Тема 10. АССОЦИАТИВНЫЕ ПРАВИЛА

10.1. Введение в ассоциативные правила

Ассоциативные правила представляют собой метод анализа взаимосвязей между переменными в больших базах данных. Часто такими базами выступали данные о покупках, совершаемых в супермаркетах, однако на данный момент сфера применения ассоциативных правил очень широка и не ограничивается анализом потребительской корзины.

Исходные данные для проведения такого вида анализа — наборы транзакций. Под транзакцией понимается множество событий, произошедших одновременно. Например, транзакцией можно считать покупку товаров клиентом супермаркета за один визит, анкету человека, подающего заявку на получение кредита в банк, профиль социально-экономических характеристик муниципального образования, перечень услуг, которыми пользуется абонент сотовой связи, перечень страниц веб-сайта, посещенных за одну сессию и т. п.

Целью анализа является установление следующих зависимостей: если в транзакции встретился некоторый набор элементов X, то на основании этого можно сделать вывод о том, что другой набор элементов Y также же должен появиться в этой транзакции. Установление таких зависимостей дает возможность находить очень простые и интуитивно понятные правила.

Алгоритмы поиска ассоциативных правил предназначены для нахождения всех правил $X \Rightarrow Y$, причем поддержка и достоверность этих правил должны быть выше некоторых заранее определенных порогов, называемых, соответственно минимальной поддержкой (minsupport) и минимальной достоверностью (minconfidence). Аналогично, поддержка и достоверность ограничиваются сверху порогами максимальной поддержки (maxsupport) и максимальной достоверности (maxconfidence). В результате получаются два окна, в которые должны попасть поддержка и достоверность правила, чтобы оно было предъявлено аналитику.

Значения для параметров минимальная (максимальная) поддержка и минимальная (максимальная) достоверность выбираются таким образом, чтобы ограничить количество найденных правил. Если поддержка имеет большое значение, то алгоритмы будут находить правила, хорошо известные аналитикам или настолько очевидные, что нет никакого смысла проводить такой анализ. С другой стороны, низкое значение поддержки ведет к генерации огромного количества правил, что, конечно, требует существенных вычислительных ресурсов. Большинство интересных правил находится именно при низком значении порога поддержки, хотя слишком низкое значение поддержки ведет к генерации статистически необоснованных правил. Ассоциативные правила с высокой поддержкой могут применяться для формализации хорошо известных правил, например, в автоматизированных системах для управления процессами или персоналом. Надо отметить, что понятия «высокая» и «низкая» поддержка или достоверность

очень сильно зависят от предметной области. Например, в торговле 1% вероятности совместного приобретения хлеба и молока не значит ничего, в то время как вероятность в 1% отказа двигателя самолета совершенно неприемлема, и такое правило становится чрезвычайно важным.

Поиск ассоциативных правил совсем не тривиальная задача, как может показаться на первый взгляд. Одна из проблем — алгоритмическая сложность при нахождении часто встречающих наборов элементов, так как с ростом числа элементов экспоненциально растет число потенциальных наборов элементов.

Обычные ассоциативные правила — это правила, в которых как в условии, так и в следствии присутствуют только элементы транзакций и при вычислении которых используется только информация о том, присутствует ли элемент в транзакции или нет.

Для поиска обычных ассоциативных правил в Loginom служит обработчик *Ассоциативные правила*, в котором необходимы следующие настройки.

Для начала необходимо указать, что является идентификатором (ID) транзакции, а что — элементом транзакции. Например, идентификатор транзакции — это номер чека или код накладной. А элемент — это наименование товара в чеке или накладной.

Затем задаются условия, по которым определяются частые предметные наборы — наборы элементов, наиболее часто встречающиеся в транзакциях. В дальнейшем только эти наборы участвуют в формировании правил:

- *минимальная поддержка, %* — минимальная частота, с которой набор встречается в транзакциях (значение 0 до 100);
- *исключать элементы с поддержкой, больше максимальной* — элементы, которые слишком часто встречаются в транзакциях, как правило, не несут информации о закономерностях сочетания с ними других элементов. Для их определения и исключения из частых наборов задается *максимальная поддержка, %* — максимальная частота, с которой элемент встречается в транзакциях (значение от 0 до 100);
- *содержащие выбранные элементы* — задает поля вспомогательного набора данных, содержащие дополнительные элементы транзакций;
- *исключать одиночные наборы* — исключает наборы из одного элемента;
- *максимальное число элементов* — задает максимальное количество элементов в наборе (максимальная мощность набора).

В результирующий набор попадают правила, удовлетворяющие следующим условиям:

- *минимальная достоверность правила, %* — позволяет отсеять наименее точные правила (значение от 0 до 100);
- *минимальный лифт правила* — значение лифта > 1 косвенно подтверждает значимость правила, поскольку говорит о положительной связи двух предметных наборов (условия и следствия правила). Значение лифта, равное или меньшее 1, говорит об отсутствии или отрицательной связи. Задавая минимальную величину лифта, можно отсеять наименее значимые правила;

- *максимальное число следствий* — максимальное количество элементов в наборе, представляющем следствие правила.

