ПОИСК АССОЦИАТИВНЫХ ПРАВИЛ

1. Основные понятия

Одним из распространенных методов Data Mining является аффинитивный анализ (affinity analysis), целью которого является исследование взаимной связи между событиями, которые происходят совместно.

Анализ рыночной корзины (market basket analysis) является разновидностью аффинитивного анализа.

Целью анализа рыночной корзины является обнаружение ассоциаций между различными событиями посредством поиска ассоциативных правил, позволяющих количественно описать взаимную связь между двумя и более событиями. В контексте задачи поиска ассоциативных правил под транзакцией понимают множество событий, которые произошли одновременно. Таким событием, например, может быть покупка того или иного товара (предмета). Тогда рыночная корзина – это набор товаров (предметов), приобретенных в рамках одной отдельно взятой транзакции.

Транзакционная база данных (база данных транзакций, БД транзакций, transaction database) – двумерная таблица, содержащая номер транзакции и перечень покупок, приобретенных во время этой транзакции.

Пусть число транзакций в базе – N, и каждая транзакция D_i представляет собой некоторый набор предметов ($i = \overline{1, N}$).

В таблице 1 приведен пример набора транзакций (базы данных транзакций) из 20 предметов (товаров). Такой набор может быть сформирован, например, на основе чеков о покупке товаров (предметов) в магазине «Канцтовары». Следует отметить, что этот набор содержит очень мало транзакций (всего 20), что не позволяет считать выводы, сделанные на его основе объективными. Однако этот набор позволяет в полной мере продемонстрировать работу по поиску ассоциативных правил.

Любое ассоциативное правило состоит из двух наборов предметов, называемых условием (antecedent) A и следствием (consequent) B.

Ассоциативное правило записывается в виде $A \rightarrow B$, что читается как «из A следует B» или «Если *условие* A, то *следствие* B».

Таблица 1. Пример набора транзакций на основе чеков магазина «Канцтовары»

Номер чека	Список предметов в чеке							
1	Карандаш	Ластик	Альбом	Линейка	Тетрадь			
2	Ручка	Тетрадь	Карандаш	Ластик	Пенал	Линейка		
3	Карандаш	Ластик	Ручка					
4	Краски	Альбом	Палитра	Карандаш	Кисть	Ластик		
5	Карандаш	Ластик	Тетрадь	Ручка				
6	Кисть	Краски	Палитра	Альбом				
7	Краски	Альбом	Карандаш	Ластик	Тетрадь			
8	Карандаш	Ластик	Линейка					
9	Тетрадь	Ручка	Пенал	Ластик	Карандаш			
10	Ластик	Карандаш	Тетрадь	Пенал	Кисть	Альбом	Ручка	Краски
11	Линейка	Карандаш	Ластик	Тетрадь				
12	Пенал	Ластик	Тетрадь	Карандаш				
13	Ручка	Пенал						
14	Ручка	Альбом	Линейка					
15	Альбом	Тетрадь	Карандаш	Ластик				
16	Карандаш	Ластик	Ручка	Тетрадь	Линейка			
17	Краски	Палитра	Альбом	Ластик	Карандаш	Тетрадь		
18	Карандаш	Краски	Альбом	Ластик				
19	Пенал	Линейка						
20	Карандаш	Ручка	Тетрадь	Ластик				

Ассоциативные правила описывают связь между наборами предметов, соответствующими условию *A* и следствию *B*.

Эта связь характеризуется такими показателями, как поддержка (support) *S* и достоверность (confidence) *C*.

Поддержка *S* – доля транзакций, содержащих одновременно условие *A* и следствие *B*:

$$S(A \to B) = P(A \cap B) = \frac{\text{число транзакций, содержащих } A \text{ и } B}{\text{общее число транзакций}}.$$

Достоверность *C* – доля транзакций, содержащий и условие *A*, и следствие *B*, к числу транзакций, содержащих только условие *A*.

$$C(A \to B) = P(B \mid A) = P(A \cap B) / P(A) =$$

$$= \frac{\text{число транзакций, содержащих } A \text{ и } B}{\text{число транзакций, содержащих } \textit{только } A}.$$

Рассмотрим ассоциацию *Карандаш* → *Ластик*.

Очевидно, что эти продукты (товары) хорошо совместимы.

Так как число транзакций, содержащих как *Карандаш*, так и *Ластик*, равно 16, а общее число транзакций – 20 (таблица 2), то поддержка этой ассоциации будет:

S(*Карандаш* → *Ластик*) = 16/20 = 4/5=0,8.

Так как число транзакций, содержащих только *Карандаш* (условие), равно 16 (таблица 2), то достоверность этой ассоциации будет:

C(*Карандаш* → *Ластик*)=16/16=1.

Все наблюдения, содержащие *Карандаш*, содержат и *Ластик*, поэтому такая ассоциация может рассматриваться как ***правило***.

Теперь рассмотрим ассоциацию *Кисть* → *Ластик*.

Очевидно, что у этих продуктов (товаров) совместимость меньше, чем у продуктов (товаров) *Карандаш* и *Ластик*.

Так как число транзакций, содержащих как *Кисть*, так и *Ластик*, равно 2, а общее число транзакций – 20 (таблица 3), то поддержка этой ассоциации будет:

S(*Кисть* → *Ластик*) = 2/20 = 1/10=0,1.

Так как число транзакций, содержащих только *Кисть* (условие), равно 3 (таблица 3), то достоверность этой ассоциации будет:

C(*Кисть* → *Ластик*)=2/3=0,67.

Таблица 2. Пример набора транзакций на основе чеков магазина «Канцтовары» с выделенными предметами Карандаш и Ластик

Номер чека	Список предметов в чеке							
1	Карандаш	Ластик	Альбом	Линейка	Тетрадь			
2	Ручка	Тетрадь	Карандаш	Ластик	Пенал	Линейка		
3	Карандаш	Ластик	Ручка					
4	Краски	Альбом	Палитра	Карандаш	Кисть	Ластик		
5	Карандаш	Ластик	Тетрадь	Ручка				
6	Кисть	Краски	Палитра	Альбом				
7	Краски	Альбом	Карандаш	Ластик	Тетрадь			
8	Карандаш	Ластик	Линейка					
9	Тетрадь	Ручка	Пенал	Ластик	Карандаш			
10	Ластик	Карандаш	Тетрадь	Пенал	Кисть	Альбом	Ручка	Краски
11	Линейка	Карандаш	Ластик	Тетрадь				
12	Пенал	Ластик	Тетрадь	Карандаш				
13	Ручка	Пенал						
14	Ручка	Альбом	Линейка					
15	Альбом	Тетрадь	Карандаш	Ластик				
16	Карандаш	Ластик	Ручка	Тетрадь	Линейка			
17	Краски	Палитра	Альбом	Ластик	Карандаш	Тетрадь		
18	Карандаш	Краски	Альбом	Ластик				
19	Пенал	Линейка						
20	Карандаш	Ручка	Тетрадь	Ластик				

Таблица 3. Пример набора транзакций на основе чеков магазина «Канцтовары» с выделенными предметами Кисть и Ластик

Номер чека	Список предметов в чеке							
1	Карандаш	Ластик	Альбом	Линейка	Тетрадь			
2	Ручка	Тетрадь	Карандаш	Ластик	Пенал	Линейка		
3	Карандаш	Ластик	Ручка					
4	Краски	Альбом	Палитра	Карандаш	Кисть	Ластик		
5	Карандаш	Ластик	Тетрадь	Ручка				
6	Кисть	Краски	Палитра	Альбом				
7	Краски	Альбом	Карандаш	Ластик	Тетрадь			
8	Карандаш	Ластик	Линейка					
9	Тетрадь	Ручка	Пенал	Ластик	Карандаш			
10	Ластик	Карандаш	Тетрадь	Пенал	Кисть	Альбом	Ручка	Краски
11	Линейка	Карандаш	Ластик	Тетрадь				
12	Пенал	Ластик	Тетрадь	Карандаш				
13	Ручка	Пенал						
14	Ручка	Альбом	Линейка					
15	Альбом	Тетрадь	Карандаш	Ластик				
16	Карандаш	Ластик	Ручка	Тетрадь	Линейка			
17	Краски	Палитра	Альбом	Ластик	Карандаш	Тетрадь		
18	Карандаш	Краски	Альбом	Ластик				
19	Пенал	Линейка						
20	Карандаш	Ручка	Тетрадь	Ластик				

Невысокая достоверность этой ассоциации дает повод усомниться в том, что она является правилом. При этом поддержка этой ассоциации является совсем маленькой.

