			■				
6										
7										
8										
9										
10										

- Замкнутые наборы: $\{C, D, E, F\}, \{C, F\}$

Пример



	A	B	C	D	E	F	G	H	I	J
1										
2										
3			■	■	■	■				
4			■	■	■	■				
5			■							
6						■				
7										
8										
9										
10										

- Замкнутые наборы: $\{C, D, E, F\}, \{C\}, \{F\}$

Замкнутые частые наборы позволяют найти поддержку остальных частых наборов

Найти $\mathcal{L}_{k_{max}}$, где k_{max} – макс мощность замкнутого частого набора

for $k := k_{max}$ **down to** 1 **do**

  Найти $\mathcal{L}_k$

  for all $I \in \mathcal{L}_k$ **do**

    if I не является замкнутым частым **then**

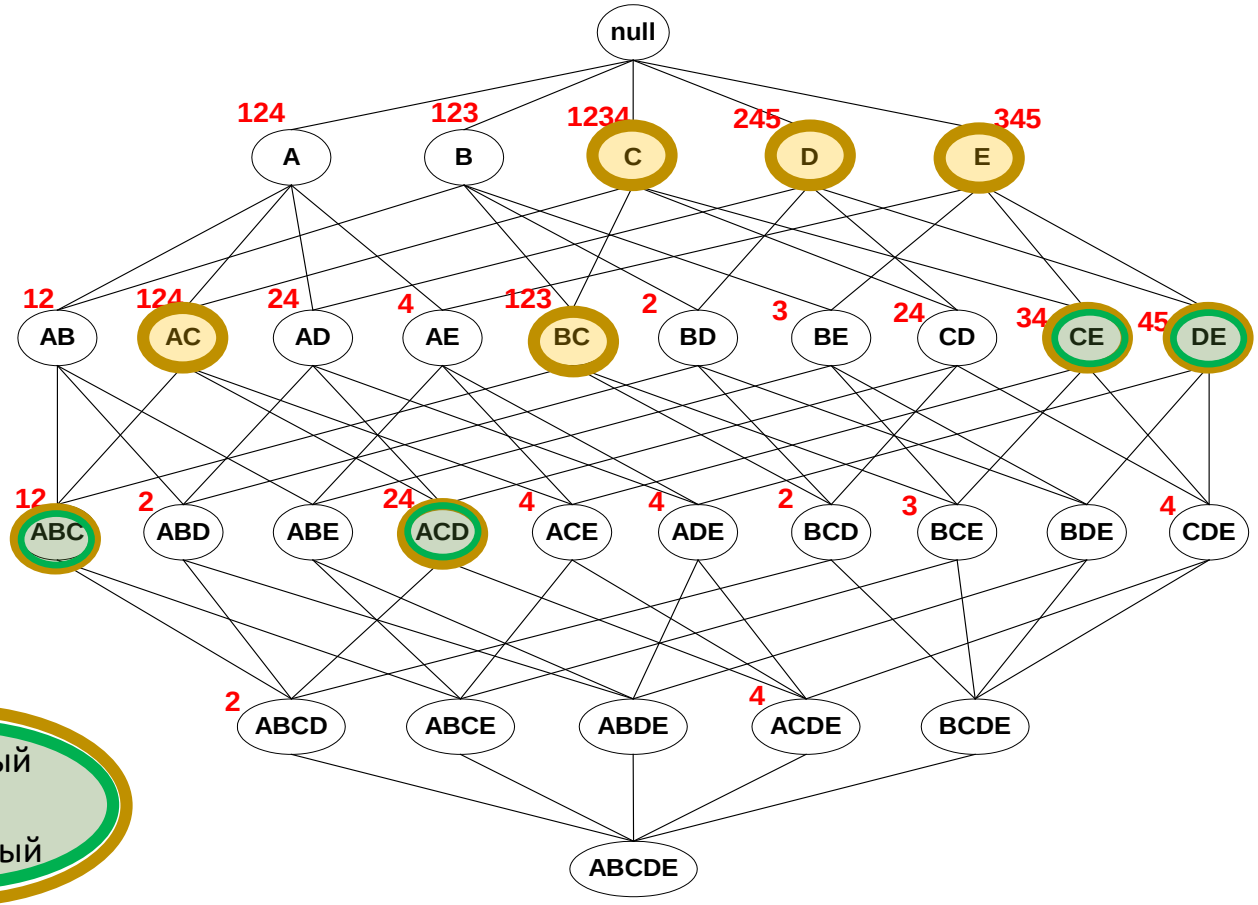
      $I.sup := \max\{J.sup \mid J \in \mathcal{L}_{k+1} \wedge I \subseteq J\}$