Выявление действительно интересных правил — это одна из главных подзадач при вычислении ассоциативных зависимостей. Для того чтобы получить действительно интересные зависимости, нужно разобраться с несколькими эмпирическими правилами:

- *уменьшение минимальной поддержки* приводит к тому, что увеличивается количество потенциально интересных правил. Одним из ограничений уменьшения порога минимальной поддержки является то, что слишком маленькая поддержка правила делает его статистически необоснованным;
- *уменьшение порога достоверности* также приводит к увеличению количества правил. Значение минимальной достоверности не должно быть слишком маленьким, так как ценность правила с достоверностью 5% чаще всего настолько мала, что это и правилом считать нельзя.

Правило со слишком большой поддержкой с точки зрения статистики представляет собой большую ценность, но, с практической точки зрения, это, скорее всего, означает то, что-либо правило всем известно либо товары, присутствующие в нем, являются лидерами продаж, откуда следует их низкая практическая ценность.

Если значение верхнего предела поддержки имеет слишком большое значение, то в обнаруженных правилах основную часть будут составлять товары – лидеры продаж. При таком раскладе не представляется возможным уменьшить минимальный порог поддержки до того значения, при котором могут появляться интересные правила. Причиной тому является просто огромное число правил и, как следствие, нехватка системных ресурсов. Причем получаемые правила примерно на 95% содержат товары — лидеры продаж.

Варьируя верхним и нижним пределами поддержки, можно избавиться от очевидных и неинтересных закономерностей. Как следствие, правила, генерируемые алгоритмом, принимают приближенный к реальности вид.

10.2. Методические указания

В файле *Задача 10.1. Предметы бытовой химии.xlsx* имеются данные для анализа потребительской корзины розничной сети, занимающейся продажей бытовой химии (рис. 10.1). Набор данных насчитывает 5000 чеков.

	A	B
1	Чек	Товар
2	SO51184	Гель для туалетов
3	SO51184	Сода кальцинированная
4	SO51184	Чистящий порошок универсальный
5000	SO60273	Средство по уходу за зеркалами и стеклами
5001	SO60288	Платки носовые

Рис. 10.1

Требуется для выявления совместно приобретаемых товаров в супермаркетах осуществить анализ потребительской корзины с помощью поиска ассоциативных правил.

Решение

Создадим новый пакет *Ассоциативные правила*. Выполним импорт исходных данных. Для этого создадим узел сценария, выполняющий действие импорта (рис. 10.2).

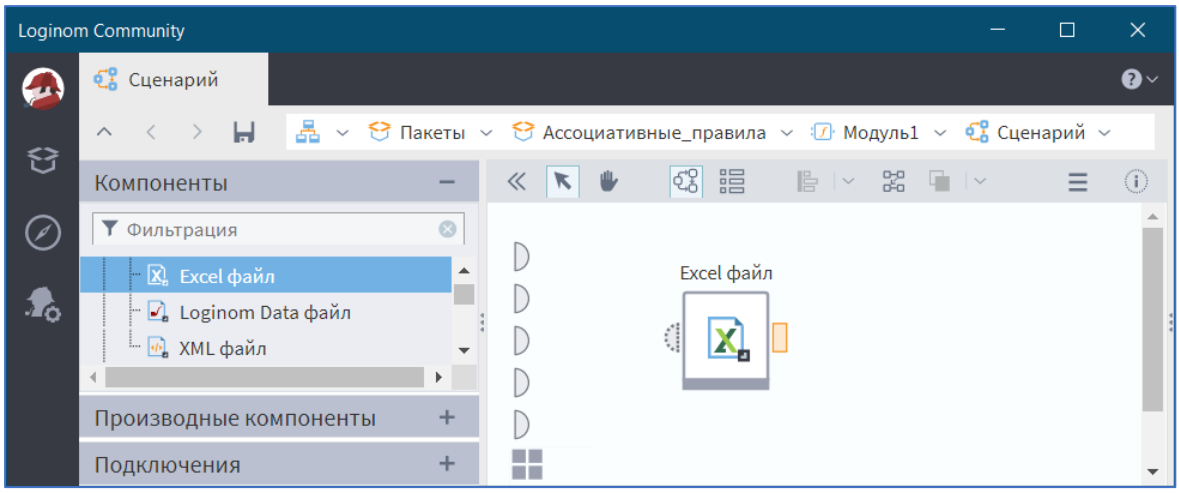


Рис. 10.2

Вызовем *Мастер настройки.* Пройдем шаги мастера. На шаге *Параметры импорта с разделителями* при необходимости укажем вид данных *Дискретный* (рис. 10.3).

