2 ЛАБОРАТОРНАЯ РАБОТА «АНАЛИТИЧЕСКАЯ ПЛАТФОРМА DEDUCTOR. АССОЦИАТИВНЫЕ ПРАВИЛА»

2.1 Теоретические сведения

Одним из распространенных аналитических методов является аффинитивный анализ (англ: affinity analysis). Название метода происходит от английского слова affinity – близость, сходство. Целью данного метода является исследование взаимной связи между событиями, которые происходят совместно. Одной из разновидностей аффинитивного анализа является анализ рыночной корзины (англ: market basket analysis), цель которого – обнаружить ассоциации между различными событиями, т.е. найти правила для количественного описания взаимной связи между двумя или более событиями. Такие правила называются ассоциативными правилами (англ.: association rules).

Примерами приложения ассоциативных правил могут быть следующие задачи: обнаружение 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 посетителя веб-ресурса и т.д.

Базовым понятием в теории ассоциативных правил является транзакция. Транзакция – некоторое множество событий, происходящих совместно.

Типичной транзакцией является приобретение клиентом некоторого товара в супермаркете. В таблице 1 представлен простой пример набора транзакций. В каждой строке таблицы содержится комбинация продуктов, приобретённых за одну покупку.

Таблица 1 – Пример набора транзакций

Номер транзакции	Товары
1	сливы, салат, помидоры
2	сельдерей, конфеты
3	конфеты
4	яблоки, морковь, помидоры, картофель, конфеты
5	яблоки, апельсины, салат, конфеты, помидор
6	персики, апельсины, сельдерей, помидоры
7	фасоль, салат, помидоры
8	апельсины, салат, помидоры
9	яблоки, сливы, морковь, помидоры, лук, конфеты
10	яблоки, картофель

Хотя на практике приходится иметь дело с миллионами транзакций, в которых участвуют десятки и сотни различных продуктов, данный пример ограничен 10 транзакциями, содержащими 13 видов продуктов, что достаточно для иллюстрации методики обнаружения ассоциативных правил. В подавляющем большинстве случаев клиент приобретает не один товар, а некоторый набор товаров, который и называется рыночной корзиной. При этом возникает вопрос: является ли покупка одного товара в корзине следствием или причиной покупки другого товара, т.е. являются ли данные события связанными? Эту связь и устанавливают ассоциативные правила. Например, может быть обнаружено ассоциативное правило, утверждающее, что покупатель, купивший молоко, с вероятностью 75% купит и хлеб.

Визуальный анализ примера (таблица 1) показывает, что все четыре транзакции, в которых фигурирует салат, также включают и помидоры, и что четыре из семи транзакций, содержащих помидоры, также содержат и 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 (англ: antecedent) и следствие (англ: consequent), записываемых в виде X →Y, что читается «из X следует Y». Таким образом, ассоциативное правило формулируется в виде «Если условие, то следствие». Условие часто ограничивают содержанием только одного предмета. Правила обычно отображаются с помощью стрелок, направленных от условия к следствию, например, (помидоры) → (салат). Условие и следствие часто называются соответственно левосторонним (LHS – left-handside) и правосторонним (RHS – right-handside) компонентом ассоциативного правила. Ассоциативные правила описывают связь между наборами предметов, соответствующим условию и следствию. Эта связь характеризуется двумя показателями – поддержкой и достоверностью. Обозначим D как базу данных транзакций, а N как число транзакций в этой базе. Каждая транзакция D_i представляет собой некоторый набор предметов. Зададим, что S (англ.: support) – поддержка, C (англ.: confidence) – достоверность.

Поддержка ассоциативного правила – это число транзакций, которые содержат как условие, так и следствие. Например, для ассоциации A→B можно записать:

$$S(A \to B) = \frac{\text{количество транзакций, содержащих } A \text{ и } B}{\text{общее количество транзакций}}$$

Достоверность ассоциативного правила – это мера точности правила, которая определяется как отношение количества транзакций, содержащих как условие, так и следствие, к количеству транзакций, содержащих только условие.

Например, для ассоциации A→B можно записать

$$C(A \to B) = \frac{\text{количество транзакций, содержащих } A \text{ и } B}{\text{количество транзакций, содержащих только } A}$$

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

Рассмотрим пример для вычисления поддержки и достоверности для ассоциаций из таблицы 1. Возьмём ассоциацию (салат)→(помидоры). Поскольку количество транзакций, содержащее как (салат), так и (помидоры), равно 4, а общее число транзакций 10, то поддержка данной ассоциации будет:

S((салат)→ (помидоры)) = 4/10 = 0,4 .

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

C((салат) →(помидоры)) = 4/4 = 1.

Иными словами, все наблюдения, содержащие салат, также содержат и помидоры, что позволяет сделать 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 планирует сделать растительное блюдо, вряд ли станет покупать конфеты, а покупатель, желающий приобрести что-нибудь к чаю, скорее всего, не станет покупать помидоры). Поддержка данной ассоциации S = 3/10 = 0,3, а достоверность C = 3/7 = 0,43. Таким образом, сравнительно невысокая достоверность данной ассоциации даёт повод усомниться в том, что она является правилом.

Аналитики могут отдавать предпочтение правилам, которые имеют только высокую поддержку или только высокую достоверность, либо, что является наиболее частым, оба эти показателя. Правила, для которых значения поддержки или достоверности превышают некоторый, заданный пользователем порог, называются сильными правилами (strongrules). Например, аналитика может интересовать, какие товары в супермаркете, покупаемые вместе, образуют ассоциации с минимальной поддержкой 20% и минимальной достоверностью 70 %.