При анализе ассоциаций наиболее часто предпочтение отдается тем, которые имеют высокие значения поддержки и достоверности. Хотя в ряде случаев могут выбираться и ассоциации, имеющие только высокую поддержку или только высокую достоверность. Правила, значения поддержки (достоверности) которых превышают пороговое значение, заданное пользователем, называются *сильными* правилами (strong rules).

Например, при анализе ассоциаций на основе чеков о покупках в магазине с целью формирования рекомендаций по формированию успешного плана его деятельности могут быть интересны ассоциации, для которых значения поддержки и достоверности не ниже пороговых значений в 25% и 75% соответственно.

При анализе ассоциаций с целью выявления мошеннических действий могут быть интересны ассоциации, значение поддержки которых невелико (например, не больше 1%), так как с такими действиями связано малое число транзакций.

Значимость ассоциации может быть вычислена как разность между поддержкой правила в целом и произведением поддержки *только* условия и поддержки *только* следствия.

Если условие и следствие независимы, то поддержка ассоциации примерно соответствует произведению поддержек условия A и следствия B:

$S_{AB} \approx S_A \cdot S_B$.

В таком случае можно будет сделать следующий вывод: хотя условие A и следствие B часто встречаются вместе, не менее часто они встречаются и по отдельности.

Например, если товар A встречался в 75 транзакциях из 100, а товар B – в 85, и в 64 транзакциях из 100 они встречаются вместе, то, несмотря на высокое значение поддержки (S_{AB} = 0,64), ассоциация $A \rightarrow B$ не обязательно является правилом: товары A и B покупаются независимо друг от друга, но по причине их популярности часто встречаются в одной транзакции. Поскольку произведение значений поддержек условия A и следствия B вычисляется как $S_A \cdot S_B$ = 0,75·0,85 = 0,6375, то есть отличается от S_{AB} = 0,64 всего на 0,0025,

предположение о независимости товаров A и B можно считать достаточно обоснованным.

Если условие A и следствие B независимы, то ассоциация скорее всего не может считаться правилом, даже если она имеет высокие значения поддержки и достоверности.

При оценке **значимости ассоциаций применяют объективные и субъективные меры значимости.**

Объективные меры – поддержка и достоверность. Такие меры значимости могут применяться в любой предметной области.

Субъективные меры – лифт (lift) и левередж (leverage, плечо, рычаг). Такие меры значимости связаны напрямую связаны с информацией, определяемой контекстом решаемой задачи анализа данных.

Лифт – отношение частоты появления условия A в транзакциях, которые также содержат и следствие B, к частоте появления следствия B в целом:

$$L(A \to B) = C(A \to B) / S(B).$$

Значения лифта, большие чем 1, показывают, что условие чаще появляется в транзакциях, содержащих следствие B, чем в остальных.

Лифт является обобщенной мерой связи двух предметных наборов, фигурирующих в числителе и знаменателе формулы: при значениях лифта, превосходящих 1, связь положительная, при значении лифта, равном 1, связь отсутствует, а при значениях лифта, меньших 1, связь отрицательная.

Рассмотрим ассоциацию *Ластик* $\to$ *Карандаш* и вычислим для нее лифт.

S(*Карандаш*)=16/20=0,8; C(*Ластик* $\to$ *Карандаш*)=16/16=1.

Следовательно, L(*Ластик* $\to$ *Карандаш*)=1/0,8=1,25.

Рассмотрим ассоциацию *Ластик* $\to$ *Кисть* и вычислим для нее лифт.

S(*Кисть*)=3/20=0,15; C(*Ластик* $\to$ *Кисть*)=2/16=0,125.

Следовательно, L(*Ластик* $\to$ *Кисть*)=0,125/0,15=0,833.

Для первой ассоциации лифт больше 1, для второй – меньше 1, значит, ластик больше влияет на покупку карандаша, чем на покупку кисти.

Следует отметить, что лифт не всегда является удачной мерой значимости ассоциации. Так, например, ассоциация с меньшим значением поддержкой и большим значением лифта может быть менее значимой, чем ассоциация с большим значением поддержки и меньшим значением лифта, так как что вторая ассоциация используется для большего числа транзакций. Следо-

вательно, увеличение числа транзакций приводит к увеличению связи между условием и следствием ассоциации.

Левередж – разность между наблюдаемой частотой, с которой условие и следствие появляются совместно, и произведением частот появления условия и следствия по отдельности, то есть разность между поддержкой ассоциации и произведением поддержек условия и следствия по отдельности:

$Lev(A \to B) = S(A \to B) - S(A) \cdot S(B)$.

Рассмотрим ассоциации *Карандаш* $\to$ *Ластик* и *Тетрадь* $\to$ *Ластик*.

Эти ассоциации имеют одинаковую *достоверность* $C(A \to B) = 1$, т.к. карандаш (*всего 16 раз*) (таблица 2) и тетрадь (*всего 12 раз*) (таблица 4) всегда продаются вместе с ластиком (16 и 12 соответственно).

Эти ассоциации имеют одинаковые *лифты*, так как в обеих ассоциациях поддержка следствия *S*(*Ластик*)=16/20=0,8 (*ластик встретился в транзациях 16 раз*), и, следовательно,

L(*Карандаш* $\to$ *Ластик*)=*L*(*Тетрадь* $\to$ *Ластик*)=1/0,8=1,25.

Вычислим левереджи для этих ассоциаций.

S(*Карандаш* $\to$ *Ластик*)=16/20=0,8.

S(*Карандаш*)=16/20=0,8. *S*(*Ластик*)=16/20=0,8.

Следовательно, *Lev*(*Карандаш* $\to$ *Ластик*)=0,8 – 0,8·0,8=0,16.

S(*Тетрадь* $\to$ *Ластик*)=12/20=0,6.

S(*Тетрадь*)=12/20=0,6. *S*(*Ластик*)=16/20=0,8.

Следовательно, *Lev*(*Тетрадь* $\to$ *Ластик*)=0,6 – 0,6·0,8=0,12.

Ассоциация *Карандаш* $\to$ *Ластик* представляет больший интерес, так как встречается чаще (то есть применяется в большем числе транзакций).

Таким образом, значимость ассоциации *Карандаш* $\to$ *Ластик* более высокая.

Иногда вместо лифта используют еще одну субъективную меру – улучшение (improvement), которую вычисляют подобно левереджу, но берут не разность, а отношение между наблюдаемой частотой, с которой условие и следствие появляются совместно, и произведением частот появления условия и следствия по отдельности:

$$I(A \to B) = \frac{S(A \to B)}{S(A) \cdot S(B)}.$$

Таблица 4. Пример набора транзакций на основе чеков магазина «Канцтовары»

Номер чека	Список предметов в чеке							
1	Карандаш	Ластик	Альбом	Линейка	Тетрадь			
2	Ручка	Тетрадь	Карандаш	Ластик	Пенал	Линейка		
3	Карандаш	Ластик	Ручка					
4	Краски	Альбом	Палитра	Карандаш	Кисть	Ластик		
5	Карандаш	Ластик	Тетрадь	Ручка				
6	Кисть	Краски	Палитра	Альбом				
7	Краски	Альбом	Карандаш	Ластик	Тетрадь			
8	Карандаш	Ластик	Линейка					
9	Тетрадь	Ручка	Пенал	Ластик	Карандаш			
10	Ластик	Карандаш	Тетрадь	Пенал	Кисть	Альбом	Ручка	Краски
11	Линейка	Карандаш	Ластик	Тетрадь				
12	Пенал	Ластик	Тетрадь	Карандаш				
13	Ручка	Пенал						
14	Ручка	Альбом	Линейка					
15	Альбом	Тетрадь	Карандаш	Ластик				
16	Карандаш	Ластик	Ручка	Тетрадь	Линейка			
17	Краски	Палитра	Альбом	Ластик	Карандаш	Тетрадь		
18	Карандаш	Краски	Альбом	Ластик				
19	Пенал	Линейка						
20	Карандаш	Ручка	Тетрадь	Ластик				

Улучшение $I(A \to B)$ показывает, полезнее ли ассоциация случайного угадывания. Если улучшение $I(A \to B)$ больше 1, то то означает, что вероятнее предсказать наличие набора *B* в следствии с помощью ассоциации, чем угадать случайно.

Такие меры, как лифт $L(A \to B)$, левередж $Lev(A \to B)$ и улучшение $I(A \to B)$, могут использоваться для последующего сокращения числа рассматриваемых ассоциаций посредством установки порогового значения для используемой меры значимости: ассоциации, у которых значение меры значимости ниже порогового, исключаются из дальнейшего рассмотрения.