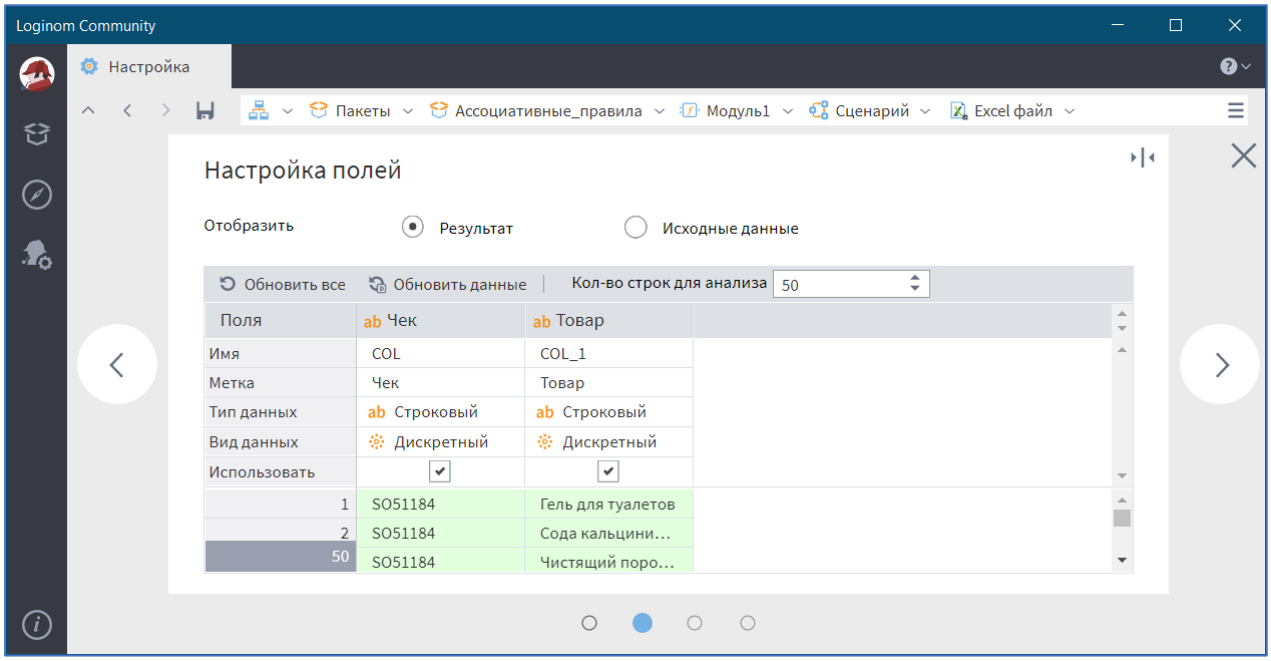


Рис. 10.3

На шаге *Описание узла* укажем в описании узла метку *Задача 10.1. Предметы бытовой химии.*

Добавим визуализатор *Таблица* к узлу сценария (рис. 10.4).

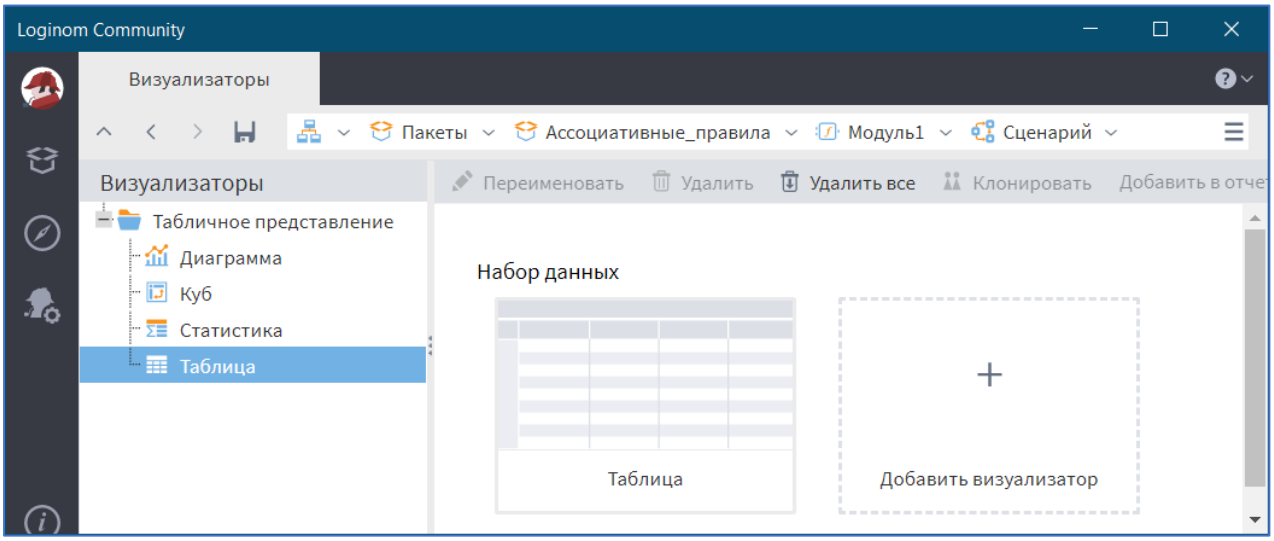


Рис. 10.4

Таблица с исходными данными имеет вид (рис. 10.5).