В Deductor Studio для решения задач ассоциации используется обработчик Ассоциативные правила. В нем реализован алгоритм Apriori. Обработчик требует на входе два поля: идентификатор транзакции и элемент транзакции. Например, идентификатор транзакции – это номер чека или код клиента. А элемент – это наименование товара в чеке или услуга, заказанная

клиентом. Оба поля (идентификатор и элемент транзакции) должны быть дискретного вида. После работы обработчика по умолчанию предлагается визуализатор Правила. Вся остальная дополнительная информация, располагается в специализированных визуализаторах Популярные наборы, Дерево правил, Что если.

2.2 Пример решения конкретной задачи ассоциации

Рассмотрим область торговли канцелярскими товарами. Канцелярские товары – неотъемлемая часть потребительской корзины. Анализ рыночной корзины позволит оптимизировать ассортимент и запасы канцелярских товаров, размещение их в торговых залах, увеличивать объемы продаж за счет предложения клиентам сопутствующих товаров и т.д.

Можно, например, выделить следующие задачи:

- выявить типичные или наиболее популярные шаблоны покупок канцелярских товаров (например, набор школьника и т.д.);
- предсказать какие товары покупатели могут выбрать в зависимости от того, что уже есть в их корзинах;
- предложить рекламные акции типа «Каждому купившему товары А и В, товар С в подарок».

Данные по транзакциям находятся в текстовом файле Товары.txt. Выборка для рассматриваемого примера состоит из 16032 реальных записей. Для проведения анализа используется версия Deductor Studio Academic, поэтому исходные данные преобразованы в формат txt.

Реализация в Deductor Studio.

Запустить программу Deductor Studio. Запустить Мастер импорта (рисунок 1).

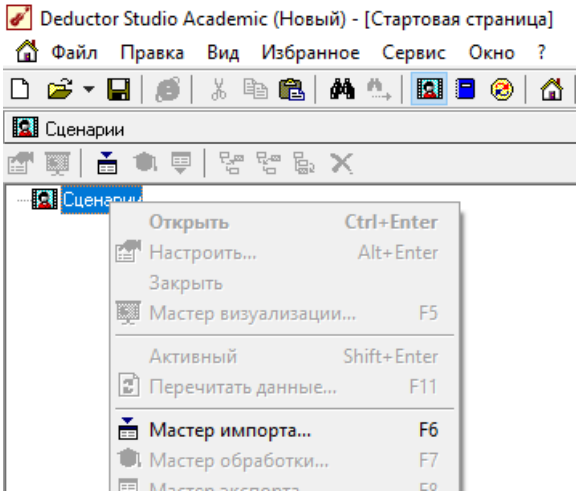


Рисунок 1 – Запуск Мастера импорта

В окне указать текстовый файл. Далее указать путь к текстовому файлу с данными (рисунок 2). Далее выбрать разделитель точка с запятой (рисунок 3). На следующем шаге проверить параметры столбцов (Столбец НомерЧека должен быть строкового типа). Внимательно проверить тип данных. Например, на рисунке 3 видно, что НомерЧека имеет неверный формат. Запустить процесс импорта данных.