2. Алгоритм Apriori

Число возможных ассоциаций растет экспоненциально с увеличением числа предметов, на основе которых могут формироваться ассоциации.

Пусть при общее число предметов в базе данных транзакций равно *k*.

Пусть все ассоциации содержат по одному предмету в условии и следствии.

В этом случае требуется проанализировать $k \cdot 2^{k-1}$ ассоциаций.

Пусть, например, *k*=100. Тогда число возможных ассоциаций равно:

$100 \cdot 2^{99} \approx 6{,}4 \cdot 10^{31}$.

Очевидно, что в условии и следствии может содержаться большее число предметов. Следовательно, число возможных ассоциаций будет еще бóльшим.

Исходя из вышесказанного, можно сделать следующий вывод: поиск ассоциативных правил посредством вычисления поддержки и достоверности для всех возможных ассоциаций и сравнения их с заданным пороговым значением будет малоэффективным из-за больших вычислительных затрат.

Обычно при поиске ассоциативных правил используется методика, предполагающая выявление частых предметных наборов. При реализации такой методики анализируются только те ассоциации, которые встречаются часто. При этом задается некоторое пороговое значение, позволяющее отнести предметных набор к частым или к нечастым.

Одним из алгоритмов, основанных на применении методики выявления частых предметных наборов, является алгоритм Apriori.

Этот алгоритм предложили ученые Ракеш Агравал (Rakesh Agrawal) и Рамакришнан Шрикант (Ramakrishnan Srikant) в 1994 году.

Цель алгоритма Apriori – сузить пространство поиска до размеров, обеспечивающих приемлемые временные затраты на поиск ассоциативных правил.

2.1. Выявление частых предметных наборов

В основе алгоритма Apriori лежит понятие частого набора (frequent itemset), который также можно назвать частым предметным набором, для которого определяется частота его появления в базе транзакций.

Под частотой понимается простое число транзакций, в которых содержится рассматриваемый предметный набор.

Частым предметным набором будет тот, который встречается чаще, чем в числе транзакций, описываемом некоторой пороговой частотой Δ.

Кроме того, можно считать, что частый предметный набор – предметный набор с поддержкой, большей заданного порогового значения либо равной ему.

Это пороговое значение называется минимальной поддержкой.

Методика поиска ассоциативных правил с использованием частых наборов содержит следующую последовательность шагов.

Шаг 1. Поиск частых предметных наборов.

Шаг 2. Генерация на основе частых предметных наборов ассоциативных правил, удовлетворяющих условиям минимальной поддержки и достоверности.

Алгоритм Apriori использует свойство антимонотонности, чтобы сузить пространство поиска.

Свойство антимонотонности: если предметный набор Z не является частым, то добавление некоторого нового предмета A к набору Z не делает его более частым.

Если набор Z не является частым, то и набор $Z \cup A$ также не будет являться частым.

Использование свойства антимонотонности позволяет значительно уменьшить пространство поиска ассоциативных правил.

Рассмотрим набор транзакций, приведенный в таблице 1, и применим к нему алгоритм Apriori.

Таблица 5. Представление данных из набора транзакций в нормализованном виде

№ чека	Альбом	Карандаш	Кисть	Краски	Ластик	Линейка	Палитра	Пенал	Ручка	Тетрадь
1	1	1	0	0	1	1	0	0	0	1
2	0	1	0	0	1	1	0	1	1	1
3	0	1	0	0	1	0	0	0	1	0
4	1	1	1	1	1	0	1	0	0	0
5	0	1	0	0	1	0	0	0	1	1
6	1	0	1	1	0	0	1	0	0	0
7	1	1	0	1	1	0	0	0	0	1
8	0	1	0	0	1	1	0	0	0	0
9	0	1	0	0	1	0	0	1	1	1
10	1	1	1	1	1	0	0	1	1	1
11	0	1	0	0	1	1	0	0	0	1
12	0	1	0	0	1	0	0	1	0	1
13	0	0	0	0	0	0	0	1	1	0
14	1	0	0	0	0	1	0	0	1	0
15	1	1	0	0	1	0	0	0	0	1
16	0	1	0	0	1	1	0	0	1	1
17	1	1	0	1	1	0	1	0	0	1
18	1	1	0	1	1	0	0	0	0	0
18	0	0	0	0	0	1	0	1	0	0
20	0	1	0	0	1	0	0	0	1	1
Сумма по столбцу	9	16	3	6	16	7	3	6	9	12

Будем считать частыми такие предметные наборы, которые встречаются в в базе транзакций более 5 раз, то есть определим пороговое значение частоты как $\Delta = 5$.

Выполним поиск ассоциативных правил, реализуя следующие шаги.

Шаг 1. Выполним поиск частых однопредметных наборов, представив набор транзакций в нормализованном виде (таблица 5).

На пересечении строки транзакции (строки с номером чека) и столбца с названием предмета поставим число «1», если предмет присутствует в транзакции, и «0» – в противном случае.

Просуммируем значения в каждом столбце для вычисления частоты появления каждого предмета в наборе транзакций.

Запишем частоту появления каждого предмета в наборе транзакций в соответствующую ячейку последней строки таблицы 5.

Как видно из таблицы 5, сумма в каждом столбце, за исключением столбцов с названиями предметов **Кисть** и **Палитра** больше порогового значения частоты $\Delta = 5$.

Все предметы, для которых частота появления в наборе транзакций больше или равна $\Delta = 5$, будем считать частыми однопредметными наборами.

Сформируем множество F_1 частых однопредметных наборов:

F_1={Альбом, Карандаш, Краски, Ластик, Линейка, Пенал, Ручка, Тетрадь}.

Шаг 2. Выполним поиск частых двухпредметных наборов (таблица 6).

В общем случае для формирования множества F_k частых *k*-предметных наборов реализуют следующий алгоритм:

– создать множество наборов-кандидатов в частные *k*-предметные наборы посредством связывания множества F_{k-1} с самим собой;

– выполнить редукцию множества F_k, используя свойство антимонотонности.

Оставшиеся предметные наборы формируют искомое множество F_k частых *k*-предметных наборов.

Так как пороговое значение частоты $\Delta = 5$, то в множество F_2 войдут только те двухпредметные наборы, которые встречаются 5 и более раз (таблица 6). Информация по найденным частным двухпредметным наборам выделена в таблице 6 жирным шрифтом.

Таблица 6. Двухпредметные наборы-кандидаты в F_2

Предметный набор		Частота появления в наборе тракзакций
Предмет 1	Предмет 2	
Альбом	**Карандаш**	**7**
Альбом	**Краски**	**6**
Альбом	**Ластик**	**7**
Альбом	Линейка	2
Альбом	Ручка	2
Альбом	**Тетрадь**	**5**
Карандаш	**Краски**	**5**
Карандаш	**Ластик**	**16**
Карандаш	**Линейка**	**5**
Карандаш	**Ручка**	**7**
Карандаш	**Тетрадь**	**12**
Краски	**Ластик**	**5**
Краски	Линейка	0
Краски	Ручка	1
Краски	Тетрадь	3
Ластик	**Линейка**	**5**
Ластик	**Ручка**	**7**
Ластик	**Тетрадь**	**12**
Линейка	Ручка	3
Линейка	Тетрадь	4
Ручка	**Тетрадь**	**6**

Сформируем множество F_2 частых двухпредметных наборов:

F_2 {{Альбом, Карандаш}, {Альбом, Краски}, {Альбом, Ластик},
{Альбом, Тетрадь}, {Карандаш, Краски}, {Карандаш, Ластик},
{Карандаш, Линейка}, {Карандаш, Ручка}, { Карандаш, Тетрадь},
{Краски, Ластик}, {Ластик, Линейка}, {Ластик, Ручка},
{Ластик, Тетрадь}, {Ручка, Тетрадь}}.

Шаг 3. Выполним поиск частых трехпредметных наборов (таблица 7).

Для получения множества F_3 свяжем множество F_2 с самим собой.

В общем случае *k*-предметные наборы являются связываемыми, если у них первые $k-1$ предметов, следующих в алфавитном порядке, общие.

Так, наборы {**Альбом**, **Карандаш**} и {**Альбом**, **Краски**}, для которых $k=2$, чтобы быть связываемыми, должны иметь один ($k-1=1$) общий первый предмет, которым и является **Альбом**.

В результате связывания получим:

{**Альбом**, **Карандаш**} + {**Альбом**, **Краски**}={**Альбом**, **Карандаш**, **Краски**}.

Так как пороговое значение частоты $\Delta = 5$, то в множество F_3 войдут только те трехпредметные наборы, которые встречаются 5 и более раз (таблица 7). Информация по найденным частным двухпредметным наборам выделена в таблице 7 жирным шрифтом.

