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 №1

на тему «Поиск частых наборов»

|  |  |
| --- | --- |
|  | Выполнил  Студент группы КЭ-120  \_\_\_\_\_\_\_\_\_\_\_\_\_ Д.А. Снегирева  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_2020 г. |
|  | Email: dashasneg@mail.ru |

Челябинск 2020

**ЗАДАНИЕ**

Выполните поиск частых наборов объектов в трех различных наборах данных с помощью следующих алгоритмов (или их модификаций): Apriori, FP-Growth, ECLAT. Наборы данных должны существенно отличаться друг от друга по количеству транзакций и/или типичной длине транзакции (количеству объектов). Варьируйте пороговое значение поддержки (например: 1%, 3%, 5%, 10%, 15%, 20%). Проверьте идентичность результатов, полученных с помощью различных алгоритмов.

1. Подготовьте список частых наборов, в которых не более семи объектов (разумное количество). Проанализируйте и изложите содержательный смысл полученного результата.

2. Выполните визуализацию полученных результатов в виде следующих диаграмм: ­ сравнение быстродействия алгоритмов на фиксированном наборе данных при изменяемом пороге поддержки; ­

* общее количество частых наборов объектов на фиксированном наборе данных при изменяемом пороге поддержки; ­
* максимальная длина частого набора объектов на фиксированном наборе данных при изменяемом пороге поддержки; ­
* количество частых наборов объектов различной длины на фиксированном наборе данных при изменяемом пороге поддержки.

СОДЕРЖАНИЕ

[ЗАДАНИЕ 2](#_Toc68233149)

[СОДЕРЖАНИЕ 3](#_Toc68233150)

[1 КРАТКИЕ СВЕДЕНИЯ О НАБОРАХ ДАННЫХ И СРЕДСТВАХ РЕАЛИЗАЦИИ 4](#_Toc68233151)

[2 СПИСОК ЧАСТЫХ НАБОРОВ 5](#_Toc68233152)

[3 ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ 7](#_Toc68233154)

# 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\_sets

# СПИСОК ЧАСТЫХ НАБОРОВ

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%:

* + {'America/Los\_Angeles', 'United States', 'A', 'OurAirports', 'airport', '-8'} (поддержка 2%)
  + {'America/Anchorage', 'United States', 'A', 'OurAirports', 'airport', '-9'} (поддержка 2%)
  + {'Europe/Paris', 'France', 'E', 'OurAirports', 'airport', '1'} (поддержка 3%)
  + {'Europe/Berlin', 'E', 'OurAirports', 'airport', 'Germany', '1'} (поддержка 3%)
  + {'America/Chicago', 'United States', 'A', 'OurAirports', 'airport', '-6'} (поддержка 4%)
  + {'America/New\_York', '-5', 'United States', 'A', 'OurAirports', 'airport'} (поддержка 6%)

В частых наборах часто встречается США, что означает, что в набор содержит много информации связанной с аэропортами США.

В наборе данных market basket optimization были найдены следующие частые наборы с минимальной поддержкой 4%:

* + {'ground beef', 'mineral water'} (поддержка 4%)
  + {'milk', 'mineral water'} (поддержка 5%)
  + {'chocolate', 'mineral water'} (поддержка 5%)
  + {'eggs', 'mineral water'} (поддержка 5%)
  + {'spaghetti', 'mineral water'} (поддержка 6%)

Из полученного списка наборов можно сделать вывод, что вода является популярным товаром и ее часто приобретают с другими продуктами.

В наборе данных retail были найдены следующие частые наборы с минимальной поддержкой 2%:

* + {'38', '48', '39'} (поддержка 7%)
  + {'32', '48', '39'} (поддержка 6%)
  + {'41', '48', '39'} (поддержка 8%)

Из данного списка нельзя сделать каких-либо выводов кроме того, что цифры 48 и 39 часто встречаются вместе.

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.

