Ассоциативные правила – алгоритм Apriori [M.102]

При практической реализации систем поиска ассоциативных правил используют различные методы, которые позволяют снизить пространство поиска до размеров, обеспечивающих приемлемые вычислительные и временные затраты, например *алгоритм Apriori* (Agrawal и Srikant, 1994).

Частые предметные наборы и их обнаружение

В основе алгоритма Apriori лежит понятие *частого набора* (frequent itemset), который также можно назвать частым предметным набором, часто встречающимся множеством (соответственно, он связан с понятием частоты). Под *частотой* понимается простое количество транзакций, в которых содержится данный предметный набор. Тогда частыми наборами будут те из них, которые встречаются чаще, чем в заданном числе транзакций.

Определение

> *Частый предметный набор — предметный набор с поддержкой больше заданного порога либо равной ему. Этот порог называется минимальной поддержкой.*

Методика поиска ассоциативных правил с использованием частых наборов состоит из двух шагов.

1. Следует найти частые наборы.
2. На их основе необходимо сгенерировать ассоциативные правила, удовлетворяющие условиям минимальной поддержки и достоверности.

Чтобы сократить пространство поиска ассоциативных правил, алгоритм Apriori использует свойство антимонотонности. Свойство утверждает, что если предметный набор Z не является частым, то добавление некоторого нового предмета A к набору Z не делает его более частым. Другими словами, если Z не является частым набором, то и набор $Z \cup A$ также не будет являться таковым. Данное полезное свойство позволяет значительно уменьшить пространство поиска ассоциативных правил.

Пусть имеется множество транзакций D, представленное в таблице 2.

Таблица 2 – Множество транзакций

№ транзакции	Предметные наборы
1	Капуста, перец, кукуруза
2	Спаржа, кабачки, кукуруза
3	Кукуруза, помидоры, фасоль, кабачки
4	Перец, кукуруза, помидоры, фасоль
5	Фасоль, спаржа, капуста
6	Кабачки, спаржа, фасоль, помидоры
7	Помидоры, кукуруза
8	Капуста, помидоры, перец
9	Кабачки, спаржа, фасоль
10	Фасоль, кукуруза
11	Перец, капуста, фасоль, кабачки
12	Спаржа, фасоль, кабачки
13	Кабачки, кукуруза, спаржа, фасоль
14	Кукуруза, перец, помидоры, фасоль, капуста

Будем считать частыми наборы, которые встречаются в выборке более чем f = 4 раза. Сначала найдем частые однопредметные наборы. Для этого представим базу данных транзакций из таблицы 2 в нормализованном виде, который демонстрируется в таблице 3.

Таблица 3 – Нормализованный вид множества транзакций

№ транзакции	Спаржа	Фасоль	Капуста	Кукуруза	Перец	Кабачки	Помидоры
1	0	0	1	1	1	0	0
2	1	0	0	1	0	1	0
3	0	1	0	1	0	1	1
4	0	1	0	1	1	0	1
5	1	1	0	0	0	0	1
6	1	1	0	0	0	1	1
7	0	0	0	1	0	0	1
8	0	0	1	0	1	0	1
9	1	1	0	0	0	1	0
10	0	1	0	1	0	0	0
11	0	1	1	0	1	1	0
12	1	1	0	0	0	1	0
13	1	1	0	1	0	1	0
14	0	1	1	1	1	0	1

На пересечении строки транзакции и столбца предмета ставится 1, если данный предмет присутствует в транзакции, и 0 — в противном случае. Тогда, просуммировав значения в

каждом столбце, мы получим частоту появления каждого предмета. Поскольку все суммы равны или превышают 4, все предметы можно рассматривать как частые однопредметные наборы. Обозначим их в виде множества $F_1 = \{спаржа, фасоль, капуста, кукуруза, перец, кабачки, помидоры\}$.

Теперь переходим к поиску частых 2-предметных наборов. Вообще, для поиска F_k, то есть k-предметных наборов, алгоритм Apriori сначала создает множество F_k кандидатов в k-предметные наборы путем связывания множества Fk – 1 с самим собой. Затем Fk сокращается с использованием свойства антимонотонности. Предметные наборы множества F_k, которые остались после сокращения, формируют F_k. Множество F_2 содержит все комбинации предметов, представленные в таблице 4.

Таблица 4 – Предметные наборы

Набор	Количество	Набор	Количество
Спаржа, фасоль	5	Капуста, кукуруза	2
Спаржа, капуста	1	Капуста, перец	4
Спаржа, кукуруза	2	Капуста, кабачки	1
Спаржа, перец	0	Капуста, помидоры	2
Спаржа, кабачки	5	Кукуруза, перец	3
Спаржа, помидоры	1	Кукуруза, кабачки	3
Фасоль, капуста	3	Кукуруза, помидоры	4
Фасоль, кукуруза	5	Перец, кабачки	1
Фасоль, перец	3	Перец, помидоры	3
Фасоль, кабачки	6	Кабачки, помидоры	2
Фасоль, помидоры	4		