Множество трехпредметных наборов-кандидатов в F_3 сокращается с помощью свойства антимонотонности: поддержка любого набора предметов не может превышать минимальной поддержки любого из его поднаборов.

С увеличением числа предметов в наборе поддержка уменьшается или остается такой же.

В общем случае любой *k*-предметный набор будет часто встречающимся тогда и только тогда, когда все его $(k-1)$-предметные поднаборы будут часто встречающимися.

Рассмотрим набор {Альбом, Карандаш, Краски}.

В нем есть двухпредметные поднаборы: {Альбом, Карандаш}, {Альбом, Краски}, {Карандаш, Краски}, частота появления которых в наборе тракзакций соответственно равна 7, 6 и 7. Следовательно, эти поднаборы частые и трехпредметный набор {Альбом, Карандаш, Краски} не подлежит изъятию из дальнейшего рассмотрения.

Таблица 7. Трехпредметные наборы-кандидаты в F_3

Предметный набор			Частота появления в наборе тракзакций
Предмет 1	Предмет 2	Предмет 3	
Альбом	**Карандаш**	**Краски**	**5**
Альбом	**Карандаш**	**Ластик**	**7**
Альбом	**Карандаш**	**Тетрадь**	**5**
Альбом	**Краски**	**Ластик**	**5**
Альбом	Краски	Тетрадь	3
Альбом	**Ластик**	**Тетрадь**	**5**
Карандаш	**Краски**	**Ластик**	**5**
Карандаш	Краски	Линейка	0
Карандаш	Краски	Ручка	1
Карандаш	Краски	Тетрадь	3
Карандаш	**Ластик**	**Линейка**	**5**
Карандаш	**Ластик**	**Ручка**	**7**
Карандаш	**Ластик**	**Тетрадь**	**12**
Карандаш	Линейка	Ручка	2
Карандаш	Линейка	Тетрадь	4
Карандаш	**Ручка**	**Тетрадь**	**6**
Ластик	Линейка	Ручка	2
Ластик	Линейка	Тетрадь	4
Ластик	**Ручка**	**Тетрадь**	**6**

Рассмотрим набор {Альбом, Краски, Тетрадь}.

В нем есть двухпредметные поднаборы: {Альбом, Краски}, {Альбом, Тетрадь}, {Краски, Тетрадь}, частота появления которых в наборе тракзакций соответственно равна 6, 5 и 3. Первые два поднабора – частые, а третий – нечастый. Следовательно, и трехпредметный набор {Альбом, Краски, Тетрадь} не является частым и подлежит изъятию из дальнейшего рассмотрения.

Сформируем множество F_3 частых трехпредметных наборов:

F_3 {{Альбом, Карандаш, Краски}, {Альбом, Карандаш, Ластик},
{Альбом, Карандаш, Тетрадь}, {Альбом, Краски, Ластик},
{Альбом, Ластик, Тетрадь}, {Карандаш, Краски, Ластик},
{Карандаш, Ластик, Линейка}, {Карандаш, Ластик, Ручка},
{Карандаш, Ластик, Тетрадь}, {Карандаш, Ручка, Тетрадь},
{Ластик, Ручка, Тетрадь}}.

Шаг 4. Выполним поиск частых четыреххпредметных наборов (таблица 8).

Для получения множества F_4 свяжем множество F_3 с самим собой и сформируем множество F_4 частых четырехпредметных наборов:

F_4 {{Альбом, Карандаш, Краски, Ластик},
{Альбом, Карандаш, Ластик, Тетрадь},
{Карандаш, Ластик, Ручка, Тетрадь}}.

Эти четырехпредметные наборы уже нельзя связать. Поэтому задача поиска частых предметных наборов на исходном множестве транзакций решена.

2.2. Генерация ассоциативных правил

Для генерации ассоциативных правил на основе частых предметных наборов set_j ($j = \overline{1,m}$, m – число выявленных частных предметных наборов) к каждому частому набору set_j необходимо применить следующую процедуру.

1. Сгенерировать все возможные поднаборы $subset_{j,l_j}$ ($j = \overline{1,m}$; $l_j = \overline{1,m_j}$, m_j – число поднаборов на основе набора set_j).

2. Рассмотреть ассоциацию *R*: $subset_{j,l_j} \rightarrow (set_j - subset_{j,l_j})$, где $subset_{j,l_j}$ – непустой поднабор набора set_j; $(set_j - subset_{j,l_j})$ –набор set_j без поднабора $subset_{j,l_j}$. Ассоциация *R* считается ассоциативным правилом, если удовлетворяет условию заданного минимума для поддержки и достоверности.

Таблица 8. Четырехпредметные наборы-кандидаты в F_4

Предметный набор				Частота появления в наборе тракзакций
Предмет 1	Предмет 2	Предмет 3	Предмет 4	
Альбом	**Карандаш**	**Краски**	**Ластик**	**5**
Альбом	Карандаш	Краски	Тетрадь	3
Альбом	**Карандаш**	**Ластик**	**Тетрадь**	**5**
Карандаш	Ластик	Линейка	Ручка	2
Карандаш	Ластик	Линейка	Тетрадь	4
Карандаш	**Ластик**	**Ручка**	**Тетрадь**	**6**

Рассмотрим предметный набор-кандидат в ассоциативные правила.

set ={Альбом, Карандаш, Краски}.

В него входят поднаборы:

{Альбом}, {Карандаш}, {Краски}, {Альбом, Карандаш}, {Альбом, Краски}, {Карандаш, Краски}.

Построим ассоциативные правила, в которых условия содержат два предмета или один предмет.

Для ассоциативного правила с двумя предметами в условии предположим, что

$subset$ ={Альбом, Карандаш}.

Тогда ($set - subset$)={Краски}.

Рассмотрим правило:

R: {Альбом, Карандаш}$\rightarrow${Краски}.

Поддержка, показывающая долю транзакций, которые содержат как условие {Альбом, Карандаш}, так и следствие {Краски}, в общем наборе из 20 транзакций составляет 25 % (5 из 20 транзакций).

Чтобы найти достоверность, нужно учесть, что набор {Альбом, Карандаш} появляется в 7 из 20 транзакций, 5 из которых также содержат {Краски}. Тогда достоверность будет равна 5/7 (71,4 %).

В таблице 9 приведены ассоциации с двумя предметами в условии для трехпредметного набора {Альбом, Карандаш, Краски}.

Если предположить, что минимальная достоверность для ассоциативного правила составляет 60%, то все ассоциации в таблице 9 будут правилами.

Если предположить, что минимальная достоверность для ассоциативного правила составляет 80%, то только вторая и третья ассоциации будут правилами.

Таблица 9. Ассоциации с двумя предметами в условии для трехпредметного набора {Альбом, Карандаш, Краски}

Если условие, то следствие	Поддержка	Достоверность
Если {Альбом и Карандаш}, то {Краски}	5/20 = 25 %	5/7 = 71,4 %
Если {Альбом и Краски}, то {Карандаш}	5/20 = 25 %	5/6 = 80 %
Если {Карандаш и Краски}, то {Альбом}	5/20 = 25 %	5/5 = 100 %

Таблица 10. Ассоциации с одним предметом в условии

Если условие, то следствие	Поддержка	Достоверность
{Альбом}→{Карандаш}	0,35	0,778
{Карандаш}→{Альбом}	0,35	0,438
{Альбом}→{Краски}	0,3	0,667
{Краски}→{Альбом}	0,3	1
{Альбом}→{Ластик}	0,35	0,778
{Ластик}→{Альбом}	0,35	0,438
{Альбом}→{Тетрадь}	0,25	0,556
{Тетрадь}→{Альбом}	0,25	0,417
{Карандаш}→{Краски}	0,25	0,313
{Краски}→{Карандаш}	0,25	0,833
{Карандаш}→{Ластик}	0,8	1
{Ластик}→{Карандаш}	0,8	1
{Карандаш}→{Линейка}	0,25	0,313
{Линейка}→{Карандаш}	0,25	0,714
{Карандаш→{Ручка}	0,35	0,438
{Ручка}→{Карандаш}	0,35	0,778
{Карандаш}→{Тетрадь}	0,6	0,75
{Тетрадь}→{Карандаш}	0,6	1
{Краски}→{Ластик}	0,25	0,833
{Ластик}→{Краски}	0,25	0,313
{Ластик}→{Линейка}	0,25	0,313
{Линейка}→{Ластик}	0,25	0,714
{Ластик}→{Ручка}	0,35	0,438
{Ручка}→{Ластик}	0,35	0,778
{Ластик}→{Тетрадь}	0,6	0,75
{Тетрадь}→{Ластик}	0,6	1
{Ручка}→{Тетрадь}	0,3	0,667
{Тетрадь}→{Ручка}	0,3	0,5