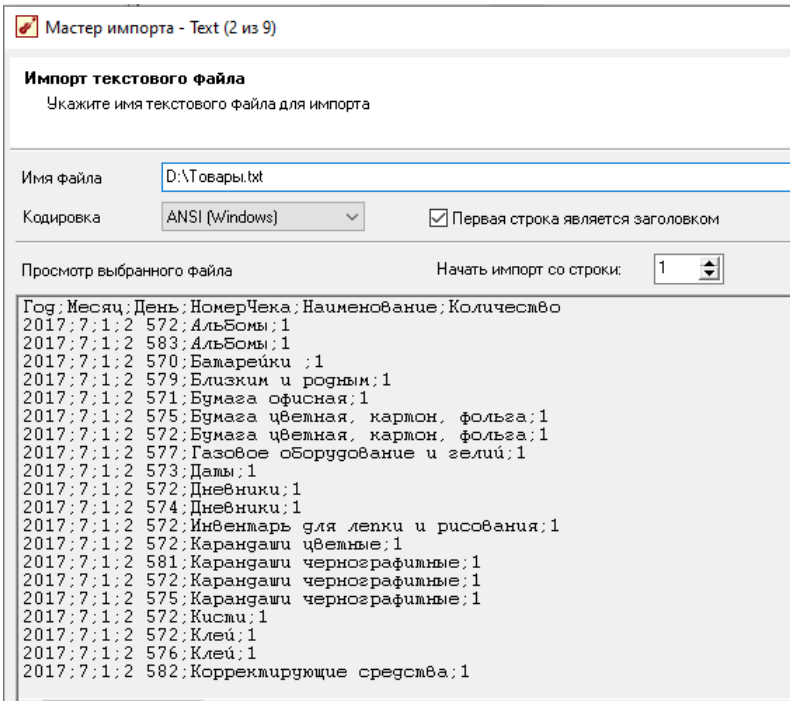


Рисунок 2 – Второй шаг Мастера импорта

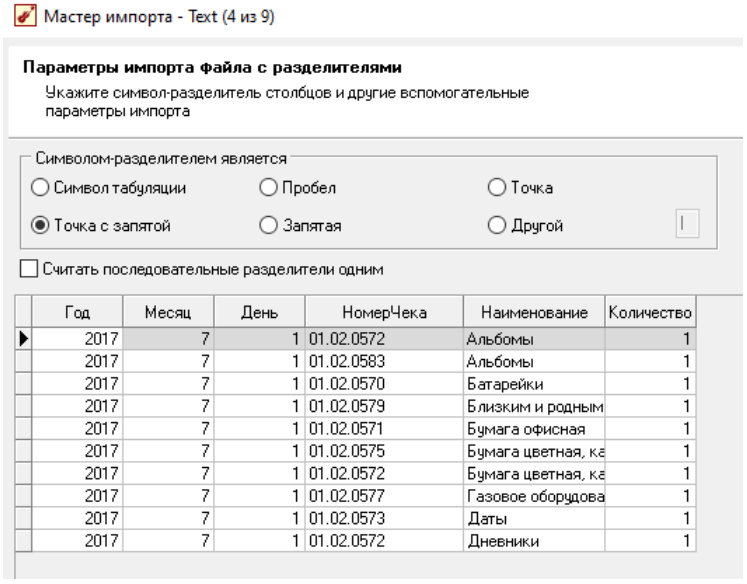
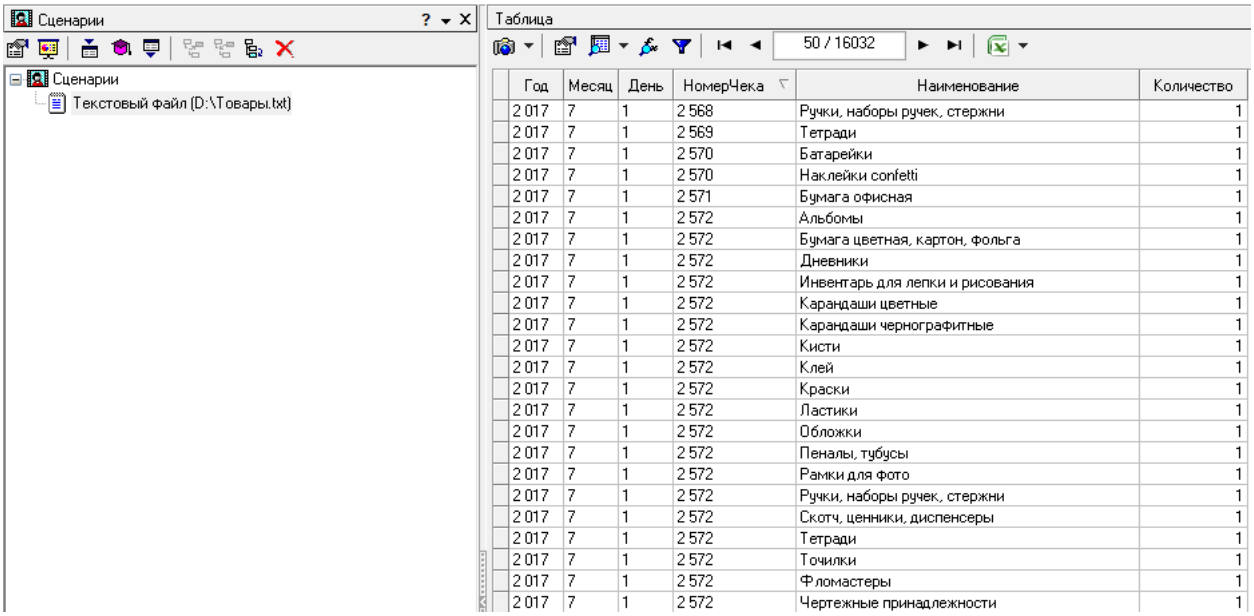


Рисунок 3 – Четвёртый шаг Мастера импорта

Результат импорта текстового файла в Deductor Studio на рисунке 4.



Год	Месяц	День	НомерЧека	Наименование	Количество
2017	7	1	2568	Ручки, наборы ручек, стержни	1
2017	7	1	2569	Тетради	1
2017	7	1	2570	Батарейки	1
2017	7	1	2570	Наклейки confetti	1
2017	7	1	2571	Бумага офисная	1
2017	7	1	2572	Альбомы	1
2017	7	1	2572	Бумага цветная, картон, фольга	1
2017	7	1	2572	Дневники	1
2017	7	1	2572	Инвентарь для лепки и рисования	1
2017	7	1	2572	Карандаши цветные	1
2017	7	1	2572	Карандаши чернографитные	1
2017	7	1	2572	Кисти	1
2017	7	1	2572	Клей	1
2017	7	1	2572	Краски	1
2017	7	1	2572	Ластики	1
2017	7	1	2572	Обложки	1
2017	7	1	2572	Пеналы, тубусы	1
2017	7	1	2572	Рамки для фото	1
2017	7	1	2572	Ручки, наборы ручек, стержни	1
2017	7	1	2572	Скотч, ценники, диспенсеры	1
2017	7	1	2572	Тетради	1
2017	7	1	2572	Точилки	1
2017	7	1	2572	Фломастеры	1
2017	7	1	2572	Чертежные принадлежности	1

Рисунок 4 – Импорт тестового файла в Deductor

Для поиска ассоциативных правил используется Мастер обработки (рисунок 5).