Поскольку $f = 4$, из таблицы 4 в множество F_2 (то есть множество 2-предметных наборов) войдут только те наборы, которые встречаются в исходной выборке 4 раза или более. Таким образом,

$$F_2 = \begin{matrix} \{спаржа, фасоль\} \\ \{спаржа, кабачки\} \\ \{фасоль, кукуруза\} \\ \{фасоль, кабачки\} \\ \{фасоль, помидоры\} \\ \{капуста, перец\} \\ \{кукуруза, помидоры\}. \end{matrix}$$

Далее мы используем частые 2-предметные наборы из множества F_2 для генерации множества F_3 3-предметных наборов. Для этого нужно связать множество F_2 с самим собой, где предметные наборы являются связываемыми, если у них первые $k-1$ предметов общие (предметы должны следовать в алфавитном порядке). Например, наборы {спаржа, фасоль} и

{спаржа, кабачки}, для которых $k = 2$, чтобы быть связываемыми, должны иметь $k - 1 = 1$ общий первый элемент, которым и является спаржа. В результате связывания пары 2-предметных наборов мы получим:

$$\{спаржа, фасоль\} + \{спаржа, кабачки\} = \{спаржа, фасоль, кабачки\}.$$

Аналогично $\{фасоль, кукуруза\}$ и $\{фасоль, кабачки\}$ могут быть объединены в 3-предметный набор $\{фасоль, кукуруза, кабачки\}$. И наконец, так же формируются остальные 3-предметные наборы $\{фасоль, кабачки, помидоры\}$ и $\{фасоль, кукуруза, помидоры\}$. Таким образом:

$$F_3 = \begin{matrix} \{спаржа, фасоль, кабачки\} \\ \{фасоль, кукуруза, кабачки\} \\ \{фасоль, кабачки, помидоры\} \\ \{фасоль, кукуруза, помидоры\}. \end{matrix}$$

Затем F_3 также сокращается с помощью свойства антимонотонности. Для каждого предметного набора s из множества F3 создаются и проверяются поднаборы размером $k - 1$. Если любой из этих поднаборов не является частым и, следовательно, наборы s также не могут быть частыми (в соответствии со свойством антимонотонности), то он должен быть исключен из рассмотрения. Например, пусть $s = \{спаржа, фасоль, кабачки\}$. Тогда поднаборы размера $k - 1 = 2$, сгенерированные на основе набора s, — $\{спаржа, фасоль\}$, $\{спаржа, кабачки\}$ и $\{фасоль, кабачки\}$.

Из таблицы 4 можно увидеть, что все эти поднаборы являются частыми, значит, и набор $s = \{спаржа, фасоль, кабачки\}$ будет частым и сокращению не подлежит. Таким же образом можно убедиться, что и набор $s = \{фасоль, кукуруза, помидоры\}$ является частым.

Рассмотрим набор $s = \{фасоль, кукуруза, кабачки\}$. Поднабор $\{кукуруза, кабачки\}$ появляется всего три раза (см. таблицу.4), поэтому не является частым. Тогда в соответствии со свойством антимонотонности и набор $s = \{фасоль, кукуруза, кабачки\}$ не будет частым — мы должны его отбросить.

Теперь рассмотрим набор {фасоль, кабачки, помидоры}. Поскольку поднабор $\{кабачки, помидоры\}$ не является частым (частота его появления — всего 2), набор $\{фасоль, кабачки, помидоры\}$ также не является частым и вследствие этого будет исключен из рассмотрения.

Таким образом, в множество F_3 3-предметных частых наборов попадают два набора — $\{спаржа, фасоль, кабачки\}$ и $\{фасоль, кукуруза, помидоры\}$. Их уже нельзя связать, поэтому задача поиска частых предметных наборов на исходном множестве транзакций решена.

Генерация ассоциативных правил

После того как все частые предметные наборы найдены, можно переходить к генерации на их основе ассоциативных правил. Для этого к каждому частому предметному набору s, полученному на основе множества транзакций D, нужно применить процедуру, состоящую из двух шагов:

1. Генерируются все возможные поднаборы s.

2 Если поднабор ss является непустым поднабором s, то рассматривается ассоциативное правило R: ss $\rightarrow (s - ss)$, где $s - ss$ представляет собой набор s без поднабора ss. R будет считаться ассоциативным правилом, если будет удовлетворять условию заданного минимума поддержки и достоверности. Данная процедура повторяется для каждого подмножества ss из s.

Рассмотрим предметные наборы – кандидаты в ассоциативные правила, содержащие два предмета в условии, например набор $s = \{спаржа, фасоль, кабачки\}$ из множества 3-компонентных предметных наборов F_3, полученных на этапе поиска частых наборов. Соответствующими поднаборами s являются: $\{спаржа\}$, $\{фасоль\}$, $\{кабачки\}$, $\{спаржа, фасоль\}$, $\{спаржа, кабачки\}$, $\{фасоль, кабачки\}$.