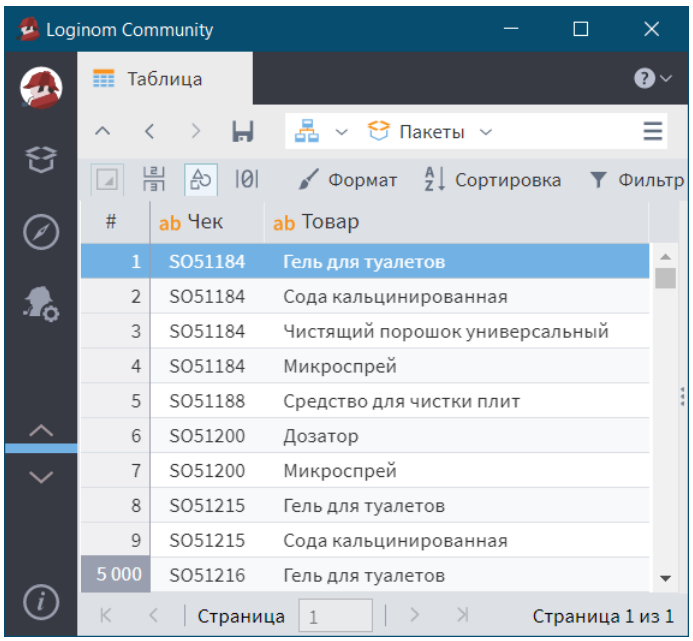


Рис. 10.5

Проведем поиск ассоциативных зависимостей в покупательской корзине на основе импортированных данных. Для этого переместим компонент *Ассоциативные правила* в рабочую область сценария. Последовательность обработки данных задается соединением выходного порта узла импорта с входным портом ассоциативных правил (рис. 10.6).

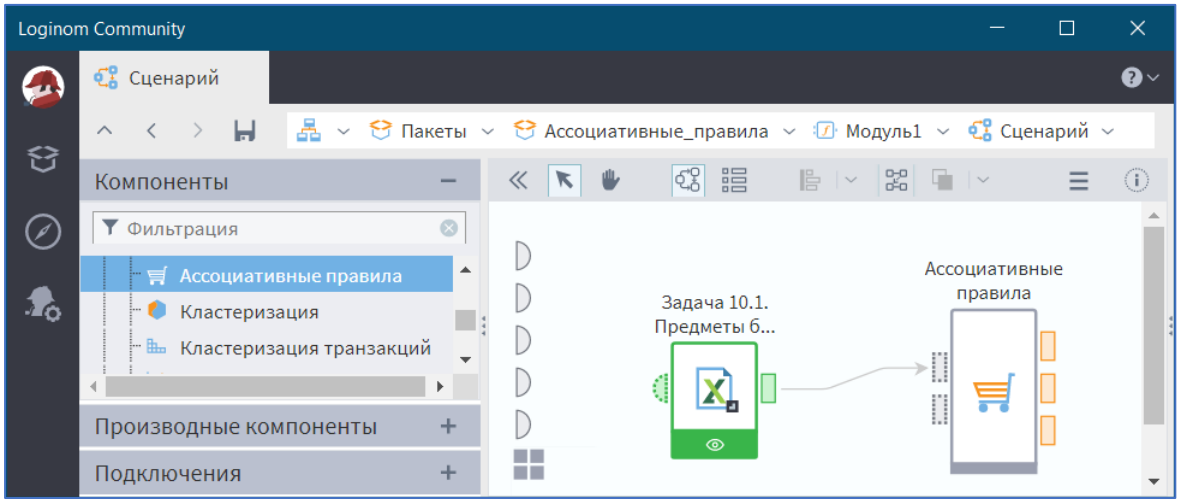


Рис. 10.6

Пройдем шаги *Мастера настройки*. На шаге *Настройка входных столбцов* настроим назначение исходных столбцов данных. Столбец *Чек* зададим как *Транзакция*, столбец *Товар* — как *Элемент* (рис. 10.7).

