Ассоциативные правила

Ассоциативные правила

Аффинитивный анализ (affinity analysis) — один из распространенных методов Data Mining. Его название происходит от английского слова affinity, которое в переводе означает «близость», «сходство». Цель данного метода — исследование взаимной связи между событиями, которые происходят совместно. Разновидностью аффинитивного анализа является *анализ рыночной корзины* (market basket analysis), цель которого — обнаружить ассоциации между различными событиями, то есть найти правила для количественного описания взаимной связи между двумя или более событиями. Такие правила называются *ассоциативными правилами* (association rules).

Примерами приложения ассоциативных правил могут быть следующие задачи:

- выявление наборов товаров, которые в супермаркетах часто покупаются вместе или никогда не покупаются вместе;
- определение доли клиентов, положительно относящихся к нововведениям в их обслуживании;
- определение профиля посетителей веб-ресурса;
- определение доли случаев, в которых новое лекарство показывает опасный побочный эффект.

Базовым понятием в теории ассоциативных правил является *транзакция* — некоторое множество событий, происходящих совместно. Типичная транзакция —приобретение клиентом товара в супермаркете. В подавляющем большинстве случаев клиент покупает не один товар, а набор товаров, который называется рыночной корзиной. При этом возникает вопрос: является ли покупка одного товара в корзине следствием или причиной покупки другого товара, то есть связаны ли данные события? Эту связь и устанавливают ассоциативные правила. Например, может быть обнаружено ассоциативное правило, утверждающее, что клиент, купивший молоко, с вероятностью 75 % купит и хлеб.

Следующее важное понятие — *предметный набор*. Это непустое множество предметов, появившихся в одной транзакции.

Анализ рыночной корзины — это анализ наборов данных для определения комбинаций товаров, связанных между собой. Иными словами, производится поиск товаров, присутствие которых в транзакции влияет на вероятность наличия других товаров или комбинаций товаров. Современные кассовые аппараты в супермаркетах позволяют собирать информацию о покупках, которая может храниться в базе данных. Затем накопленные данные могут использоваться для построения систем поиска ассоциативных правил.

Таблица 1– Пример набора транзакций

№	Транзакция
1	Сливы, салат, помидоры
2	Сельдерей, конфеты
3	Конфеты
4	Яблоки, морковь, помидоры, картофель, конфеты
5	Яблоки, апельсины, салат, конфеты, помидоры
6	Персики, апельсины, сельдерей, помидоры
7	Фасоль, салат, помидоры
8	Апельсины, салат, морковь, помидоры, конфеты
9	Яблоки, бананы, сливы, морковь, помидоры, лук, конфеты

Визуальный анализ примера показывает, что все четыре транзакции, в которых фигурирует салат, также включают помидоры и что четыре из семи транзакций, содержащих помидоры, также содержат салат. Салат и помидоры в большинстве случаев покупаются вместе. Ассоциативные правила позволяют обнаруживать и количественно описывать такие совпадения.

Ассоциативное правило состоит из двух наборов предметов, называемых условие (antecedent) и следствие (consequent), записываемых в виде X → Y, что читается следующим образом: «Из X следует Y». Таким образом, ассоциативное правило формулируется в виде: «Если условие, то следствие».

Условие может ограничиваться только одним предметом. Правила обычно отображаются с помощью стрелок, направленных от условия к следствию, например, помидоры → салат. Условие и следствие часто называются соответственно левосторонним (left-hand side — LHS) и правосторонним (right-hand side — RHS) компонентами ассоциативного правила.

Ассоциативные правила описывают связь между наборами предметов, соответствующими условию и следствию. Эта связь характеризуется двумя показателями — поддержкой (support) и достоверностью (confidence).

Поддержка ассоциативного правила — это число транзакций, которые содержат как условие, так и следствие. Например, для ассоциации $A \rightarrow B$ можно записать:

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

Достоверность ассоциативного правила $A \rightarrow B$ представляет собой меру точности правила и определяется как отношение количества транзакций, содержащих и условие, и следствие, к количеству транзакций, содержащих только условие:

$$C(A \rightarrow B) = P(B \mid A) = P(B \cap A)/P(A) = \frac{\text{количество транзакций, содержащих A и B}}{\text{количество транзакций, содержащих только A}}.$$

Если поддержка и достоверность достаточно высоки, можно с большой вероятностью утверждать, что любая будущая транзакция, которая включает условие, будет также содержать и следствие.