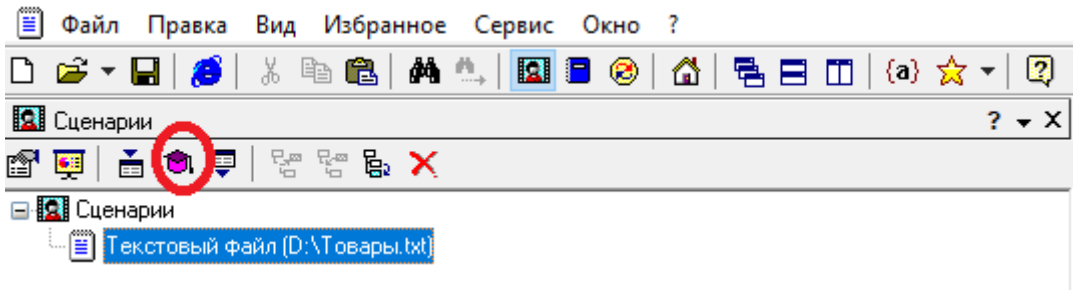


Рисунок 5 – Кнопка вызова Мастера обработки

Указать тип обработки «Ассоциативные правила» (рисунок 6).

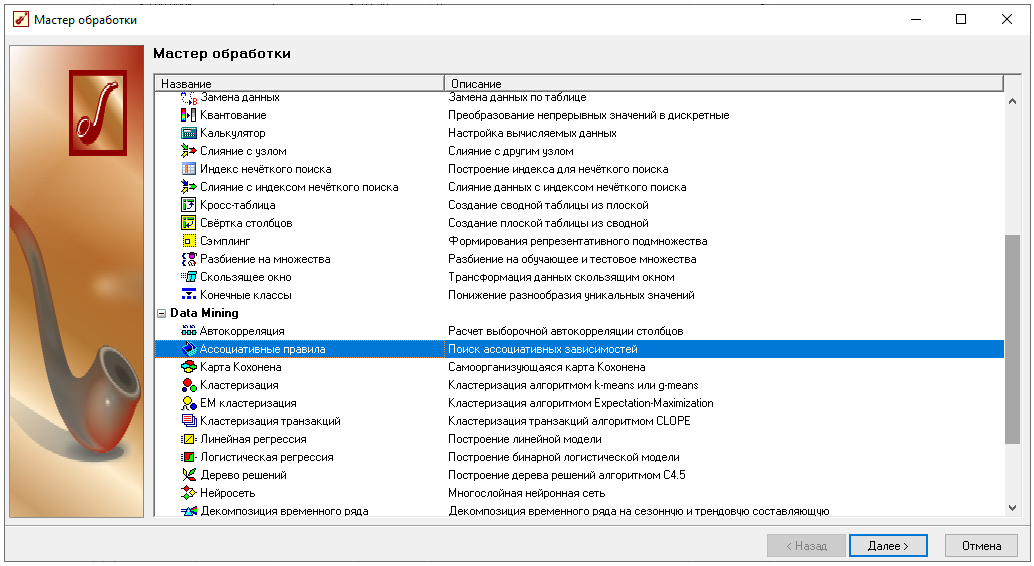


Рисунок 6 – Шаг Мастера обработки

Далее настроить назначения столбцов (рисунок 7). Идентификатор транзакции (НомерЧека), а элемент транзакции (Наименование). Тип поля должен быть строковый.

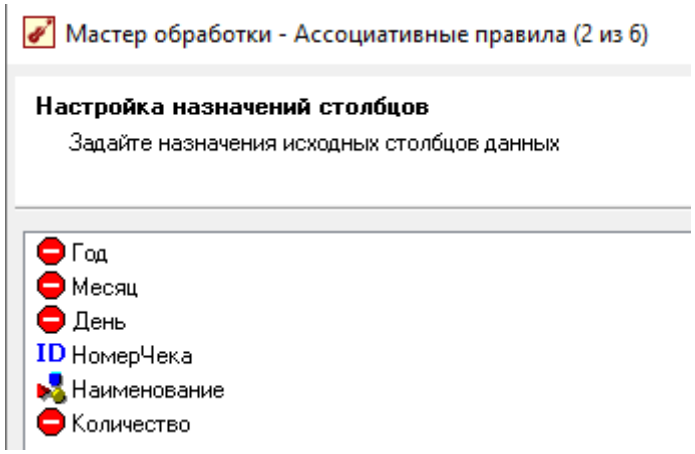


Рисунок 7 – Настройка назначений столбцов

На следующем шаге мастера нужно настроить параметры построения ассоциативных правил, что, по сути, есть параметры алгоритма Apriori. Здесь для изменения доступны следующие параметры.

Минимальная и максимальная поддержка в % – ограничивают пространство поиска часто встречающихся предметных наборов. Эти границы определяют множество популярных наборов, из которых и будут создаваться ассоциативные правила.

Минимальная и максимальная достоверность в % – в результирующий набор попадут только те ассоциативные правила, которые удовлетворяют условиям минимальной и максимальной достоверности.

Максимальная мощность искомых часто встречающихся множеств – параметр ограничивает длину k-предметного набора.

Например, при установке значения 4 шаг генерации популярных наборов будет остановлен после получения множества 4-предметных наборов. В конечном итоге это позволяет избежать появления длинных ассоциативных правил, которые трудно интерпретируются.