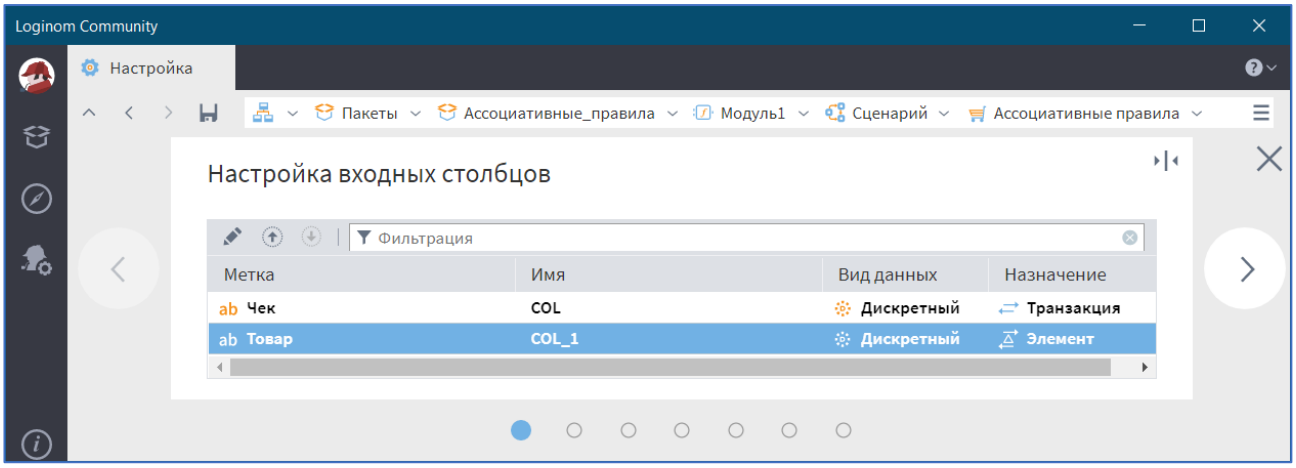


Рис. 10.7

На шаге *Ассоциативные правила* зададим параметры в соответствии с рис. 10.8.

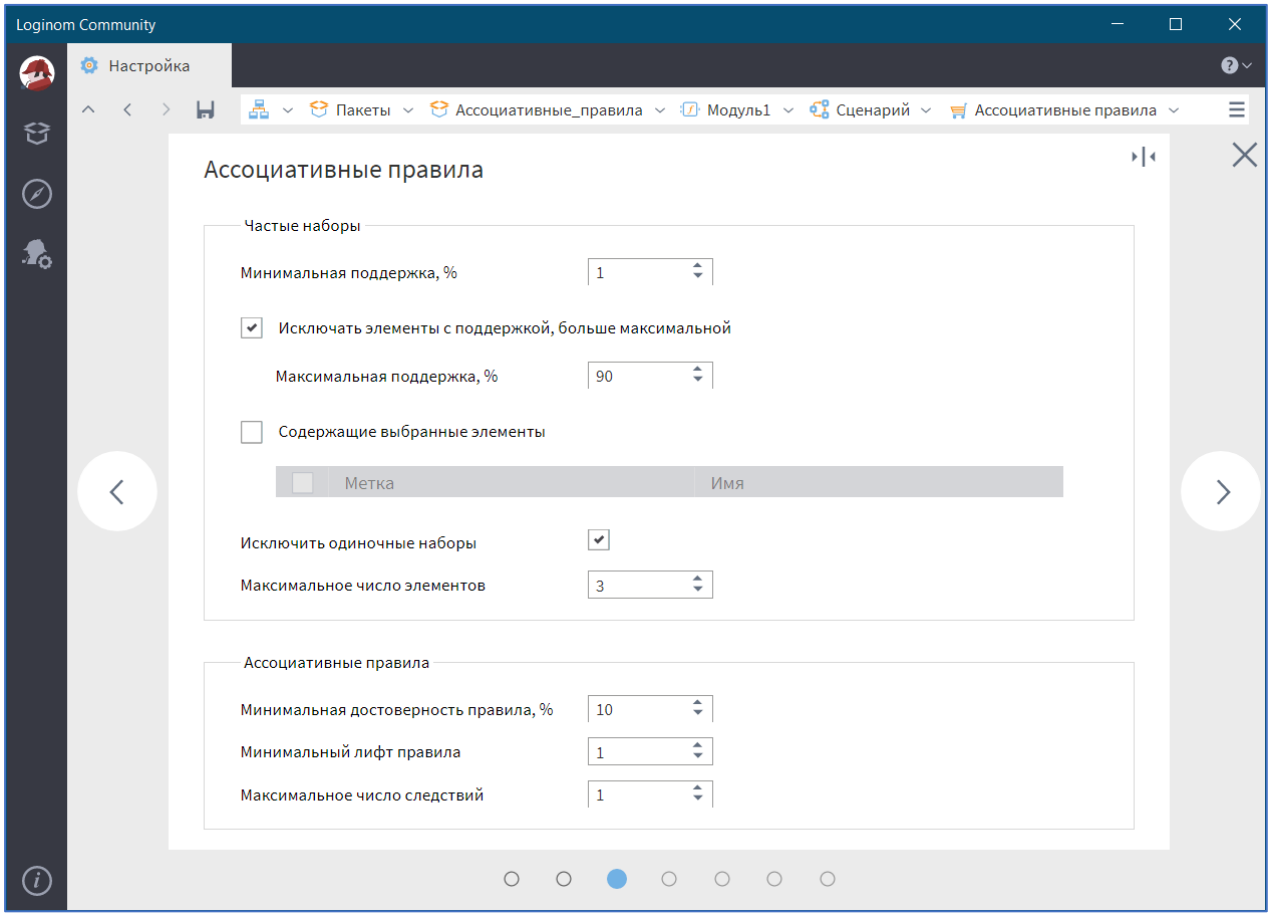


Рис. 10.8

Переобучим узел *Ассоциативные правила* и перейдем к настройкам визуализаторов (рис. 10.9).