Возьмем ассоциацию салат → помидоры. Поскольку количество транзакций, содержащих как салат, так и помидоры, равно 4, а общее число транзакций — 9, то поддержка данной ассоциации будет:

$$S(салат \rightarrow помидоры) = 4 / 9 = 0{,}444.$$

Поскольку количество транзакций, содержащих только салат (условие), равно 4, то достоверность данной ассоциации будет:

$$C(салат \rightarrow помидоры) = 4 / 4 = 1.$$

Иными словами, все наблюдения, содержащие салат, также содержат и помидоры, из чего делаем вывод о том, что данная ассоциация может рассматриваться как правило. С точки зрения интуитивного поведения такое правило вполне объяснимо, поскольку оба продукта широко используются для приготовления растительных блюд и часто покупаются вместе.

Теперь рассмотрим ассоциацию конфеты → помидоры, в которой содержатся, в общем-то, слабо совместимые в гастрономическом плане продукты: тот, кто решил приготовить растительное блюдо, вряд ли станет покупать конфеты, а тот, кто желает приобрести что-нибудь к чаю, скорее всего, не станет покупать помидоры. Поддержка данной ассоциации: $S = 4 / 9 = 0{,}444$, а достоверность: $C = 4 / 6 = 0{,}67$. Таким образом, сравнительно невысокая достоверность данной ассоциации дает повод усомниться в том, что она является правилом.

Значимость ассоциативных правил

Методики поиска ассоциативных правил обнаруживают все ассоциации, которые удовлетворяют ограничениям на поддержку и достоверность, наложенным пользователем. Это приводит к необходимости рассматривать десятки и сотни тысяч ассоциаций, что делает невозможным обработку такого количества данных вручную. Число правил желательно уменьшить таким образом, чтобы проанализировать только наиболее значимые из них. Значимость часто вычисляется как разность между поддержкой правила в целом и произведением поддержки только условия и поддержки только следствия.

Если условие и следствие независимы, то поддержка правила примерно соответствует произведению поддержек условия и следствия, то есть $S_{AB} \approx S_A S_B$. Это значит, что хотя условие и следствие часто встречаются вместе, не менее часто они встречаются и по отдельности. Например, если товар A встречался в 70 транзакциях из 100, а товар B — в 80 и в 50 транзакциях из 100 они встречаются вместе, то несмотря на высокую поддержку ($S_{AB} = 0{,}5$) это не обязательно правило. Просто эти товары покупаются независимо друг от друга, но в силу их популярности часто встречаются в одной транзакции. Поскольку произведение поддержек условия и следствия $S_A S_B = 0{,}7 \cdot 0{,}8 = 0{,}56$, то есть отличается от $S_{AB} = 0{,}5$ всего на $0{,}06$, предположение о независимости товаров A и B достаточно обоснованно.

Однако если условие и следствие независимы, то правило вряд ли представляет интерес независимо от того, насколько высоки его поддержка и достоверность. Например, если статистика дорожно-транспортных происшествий показывает, что из 100 аварий в 80 участвуют автомобили марки ВАЗ, то, на первый взгляд, это выглядит как правило «если $авария$, то $ВАЗ$». Но если учесть, что парк автомобилей ВАЗ составляет, скажем, 80 % от общего числа легковых автомобилей, то такое правило вряд ли можно назвать значимым.

Пример

Рассмотрим проект Data Mining в области продаж, который призван объяснить связь между покупками женского белья и других товаров. Простой ассоциативный анализ (рассматривающий отдельные корзины закупок всех клиентов за год) обнаруживает с очень высокой степенью поддержки и достоверности, что женское белье ассоциировано с конфетами. Сначала это вызывает некоторое недоумение, но последующий анализ показывает, что все основные категории товаров ассоциированы с конфетами с высоким уровнем достоверности. Большинство клиентов покупают конфеты в любом случае. Ассоциации с конфетами как следствием с поддержкой $S = 0{,}87$ оказываются малоинтересны, поскольку $87\ \%$ покупателей всегда приобретают конфеты. Чтобы быть значимой, ассоциация с конфетами в качестве следствия должна иметь поддержку большую, чем $0{,}87$.

