

Data Mining

Ассоциативные правила

- 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.

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.

Базовые понятия

- *транзакция* — некоторое множество событий, происходящих совместно.
- *предметный набор* — это непустое множество предметов, появившихся в одной транзакции.

№	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

- ассоциативное правило формулируется в виде: «**Если условие, то следствие**».
- Условие может ограничиваться только одним предметом
- (left-hand side — LHS) и (right-hand side — RHS) компоненты

Показатели

- *Поддержка ассоциативного правила — это число транзакций, которые содержат как условие, так и следствие.*

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

Показатели

- *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 условие:*

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

Значимость ассоциативных правил

Если условие и следствие независимы, то поддержка правила примерно соответствует произведению поддержек условия и следствия, то есть $S_{AB} \approx S_A S_B$

Пример с товарами и автомобилем ВАЗ

Дополнительные показатели

- *Лифт* — это отношение частоты появления условия в транзакциях, которые также содержат и следствие, к частоте появления следствия в целом.
- $L(A \rightarrow B) = C(A \rightarrow B) / S(B)$.
 - $L > 1$, связь положительная
 - $L = 1$ связь отсутствует
 - $L < 1$ связь отрицательная

НО!

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.

L > , S <	L < , S >
<p>Поддержка мала</p> <p>количество транзакций, содержащих A и B < общее количество транзакций</p>	<p>Поддержка велика</p> <p>количество транзакций, содержащих A и B \approx общее количество транзакций</p>
<p>Лифт большой</p> <p>Мера > поддержки следствия(B)</p>	<p>Лифт малый</p> <p>Мера < поддержки следствия(B)</p>
<p>количество транзакций, содержащих A и B > количество транзакций, содержащих только A, (условие) * количество транзакций, содержащих только B, (следствие) /общее количество транзакций</p>	<p>количество транзакций, содержащих A и B < количество транзакций, содержащих только A, (условие) * количество транзакций, содержащих только B, (следствие)/общее количество транзакций</p>

Чего не хватает для формулы значимости ???

Дополнительные показатели

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. Предложена Г. Пятецким-Шапиро.

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

- 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.

Алгоритм Apriori

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

Методика поиска

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

СВОЙСТВО АНТИМОНОТОННОСТИ

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

Т.е. , если Z не является частым набором, то и набор $Z \cup A$ также не будет являться таковым.

Набор транзакций D

№ 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

№ 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

$F1 = \{\text{спаржа, фасоль, капуста, кукуруза, перец, кабачки, помидоры}\}$

Создание множеств F_k

- 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.

Множества F_2

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		

$F_2 =$

{спаржа, фасоль}

{спаржа, кабачки}

{фасоль, кукуруза}

{фасоль, кабачки}

{фасоль, помидоры}

{капуста, перец}

{кукуруза, помидоры}

Генерация множеств $F3$

Для этого нужно связать наборы из множества $F2$ между собой, если у них первые $k - 1$ предметов общие.

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

$$F_3 = \begin{aligned} & \{ \text{спаржа, фасоль, кабачки} \} \\ & \{ \text{фасоль, кукуруза, кабачки} \} \\ & \{ \text{фасоль, кабачки, помидоры} \} \\ & \{ \text{фасоль, кукуруза, помидоры} \} \end{aligned}$$

Теперь F_3 также сокращается с помощью свойства антимонотонности. Для каждого предметного набора s из множества F_3 создаются и проверяются поднаборы размером $k - 1$.

