Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Высшая школа электроники и компьютерных наук
Кафедра «Системное программирование»

ОТЧЁТ

по лабораторной работе №2

на тему «Поиск ассоциативных правил»

Выполнил
Студент группы КЭ-120
_____ Д.А. Снегирева
«____»_____2020 г.
Email: dashasneg@mail.ru

Челябинск 2020

ЗАДАНИЕ

Выполните поиск ассоциативных правил для наборов данных из задания 1. Зафиксируйте значение пороговое значение поддержки (например, 10%), варьируйте пороговое значение достоверности (например, от 70% до 95% с шагом 5%). Получите список результирующих правил в удобочитаемом виде (антецедент консеквент).

1. Подготовьте список правил, в которых антецедент и консеквент суммарно включают в себя не более семи объектов (разумное количество). Проанализируйте и изложите содержательный смысл полученного результата.

2. Выполните визуализацию полученных результатов в виде следующих диаграмм:

– сравнение быстродействия поиска правил на фиксированном наборе данных при изменяемом пороге достоверности;

– общее количество найденных правил на фиксированном наборе данных при изменяемом пороге достоверности;

– максимальное количество объектов в правиле на фиксированном наборе данных при изменяемом пороге достоверности;

– количество правил, в которых антецедент и консеквент суммарно включают в себя не более семи объектов, на фиксированном наборе данных при изменяемом пороге достоверности.

СОДЕРЖАНИЕ

1 КРАТКИЕ СВЕДЕНИЯ О НАБОРАХ ДАННЫХ И СРЕДСТВАХ РЕАЛИЗАЦИИ

В данной работе были использованы следующие наборы данных:

1) airports (https://openflights.org/data.html) – датасет, содержащий информацию об аэропортах по всему миру.

2) market basket optimization (https://www.kaggle.com/roshansharma/market-basket-optimization/version/1) – датасет, содержащий информацию о покупках. Данные сгруппированы в транзакции.

3) retail (http://fimi.uantwerpen.be/data/) – датасет, также содержащий информацию о покупках, но все значения представлены цифрами.

В работе была использована библиотека Кристиана Боргельта (Christian Borgelt) PyFIM (https://borgelt.net/pyfim.html). Конкретно в работе были использованы реализации алгоритмов apriori, eclat и fpgrowth,

Репозиторий задания: https://github.com/DasHaSneg/BigDataMiningCourse

Каталог задания: 1_search_rules

2 СПИСОК ПРАВИЛ

В ходе выполнения работы был подготовлен список частых наборов для каждого набора данных.

В наборе данных airports были найдены следующие частые наборы с минимальной поддержкой 2% и достоверностью выше 80%:

- {America/New_York + -5 + United States + A + airport} -> OurAirports
- {America/New_York + -5 + United States + A + OurAirports} -> airport
- {America/New_York + -5 + United States + airport + OurAirports} -> A
- {America/New_York + -5 + A + airport + OurAirports} -> United States
- {America/New_York + United States + A + airport + OurAirports} -> -5
- {-5 + United States + A + airport + OurAirports} -> America/New_York

Из списка можно сделать вывод, что в наборе данных много информации о аэропорте в США в Нью-Йорке.

В наборе данных market basket optimization были найдены следующие частые наборы с минимальной поддержкой 1% и достоверностью выше 53%:

- {frozen vegetables + ground beef} -> mineral water
- {olive oil + frozen vegetables} -> mineral water
- {turkey + milk} -> mineral water
- {cooking oil + eggs} -> mineral water
- {soup + chocolate} -> mineral water
- {soup + milk} -> mineral water

Из полученного списка наборов можно сделать вывод, что вода является популярным товаром.

В наборе данных retail были найдены следующие частые наборы с минимальной поддержкой 2% и достоверностью выше 70%:

- {41} -> 39
- {41 + 48} -> 39
- {38 + 48} -> 39
- {38 + 41} -> 39

- {89} -> 39
- {89} -> 48

Из данного списка нельзя сделать каких-либо выводов кроме того, цифра 39 часто встречается с цифрами 38, 48 и 41.