Таблица 5 – Ассоциативные правила с двумя предметами в условии

Если условие, то следствие	Поддержка	Достоверность	Если условие, то следствие
Если $\{спаржа$ и $фасоль\}$, то $\{кабачки\}$	$4/14 = 28{,}6\ \%$	$4/5 = 80\ \%$	Если $\{спаржа$ и $фасоль\}$, то $\{кабачки\}$
Если $\{спаржа$ и $кабачки\}$, то $\{фасоль\}$	$4/14 = 28{,}6\ \%$	$4/5 = 80\ \%$	Если $\{спаржа$ и $кабачки\}$, то $\{фасоль\}$

Для первого ассоциативного правила в таблицы 5 предположим, что $ss = \{спаржа, фасоль\}$, и тогда $(s - ss) = \{кабачки\}$.

Рассмотрим правило R: $\{спаржа, фасоль\} \rightarrow \{кабачки\}$.

Поддержка, показывающая долю транзакций, которые содержат как условие $\{спаржа, фасоль\}$, так и следствие $\{кабачки\}$, в общем наборе транзакций, имеющихся в базе данных, составляет $28{,}6$ % (4 из 14 транзакций). Чтобы найти достоверность, мы должны учесть, что набор $\{спаржа, фасоль\}$ появляется в 5 из 14 транзакций, 4 из которых также содержат {кабачки}. Тогда достоверность будет $4/5 = 80$ %. Аналогично определяются поддержка и достоверность для остальных правил в таблице 5.

Если предположить, что минимальная достоверность для правила составляет 60 %, то все ассоциации, представленные в таблице 5, будут правилами. Если порог установить равным 80 %, то правилами будут считаться только первые две ассоциации.

Наконец, рассмотрим кандидатов в правила, содержащих одно условие и одно следствие. Для этого применим описанную выше методику генерации ассоциативных правил к множеству F_2 2-компонентных предметных наборов, и результаты представим в таблице 6.

Таблица 6 – Ассоциативные правила с одним предметом в условии

Если условие, то следствие	Поддержка	Достоверность
Если {*спаржа*}, то {*фасоль*}	5/14 = 35,7 %	5/6 = 83,3 %
Если {*фасоль*}, то {*спаржа*}	5/14 = 35,7 %	5/10 = 50 %
Если {*спаржа*}, то {*кабачки*}	5/14 = 35,7 %	5/6 = 83,3 %
Если {*кабачки*}, то {*спаржа*}	5/14 = 35,7 %	5/7 = 71,4 %
Если {*фасоль*}, то {*кукуруза*}	5/14 = 35,7 %	5/10 = 50 %
Если {*кукуруза*}, то {*фасоль*}	5/14 = 35,7 %	5/8 = 62,5 %
Если {*фасоль*}, то {*кабачки*}	6/14 = 42,9 %	6/10 = 60 %
Если {*кабачки*}, то {*фасоль*}	6/14 = 42,9 %	6/7 = 85,7 %
Если {*фасоль*}, то {*помидоры*}	4/14 = 28,6 %	4/10 = 40 %
Если {*помидоры*}, то {*фасоль*}	4/14 = 28,6 %	4/6 = 66,7 %
Если {*капуста*}, то {*перец*}	4/14 = 28,6 %	4/5 = 80 %
Если {*перец*}, то {*капуста*}	4/14 = 28,6 %	4/5 = 80 %
Если {*кукуруза*}, то {*помидоры*}	4/14 = 28,6 %	4/8 = 50 %
Если {*помидоры*}, то {*кукуруза*}	4/14 = 28,6 %	4/6 = 66,7 %

Чтобы проверить значимость сгенерированных правил, обычно перемножают их значения поддержки и достоверности, что позволяет аналитику ранжировать правила в соответствии с их значимостью и достоверностью. В таблице 7 представлен список правил, сгенерированных на основе исходного множества транзакций (см. таблицу 2) при заданном уровне минимальной достоверности 80 %.

Таблица 7 – Ассоциативные правила

Если условие, то следствие	Поддержка, S	Достоверность, C	C x S
Если {*кабачки*}, то {*фасоль*}	6/14 = 42,9 %	6/7 = 85,7 %	0,3677
Если {*спаржа*}, то {*фасоль*}	5/14 = 35,7 %	5/6 = 83,3 %	0,2974
Если {*спаржа*}, то {*кабачки*}	5/14 = 35,7 %	5/6 = 83,3 %	0,2974
Если {*капуста*}, то {*перец*}	4/14 = 28,6 %	4/5 = 80 %	0,2288
Если {*перец*}, то {*капуста*}	4/14 = 28,6 %	4/5 = 80 %	0,2288
Если {*спаржа* и *фасоль*}, то {*кабачки*}	4/14 = 28,6 %	4/5 = 80 %	0,2288
Если {*спаржа* и *кабачки*}, то {*фасоль*}	4/14 = 28,6 %	4/5 = 80%	0,2288

Таким образом, в результате применения алгоритма Apriori нам удалось обнаружить 7 ассоциативных правил, с достоверностью не менее 80 % показывающих, какие продукты из исходного набора чаще всего продаются вместе. Это знание позволит разработать более совершенную маркетинговую стратегию, оптимизировать закупки и размещение товара на прилавках и витринах.