Замкнутые и максимально частые наборы

- *Максимально частый набор:* частый и не является непосредственным подмножеством частого набора
- *Замкнутый набор:* отсутствует его непосредственное надмножество с той же поддержкой

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

$minsup = 2$



макс частый

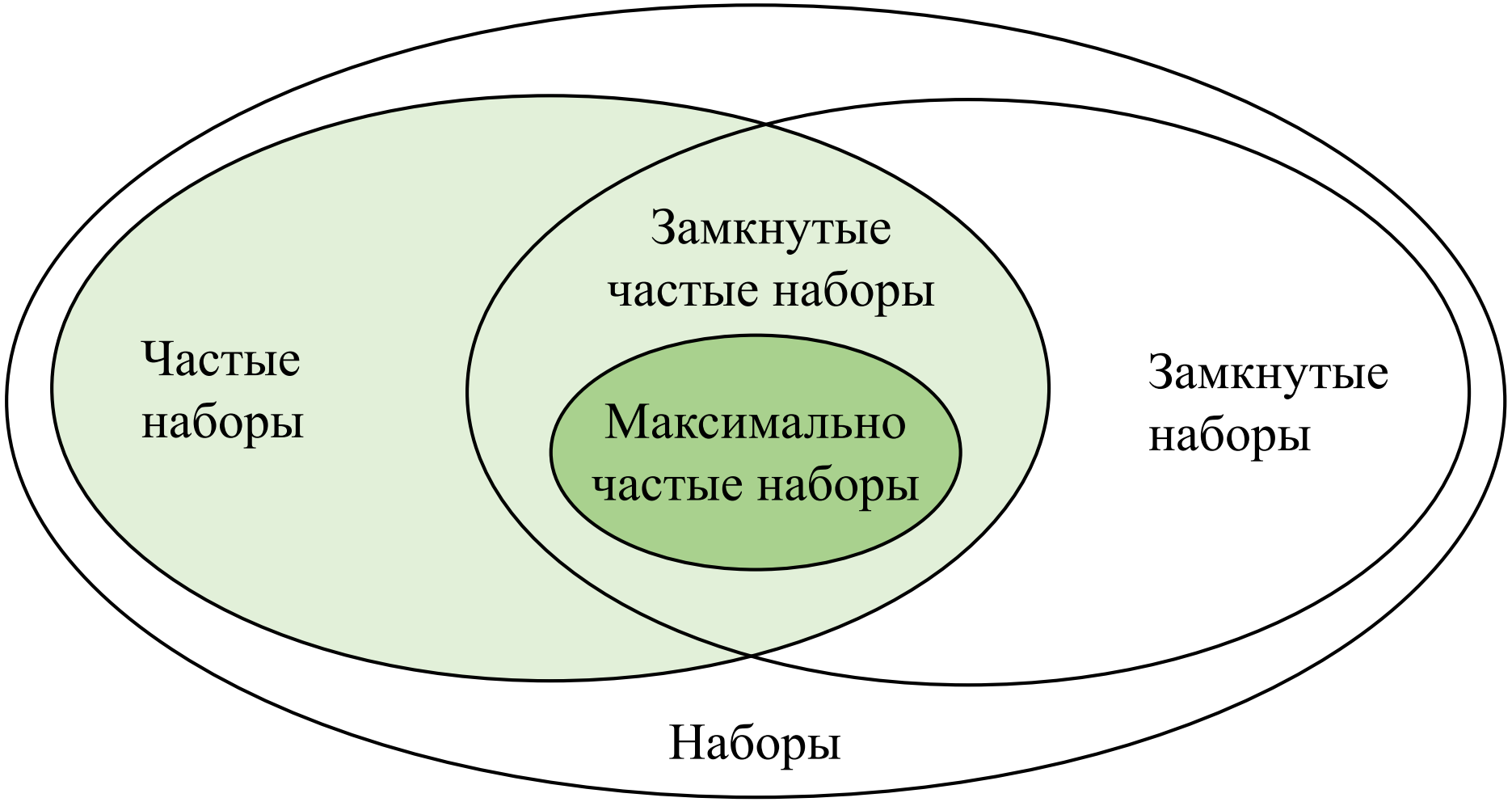
замкнутый
и
макс частый

Замкнутые частые наборы

TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

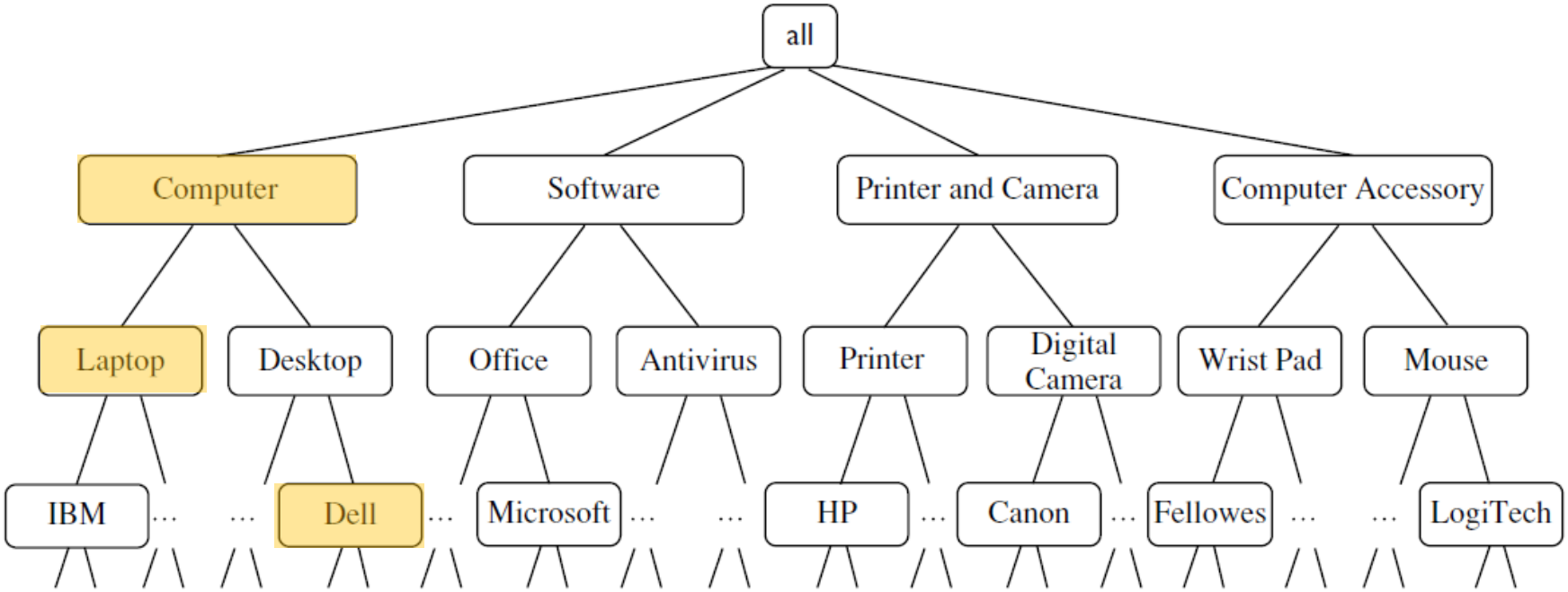
- $minsup = 5$
- Замкнутые частые:
  $\{A1, \dots, A10\}, \{B1, \dots, B10\}, \{C1, \dots, C10\}$
- Все частые: $|\mathcal{L}| = 3069$

