Министерство науки и высшего образования Российской Федерации

Новосибирский Государственный технический университет

Кафедра автоматизированных систем управления



**ОТЧЁТ**

**Отчет по лабораторной работе 2**

**по дисциплине «Методы анализа данных»**

**«Ассоциативные правила»**

Выполнили

студенты группы АВТ-812:

Березин Дмитрий

Бородина Алина

Глинин Евгений

Преподаватель:

Ганелина Наталья Давидовна,

к.т.н., доцент кафедры АСУ

г. Новосибирск

2021 г.

Содержание

[Цель работы 3](#_Toc86425729)

[Текст задания 3](#_Toc86425730)

[Описание массива данных 4](#_Toc86425731)

[Процедура очистки данных 5](#_Toc86425732)

[Параметры и результаты эксперимента 6](#_Toc86425733)

[Сравнительный анализ и интерпретация результатов 9](#_Toc86425734)

# Цель работы

Знакомство с алгоритмами поиска ассоциативных правил.

# Текст задания

1. Познакомиться с интерфейсом и возможностями платформы Deductor Academic.

2. Загрузить демо-пример («Анализ покупательской корзины»). Можно использовать другой массив подходящего формата.

3. Последовательно меняя параметры (минимальное и максимальное значение поддержки и достоверности) получить наборы ассоциативных правил.

4. Проанализировать полученные результаты: выделить полезные и тривиальные правила; оценить значение достоверности, лифта правил; оценить влияние параметров на получаемые правила.

5. Интерпретировать полученные результаты.

6. Изучить алгоритм FPG. Сравнить с алгоритмом Apriori.

7. Оформить отчет.

# Описание массива данных

Для выделения ассоциативных правил нами был использован массив чеков с покупками в магазине хозяйственных товаров. Массив состоит из 5000 элементов, которые содержат один из представленных в магазине товаров и идентификатор чека. Всего этих товаров 38. Каждый отдельно взятый чек мы будем рассматривать как транзакцию. Одна транзакция может содержать до 6 элементов. Общее количество транзакций – 2045.

# Процедура очистки данных

Для выбранных данных процедуру очистки проводить не надо. Это объясняется тем, что массив не содержит каких-либо пробелов, выбросы зафиксировать также невозможно. Фильтрацию по номерам чеков или названиям товаров мы также не проводим, потому что хотим выделить ассоциативные правила для всего ассортимента товаров. Дубликаты в массиве данных также могут встречаться, это означает, что в одном чеке было куплено несколько одинаковых товаров.

# Параметры и результаты эксперимента

Для выявления ассоциативных правил выполним эксперимент со следующими параметрами:

* Минимальная поддержка: 1%
* Максимальная поддержка: 20%
* Минимальная достоверность: 50%
* Максимальная достоверность: 100%

Причины выбора таких значений следующие:

* Минимальная поддержка меньше 1% нас не интересует, потому что полученные правила с поддержкой меньше 1% будут случаться крайне редко. Так мы отбросим большое количество правил, которые могут дать 100% достоверность, но которые были выведены на основе единственной транзакции из нескольких тысяч. Их можно принять за информационный шум, перегружающий список всех доступных правил и мешающий поиску полезных правил.
* Уровень поддержки больше 20% задавать не имеет смысла, правил с таким уровнем поддержки из нашей выборки выделить нельзя.
* Достоверность больше 50% говорит нам о том, что выделенные правила будут больше подтверждаться, чем опровергаться. Мы хотим выделить такие правила, которые показывали бы совместимые товары. Если бы нам было необходимо выделить правила, которые помогли бы разделить товары в несовместимые группы, то мы бы выставляли уровень достоверности в ассоциативных правилах от 0% до 50%. Потому что такие правила на практике редко подтверждаются, что в контексте нашей задачи говорит о том, что выбранные товары несовместимы между собой.

Результаты проведенного эксперимента представлены ниже в виде набора ассоциативных правил. Правила в таблице отсортированы по убыванию значения достоверности.