Далее нажать на кнопку Пуск, что приведёт к запуску процесса построения ассоциативных правил. Процесс поиска ассоциативных правил (рисунок 8).

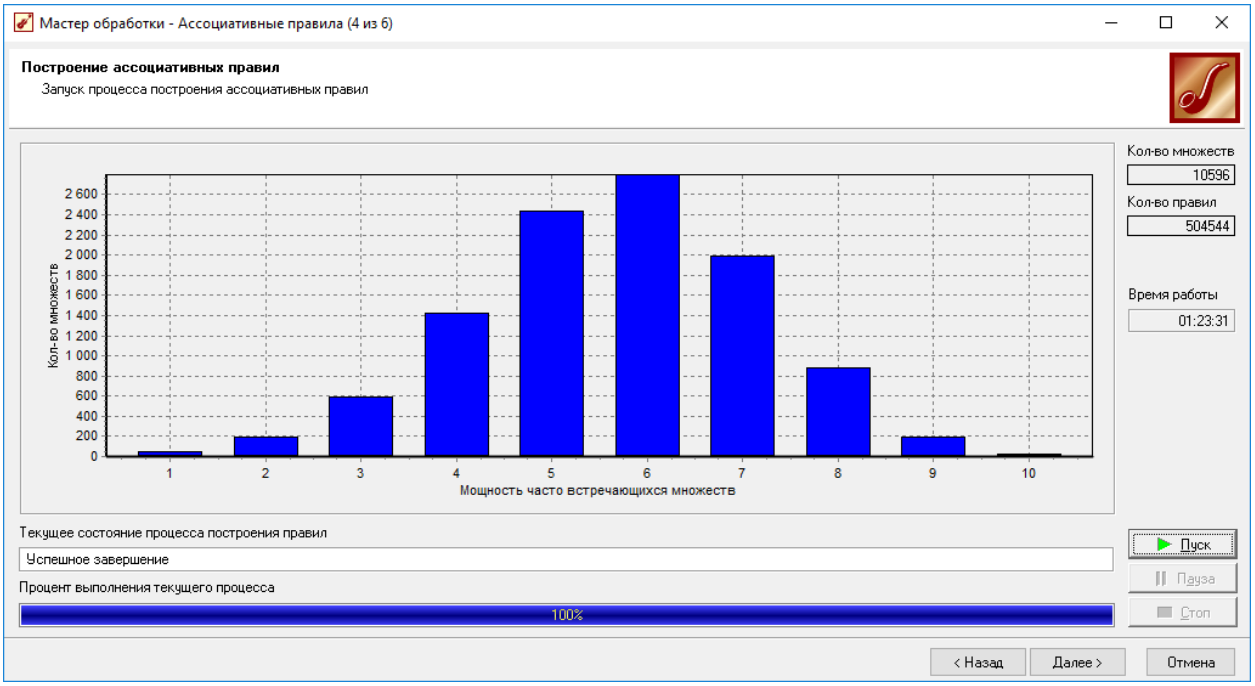


Рисунок 8 – Процесс поиска ассоциативных правил

Далее выбираем все доступные специализированные визуализаторы. Просмотр полученных результатов, визуализаторы «Правила» (рисунок 9), «Популярные наборы» (рисунок 10), «Дерево правил», «Что-если» (рисунок 11).

Можно сохранить сценарий под именем *.ded.

Визуализатор «Правила» отображает ассоциативные правила в виде списка правил. Эксперту предоставляется набор правил, которые описывают поведение покупателей.