Взаимосвязь понятий

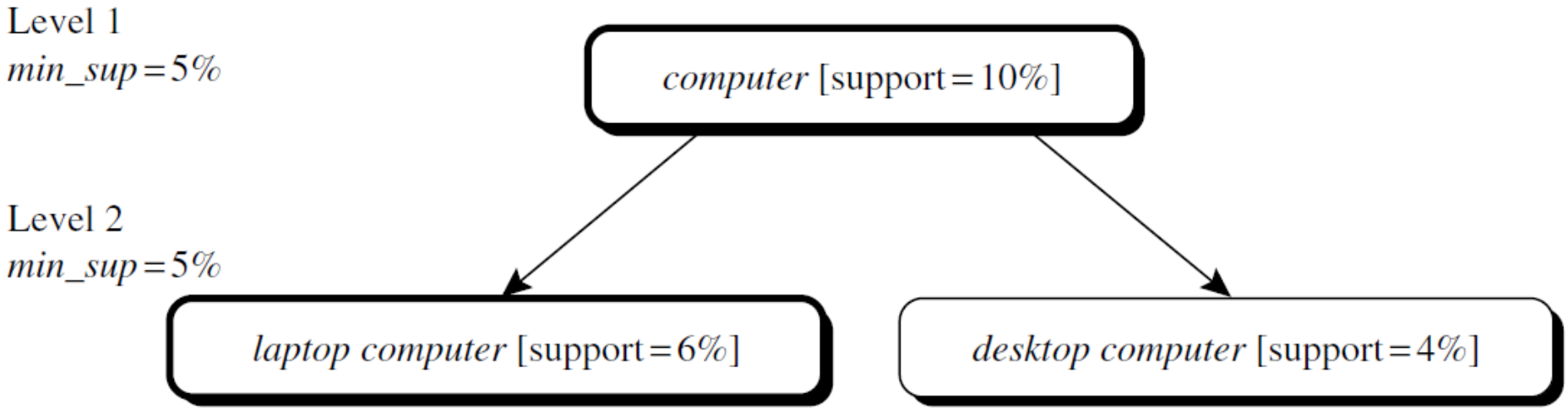


Иерархии в объектах

TID	Items Purchased
T100	Apple 17″ MacBook Pro Notebook, HP Photosmart Pro b9180
T200	Microsoft Office Professional 2010, Microsoft Wireless Optical Mouse 5000
T300	Logitech VX Nano Cordless Laser Mouse, Fellowes GEL Wrist Rest
T400	Dell Studio XPS 16 Notebook, Canon PowerShot SD1400
T500	Lenovo ThinkPad X200 Tablet PC, Symantec Norton Antivirus 2010

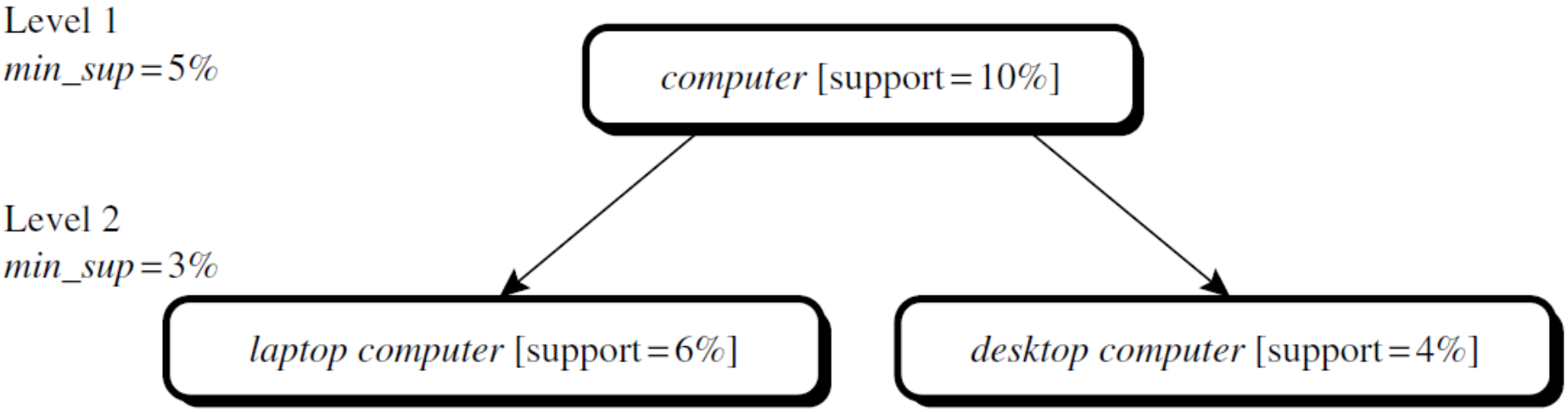


Поиск с унифицированной поддержкой



- Упрощение поиска с одним порогом $minsup$
  - отбрасывание наборов с редким «родителем»
- «Потомок» не обязательно настолько частый, как «предок»
  - слишком большой порог $minsup$ – потеря устойчивых шаблонов на нижних уровнях иерархии,
  - слишком маленький порог $minsup$ – много малозначимых шаблонов на верхних уровнях иерархии