3 ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ

В ходе выполнения работы была выполнена визуализация полученных результатов для каждого набора данных. На рис. с 1 по 3 представлены диаграммы для airports, market basket и retail соответственно.

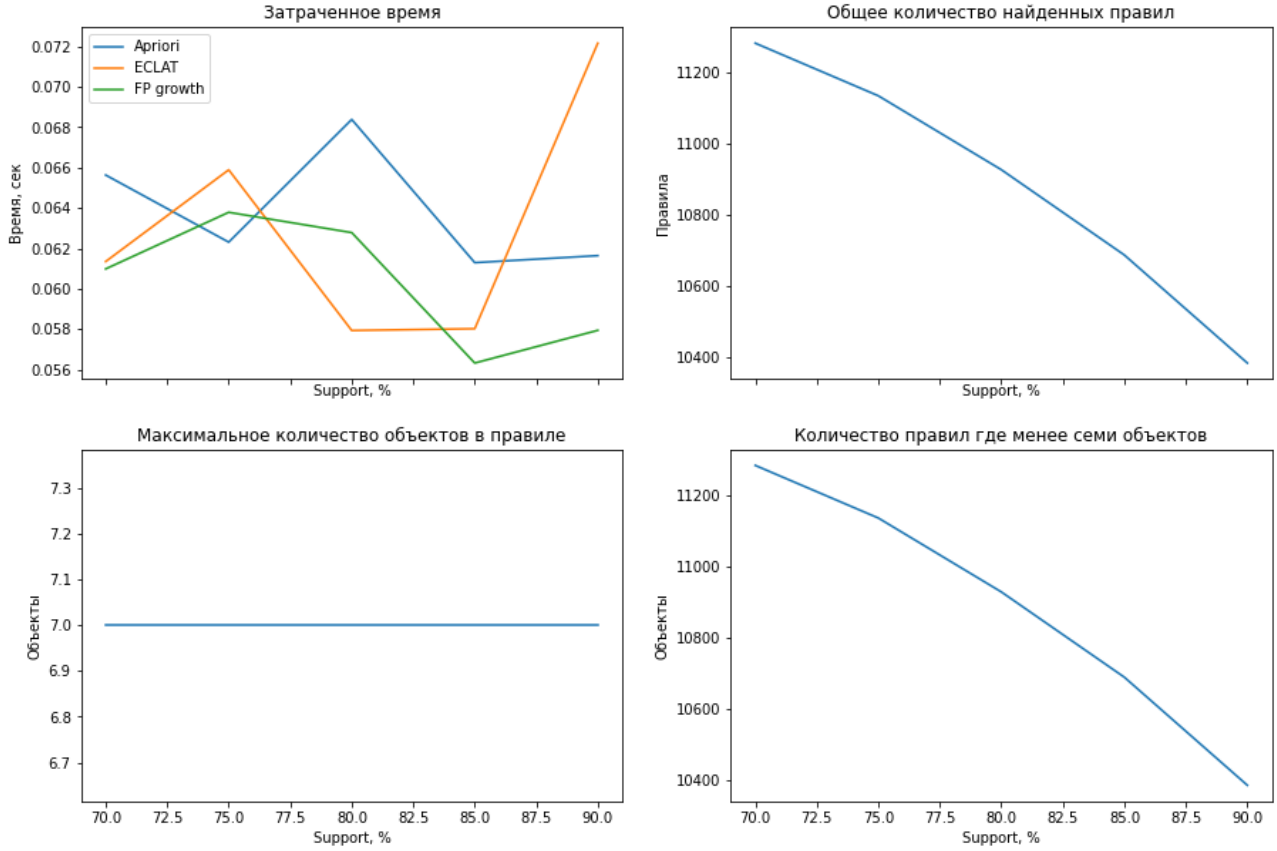

Рис. 1. Результаты для набора данных airports

На каждом рисунке представлены следующие диаграммы:

– сравнение быстродействия поиска правил на фиксированном наборе данных при изменяемом пороге достоверности;

– общее количество найденных правил на фиксированном наборе данных при изменяемом пороге достоверности;

– максимальное количество объектов в правиле на фиксированном наборе данных при изменяемом пороге достоверности;

– количество правил, в которых антецедент и консеквент суммарно включают в себя не более семи объектов, на фиксированном наборе данных при изменяемом пороге достоверности.