По этой причине при поиске ассоциативных правил используются дополнительные показатели, позволяющие оценить значимость правила. Можно выделить объективные и субъективные меры значимости правил. Объективными являются такие меры, как поддержка и достоверность, которые могут применяться независимо от конкретного приложения. Субъективные меры связаны со специальной информацией, определяемой пользователем в контексте решаемой задачи. Такими субъективными мерами являются лифт (lift) и левередж (от англ. leverage — плечо, рычаг).

Лифт (оригинальное название – интерес, также встречается термин «улучшение») вычисляется следующим образом:

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

Лифт — это отношение частоты появления условия в транзакциях, которые также содержат и следствие, к частоте появления следствия в целом. Значения лифта большие, чем единица, показывают, что условие чаще появляется в транзакциях, содержащих следствие, чем в остальных. Можно сказать, что лифт является обобщенной мерой связи двух предметных наборов: при значениях лифта > 1 связь положительная, при 1 она отсутствует, а при значениях < 1 — отрицательная.

Рассмотрим ассоциацию помидоры → салат из таблицы 1.

$S(салат) = 4/9 = 0{,}444$; $C(помидоры \rightarrow салат) = 4/7 = 0{,}57$.

Следовательно, $L(помидоры \rightarrow салат) = 0{,}57/0{,}444 = 1{,}285$.

Теперь рассмотрим ассоциацию помидоры → конфеты.

$S(конфеты) = 0{,}67$; $C(помидоры \rightarrow конфеты) = 4/7 = 0{,}57$.

Тогда $L(помидоры \rightarrow конфеты) = 0{,}57/0{,}67 = 0{,}85$.

Большее значение лифта для первой ассоциации показывает, что помидоры больше влияют на частоту покупок салата, чем конфет.

Хотя лифт используется широко, он не всегда оказывается удачной мерой значимости правила. Правило с меньшей поддержкой и большим лифтом может быть менее значимым, чем альтернативное правило с большей поддержкой и меньшим лифтом, потому что последнее применяется для большего числа покупателей. Значит, увеличение числа покупателей приводит к возрастанию связи между условием и следствием. Другой мерой значимости правила, предложенной Г. Пятецким-Шапиро, является левередж:

$$T(A \rightarrow B) = S(A \rightarrow B) - S(A)S(B).$$

Левередж — это разность между наблюдаемой частотой, с которой условие и следствие появляются совместно (то есть поддержкой ассоциации), и произведением частот появления (поддержек) условия и следствия по отдельности.

Рассмотрим ассоциации морковь → помидоры и салат → помидоры, которые имеют одинаковую поддержку $C = 1$, поскольку салат и морковь всегда продаются вместе с помидорами (см. таблицу 1). Лифты для данных ассоциаций также будут одинаковыми, поскольку в обеих ассоциациях поддержка следствия $S(помидоры) = 7/9 = 0{,}78$.

Тогда $L(морковь \rightarrow помидоры) = L(салат \rightarrow помидоры) = 1/0{,}78 = 1{,}285$.

Последняя ассоциация представляет больший интерес, так как она встречается чаще, то есть применяется для большего числа покупателей.

$S(морковь \rightarrow помидоры) = 3/9 = 0{,}333$; $S(морковь) = 0{,}333$; $S(помидоры) = 0{,}78$.

Таким образом, $T(морковь \rightarrow помидоры) = 0{,}333 - 0{,}333 \cdot 0{,}78 = 0{,}073$.

S(салат → помидоры) = 0,444; S(салат) = 0,444; S(помидоры) = 0,78.

Следовательно, $T(салат \rightarrow помидоры) = 0{,}444 - 0{,}444 \cdot 0{,}78 = 0{,}098$.

Итак, значимость второй ассоциации больше, чем первой.

Такие меры, как лифт и левередж могут использоваться для последующего ограничения набора рассматриваемых ассоциаций путем установки порога значимости, ниже которого ассоциации отбрасываются.