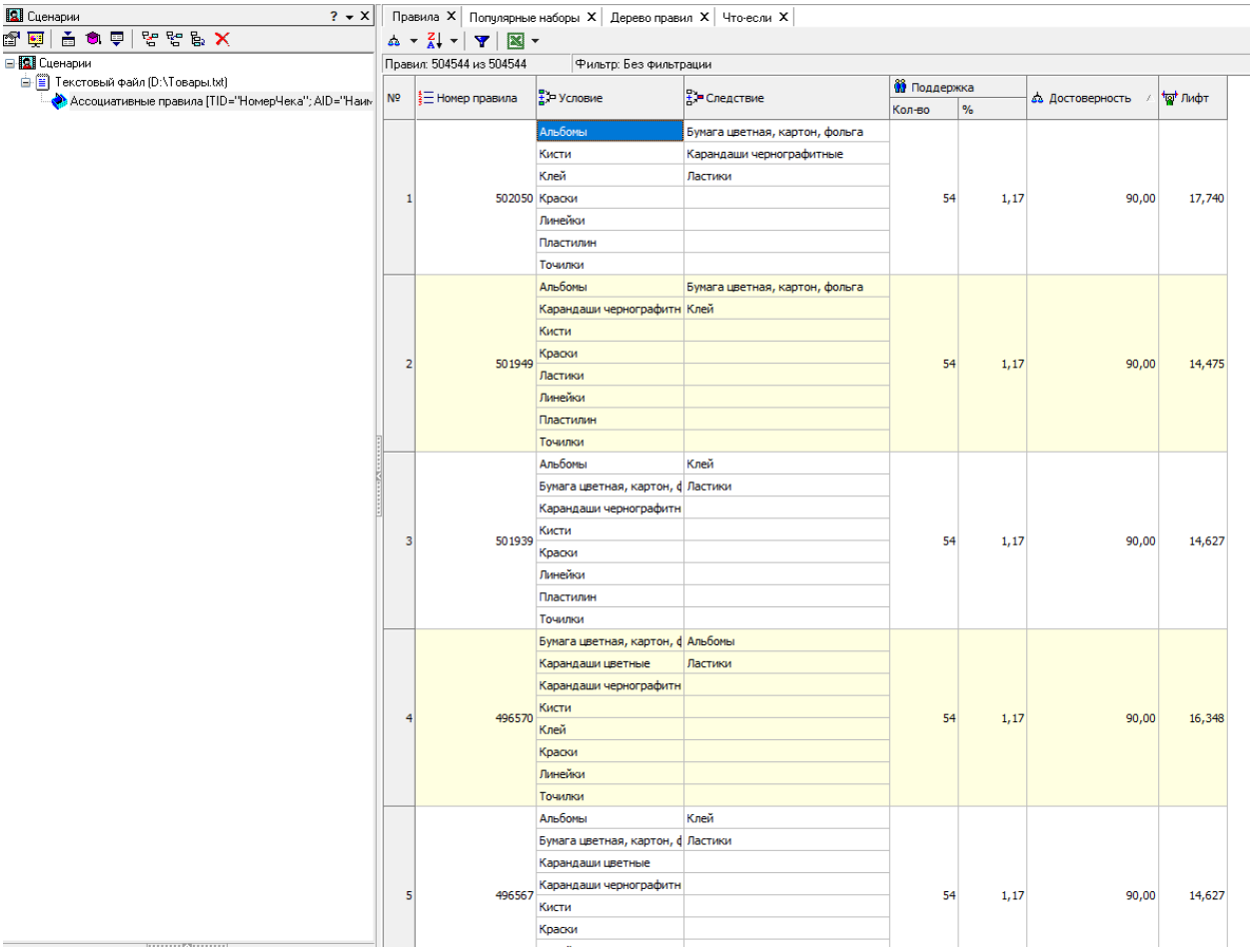
Популярные наборы – это множества, состоящие из одного и более элементов, которые наиболее часто встречаются в транзакциях одновременно.

Как видно из рисунка 10, графитные карандаши – наиболее часто покупаемый товар.

Анализ «Что-если» в ассоциативных правилах позволяет ответить на вопрос, что получим в качестве следствия, если выберем данные условия?

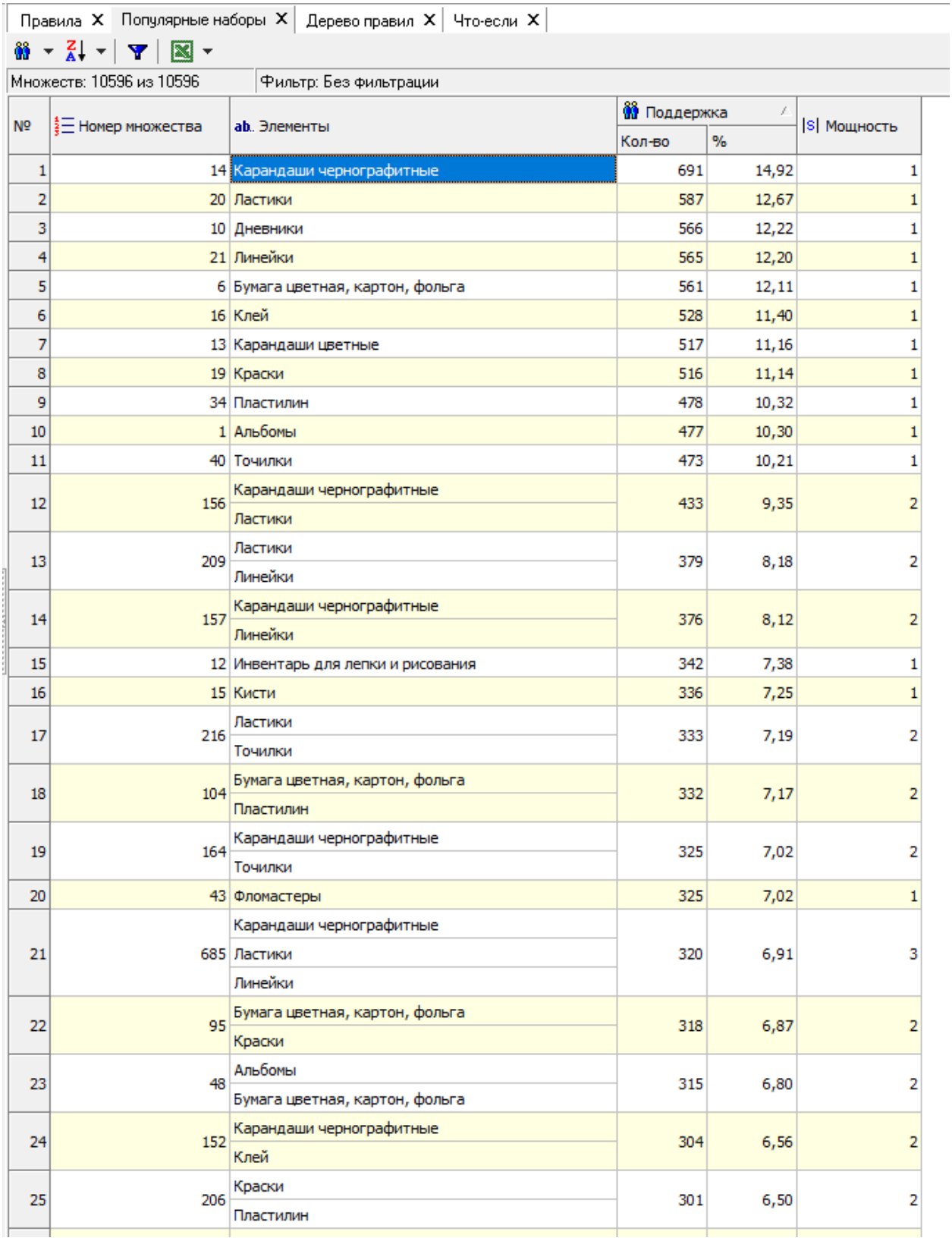
Например, какие товары приобретаются совместно с выбранными товарами. На рисунке 6 приведен пример. При покупке альбома и красок, часто покупают и цветную бумагу.