Таблица 11. Ассоциации с двумя предметами **в условии**

Если условие, то следствие	Поддержка	Достоверность
{Альбом, Карандаш}→{Краски}	0,25	0,714
{Альбом, Краски}→{Карандаш}	0,25	0,833
{Карандаш, Краски}→{Альбом}	0,25	1
{Альбом, Карандаш}→{Ластик}	0,35	1
{Альбом, Ластик}→{Карандаш}	0,35	1
{Карандаш, Ластик}→{Альбом}	0,35	0,438
{Альбом, Карандаш}→{Тетрадь}	0,25	0,714
{Альбом, Тетрадь}→{Карандаш}	0,25	1
{Карандаш, Тетрадь}→{Альбом}	0,25	0,417
{Альбом, Краски}→{Ластик}	0,25	0,833
{Альбом, Ластик}→{Краски}	0,25	0,714
{Краски, Ластик}→{Альбом}	0,25	1
{Альбом, Ластик}→{Тетрадь}	0,25	0,714
{Альбом, Тетрадь}→{Ластик}	0,25	1
{Ластик, Тетрадь}→{Альбом}	0,25	0,417
{Карандаш, Краски}→{Ластик}	0,25	1
{Карандаш, Ластик}→{Краски}	0,25	0,313
{Краски, Ластик}→{Карандаш}	0,25	1
{Карандаш, Ластик}→{Линейка}	0,25	0,313
{Карандаш, Линейка}→{Ластик}	0,25	1
{Ластик, Линейка}→{Карандаш}	0,25	1
{Карандаш, Ластик}→{Ручка}	0,35	0,438
{Карандаш, Ручка}→{Ластик}	0,35	1
{Ластик, Ручка}→{Карандаш}	0,35	1
{Карандаш, Ластик}→{Тетрадь}	0,6	0,75
{Карандаш, Тетрадь}→{Ластик}	0,6	1
{Ластик, Тетрадь}→{Карандаш}	0,6	1
{Карандаш, Ручка}→{Тетрадь}	0,3	0,857
{Карандаш, Тетрадь}→{Ручка}	0,3	0,5
{Ручка, Тетрадь}→{Карандаш}	0,3	1
{Ластик, Ручка}→{Тетрадь}	0,3	0,857
{Ластик, Тетрадь}→{Ручка}	0,3	0,5
{Ручка, Тетрадь}→{Ластик}	0,3	1

Таблица 12. Ассоциации с тремя предметами **в условии**

Если условие, то следствие	Поддержка	Достоверность
{Карандаш, Ластик,Ручка}→{Тетрадь}	0,3	0,857
{Ластик, Ручка,Тетрадь}→{Карандаш}	0,3	1
{Ручка, Тетрадь,Карандаш}→{Ластик}	0,3	1
{Тетрадь, Карандаш,Ластик}→{Ручка}	0,3	0,5
{Альбом, Карандаш,Краски}→{Ластик}	0,25	1
{Карандаш, Краски,Ластик}→{Альбом}	0,25	1
{Краски, Ластик,Альбом}→{Карандаш}	0,25	1
{Ластик, Альбом,Карандаш}→{Краски}	0,25	0,714
{Альбом, Карандаш,Ластик}→{Тетрадь}	0,25	0,714
{Карандаш, Ластик,Тетрадь}→{Альбом}	0,25	0,417
{Ластик, Тетрадь,Альбом}→{Карандаш}	0,25	1
{Тетрадь, Альбом,Карандаш}→{Ластик}	0,25	1

Получим все ассоциативные правила на основе выявленных частых предметных наборов *с одним предметом в следствии*. Ассоциативные правила с двумя и более предметами в следствии могут быть построены аналогичным образом.

В таблице 10 приведены все ассоциации с одним предметом в условии и одним предметом (всего – 28 ассоциаций).

В таблице 11 приведены все ассоциации с двумя предметами в условии (всего – 33 ассоциации).

В таблице 12 приведены все ассоциации с тремя предметами в условии (всего – 12 ассоциаций).

Чтобы оценить значимость сгенерированных ассоциаций необходимо перемножить соответствующие им значения поддержки и достоверности.

Выберем в качестве ассоциативных правил те ассоциации из таблиц 10 – 12, у которых достоверность не ниже 80% (таблица 13).

В результате применения алгоритма Apriori удалось найти 34 ассоциативных правил, которые с достоверностью не ниже 80% показывают, какие продукты из исходного набора предметов чаще всего продаются вместе.

Если понизить пороговое значение достоверности до 60%, то число ассоциативных правил увеличиться на 17 штук и будет равно 51.

Таблица 13. Ассоциациативные правила при минимальной достоверности 80%

Если условие, то следствие	Поддержка S	Достоверность C	$S \cdot C \cdot$
{Карандаш, Ластик, Ручка}→{Тетрадь}	0,3	0,857	0,257
{Ластик, Ручка ,Тетрадь}→{Карандаш}	0,3	1	0,3
{Ручка, Тетрадь, Карандаш}→{Ластик}	0,3	1	0,3
{Альбом, Карандаш, Краски}→{Ластик}	0,25	1	0,25
{Карандаш, Краски, Ластик}→{Альбом}	0,25	1	0,25
{Краски, Ластик, Альбом}→{Карандаш}	0,25	1	0,25
{Ластик, Тетрадь, Альбом}→{Карандаш}	0,25	1	0,25
{Тетрадь, Альбом, Карандаш}→{Ластик}	0,25	1	0,25
{Альбом, Краски}→{Карандаш}	0,25	0,833	0,208
{Карандаш, Краски}→{Альбом}	0,25	1	0,25
{Альбом, Карандаш}→{Ластик}	0,35	1	0,35
{Альбом, Ластик}→{Карандаш}	0,35	1	0,35
{Альбом, Краски}→{Ластик}	0,25	0,833	0,208
{Краски, Ластик}→{Альбом}	0,25	1	0,25
{Альбом, Тетрадь}→{Ластик}	0,25	1	0,25
{Карандаш, Краски}→{Ластик}	0,25	1	0,25
{Краски, Ластик}→{Карандаш}	0,25	1	0,25
{Карандаш, Линейка}→{Ластик}	0,25	1	0,25
{Ластик, Линейка}→{Карандаш}	0,25	1	0,25
{Карандаш, Ручка}→{Ластик}	0,35	1	0,35
{Ластик, Ручка}→{Карандаш}	0,35	1	0,35
{Карандаш, Тетрадь}→{Ластик}	0,6	1	0,6
{Ластик, Тетрадь}→{Карандаш}	0,6	1	0,6
{Карандаш, Ручка}→{Тетрадь}	0,3	0,857	0,257
{Ручка, Тетрадь}→{Карандаш}	0,3	1	0,3
{Ластик, Ручка}→{Тетрадь}	0,3	0,857	0,257
{Ручка, Тетрадь}→{Ластик}	0,3	1	0,3
{Краски}→{Альбом}	0,3	1	0,3
{Краски}→{Карандаш}	0,25	0,833	0,208
{Карандаш}→{Ластик}	0,8	1	0,8
{Ластик}→{Карандаш}	0,8	1	0,8
{Тетрадь}→{Карандаш}	0,6	1	0,6
{Краски}→{Ластик}	0,25	0,833	0,208
{Тетрадь}→{Ластик}	0,6	1	0,6

Очевидно, что по результатам выявления ассоциативных правил можно разработать адекватную маркетинговую стратегию, оптимизировать закупки и размещение товаров (продуктов) на прилавках и витринах.

4. Алгоритм Apriori на языке Python

Для демонстрации принципов работы библиотек, реализующих поиск ассоциативных правил, будем использовать Jupyter Notebook.

На языке Python имеются различные реализации алгоритма Apriori, предполагающие, в том числе, использование отдельных функций, реализованных на других языках программирования, например, на C++.

Рассмотрим, в частности, работу с двумя реализациями алгоритма Apriori, предложенными по ссылкам

https://pypi.org/project/apriori-python/

и

https://pypi.org/project/efficient-apriori/.

4.1. Программная реализация apriori-python

Рассмотрим работу с программной реализацией алгоритма Apriori, предложенной по ссылке:

https://pypi.org/project/apriori-python/.