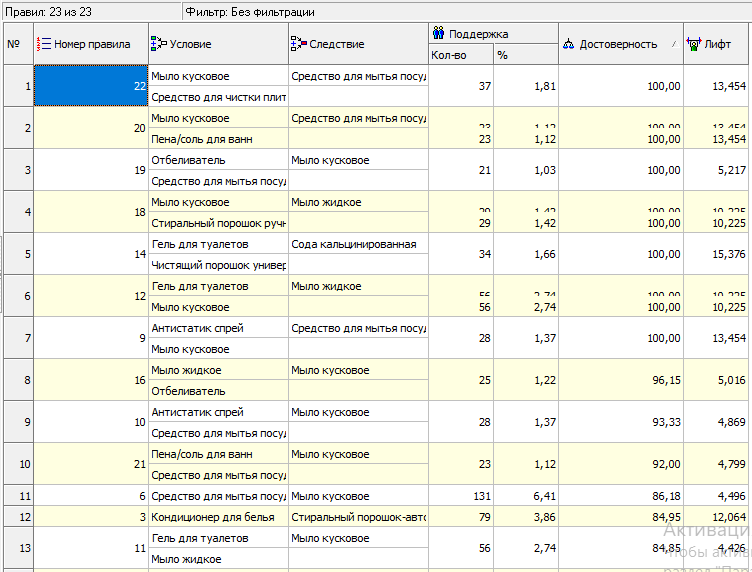


Рисунок 1 – результат эксперимента (часть 1)

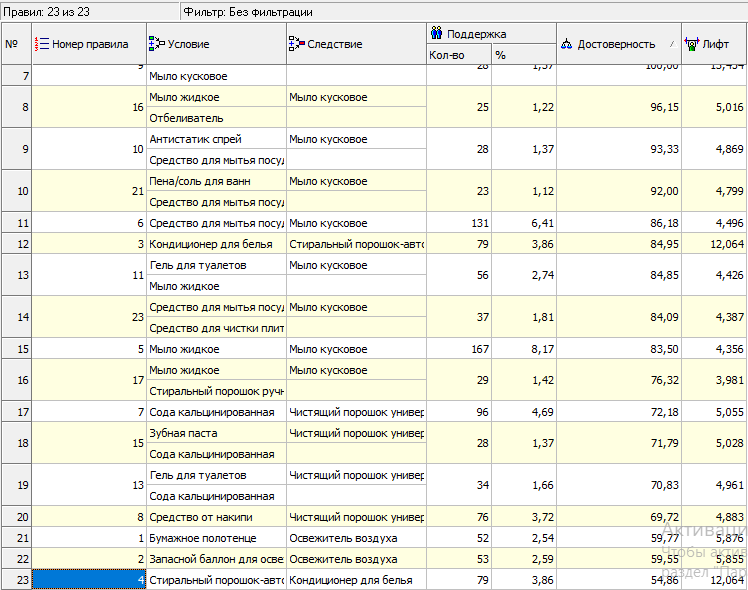


Рисунок 2 – результат эксперимента (часть 2)

# Сравнительный анализ и интерпретация результатов

Из полученных ассоциативных правил мы можем отметить несколько наиболее значимых:

1. Согласно правилу №6, если покупатель берет «средство для мытья посуды», то он скорее всего возьмет еще и «мыло кусковое», так как процент поддержки у данного правила довольно высок и достоверность правила выше 85%
2. Согласно правилу №5, если покупатель берет «мыло жидкое», то он скорее всего возьмет еще и «мыло кусковое», потому что среди полученных правил процент поддержки у этого правила максимальный. Кроме этого, достоверность правила находится на уровне 84%, что очень много.
3. Согласно правилам №20 и №22, если покупатель берет «мыло кусковое» вместе с «пена/соль для ванн» или «средство для чистки плит», то он скорее всего возьмет еще и «средство для мытья посуды», потому что достоверность у обоих правил равна 100%. Но стоит отметить, что процент поддержки довольно низок и у обоих правил он составляет чуть больше процента.

Исходя из всех полученных ассоциативных правил можно сделать вывод, что для магазина наиболее правильным решением было бы разместить рядом «средство для мытья посуды», «мыло жидкое» и «мыло кусковое», так как 14 ассоциативных правил из 23 полученных содержат в условии или следствии хотя бы один из этих товаров. Также некоторые комбинации товаров, например, те, что были отмечены в списке, дают очень высокую достоверность. Это еще раз подтверждает предложение о том, что размещать эти товары рядом было бы наиболее верным решением.

Перейдем к рассмотрению полезных ассоциативных правил из полученных нами. Полезные правила должны обладать как можно большей достоверностью и поддержкой. Для этого введем ограничение на минимальные значения поддержки и достоверности. Пускай полезными правилами будут те правила, у которых уровень поддержки больше 5%, а уровень достоверности превосходит 80%. Тогда в качестве полезных ассоциативных правил мы бы могли выбрать правила №5 и №6. Это как раз те правила, которые были приняты за наиболее значимые чуть выше.