Иерархические ассоциативные правила [M.105]

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* → *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.

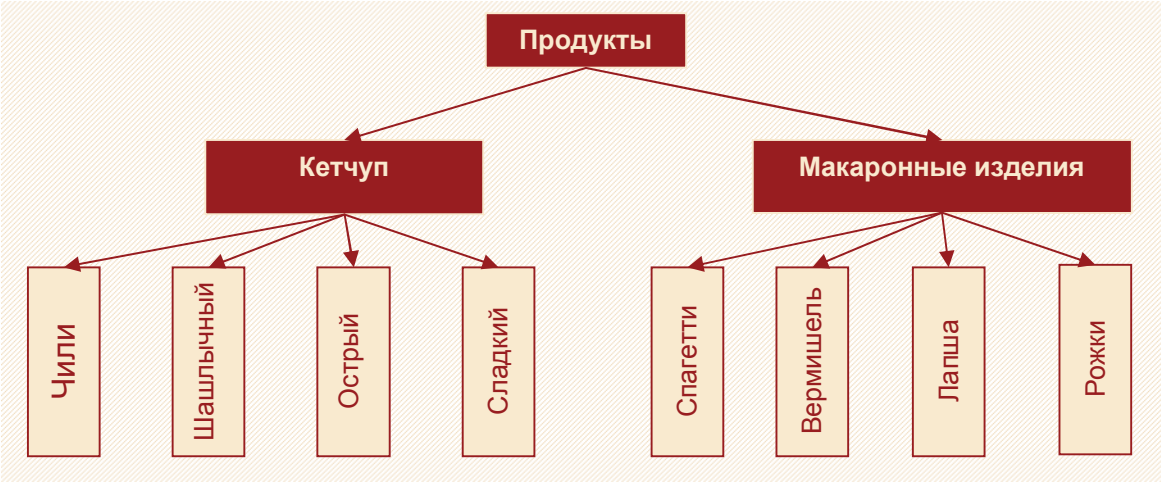

Рисунок 1 – Иерархия продуктов

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

Если структура базы данных транзакций позволяет отражать иерархию товаров, то можно исследовать все образованные ими иерархические уровни. 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).

Пусть имеется множество транзакционных данных о продажах компьютерной фирмы (таблица 1).

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.

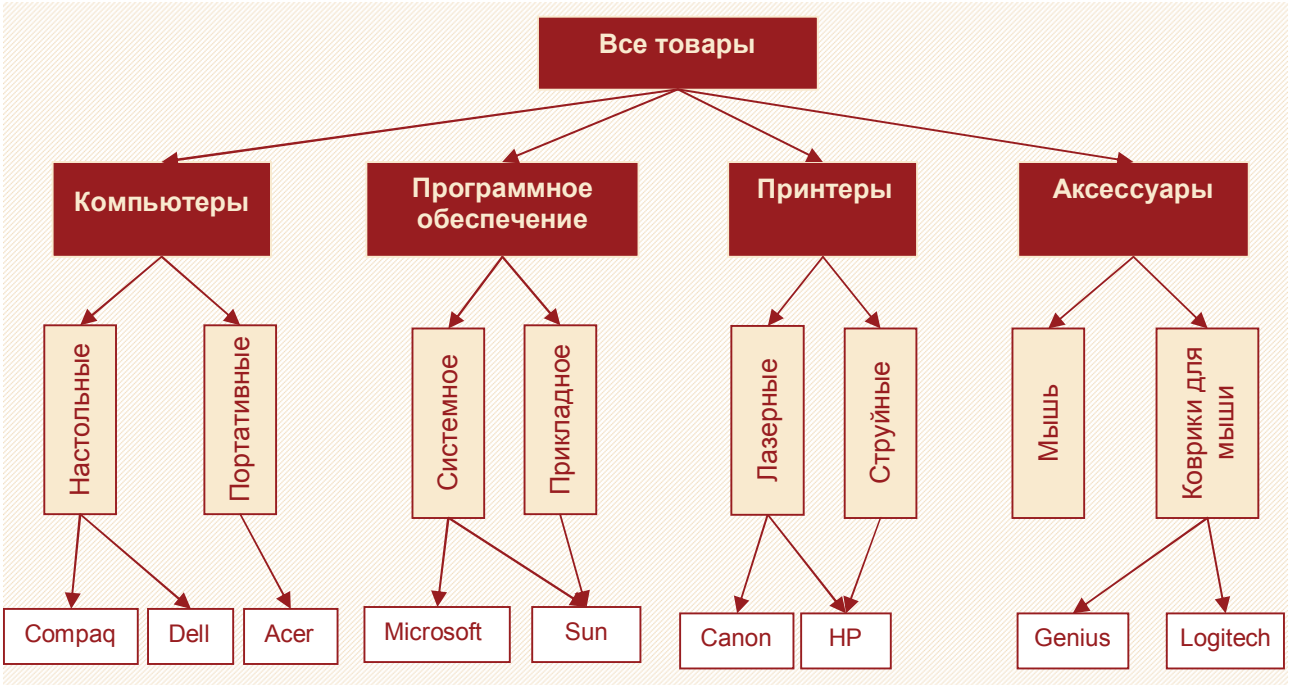

Рисунок 2 –. Иерархия товаров, связанных с компьютерной техникой

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

Стратегия, когда минимальная поддержка устанавливается для каждого уровня независимо, 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* → *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