Для установки библиотеки, реализующей работу с алгоритмом Apriori, необходимо выполнить команду:

```
pip install apriori_python
```

На рисунке 1 приведен пример программного кода, реализующего поиск ассоциативных правил с применением алгоритма Apriori для демонстрационного набора из 3 транзакций:

```
transactions= [['Тетрадь', 'Ручка', 'Краски'],
               ['Тетрадь', 'Ручка', 'кисть'],
               ['Краски', 'Ручка', 'Ластик']]
```

При этом предварительно осуществляется импортирование функции `apriori`, реализующей алгоритм.

При вызове функции `apriori` необходимо указать в списке её параметров список `transactions`, содержащий набор транзакций, а также параметры, определяющие пороговые (минимальные) значения поддержки `minSup` и достоверности `minConf`, присвоив им некоторые значения, например,

```
minSup=0.5, minConf=0.5.
```

```
from apriori_python import apriori
transactions = [['Тетрадь', 'Ручка', 'Краски'],
                ['Тетрадь', 'Ручка', 'Кисть'],
                ['Краски', 'Ручка', 'Ластик']]
freqItemSet, rules = apriori(transactions, minSup=0.5, minConf=0.5)
print(rules)
```

```
[[{'Ручка'}, {'Тетрадь'}, 0.6666666666666666], [{'Ручка'}, {'Краски'}, 0.6666666666666666],
[{'Тетрадь'}, {'Ручка'}, 1.0], [{'Краски'}, {'Ручка'}, 1.0]]
```

Рисунок 1. Пример программного кода, реализующего поиск ассоциативных правил с применением алгоритма Apriori

```
type(freqItemSet)
```

```
dict
```

```
freqItemSet
```

```
{1: {frozenset({'Краски'}), frozenset({'Ручка'}), frozenset({'Тетрадь'})},
 2: {frozenset({'Ручка', 'Тетрадь'}), frozenset({'Краски', 'Ручка'})}}
```

```
type(rules)
```

```
list
```

```
rules
```

```
[[{'Ручка'}, {'Тетрадь'}, 0.6666666666666666],
 [{'Ручка'}, {'Краски'}, 0.6666666666666666],
 [{'Тетрадь'}, {'Ручка'}, 1.0],
 [{'Краски'}, {'Ручка'}, 1.0]]
```

Рисунок 2. Примеры, позволяющие определить тип и содержимое выходных параметров алгоритма Apriori

```
rules[0]
```

```
[{'Ручка'}, {'Тетрадь'}, 0.6666666666666666]
```

```
rules[0][0]
```

```
{'Ручка'}
```

```
rules[0][1]
```

```
{'Тетрадь'}
```

```
rules[0][2]
```

```
0.6666666666666666
```

Рисунок 3. Пример обращения к компонентам списка, определяющим первое правило в алгоритме Apriori

В результате в качестве выходных параметров будут сформированы частые предметные наборы `freqItemSet`, а также ассоциативные правила `rules`, удовлетворяющие заданным пороговым значениям поддержки и достоверности.

На рисункке 2 приведены примеры, позволяющие определить тип и содержимое выходных параметров алгоритма Apriori. Так, `freqItemSet` – словарь, а `rules` – список.

На рисунке 3 приведен пример обращения к компонентам списка, определяющим первое ассоциативное правило. При этом можно обратиться к условию, следствию и достоверности правила как к отдельному элементу.

Найденные ассоциативные правила можно, например, вывести на печать командой

```
print(rules)
```

При этом для каждого ассоциативного правила будет выведено значение его достоверности.

4.2. Программная реализация алгоритма efficient-apriori

Рассмотрим работу с программной реализацией алгоритма Apriori, предложенной по ссылке:

https://pypi.org/project/efficient-apriori/.

Эта программная реализация алгоритма Apriori позиционируется как эффективная реализация авторского алгоритма Apriori *чисто* на языке Python и именуется авторами как Efficient Apriori.

Программная реализация алгоритма Efficient Apriori доступна для установки по ссылке:

https://pypi.org/project/efficient-apriori/.

Для установки библиотеки, реализующей работу с алгоритмом Efficient Apriori, необходимо выполнить команду:

```
pip install efficient-apriori
```

На рисунке 4 приведен пример программного кода, реализующего поиск ассоциативных правил с применением алгоритма Efficient Apriori для демонстрационного набора из 3 транзакций. При этом можно заменить некоторые отличия в названии входных параметров, а также в оформлении выведенных правил (по сравнению с алгоритмом Aproiri).

```
from efficient_apriori import apriori
transactions = [['Тетрадь', 'Ручка', 'Краски'],
                ['Тетрадь', 'Ручка', 'Кисть'],
                ['Краски', 'Ручка', 'Ластик']]
freqItemSet, rules = apriori(transactions, min_support=0.5, min_confidence=0.5)
print(rules)
```

```
[{Ручка} -> {Краски}, {Краски} -> {Ручка}, {Тетрадь} -> {Ручка}, {Ручка} -> {Тетрадь}]
```

Рисунок 4. Пример программного кода, реализующего поиск ассоциативных правил с применением алгоритма Efficient Apriori.

```
type(freqItemSet)
```

```
dict
```

```
freqItemSet
```

```
{1: {('Тетрадь',): 2, ('Краски',): 2, ('Ручка',): 3},
 2: {('Краски', 'Ручка'): 2, ('Ручка', 'Тетрадь'): 2}}
```

```
type(rules)
```

```
list
```

```
rules
```

```
[{Ручка} -> {Краски},
 {Краски} -> {Ручка},
 {Тетрадь} -> {Ручка},
 {Ручка} -> {Тетрадь}]
```

Рисунок 5. Примеры, позволяющие определить тип и содержимое выходных параметров алгоритма Efficient Apriori

```
rules[0]
```

```
{Ручка} -> {Краски}
```

Рисунок 6. Пример обращения к компоненте списка, определяющей первое ассоциативное правило в алгоритме Efficient Apriori

```
rules_rhs = filter(lambda rule: len(rule.lhs) == 1 and len(rule.rhs) == 1, rules)
for rule in sorted(rules_rhs, key=lambda rule: rule.confidence):
  print(rule)
```

```
{Ручка} -> {Краски} (conf: 0.667, supp: 0.667, lift: 1.000, conv: 1.000)
{Ручка} -> {Тетрадь} (conf: 0.667, supp: 0.667, lift: 1.000, conv: 1.000)
{Краски} -> {Ручка} (conf: 1.000, supp: 0.667, lift: 1.000, conv: 0.000)
{Тетрадь} -> {Ручка} (conf: 1.000, supp: 0.667, lift: 1.000, conv: 0.000)
```

Рисунок 7. Пример вывода правил, удовлетворяющих заданным условиям, в алгоритме Efficient Apriori

На рисунке 5 приведены примеры, позволяющие определить тип и содержимое выходных параметров алгоритма Apriori. Так, `freqItemSet` – словарь, а `rules` – список, но их организация несколько отлична от организации аналогичных параметров у алгоритма Apriori. В частности, значение достоверности для правила не хранится в выходном параметре `rules`.

На рисунке 6 приведен пример обращения к компонентам списка, определяющим первое ассоциативное правило.

При этом не возможно обратиться напрямую к условию или следствию ассоциативного правила как к отдельному элементы (в отличии от алгоритма Apriori).

Найденные ассоциативные правила можно, например, вывести на печать командой

```
print(rules)
```

При этом для каждого ассоциативного правила значение его достоверности не выводится, так как оно не хранится в выходном параметре `rules` (в отличии от алгоритма Apriori).

При необходимости можно выполнить вывод только тех ассоциативных правил, которые удовлетворяют тем или иным условиям, осуществив фильтрацию и сортировку.

Например, можно наложить ограничения на длину левой (`lhs` – left hand side) и правой (rhs – right hand side) частей ассоциативного правила, то есть на число предметов в них.

Кроме того, можно осуществить, например, вывод ассоциативных правил по убыванию значений некоторого показателя, например, по убыванию значений достоверности ассоциативных правил.

На рисунке 7 приведен пример вывода правил, упорядоченных по убыванию значений достоверности правил (`rule.confidence`).

При этом число предметов в условии и следствии правила равно ровно 1 («`len(rule.lhs) == 1 and len(rule.rhs) == 1`»).

При необходимости, можно для каждого правила посмотреть, что хранится в `rule.lhs`, `rule.rhs`, `rule.confidence` (достоверность), `rule.support` (поддержка), `rule.lift` (лифт), `rule.conviction` (убежденность).