Генерация ассоциативных правил

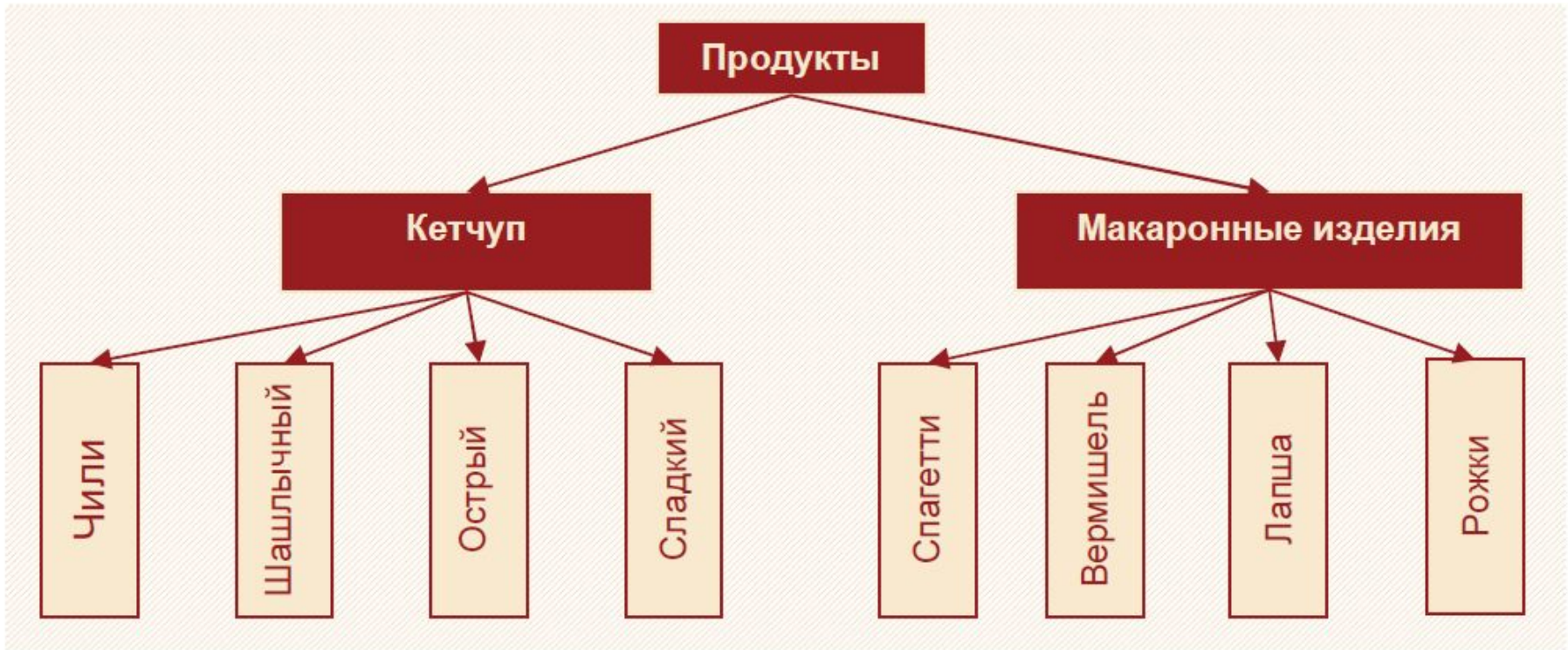
- 1 Генерируются все возможные поднаборы s
- 2 Если поднабор ss является непустым поднабором s , то рассматривается ассоциативное правило $R: ss \rightarrow (s - ss)$, где $s - ss$ представляет собой набор s без поднабора ss .

{спаржа, фасоль, кабачки}

и {фасоль, кукуруза, помидоры}

- Для первого ассоциативного правила $ss = \{спаржа, фасоль\}$, и тогда $(s - ss) = \{кабачки\}$