Поиск с уменьшенной поддержкой для нижних уровней



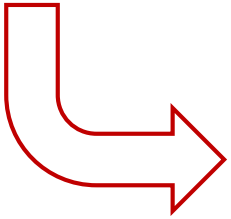
- Уменьшение порога *minsup* при продвижении от «предков» к «потомкам»

Поиск с различной поддержкой, фильтрация шаблонов

- Ценность *vs.* частота объектов
  - Порог $minsup$ назначается разным для различных семантических групп объектов
  - Пример: $minsup = 0.0005$ для $\{diamond, caviar, \dots\}$, $minsup = 0.05$ для $\{bread, butter, \dots\}$
- Отбрасывание избыточных шаблонов
  - шаблон избыточен, если его антецедент (консеквент) является потомком антецедента (консеквента) другого шаблона и имеет сравнительно одинаковые с ним поддержку и достоверность
  - Пример:
    - $laptop \rightarrow printer$ [sup $= 0.08, conf = 0.70$]
    - $Dell\ laptop \rightarrow HP\ printer$ [sup $= 0.02, conf = 0.72$] **избыточно**

Поиск шаблонов в категориальных данных

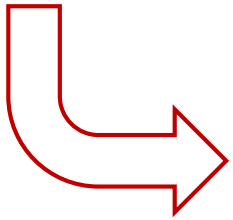
Gender	Level of Education	State	Computer at Home	Online Auction	Chat Online	Online Banking	Privacy Concerns
Female	Graduate	Illinois	Yes	Yes	Daily	Yes	Yes
Male	College	California	No	No	Never	No	No
Male	Graduate	Michigan	Yes	Yes	Monthly	Yes	Yes
Female	College	Virginia	No	Yes	Never	Yes	Yes
Female	Graduate	California	Yes	No	Never	No	Yes
Male	College	Minnesota	Yes	Yes	Weekly	Yes	Yes
Male	College	Alaska	Yes	Yes	Daily	Yes	No
Male	High School	Oregon	Yes	No	Never	No	No
Female	Graduate	Texas	No	No	Monthly	No	No
...	...	...	...	...	...	...	...



Male	Female	Education = Graduate	Education = College	Education = High School	...	Privacy = Yes	Privacy = No
0	1	1	0	0	...	1	0
1	0	0	1	0	...	0	1
1	0	1	0	0	...	1	0
0	1	0	1	0	...	1	0
0	1	1	0	0	...	1	0
1	0	0	1	0	...	1	0
1	0	0	0	0	...	0	1
1	0	0	0	1	...	0	1
0	1	1	0	0	...	0	1
...	...	...	...	...	...	...	...