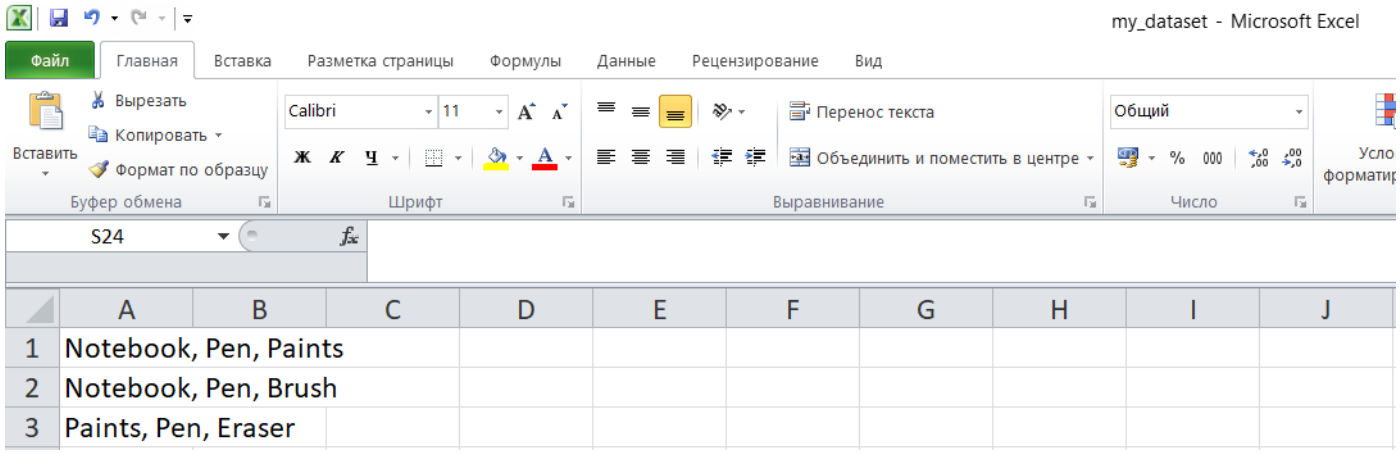

Рисунок 8. Фрагмент тестового cvs-файла с данными

```
def data_generator(filename):
    """
    Data generator
    """
    def data_gen():
        with open(filename) as file:
            for line in file:
                yield tuple(k.strip() for k in line.split(','))

    return data_gen

transactions = data_generator('my_dataset.csv')
freqItemSet, rules = apriori(transactions, min_support=0.6, min_confidence=0.6)
```

Рисунок 9. Фрагмент программного кода, использующего в своей работе генератор

```
type(transactions)
```

```
function
```

```
print(freqItemSet)
```

```
{1: {('Pen',): 3, ('Paints',): 2, ('Notebook',): 2}, 2: {('Notebook', 'Pen'): 2, ('Paints', '
Pen'): 2}}
```

```
print(rules)
```

```
[{Pen} -> {Notebook}, {Notebook} -> {Pen}, {Pen} -> {Paints}, {Paints} -> {Pen}]
```

Рисунок 10. Результаты работы алгоритма Efficient Apriori в случае применения генератора

Если необходимо работать с данными больших объемов, то есть с данными, которые слишком велики для размещения в памяти, целесообразно использовать функцию, возвращающую генератор вместо списка.

На рисунке 8 приведен фрагмент тестового cvs-файла с данными.

На рисунке 9 приведен фрагмент программного кода, использующего в своей работе генератор.

На рисунке10 приведены результаты работы алгоритма Efficient Apriori в случае применения генератора. При этом можно заменить, что тип у переменной `transactions` стал `function` (в то время как раньше тип определялся как `list`).

Хотя алгоритм Efficient Apriori работает со многими функциями, наибольший интерес представляют три функции:

`apriori()`,

`itemsets_from_transactions()`,

`generate_rules_apriori()`.

Функция `apriori()` возвращает предметные наборы и ассоциативные правила, которые получаются посредством вызова функций `itemsets_from_transactions()` и `generate_rules_apriori()` соответственно. Ассоциативные правила возвращаются как экземпляры класса `Rule`.

Ниже приведена краткая информация по этим функциям и классу.

1. Функция `apriori`.

`efficient_apriori.apriori`(*transactions: Union[List[tuple], Callable], min_support: float = 0.5, min_confidence: float = 0.5, max_length: int = 8, verbosity: int = 0, output_transaction_ids: bool = False*)

Функция реализует классический алгоритм Apriori.

Функция работает в два этапа. На этапе 1 формируются предметные наборы с заданным значением *минимальной поддержки*. На этапе 2 строятся ассоциативные правила с заданным значением *минимальной достоверности* на основе предметных наборов, найденных на этапе 1.

Оба этапа реализуются посредством исследования пространства поиска, то есть создания всех возможных предметных наборов и оценки их поддержки. Алгоритм Apriori эффективно сокращает пространство поиска, решая предварительно, имеет ли предметный набор желаемую поддержку, прежде чем выполнять итерацию по всему набору данных и проверку.

```
Parameters
----------
transactions : list of tuples, list of itemsets.TransactionWithId,
    or a callable returning a generator. Use TransactionWithId's when
    the transactions have ids which should appear in the outputs.
    The transactions may be either a list of tuples, where the tuples must
    contain hashable items. Alternatively, a callable returning a generator
    may be passed. A generator is not sufficient, since the algorithm will
    exhaust it, and it needs to iterate over it several times. Therefore,
    a callable returning a generator must be passed.
min_support : float
    The minimum support of the rules returned. The support is frequency of
    which the items in the rule appear together in the data set.
min_confidence : float
    The minimum confidence of the rules returned. Given a rule X -> Y, the
    confidence is the probability of Y, given X, i.e. P(Y|X) = conf(X -> Y)
max_length : int
    The maximum length of the itemsets and the rules.
verbosity : int
    The level of detail printing when the algorithm runs. Either 0, 1 or 2.
output_transaction_ids : bool
    If set to true, the output contains the ids of transactions that
    contain a frequent itemset. The ids are the enumeration of the
    transactions in the sequence they appear.
```

Рисунок 11. Информация по параметрам функции `apriori`

```
Parameters
----------
transactions : a list of itemsets (tuples with hashable entries),
               or a function returning a generator
    A list of transactions. They can be of varying size. To pass through
    data without reading everything into memory at once, a callable
    returning a generator may also be passed.
min_support : float
    The minimum support of the itemsets, i.e. the minimum frequency as a
    percentage.
max_length : int
    The maximum length of the itemsets.
verbosity : int
    The level of detail printing when the algorithm runs. Either 0, 1 or 2.
output_transaction_ids : bool
    If set to true, the output contains the ids of transactions that
    contain a frequent itemset. The ids are the enumeration of the
    transactions in the sequence they appear.
```

Рисунок 12. Информация по параметрам функции `itemsets_from_transactions`

```
Parameters
----------
itemsets : dict of dicts
    The first level of the dictionary is of the form (length, dict of item
    sets). The second level is of the form (itemset, count_in_dataset)).
min_confidence :  float
    The minimum confidence required for the rule to be yielded.
num_transactions : int
    The number of transactions in the data set.
verbosity : int
    The level of detail printing when the algorithm runs. Either 0, 1 or 2.
```

Рисунок 13. Информация по параметрам функции `generate_rules_apriori`

```
Parameters
----------
lhs : tuple
    The left hand side (antecedent) of the rule. Each item in the tuple
    must be hashable, e.g. a string or an integer.
rhs : tuple
    The right hand side (consequent) of the rule.
count_full : int
    The count of the union of the lhs and rhs in the dataset.
count_lhs : int
    The count of the lhs in the dataset.
count_rhs : int
    The count of the rhs in the dataset.
num_transactions : int
    The number of transactions in the dataset.
```

Рисунок 14. Информация по параметрам класса `Rule`

На рисунке 11 приведена информация по параметрам функции `apriori`.

2. Функция `itemsets_from_transactions`.

`efficient_apriori.itemsets_from_transactions`(*transactions: Union[List[tuple], Callable], min_support: float, max_length: int = 8, verbosity: int = 0, output_transaction_ids: bool = False*)

Функция реализует формирование предметных наборов.

На рисунке 12 приведена информация по параметрам функции `itemsets_from_transactions`.

```
from efficient_apriori import apriori
transactions = [['Тетрадь', 'Ручка', 'Краски'],
                ['Тетрадь', 'Ручка', 'Кисть'],
                ['Краски', 'Ручка', 'Ластик']]
freqItemSet, rules = apriori(transactions, output_transaction_ids=True)
print(freqItemSet)
```

```
{1: {('Тетрадь',): ItemsetCount(itemset_count=2, members={0, 1}), ('Краски',): ItemsetCount(i
temset_count=2, members={0, 2}), ('Ручка',): ItemsetCount(itemset_count=3, members={0, 1,
2})}, 2: {('Краски', 'Ручка'): ItemsetCount(itemset_count=2, members={0, 2}), ('Ручка', 'Тетр
адь'): ItemsetCount(itemset_count=2, members={0, 1})}}
```

Рисунок 15. Пример программного кода с выводом соответствующих частых предметных наборов, а также – информации о том, сколько раз они встретились и в каких транзакциях

3. Функция `generate_rules_apriori`.