На вкладке Дерево правил предлагается еще один удобный способ отображения множества ассоциативных правил, которое строится либо по условию, либо по следствию. При построении дерева правил по условию, на первом (верхнем) уровне находятся узлы с условиями, а на втором уровне – узлы со следствием. В дереве, построенном по следствию, наоборот, на первом уровне располагаются узлы со следствием.



№	Номер правила	Условие	Следствие	Поддержка Кол-во	Поддержка %	Достоверность	Лифт
1	502050	Альбомы	Бумага цветная, картон, фольга	54	1,17	90,00	17,740
		Кисти	Карандаши чернографитные				
		Клей	Ластики				
		Краски					
		Линейки					
		Пластилин					
		Точилки					
2	501949	Альбомы	Бумага цветная, картон, фольга	54	1,17	90,00	14,475
		Карандаши чернографитн	Клей				
		Кисти					
		Краски					
		Ластики					
		Линейки					
		Пластилин					
		Точилки					
3	501939	Альбомы	Клей	54	1,17	90,00	14,627
		Бумага цветная, картон, ф	Ластики				
		Карандаши чернографитн					
		Кисти					
		Краски					
		Линейки					
		Пластилин					
		Точилки					
4	496570	Бумага цветная, картон, ф	Альбомы	54	1,17	90,00	16,348
		Карандаши цветные	Ластики				
		Карандаши чернографитн					
		Кисти					
		Клей					
		Краски					
		Линейки					
		Точилки					
5	496567	Альбомы	Клей	54	1,17	90,00	14,627
		Бумага цветная, картон, ф	Ластики				
		Карандаши цветные					
		Карандаши чернографитн					
		Кисти					
		Краски					

Рисунок 9 – Визуализатор «Правила»



Правила | Популярные наборы | Дерево правил | Что-если

Множеств: 10596 из 10596 | Фильтр: Без фильтрации

№	Номер множества	Элементы	Поддержка		Мощность
			Кол-во	%	
1	14	Карандаши черно графитные	691	14,92	1
2	20	Ластики	587	12,67	1
3	10	Дневники	566	12,22	1
4	21	Линейки	565	12,20	1
5	6	Бумага цветная, картон, фольга	561	12,11	1
6	16	Клей	528	11,40	1
7	13	Карандаши цветные	517	11,16	1
8	19	Краски	516	11,14	1
9	34	Пластилин	478	10,32	1
10	1	Альбомы	477	10,30	1
11	40	Точилки	473	10,21	1
12	156	Карандаши черно графитные Ластики	433	9,35	2
13	209	Ластики Линейки	379	8,18	2
14	157	Карандаши черно графитные Линейки	376	8,12	2
15	12	Инвентарь для лепки и рисования	342	7,38	1
16	15	Кисти	336	7,25	1
17	216	Ластики Точилки	333	7,19	2
18	104	Бумага цветная, картон, фольга Пластилин	332	7,17	2
19	164	Карандаши черно графитные Точилки	325	7,02	2
20	43	Фломастеры	325	7,02	1
21	685	Карандаши черно графитные Ластики Линейки	320	6,91	3
22	95	Бумага цветная, картон, фольга Краски	318	6,87	2
23	48	Альбомы Бумага цветная, картон, фольга	315	6,80	2
24	152	Карандаши черно графитные Клей	304	6,56	2
25	206	Краски Пластилин	301	6,50	2

Рисунок 10 – Визуализатор «Популярные наборы»