Иерархические ассоциативные правила

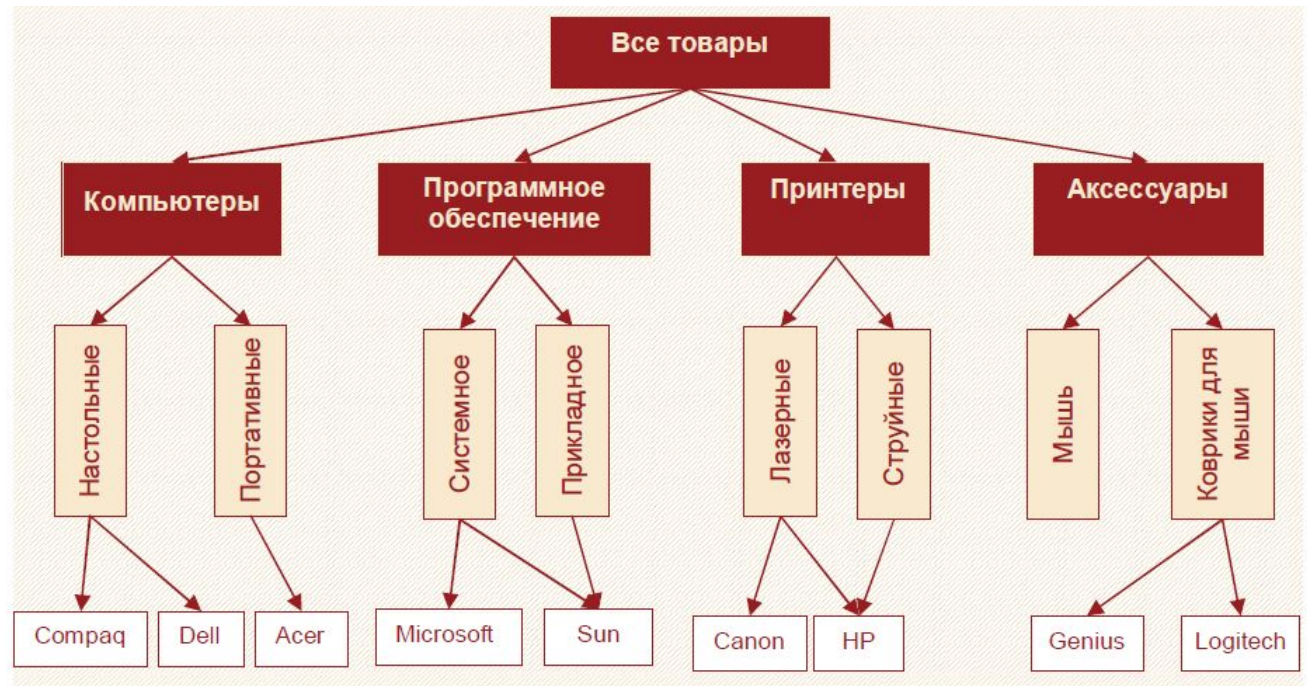


$$S(I) \geq S(i_j),$$

- где I — группа в иерархии,
- i_j — предмет, входящий в данную группу.

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

- Ассоциативные правила, обнаруженные для предметов, расположенных на различных иерархических уровнях, получили название *иерархические ассоциативные правила*.
- *В зарубежной литературе они также известны как многоуровневые правила (multilevel rules) или обобщенные правила (generalized rules).*



№ 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
...	...

- **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.**

Методы поиска иерархических ассоциативных правил

Вариант 1 — использование одинакового порога минимальной поддержки S_{min} на всех иерархических уровнях.



- Маловероятно, что предметы нижних уровней продаются так же часто, как предметы более высоких уровней.
- Если S_{min} слишком большой, это может привести к потере полезных ассоциаций между предметами низких уровней.
- Если S_{min} слишком низкий, это может породить много неинтересных ассоциаций между предметами высоких уровней.

Вариант 2 — использование пониженного порога минимальной поддержки для нижних уровней иерархии



Вариант 3 — независимая установка порога.

- 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.



- **Вариант 3** — независимая установка порога.
- 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 поддается критерию

Уровень 1 ($S_{min}=0,05$)

Компьютер, принтер ($S=0,07$)

Уровень 2
($S_{min}=0,02$)

Портативный компьютер и лазерный принтер $S=0,01$

Портативный компьютер и струйный принтер $S=0,02$

Настольный компьютер и лазерный принтер $S=0,01$

Настольный компьютер и струйный принтер $S=0,03$