Поиск ассоциативных правил

В процессе поиска ассоциативных правил может производиться обнаружение всех ассоциаций, поддержка и достоверность для которых превышают заданный минимум. Простейший алгоритм поиска ассоциативных правил рассматривает все возможные комбинации условий и следствий, оценивает для них поддержку и достоверность, а затем исключает все ассоциации, которые не удовлетворяют заданным ограничениям. Число возможных ассоциаций с увеличением числа предметов растет экспоненциально. Если в базе данных транзакций присутствует k предметов и все ассоциации являются бинарными (то есть содержат по одному предмету в условии и следствии), то потребуется проанализировать $k \cdot 2^{k-1}$ ассоциаций. Поскольку реальные базы данных транзакций, рассматриваемые при анализе рыночной корзины, обычно содержат тысячи предметов, вычислительные затраты при поиске ассоциативных правил огромны. Например, если рассматривать выборку, содержащую всего 100 предметов, то количество ассоциаций, образуемых этими предметами, составит $100 \cdot 2^{99} \approx 6{,}4 \cdot 10^{31}$. Поиск ассоциативных правил путем вычисления поддержки и достоверности для всех возможных ассоциаций и сравнения их с заданным пороговым значением малоэффективен из-за больших вычислительных затрат.

Поэтому в процессе генерации ассоциативных правил широко используются методики, позволяющие уменьшить количество ассоциаций, которое требуется проанализировать. Одной из наиболее распространенных является методика, основанная на обнаружении так называемых частых наборов, когда анализируются только те ассоциации, которые встречаются достаточно часто. На этой концепции основан известный алгоритм поиска ассоциативных правил *Apriori*.

Ассоциативные правила – алгоритм Apriori

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.

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

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 $Z È A$ также не будет являться таковым. Данное полезное свойство позволяет значительно уменьшить пространство поиска ассоциативных правил.

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 F_k – 1 с самим собой. Затем F_k 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-х предметных наборов мы получим:

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

Аналогично $\{фасоль, кукуруза\}$ и $\{фасоль, кабачки\}$ могут быть объединены в 3предметный 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.

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*}

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 3компонентных предметных наборов F_3, полученных на этапе поиска частых наборов. Соответствующими поднаборами s являются: $\{спаржа\}$, $\{фасоль\}$, $\{кабачки\}$, $\{спаржа, фасоль\}$, $\{спаржа, кабачки\}$, $\{фасоль, кабачки\}$.

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 2компонентных 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 %

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

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 %.

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.

Иерархические ассоциативные правила

Если производить поиск ассоциативных правил среди отдельных предметов (например, товаров в магазине), то можно обнаружить, что во многих случаях ассоциации с высокой поддержкой для отдельных товаров практически отсутствуют. Особенно это характерно для супермаркетов, где ассортимент товаров каждого вида очень велик. Например, в продаже могут иметься десятки разновидностей таких товаров, как $кетчуп$ и $макароны$. Поэтому, несмотря на то что в целом поддержка ассоциации $макароны \rightarrow кетчуп$ может быть очень высока, поддержка ассоциаций между отдельными видами этих товаров, скорее всего, будет низкой. Следовательно, такие ассоциации, хотя и могут представлять интерес, окажутся исключенными из рассмотрения, поскольку не будут удовлетворять некоторому минимальному порогу поддержки S_{min}.

Для решения данной проблемы при поиске ассоциативных правил рассматривают не отдельные предметы, а их иерархию. Если на нижних иерархических уровнях интересные ассоциации отсутствуют, то на более высоких они могут иметь место. Иными словами, поддержка отдельного предмета всегда будет меньше, чем поддержка группы, в которую он входит:

$S(I) \geq S(i_j)$, где I — группа в иерархии, i_j

— предмет, входящий в данную группу.

Причины этого очевидны: общая поддержка группы равна сумме поддержек всех входящих в нее предметов:

$$S(I) = \sum_{j=1}^{N} i_j \ ,$$

где N — число предметов в группе.

Например, рассмотрим иерархию продуктов, представленную на рисунке 1.