`efficient_apriori.generate_rules_apriori`(*itemsets: Dict[int, Dict[tuple, int]], min_confidence: float, num_transactions: int, verbosity: int = 0*)

Функция реализует формирование ассоциативных правил.

На рисунке 13 приведена информация по параметрам функции `generate_rules_apriori`.

3. Класс `Rule`.

class `efficient_apriori.Rule`(*lhs: tuple, rhs: tuple, count_full: int = 0, count_lhs: int = 0, count_rhs: int = 0, num_transactions: int = 0*)

Класс предназначен для описания ассоциативного правила.

На рисунке 14 приведена информация по параметрам класса `Rule`.

Если при выявлении ассоциативных правил необходимо знать, сколько раз встретился тот или иной частый предметный набор и в каких транзакциях от встретился, необходимо установить значение параметра `output_transaction_ids` равным `True`:

`output_transaction_ids=True`.

В этом случае для каждого предметного набора выводится информация об объекте `ItemsetCount`, в свойстве `itemset_count` которого хранится информация о числе транзакций, на основе которых сформирован частый предметный набор, а в свойстве `members` – информация о номерах `ids` таких транзакций.

На рисунке 15 приведен пример программного кода с выводом соответствующих частых предметных наборов, а также – информации о том, сколько раз они встретились (значение свойства `itemset_count`) и в каких транзакциях (значение свойства `members`).

Подробная документация по программной реализации алгоритма Efficient Apriori приведена по ссылке:

https://efficient-apriori.readthedocs.io/en/latest/.

Варианты и задания

Необходимо выполнить задания по поиску ассоциативных правил в соответствии с предложенным вариантом.

Варианты

1. Тестовые варианты.

Вариант 1. Сформировать набор из 30 транзакций на основе чеков покупок в продуктовом магазине.

Вариант 2. Сформировать набор из 30 транзакций на основе чеков покупок в аптеке.

Вариант 3. Сформировать набор из 30 транзакций на основе чеков покупок в магазине компьютерной техники.

Вариант 4. Сформировать набор из 30 транзакций на основе чеков покупок в магазине одежды.

Вариант 5. Сформировать набор из 30 транзакций на основе продуктов, употребляемых на завтрак, обед и ужин.

Вариант 6. Сформировать набор из 30 транзакций на основе чеков покупок в магазине канцтоваров.

Вариант 7. Сформировать набор из 30 транзакций на основе чеков покупок в магазине бытовой техники.

Вариант 8. Сформировать набор из 30 транзакций на основе чеков покупок в кондитерском отделе.

Вариант 9. Сформировать набор из 30 транзакций на основе чеков покупок в магазине косметики.

Вариант 10. Сформировать набор из 30 транзакций на основе чеков покупок в магазине спорттоваров.

2. Варианты из репозиториев.

Вариант 1.

Набор данных:

http://archive.ics.uci.edu/ml/datasets/Online+Retail

Вариант 2.

Набор данных:

https://github.com/viktree/curly-octo-chainsaw/blob/master/BreadBasket_DMS.csv

P.S. Можно предложить собственные «большие» наборы данных.

Задания

1. Применить для тестовых вариантов и вариантов из репозиториев различные алгоритмы поиска ассоциативных правил при одинаковых начальных условиях (при одинаковых пороговых значениях для поддержки и достоверности) и сравнить полученные результаты. Для тестовых вариантов выполнить ручные расчеты (например, с применением MS Excel) и расчеты с применением программных библиотек на языке Python. Для вариантов из репозиториев выполнить расчеты с применением программных библиотек на языке Python.

В качестве алгоритмов поиска ассоциативных правил использовать алгоритмы:

- Apriori (https://pypi.org/project/apriori-python/);
- Efficient Apriori (https://pypi.org/project/efficient-apriori/);
- FPGrowth (https://pypi.org/project/fpgrowth-py/).

2. Сформировать базы ассоциативных правил с уровнем минимальной достоверности 60% и 80%. Вычислить для ассоциативных правил поддержку, достоверность, значимость.

3. Оценить время формирования искомых ассоциативных правил с применением различных алгоритмов и построить диаграммы, позволяющие выполнить сравнительный анализ.

4. Выполнить визуализацию ассоциативных правил (https://pypi.org/project/pyarmviz/).

СПИСОК ЛИТЕРАТУРЫ

1. Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules // Proceedings of the 20th International Conference on Very Large Databases, 1994. P. 487 – 499.
2. Паклин Н., Орешков В. Бизнес-аналитика: от данных к знаниям: Учебное пособие. – 2-е изд., испр. – СПб.: Питер, 2013. – 704 с.
3. Apriori – масштабируемый алгоритм поиска ассоциативных правил [Электронный ресурс]. – Режим доступа: https://loginom.ru/blog/apriori, свободный (дата обращения 29.08.2021).
4. Tan J., Bu Y., Yang B. An Efficient Close Frequent Pattern Mining Algorithm // 2009 Second International Conference on Intelligent Computation Technology and Automation, 2009. – P. 528 – 531.
5. FPG – альтернативный алгоритм поиска ассоциативных правил [Электронный ресурс]. – Режим доступа: https://loginom.ru/blog/fpg, свободный (дата обращения 29.08.2021).
6. A simple apriori algorithm python implementation [Электронный ресурс]. – Режим доступа: https://pypi.org/project/apriori-python/, свободный (дата обращения 29.08.2021).
7. Apriori_Python [Электронный ресурс]. – Режим доступа: https://github.com/chonyy/apriori_python, свободный (дата обращения 29.08.2021).
8. An efficient Python implementation of the Apriori algorithm [Электронный ресурс]. – Режим доступа: https://pypi.org/project/efficient-apriori/, свободный (дата обращения 29.08.2021).
9. Apriori: Association Rule Mining In-depth Explanation and Python Implementation [Электронный ресурс]. – Режим доступа: https://towardsdatascience.com/apriori-association-rule-mining-explanation-and-python-implementation-290b42afdfc6, свободный (дата обращения 29.08.2021).
10. Efficient-Apriori [Электронный ресурс]. – Режим доступа: https://efficient-apriori.readthedocs.io/en/latest/, свободный (дата обращения 29.08.2021).

11. Efficient-Apriori [Электронный ресурс]. – Режим доступа: https://github.com/tommyod/Efficient-Apriori, свободный (дата обращения 29.08.2021).

12. Python implementation of FP Growth algorithm [Электронный ресурс]. – Режим доступа: https://pypi.org/project/fpgrowth-py/, свободный (дата обращения 29.08.2021).

13. FP Growth: Frequent Pattern Generation in Data Mining with Python Implementation [Электронный ресурс]. – Режим доступа: https://towardsdatascience.com/fp-growth-frequent-pattern-generation-in-data-mining-with-python-implementation-244e561ab1c3, свободный (дата обращения 29.08.2021).

14. FPGrowth_py [Электронный ресурс]. – Режим доступа: https://github.com/chonyy/fpgrowth_py, свободный (дата обращения 29.08.2021).

15. Mlxtend [Электронный ресурс]. – Режим доступа: https://github.com/rasbt/mlxtend/tree/master/mlxtend/frequent_patterns, свободный (дата обращения 29.08.2021).

16. FP-Growth [Электронный ресурс]. – Режим доступа: https://fp-growth.readthedocs.io/en/latest/, свободный (дата обращения 29.08.2021).

17. Understand and Build FP-Growth Algorithm in Python [Электронный ресурс]. – Режим доступа: https://towardsdatascience.com/understand-and-build-fp-growth-algorithm-in-python-d8b989bab342, свободный (дата обращения 29.08.2021).

18. Frequent Pattern (FP) Growth Algorithm In Data Mining [Электронный ресурс]. – Режим доступа: https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/, свободный (дата обращения 29.08.2021).

19. ML | Frequent Pattern Growth Algorithm [Электронный ресурс]. – Режим доступа: https://www.geeksforgeeks.org/ml-frequent-pattern-growth-algorithm/, свободный (дата обращения 29.08.2021).

20. PyARMViz [Электронный ресурс]. – Режим доступа: https://pypi.org/project/pyarmviz/, свободный (дата обращения 29.08.2021).