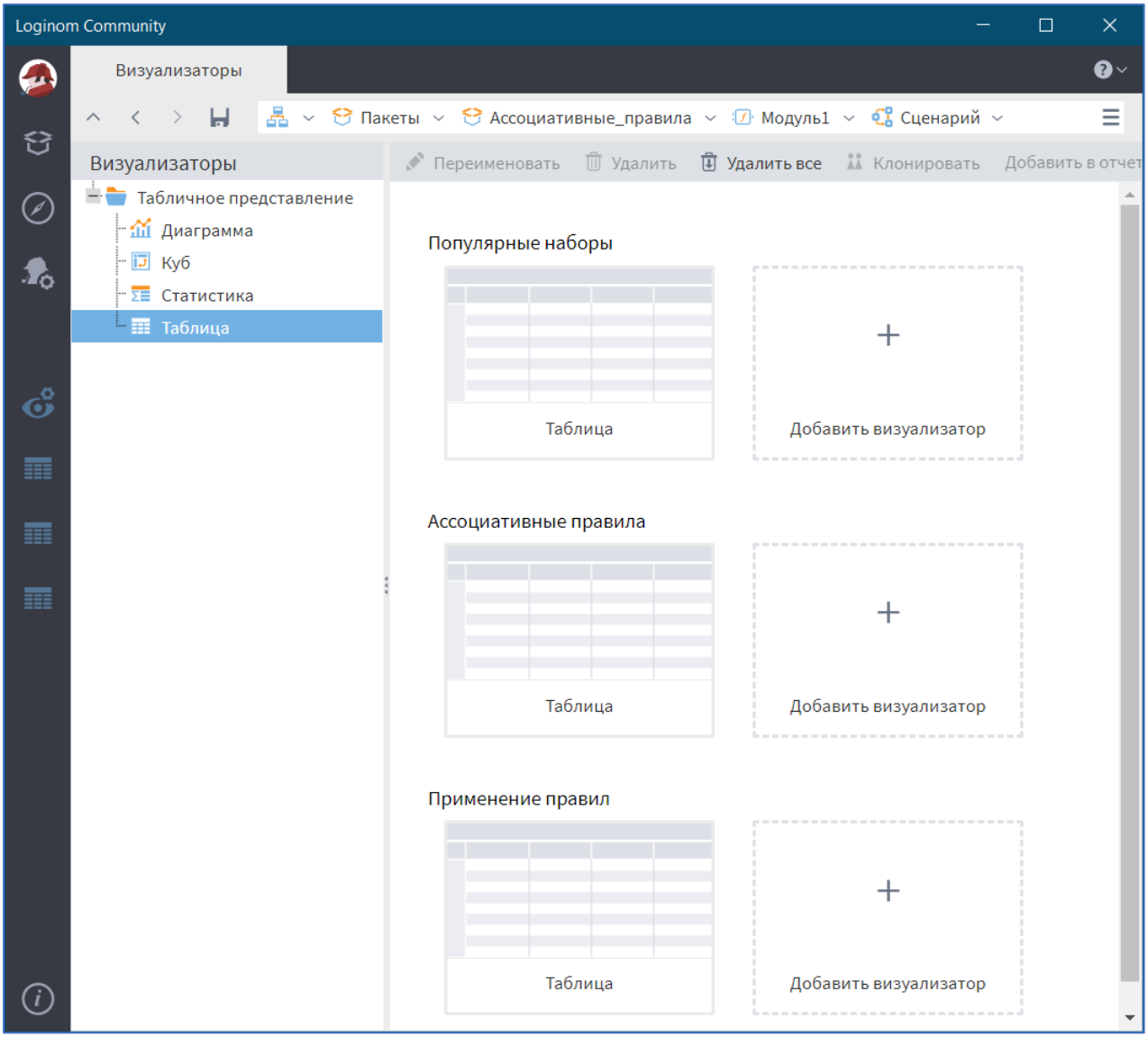


Рис. 10.9

Сначала откроем визуализатор *Ассоциативные правила* (рис. 10.10). В нем представлен набор полученных простых правил, указаны значения поддержки, достоверности и лифта.

