**Министерство образования и науки Российской Федерации**

**ФГАОУ ВО «Севастопольский государственный университет»**

**Институт информационных технологий**

**и управления в технических системах**

# Практическая работа №5

# «Аналитическая система Deductor: Поиск ассоциативных правил»

для студентов всех форм обучения направления подготовки

09.03.02 «Информационные системы и технологии»


Севастополь
**2017**

**Цель:** научится проводить поиск ассоциативных правил в Deductor Studio

**Время:** 2 часа

## Краткие теоретические сведения

Ассоциативные правила позволяют находить закономерности между связанными событиями. Примером такого правила служит утверждение, что покупатель, приобретающий "Хлеб", приобретет и "Молоко". Впервые эта задача была предложена для поиска ассоциативных правил для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis)

Транзакция – это множество событий, произошедших одновременно. Пусть имеется база данных, состоящая из покупательских транзакций. Каждая транзакция – это набор товаров, купленных покупателем за один визит.

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 должен появиться в этой транзакции. Установление таких зависимостей дает возможность находить очень простые и интуитивно понятные правила.

Основными характеристиками таких правил являются **поддержка** и **достоверность.** Правило "Из *X* следует *Y*" имеет поддержку *s*, если *s*% транзакций из всего набора содержат наборы элементов *X* и *Y.* Достоверность правила показывает, какова вероятность того, что из *X* следует *Y.* Правило "Из *X* следует *Y*" справедливо с достоверностью *c*, если *c*% транзакций из всего множества, содержащих набор элементов *X*, также содержат набор элементов *Y.*

Пример: пусть 75% транзакций, содержащих хлеб, также содержат молоко, а 3% от общего числа всех транзакций содержат оба товара. 75% – это достоверность правила, а 3% –это поддержка.

Лифт – это отношение частоты появления условия в транзакциях, которые также содержат и следствие, к частоте появления следствия в целом. Значения лифта, большие единицы, показывают, что условие появляется более часто в транзакциях, содержащих и следствие, чем в остальных.

Алгоритмы поиска ассоциативных правил предназначены для нахождения всех правил вида "из *X* следует *Y*", причем поддержка и достоверность этих правил должны находиться в рамках некоторых наперед заданных границ, называемых соответственно минимальной и максимальной поддержкой и минимальной и максимальной достоверностью. Границы значений параметров поддержки и достоверности 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. Тем не менее, большинство интересных правил находится именно при низком значении порога поддержки, хотя слишком низкое значение поддержки ведет к генерации статистически необоснованных правил. Таким образом, необходимо найти компромисс, обеспечивающий, во-первых, интересность правил и, во-вторых, их статистическую обоснованность. Поэтому значения этих границ напрямую зависят от характера анализируемых данных и подбираются индивидуально. Еще одним параметром, ограничивающим Количество найденных правил является максимальная мощность часто встречающихся множеств. Если этот параметр указан, то при поиске правил будут рассматриваться только множества, количество элементов которых будет не больше данного параметра.

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.

Все множество ассоциативных правил можно разделить на три вида:

Полезные правила – содержат действительную информацию, которая ранее была неизвестна, но имеет логичное объяснение. Такие правила могут быть использованы для принятия решений, приносящих выгоду.

Тривиальные правила – содержат действительную и легко объяснимую информацию, которая уже известна. Такие правила, хотя и объяснимы, но не могут принести какой-либо пользы, т.к. отражают или известные законы в исследуемой области, или результаты прошлой деятельности. При анализе рыночных корзин в правилах с самой высокой поддержкой и достоверностью окажутся товары-лидеры продаж. Практическая ценность таких правил крайне низка.

Непонятные правила – содержат информацию, которая не может быть объяснена. Такие правила могут быть получены или на основе аномальных значений, или глубоко скрытых знаний. Напрямую такие правила нельзя использовать для принятия решений, т.к. их необъяснимость может привести к непредсказуемым результатам. Для лучшего понимания требуется дополнительный анализ.

**Поиск ассоциативных правил в Deductor Studio**

Для поиска обычных ассоциативных правил в программе служит обработчик «Ассоциативные правила».

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.

Затем следует настройка параметров поиска правил. Всего четыре параметра:

**Минимальная и максимальная поддержка**. Ассоциативные правила ищутся только в некотором множестве всех транзакций. Для того чтобы транзакция вошла в это множество, она должна встретиться в исходной выборке количество раз, больше минимальной поддержки и меньше максимальной. Например, минимальная поддержка равна 1%, а максимальная – 20%. Количество элементов «Хлеб» и «Молоко» столбца «Товар» с одинаковым значением столбца «Номер чека» встречаются в 5% всех транзакций (номеров чека). Тогда эти две строки войдут в искомое множество.