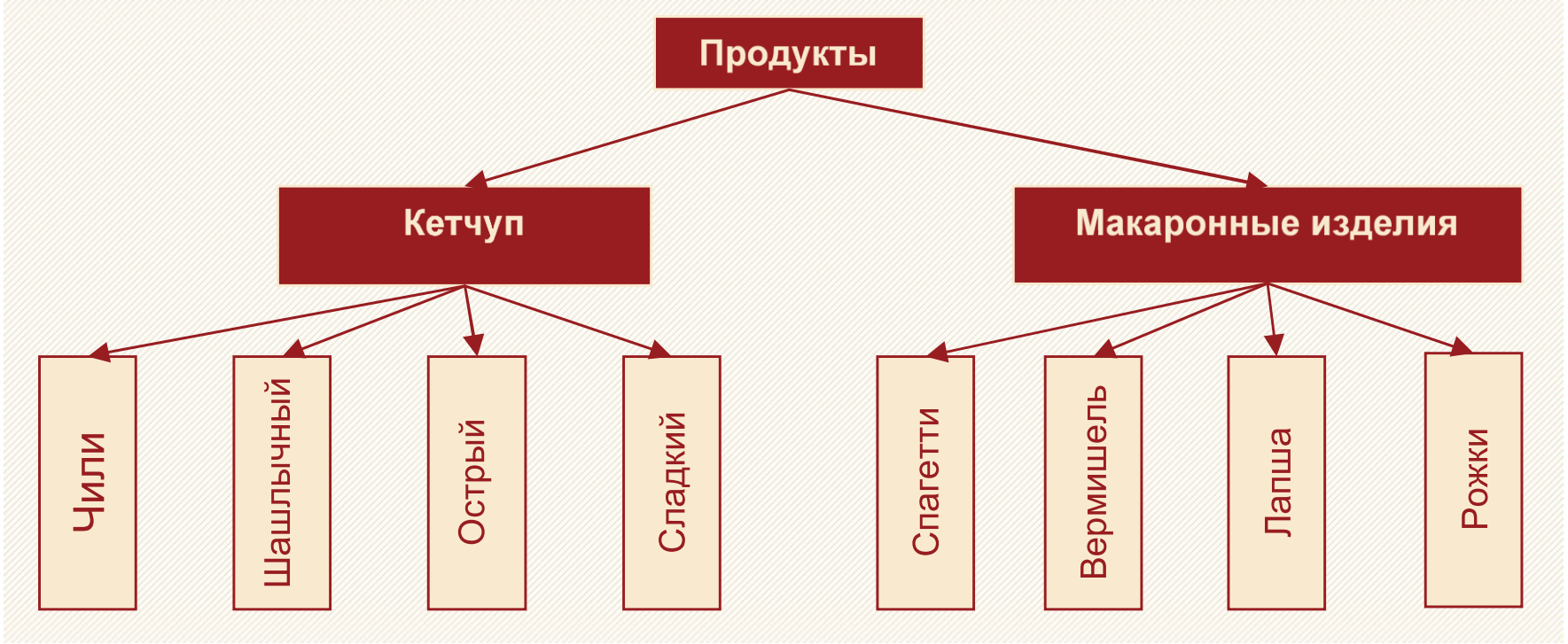


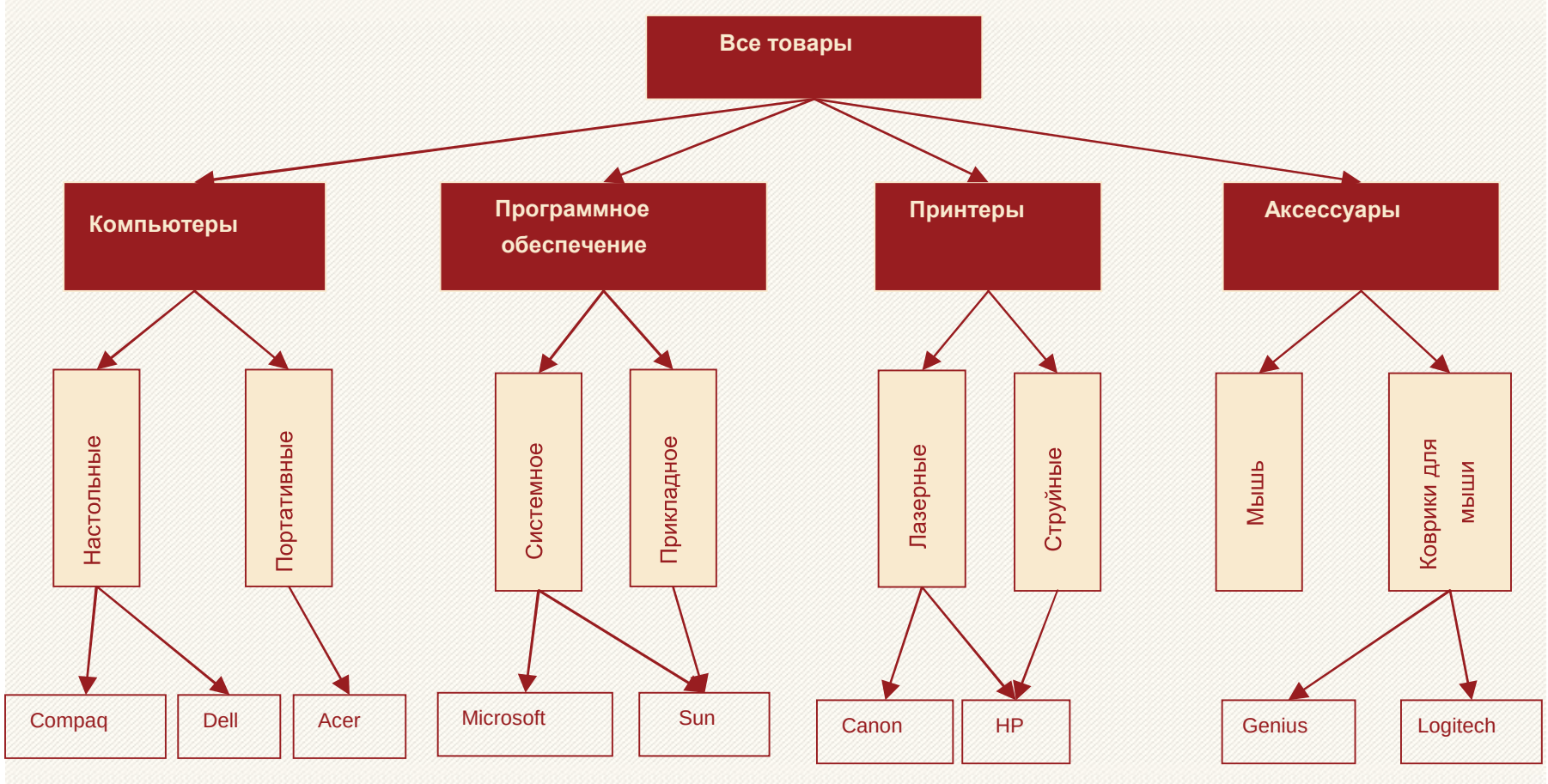
Рисунок 1 – Иерархия продуктов

В этой иерархической схеме кетчуп и макаронные изделия имеют множество подвидов. Транзакции в базе данных являются результатом сканирования штрихкодов на контрольнокассовых пунктах, поэтому содержат информацию только о конкретных товарах. Следовательно, при анализе таких максимально детализованных данных интересные ассоциации могут отсутствовать

Таблица 1 – Пример транзакций

№ транзакции	Предметные наборы
1	Настольный компьютер Acer, лазерный принтер HP
2	OS MS Windows XP, ПО MS Office
3	Мышь Genius, коврик для мыши Logitech
4	Портативный компьютер Dell, ПО MS Office
5	Настольный компьютер Compaq
…	…

Иерархия предметов, связанных с компьютерной техникой, представлена на рисунке .2. Она содержит 4 уровня. Обычно иерархические уровни нумеруются сверху вниз, начиная с нулевого. Узел, соответствующий нулевому уровню, называется корневым (root node). В нашей иерархической схеме уровень 2 включает виды компьютерного обеспечения, уровень 3 — конкретные предметы (товары), а уровень 4 — фирму-производителя.



Предметы из таблицы 1 принадлежат самому низкому уровню иерархии, то есть содержат конкретные товары конкретных производителей. Обнаружить интересные модели покупок на столь низком уровне трудно. Например, если каждый из предметов настольный компьютер *Compaq* и лазерный принтер *Canon* появляется в очень небольшом числе транзакций, то трудно будет обнаружить ассоциацию настольный компьютер

Compaq → лазерный принтер Canon

с высоким уровнем поддержки и достоверности, так как лишь небольшое количество клиентов приобретают их совместно. Хотя в целом поддержка ассоциации

компьютер → принтер,

скорее всего, будет достаточно высока.

Иными словами, ассоциации, которые содержат предметы более высоких уровней иерархии, будут с большей вероятностью удовлетворять условию минимальной поддержки и достоверности, чем предметы уровня с максимальной детализацией. Следовательно, искать интересные ассоциации в многоуровневой иерархии удобнее, чем только среди предметов самого низкого уровня. Более того, можно ожидать, что чем выше уровень иерархии, тем больше вероятность обнаружить ассоциации с высокой поддержкой.