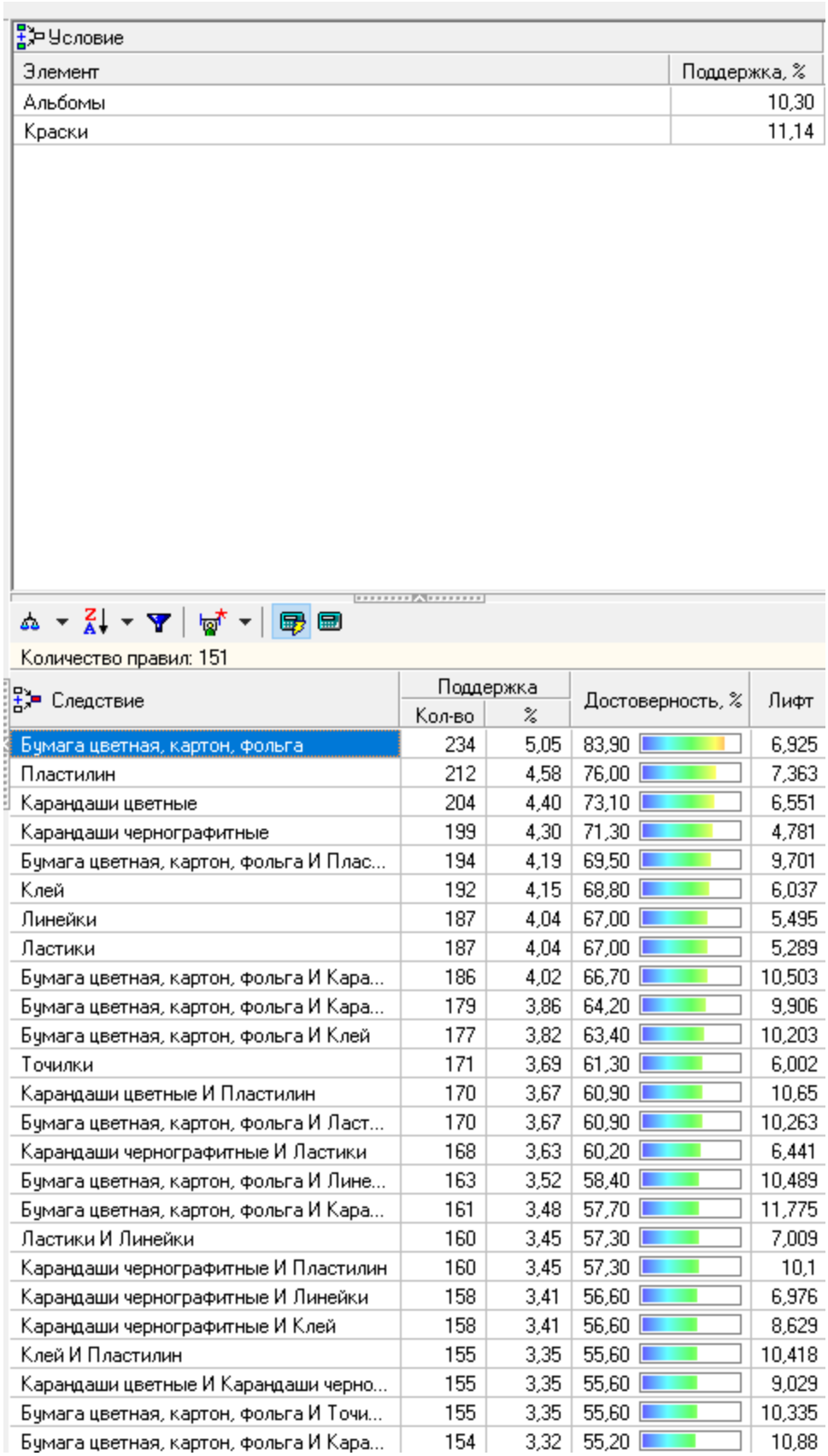


Рисунок 11 – Визуализатор «Что-если»

2.3 Задание и рекомендации

1. Изучить материал, представленный в лабораторной работе 2, а также материал в учебном пособии со страницы 75 по страницу 90, указанном ниже.

Прокопенко Н.Ю. Системы поддержки принятия решений [Электронный ресурс]: учеб. пособие /Н. Ю. Прокопенко; Нижегор. гос. архитектур.-строит. ун-т. – Н. Новгород: ННГАСУ, 2017. – 188 с. ISBN 978-5-528-00202-6.

2. Скачать с официального сайта фирмы разработчика аналитическую платформу Deductor Studio Academic 5.3 по ссылке https://basegroup.ru/deductor/download.

3. Для выполнения лабораторной работы использовать файлы с исходными данными в формате txt. Файлы загружены в СДО. Выбор файла с исходными данными осуществляется по вариантам. Вариант выбирается по номеру в списке группы.

Имя файла	Вариант
Transactions1.txt	1, 6
Transactions2.txt	2, 7
Transactions3.txt	3, 8
Transactions4.txt	4, 9
Transactions5.txt	5, 10

4. Осуществить поиск ассоциативных правил в Deductor Studio Academic по исходным данным. С помощью различных визуализаторов проанализировать полученные результаты, сделать выводы.

5. Подготовить отчёт и сдать преподавателю в электронной форме. Отчёт должен содержать краткое описание выполняемых действий, скриншоты, выводы. Наличие выводов является обязательным требованием.

Дополнение.

С 2018 года на смену АП Deductor пришла АП Loginom. https://loginom.ru/download.

Дополнительный материал по изучению Loginom в учебном пособии, указанном ниже.

Прокопенко Н. Ю. Аналитические информационные системы поддержки принятия решений [Текст]: учеб. пособие / Н.Ю. Прокопенко; Нижегор. гос. архитектур.- строит. ун-т – Н. Новгород: ННГАСУ, 2020. – 142 с.

2.4 Вопросы для защиты лабораторной работы

1. Что такое транзакция в теории ассоциативных правил? Привести пример.
2. Что такое поддержка ассоциативного правила?
3. Что такое достоверность ассоциативного правила?
4. Какой алгоритм генерации ассоциативных правил имеется в Deductor Studio?
5. Какие входные поля набора данных необходимы для запуска обработчика Ассоциативные правила в Deductor?
6. Какие специализированные визуализаторы используются для отображения результатов поиска ассоциативных правил? Описать каждый визуализатор.