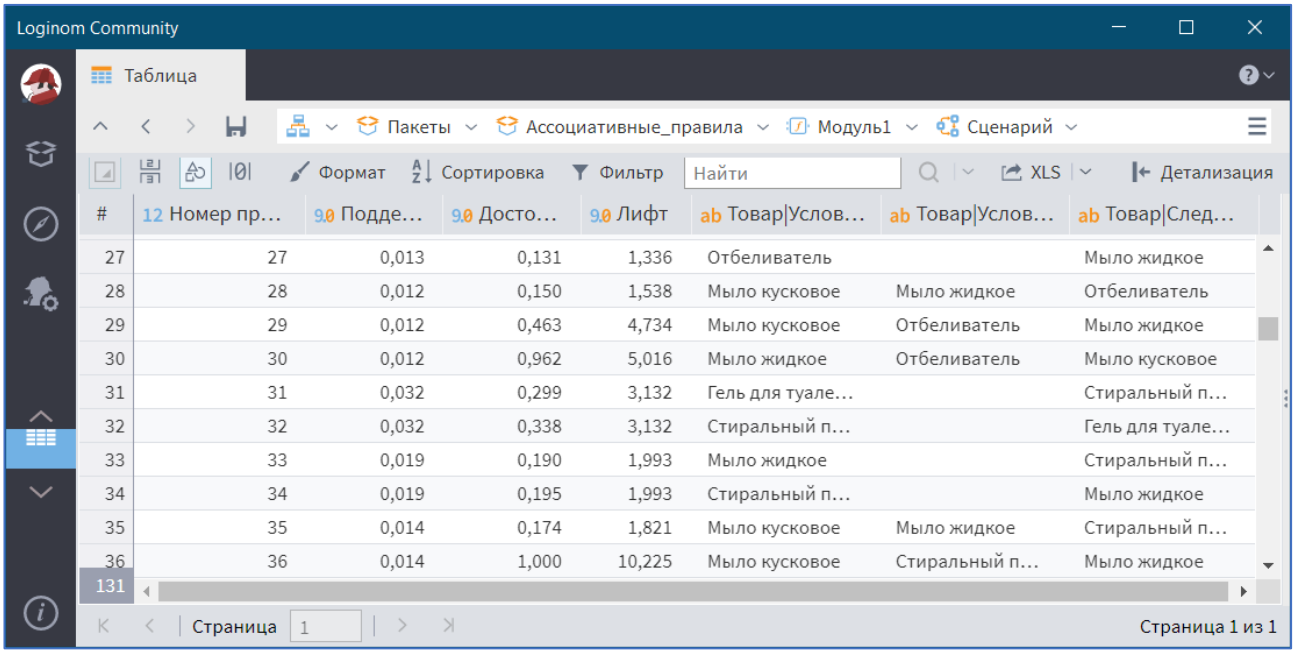


Рис. 10.10

Аналитику необходимо проанализировать каждое полученное правило и выбрать из них по-настоящему ценные. Например, правило под номером 30 показывает, что, если покупается мыло жидкое, то будут куплены отбеливатель и мыло кусковое с достоверностью 96,2%. Аналогично интерпретируются и остальные правила.

В визуализаторе *Популярные наборы* представлены наборы наиболее часто приобретаемых товаров, указаны значения поддержки и мощности (рис. 10.11).