Таким образом, иерархия предметов может использоваться для сокращения числа рассматриваемых предметных наборов.

1. Сначала ищутся ассоциации с высокой поддержкой для верхних уровней иерархии.
2. Анализируются потомки только тех предметов верхних уровней, которые удовлетворяют заданному минимуму поддержки S_{min}. Анализ потомков тех предметов, которые сами по себе являются редкими, не имеет смысла, поскольку они будут встречаться еще реже, чем их предки.

Перемещаясь вверх или вниз по иерархии, мы можем регулировать количество данных, которое нужно обработать. Так, при поиске на высоких уровнях иерархии уменьшается объем обрабатываемых данных, но увеличивается степень обобщенности полученных результатов. При спуске на нижние уровни количество обрабатываемых данных увеличивается, но при этом увеличивается и степень детальности анализа.

В качестве недостатка иерархического подхода к поиску ассоциативных правил иногда указывают на то, что полученные правила в большинстве случаев относятся не к отдельным предметам, а к их группам, что не всегда соответствует требованиям анализа. Следует заметить, что в некоторых приложениях количество отдельных предметов может быть так велико, что обнаружение ассоциативных правил даже на некотором уровне обобщения — это уже удача. Кроме того, бизнес-аналитические технологии в большей мере ориентированы на обобщенные данные, поэтому во многих случаях использование иерархий предметов при поиске ассоциативных правил открывает принципиально новые возможности, делает анализ более гибким и позволяет получать дополнительные знания.

Методы поиска иерархических ассоциативных правил

Существует несколько подходов к поиску иерархических ассоциативных правил. Большинство из них, как и классический алгоритм Apriori, основаны на вычислении поддержки и достоверности. Чаще всего используются нисходящие методы, когда частые предметные наборы исследуются на каждом иерархическом уровне, начиная с первого и заканчивая уровнем с наибольшей детализацией. Проще говоря, как только обнаруживаются все популярные предметные наборы на первом уровне, начинается поиск популярных предметных наборов на втором и т. д. На каждом уровне для открытия популярных наборов может использоваться любой алгоритм, например Apriori и его модификации. Приведем несколько вариантов таких подходов.

Вариант 1 — использование одинакового порога минимальной поддержки S_{min} на всех иерархических уровнях. При поиске правил на каждом уровне иерархии задается некоторый порог минимальной поддержки. Если поддержка предмета превышает данный порог, то он появляется достаточно часто, чтобы для него имело смысл искать ассоциации с другими предметами. Можно указать некоторый порог поддержки (например, 5 %) и использовать его на всех иерархических уровнях. На рисунке 3 приведен пример, где однопредметные наборы *компьютер* и *настольный компьютер* будут обнаружены как популярные, а портативный *компьютер* — нет, поскольку он не преодолел заданный порог поддержки.

Рисунок 3 – Иллюстрация первого варианта

При использовании одинаковой минимальной поддержки на всех уровнях процедура поиска частых наборов упрощается. Ее можно дополнительно оптимизировать, если известно, что родительский узел иерархии не содержит частых наборов. В этом случае проверка дочерних узлов может не производиться, поскольку они также не будут содержать частых наборов. Например, если известно, что компьютеры продаются редко, то отдельные их разновидности будут продаваться еще реже.

Однако подход, при котором для поиска кандидатов используется одинаковый порог поддержки на всех уровнях иерархии, имеет ряд недостатков. Так, маловероятно, что предметы нижних уровней продаются так же часто, как предметы более высоких уровней. Если порог минимальной поддержки слишком большой, это может привести к потере полезных ассоциаций между предметами низких уровней. Если порог слишком низкий, это может породить много неинтересных ассоциаций между предметами высоких уровней. Чтобы избежать данных проблем, используется методика адаптивного порога, который будет отличаться для разных уровней иерархии.

Вариант 2 — использование пониженного порога минимальной поддержки для нижних уровней иерархии. Данный подход предполагает, что на каждом иерархическом уровне задается свой порог минимальной поддержки для отбора кандидатов. При этом чем ниже уровень, тем ниже порог. Например, на рисунке 4 порог минимальной поддержки для уровней 1 и 2 составляет 0,5 и 0,3 соответственно. Таким образом, и *компьютер*, и *настольный компьютер*, и *портативный компьютер* окажутся популярными.