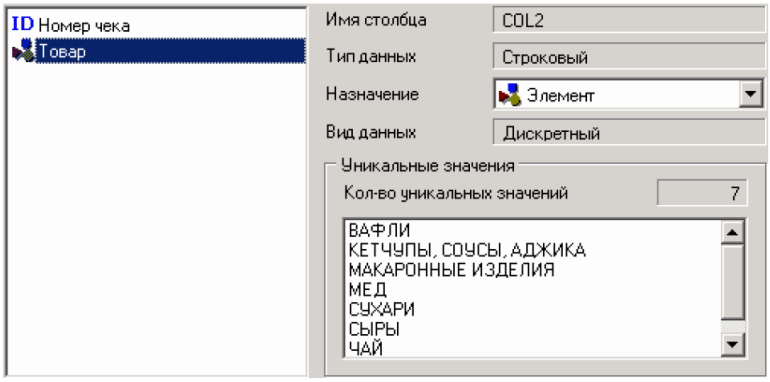
**Минимальная и максимальная достоверность.** Это процентное отношение количества транзакций, содержащих все элементы, которые входят в правило, к количеству транзакций, содержащих элементы, которые входят в условие. Если транзакция – это заказ, а элемент – товар, то достоверность характеризует, насколько часто покупаются товары, входящие в следствие, если заказ содержит товары, вошедшие во всё правило.

**Пример**:

Рассмотрим механизм поиска ассоциативных правил на примере данных о продажах товаров в некоторой торговой точке. Данные представляются в виде таблицы, в которой представлена информация по покупкам продуктов нескольких групп. Она имеет всего два поля "Номер чека" и "Товар". Необходимо решить задачу анализа потребительской корзины с целью последующего применения результатов для стимулирования продаж. применения результатов для стимулирования продаж.

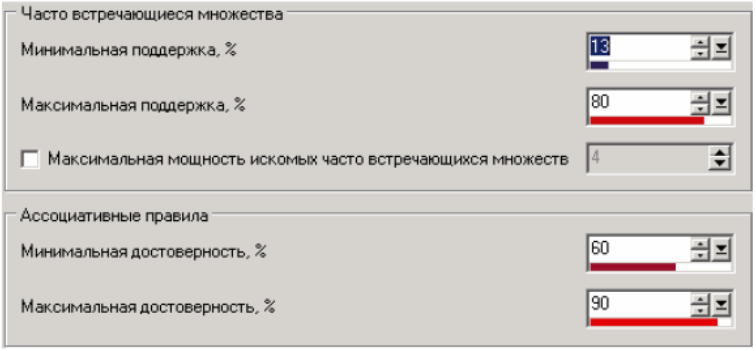
Для поиска ассоциативных правил необходимо запустить **Мастер обработки**. В нем выберать тип обработки "Ассоциативные правила".

На втором шаге Мастера следует указать, какой столбец является идентификатором транзакции (чек), который должен быть дискретным, а какой элементом транзакции (товар).




Следующий шаг позволяет настроить параметры построения ассоциативных правил: минимальную и максимальную поддержку, минимальную и максимальную достоверность, а также максимальную мощность множества. Эти параметры необходимо выставлять исходя из характера имеющихся данных. Например, границы поддержки следует указать – 13% и 80% и достоверности 60% и 90%. Можно оставить по умолчанию.

**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.




Следующий шаг позволяет запустить процесс поиска ассоциативных правил. На экране отображается информация о количестве множеств и найденных правил, а также числе часто встречающихся множеств.


После завершения процесса поиска полученные результаты можно посмотреть, используя появившиеся специальные визуализаторы "Популярные наборы", "Правила", "Дерево правил", "Что-если".

**Популярные наборы** - это множества, состоящие из одного и более элементов, которые наиболее часто встречаются в транзакциях одновременно. На сколько часто встречается множество в исходном наборе транзакций, можно судить по поддержке. Данный визуализатор отображает множества в виде списка.

| № | Множество | Поддержка | |
|---|---|---|---|
| | | % | Кол-во |
| 7 | ЧАЙ | 75,00 | 33 |
| 3 | МАКАРОННЫЕ ИЗДЕЛИЯ | 54,55 | 24 |
| 2 | КЕТЧУПЫ, СОУСЫ, АДЖИКА | 52,27 | 23 |
| 4 | МЕД | 50,00 | 22 |

Получившиеся наборы товаров наиболее часто покупают в данной торговой точке, следовательно можно принимать решения о поставках товаров, их размещении и т.д

Визуализатор "**Правила**" отображает ассоциативные правила в виде списка правил. Этот список представлен таблицей со столбцами: "Номер правила", "Условие", "Следствие", "Поддержка, %", "Поддержка, Количество", "Достоверность".

Таким образом, эксперту предоставляется набор правил, которые описывают поведение покупателей.