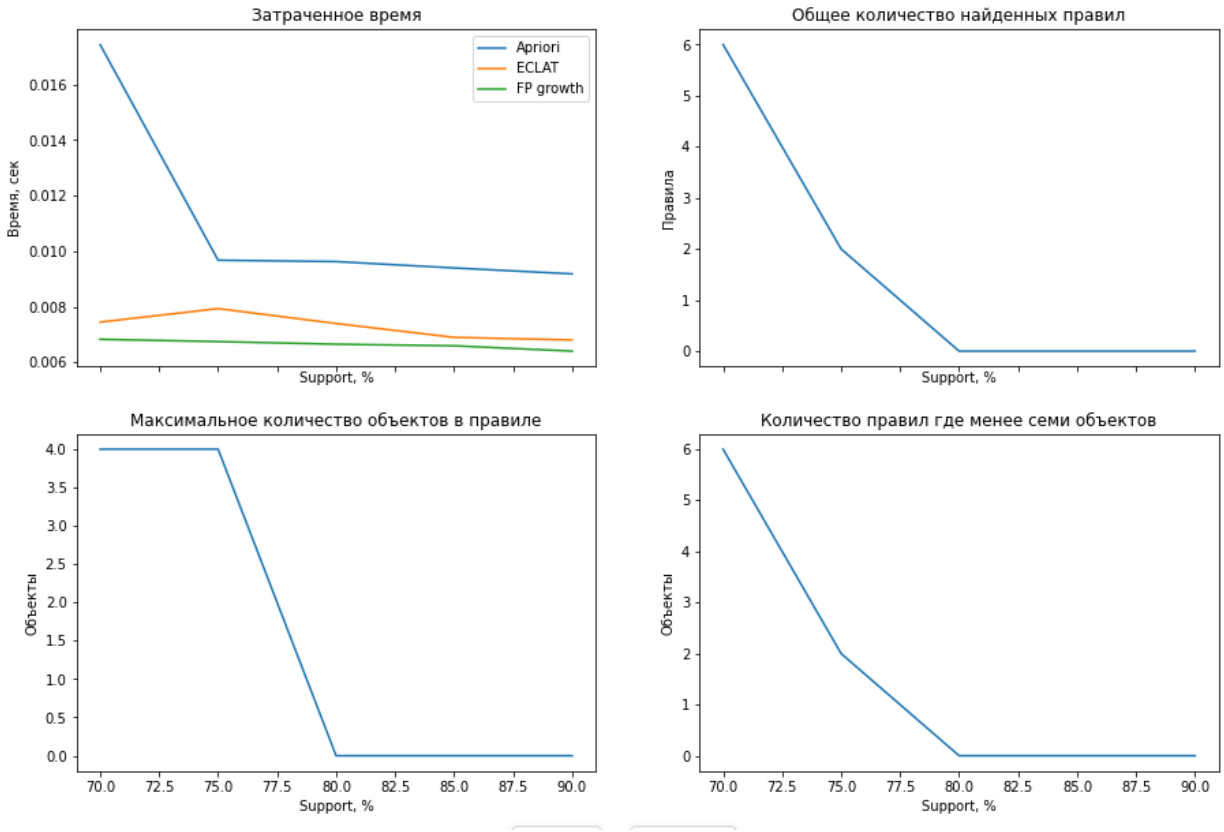

Рис. 1. Результаты для набора данных market basket

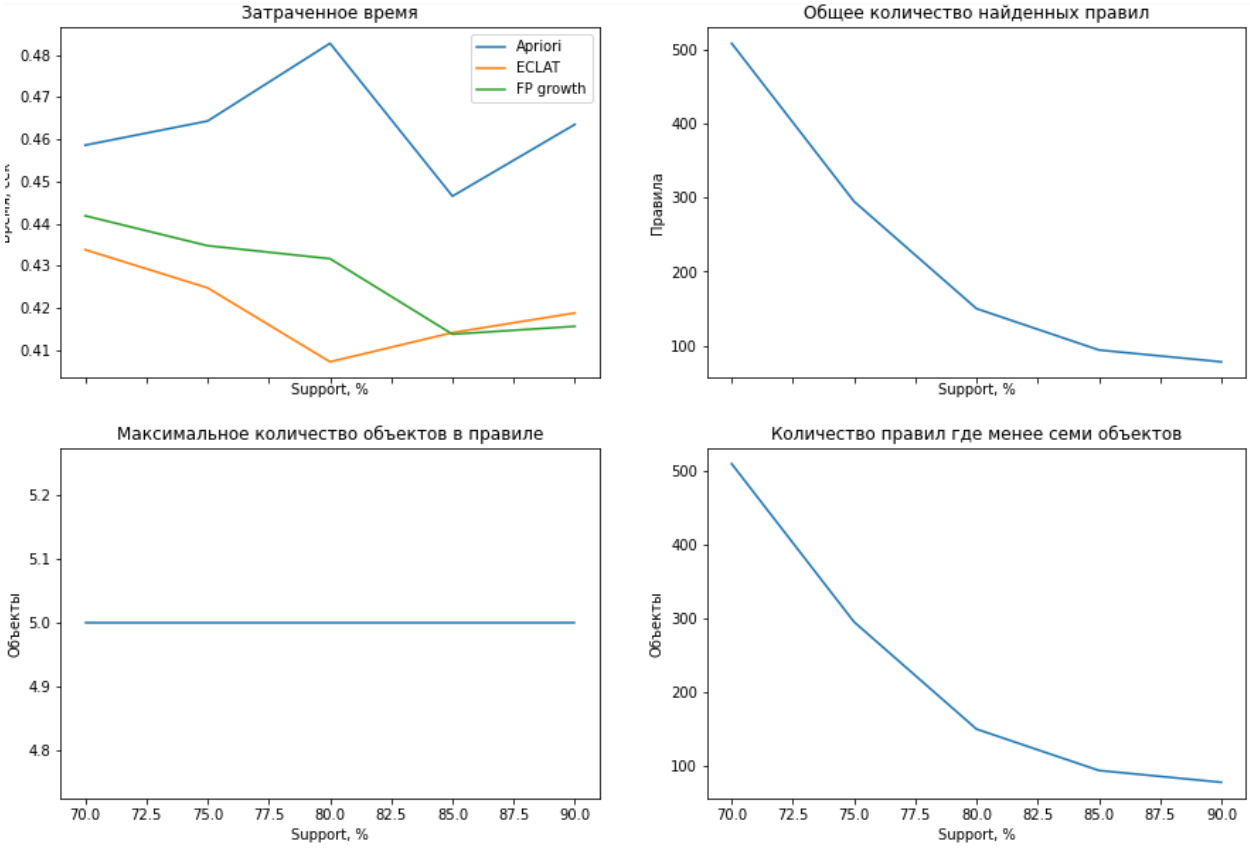

Рис. 1. Результаты для набора данных retail

Для визуализации были использованы пороговое значение поддержки 0.3 и пороговые значения достоверности от 70 до 95 с шагом 0.1.

На всех диаграммах с временем выполнения алгоритмов Apriori показывает худший результат чем два других.

Общее количество правил понижается у всех при увеличении достоверности, при этом у первого набора найдено огромное количество правил по сравнению с другими и при достижении порога достоверности 90% не достигает нуля.

Только у второго набора данных разные значения максимальных объектов в зависимости от достоверности.