Рисунок 4 – Иллюстрация второго варианта

Для поиска иерархических ассоциативных правил с уменьшением минимальной поддержки на нижних уровнях можно использовать несколько альтернативных стратегий поиска.

Вариант 3 — независимая установка порога. Осуществляется полный поиск, когда отсутствуют априорные сведения, которые могут использоваться для сокращения числа рассматриваемых предметных наборов. Проверяется каждый узел независимо от того, содержит ли его родительский узел частые предметные наборы. Здесь возможны следующие ситуации.

- Межуровневая (cross-level) фильтрация по одному предмету. Предмет на i-м уровне проверяется тогда и только тогда, когда его родительский узел на уровне $i-1$ содержит частые наборы. Например, на рисунке 5 потомки узла *компьютер* (то есть *настольный компьютер* и *портативный компьютер*) не проверяются, поскольку узел компьютер не является частым: его поддержка $S = 0{,}1$, а порог $S_{min} = 0{,}12$.

Рисунок 5 – Межуровневая фильтрация по одному предмету

- Межуровневая фильтрация по k-предметному набору. k-предметный набор на i-м уровне проверяется тогда и только тогда, когда его родительский k-предметный набор на уровне $i-1$ является частым. Например, на рисунке 6 двухпредметный набор $\{компьютер, принтер\}$ является частым, а предметные наборы $\{портативный\ компьютер, струйный\ принтер\}$, $\{портативный\ компьютер, лазерный\ принтер\}$, $\{настольный\ компьютер, струйный\ принтер\}$ и $\{настольный\ компьютер, лазерный\ принтер\}$ должны проверяться.

Рисунок 6 – Межуровневая фильтрация по k предметам

может привести к проверке огромного количества редко встречающихся предметов на нижних уровнях с обнаружением ассоциаций между предметами с низкой значимостью. Например, если *компьютерная мебель* сама по себе покупается редко, то проверка, является ли частым набор {*компьютерное кресло*, *портативный компьютер*} не имеет смысла. В то же время, если аксессуары для компьютера покупаются часто, то проверка ассоциаций *настольный компьютер* $\rightarrow$ *мышь* имеет смысл.

Стратегия межуровневой фильтрации по k-предметному набору позволяет ограничить число проверяемых наборов теми, которые являются потомками частых k-предметных наборов. Данное ограничение весьма эффективно, поскольку обычно не существует большого количества частых k-предметных наборов (особенно при $k > 2$). Следовательно, этот подход дает возможность значительно сократить число рассматриваемых наборов.

Одна из проблем такой стратегии заключается в том, что предметные наборы, которые на нижних уровнях иерархии могли бы оказаться частыми, «выпадут» из рассмотрения, поскольку их предки не смогли преодолеть более высокий порог, применяемый на верхних уровнях. Например, если предмет *монитор* $17''$ на уровне i является частым в соответствии с порогом минимальной поддержки для данного уровня, то его предок *монитор* на уровне $i-1$ может не быть частым, так как порог на этом уровне будет выше. В результате такие часто встречающиеся ассоциации, как *настольный компьютер* $\rightarrow$ *монитор* $17''$, могут быть утеряны.

Модифицированной версией межуровневой фильтрации по одному предмету является управляемая межуровневая фильтрация. В этом методе вводится еще один порог, называемый *уровнем прохода* (level passage). Он может быть установлен для относительно часто встречающихся предметов на нижних уровнях. Иными словами, данный подход позволяет потомкам предметов, которые не удовлетворяют основному порогу минимальной поддержки, подвергаться проверке, но только если эти предметы удовлетворяют проходному порогу. Каждый уровень может иметь свой собственный проходной порог. Он обычно выбирается между значениями порогов минимальной поддержки для следующего уровня и данного уровня.

Например, на рисунке 7 установка на уровне 1 проходного порога, равного 0,08, позволяет проверить и отнести к частым узлы *портативный компьютер* и *настольный компьютер* на уровне 2 даже в том случае, если их родительский узел *компьютер* не является таковым. Эта методика дает возможность сделать процесс поиска иерархических ассоциативных правил более гибким, а также уменьшить число ассоциаций с низкой значимостью.

Рисунок 7 – Управляемая межуровневая фильтрация