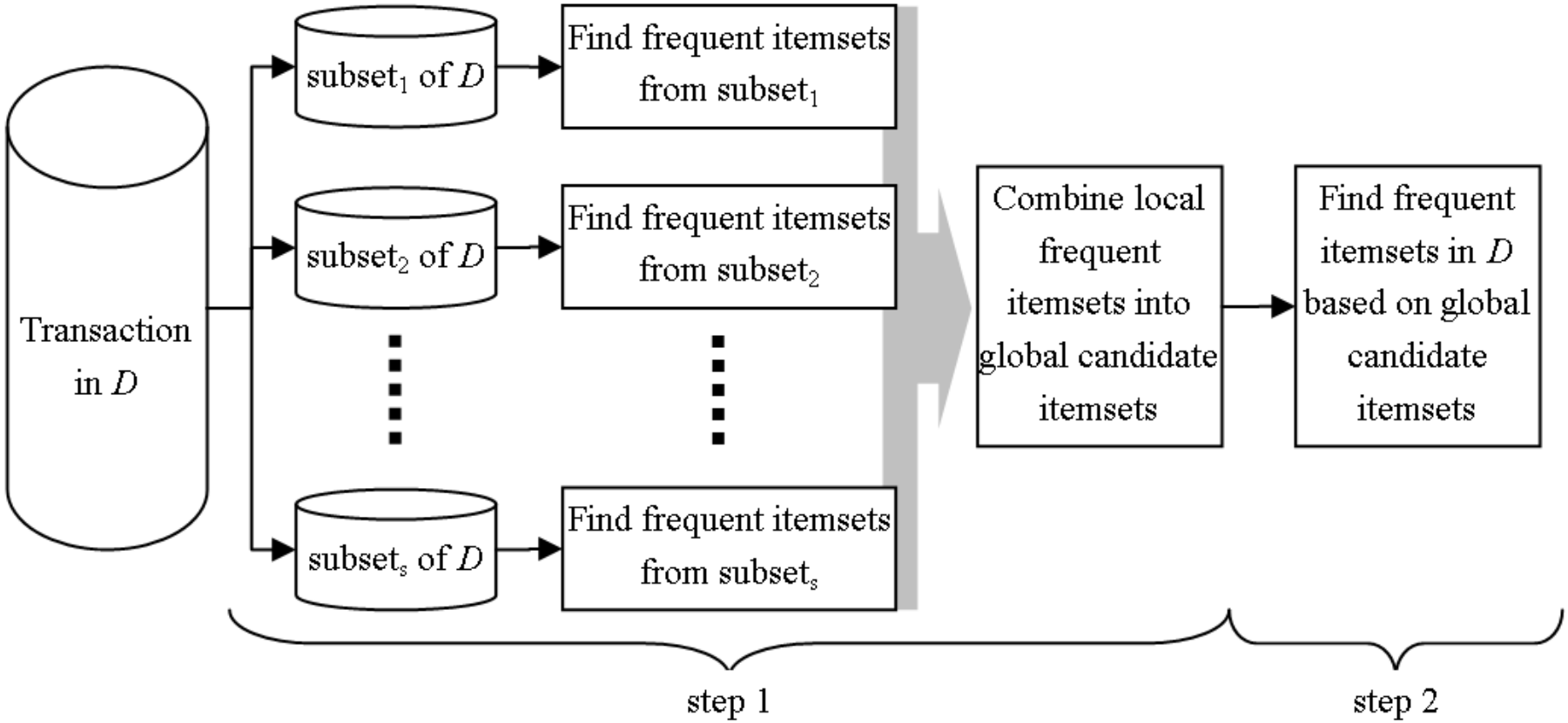
Поиск шаблонов в непрерывных атрибутах

Gender	···	Age	Annual Income	No of hours spent online per week	No of email accounts	Privacy Concern
Female	···	26	90K	20	4	Yes
Male	···	51	135K	10	2	No
Male	···	29	80K	10	3	Yes
Female	···	45	120K	15	3	Yes
Female	···	31	95K	20	5	Yes
Male	···	25	55K	25	5	Yes
Male	···	37	100K	10	1	No
Male	···	41	65K	8	2	No
Female	···	26	85K	12	1	No
···	···	···	···	···	···	···



Male	Female	··· ···	Age < 13	Age $\in [13, 21)$	Age $\in [21, 30)$	··· ···	Privacy = Yes	Privacy = No
0	1	···	0	0	1	···	1	0
1	0	···	0	0	0	···	0	1
1	0	···	0	0	1	···	1	0
0	1	···	0	0	0	···	1	0
0	1	···	0	0	0	···	1	0
1	0	···	0	0	1	···	1	0
1	0	···	0	0	0	···	0	1
1	0	···	0	0	0	···	0	1
0	1	···	0	0	1	···	0	1
···	···	···	···	···	···	···	···	···

Фрагментация базы транзакций



- $minsup\_count_i = minsup \cdot \frac{|D|}{s}$

Сэмплинг базы транзакций

- Выбрать сэмпл $S \subseteq D$, $|S| \leq |RAM|$
- Взять $minsup_S < minsup$ (чтобы пропустить меньше частых наборов из D) и найти $\mathcal{L}^S$
- $\forall I \in \mathcal{L}^S$ найти в $D \setminus S$ $\sup(I)$ для $minsup$
- Если $\mathcal{L}^S$ не содержит все частые наборы из D, выполнить сэмплинг повторно

Литература

- Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, 2011. 740 p. ISBN 978-0123814791
  - Chapter 7. Advanced Pattern Mining, pp. 279-326
- Tan P.-N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd Edition. Pearson, 2019. 839 p. ISBN-13: 978-0-13-312890-1
  - 6. Association Analysis: Advanced Concepts, pp. 451-524