Правил: 63 из 63 | Фильтр: Без фильтрации

| № | Номер правила | Условие | Следствие | Поддержка | | Достоверность | Лифт |
|---|---|---|---|---|---|---|---|
| | | | | Кол-во | % | | |
| 1 | 60 | Клей - ж. гвозди<br>Шпатлёвка | Герметики<br>Пена монтажная | 2 | 4.55 | 40.00 | 2.933 |
| 2 | 57 | Герметики<br>Пена монтажная | Клей - ж. гвозди<br>Шпатлёвка | 2 | 4.55 | 33.33 | 2.933 |
| 3 | 59 | Герметики<br>Шпатлёвка | Клей - ж. гвозди<br>Пена монтажная | 2 | 4.55 | 40.00 | 2.514 |

Например, в правиле с номером 60 в условии присутствуют два элемента: Клей-ж. гвозди и Шпатлёвка. Это правило показывает, что человек, купивший Клей-ж.гвозди и Шпатлёвка с вероятностью 40% купит ещё и Герметики и Пену монтажная.

Визуализатор "**Дерево правил**" – это всегда двухуровневое дерево. Оно может быть построено 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.

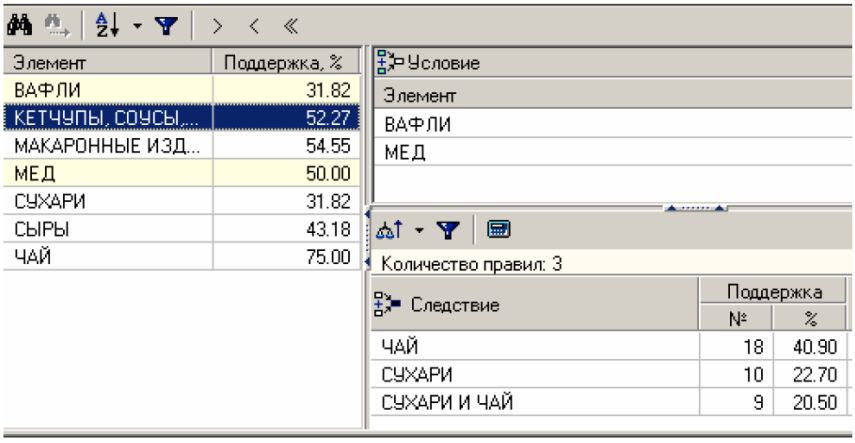

Узлы Герметики, Клей-ж. гвозди, Пена монтажная находятся на верхнем уровне дерева и представляют собой условия. А Герметики, Пена монтажная, Шпатлёвка и т.д. – следствия. Это означает, что человек, купивший Клей-ж. гвозди, купит еще и Герметик с достоверностью 71,40%, пену монтажную с достоверностью 50,00% и т.д. В окне слева расположен список со следствиями для конкретного узла с условием. Для каждого следствия указана поддержка, достоверность и лифт. Например, в исходной выборке данных герметики встретились в 10 транзакциях (чеках).

Второй вариант дерева правил – дерево, построенное по следствию. Здесь на первом уровне располагаются узлы со следствием.


Например, для того чтобы человек приобрел Герметик, он должен купить хотя бы один предмет из следующего списка: Клей-ж.гвозди, Пена монтажная, Шпатлёвка и т.д. И для каждого из этих правил отображены поддержка, достоверность и лифт.

Справа от дерева находится список правил, построенный по выбранному узлу дерева. Для каждого правила отображаются поддержка и достоверность. Если дерево построено по условию, то вверху списка отображается условие правила, а список состоит из его следствий. Тогда правила отвечают на вопрос, что будет при таком условии. Если же дерево построено по следствию, то вверху списка отображается следствие правила, а список состоит из его условий. Эти правила отвечают на вопрос, что нужно, чтобы было заданное следствие. Данный визуализатор отображает те же самые правила, что и предыдущий, но в более удобной для анализа форме.

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. В окне слева расположен список всех элементов транзакций. Справа от каждого элемента указана поддержка: сколько раз данный элемент встречается в транзакциях. В правом верхнем углу расположен список элементов, входящих в условие. Это, например, список товаров, которые приобрел покупатель. Для них нужно найти следствие. Например, товары, приобретаемые совместно с ними. Чтобы предложить человеку то, что он возможно забыл купить. В правом нижнем углу расположен список следствий. Справа от элементов списка отображается поддержка и достоверность. Пусть необходимо проанализировать, что, возможно, забыл покупатель приобрести, если он уже взял вафли и мед. Для этого следует добавить в список условий эти товары (например, с помощью двойного щелчка мыши) и затем нажать на кнопку "Вычислить правила". При этом в списке следствий появятся товары, совместно приобретаемые с данными. В данном случае появятся "Сухари", "Чай", "Сухари и чай", т. е., может быть, покупатель забыл приобрести сухари, чай или и то и другое.




Результаты анализа можно применить и для сегментации покупателей по поведению при покупках, и для анализа предпочтений клиентов, и для планирования расположения товаров в

супермаркетах, кросс-маркетинге. Предлагаемый набор визуализаторов позволяет эксперту найти интересные, необычные закономерности, понять, почему так происходит, и применить их на практике.

## Задание и порядок выполнения практической работы №5

**Задание 1:**
Подготовить данные для поиска ассоциативных данных.

**Задание 2:**
Провести поиск ассоциативных правил, проанализировать полученные результаты.