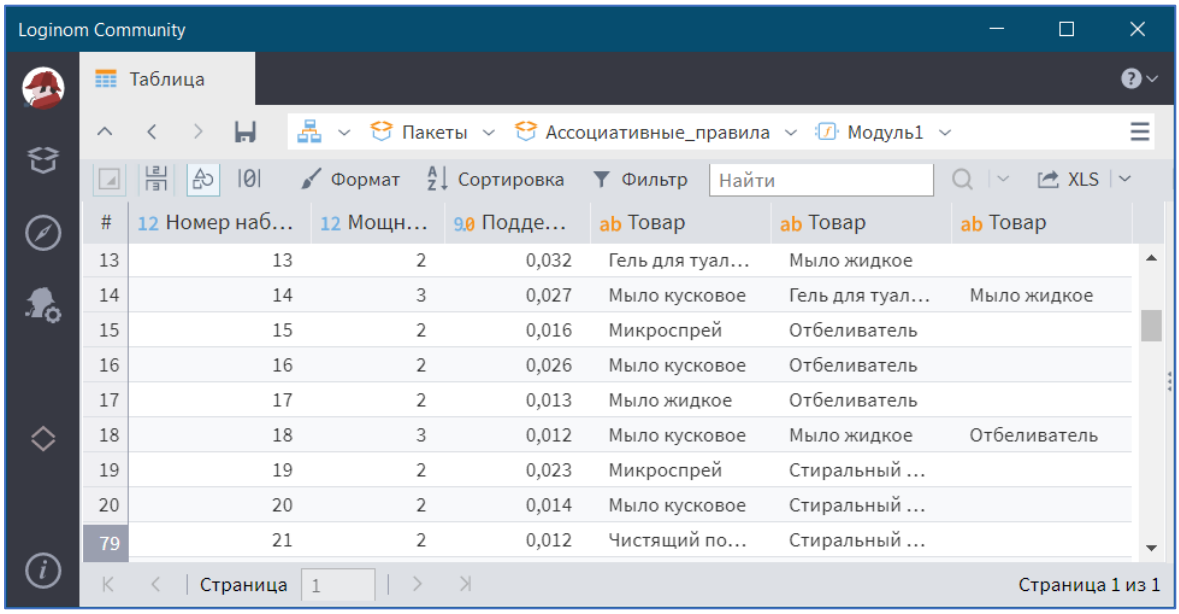
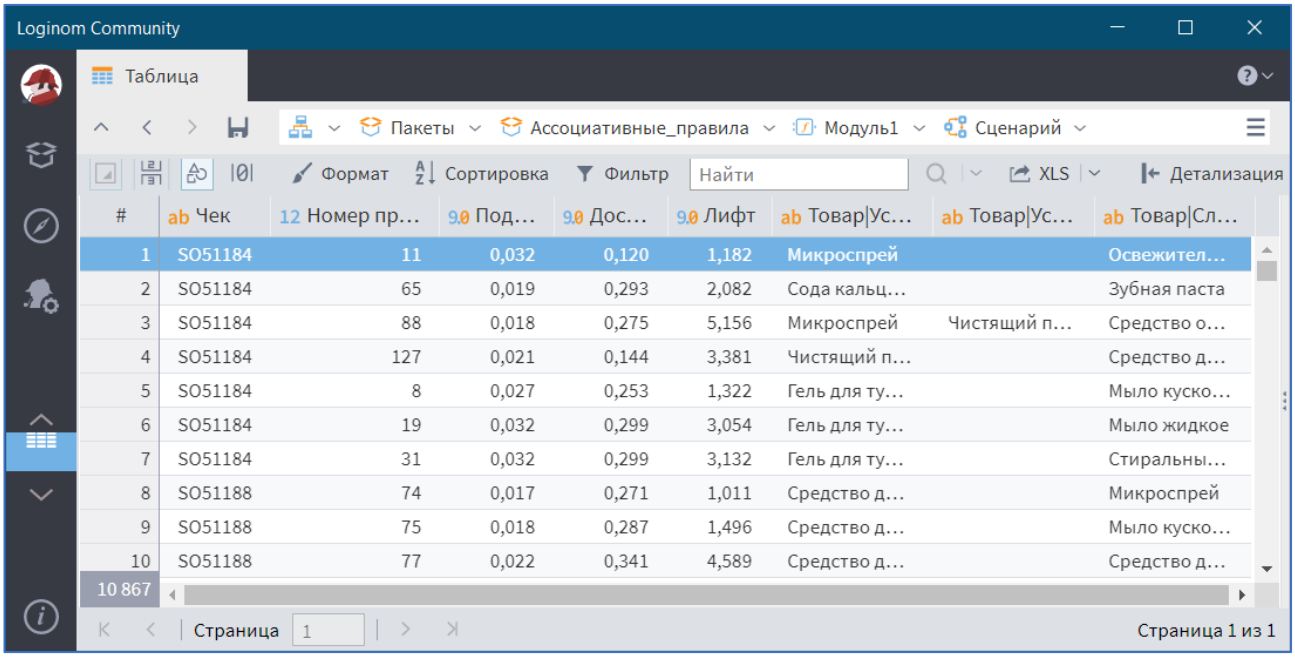


Рис. 10.11

В визуализаторе *Применение правил* по каждой транзакции описана ассоциативная связь между отдельными товарами и их наборами (рис. 10.12).



#	ab Чек	12 Номер пр…	9.0 Под…	9.0 Дос…	9.0 Лифт	ab Товар|Ус…	ab Товар|Ус…	ab Товар|Сл…
1	SO51184	11	0,032	0,120	1,182	Микроспрей		Освежител…
2	SO51184	65	0,019	0,293	2,082	Сода кальц…		Зубная паста
3	SO51184	88	0,018	0,275	5,156	Микроспрей	Чистящий п…	Средство о…
4	SO51184	127	0,021	0,144	3,381	Чистящий п…		Средство д…
5	SO51184	8	0,027	0,253	1,322	Гель для ту…		Мыло куско…
6	SO51184	19	0,032	0,299	3,054	Гель для ту…		Мыло жидкое
7	SO51184	31	0,032	0,299	3,132	Гель для ту…		Стиральны…
8	SO51188	74	0,017	0,271	1,011	Средство д…		Микроспрей
9	SO51188	75	0,018	0,287	1,496	Средство д…		Мыло куско…
10	SO51188	77	0,022	0,341	4,589	Средство д…		Средство д…

10 867

Рис. 10.12

10.3. Задание для самостоятельной работы

В файле *Задача 10.2. Продовольственные товары.xlsx* имеются данные для анализа потребительской корзины розничной сети, занимающейся продажей продовольственных товаров (рис. 10.13). Набор данных насчитывает 2615 чеков.

	A	B
1	Чек	Товар
2	85623	вода
3	85623	сигареты
4	92051	чай
2615	21369	хлеб
2616	21369	йогурт

Рис. 10.13

Требуется для выявления совместно приобретаемых товаров в розничной сети осуществить анализ потребительской корзины с помощью поиска ассоциативных правил.