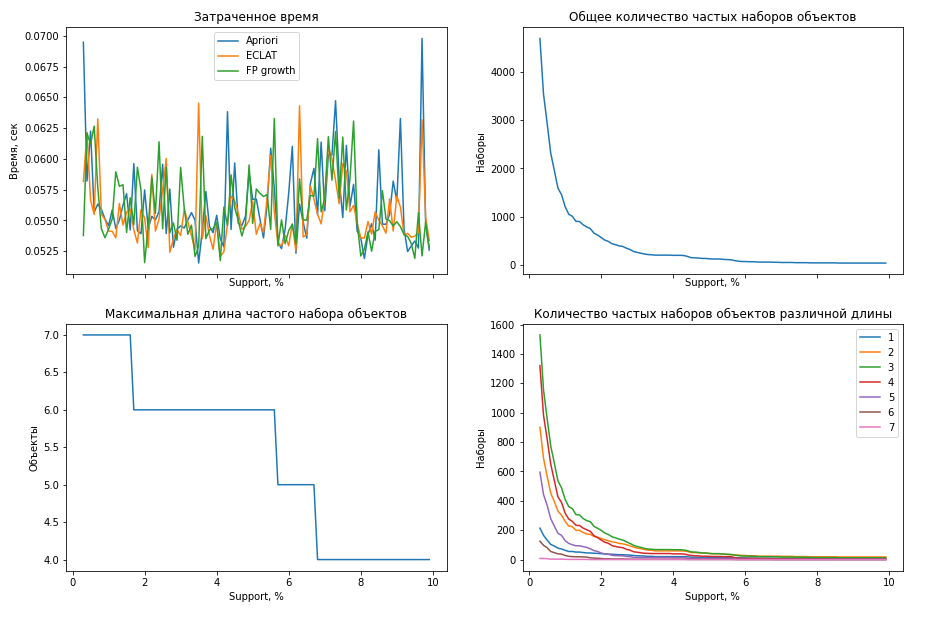


Рис. 1. Результаты для датасета airports

На каждом рисунке представлены следующие диаграммы:

* сравнение быстродействия алгоритмов на фиксированном наборе данных при изменяемом пороге поддержки;
* ­общее количество частых наборов объектов на фиксированном наборе данных при изменяемом пороге поддержки; ­
* максимальная длина частого набора объектов на фиксированном наборе данных при изменяемом пороге поддержки; ­
* количество частых наборов объектов различной длины на фиксированном наборе данных при изменяемом пороге поддержки.

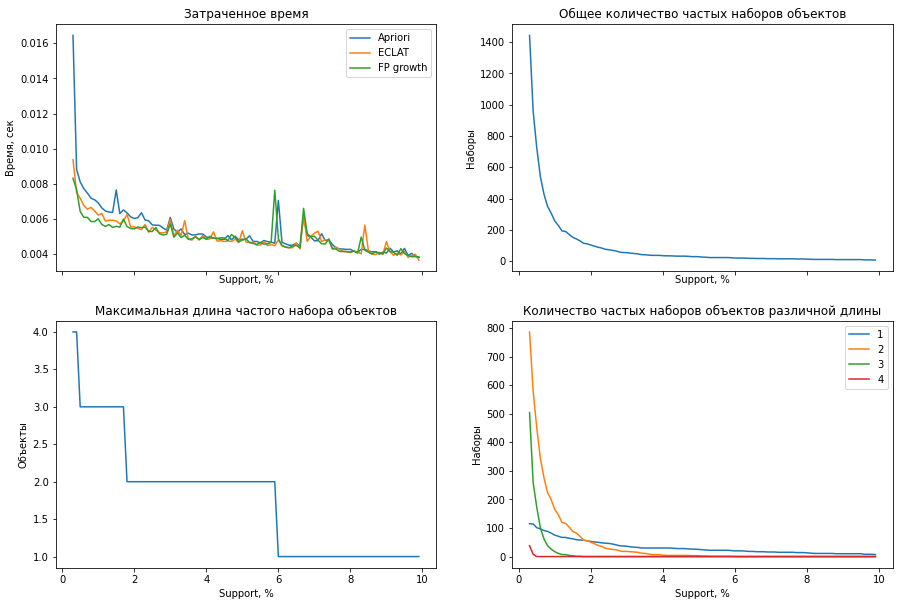


Рис. 1. Результаты для датасета market basket

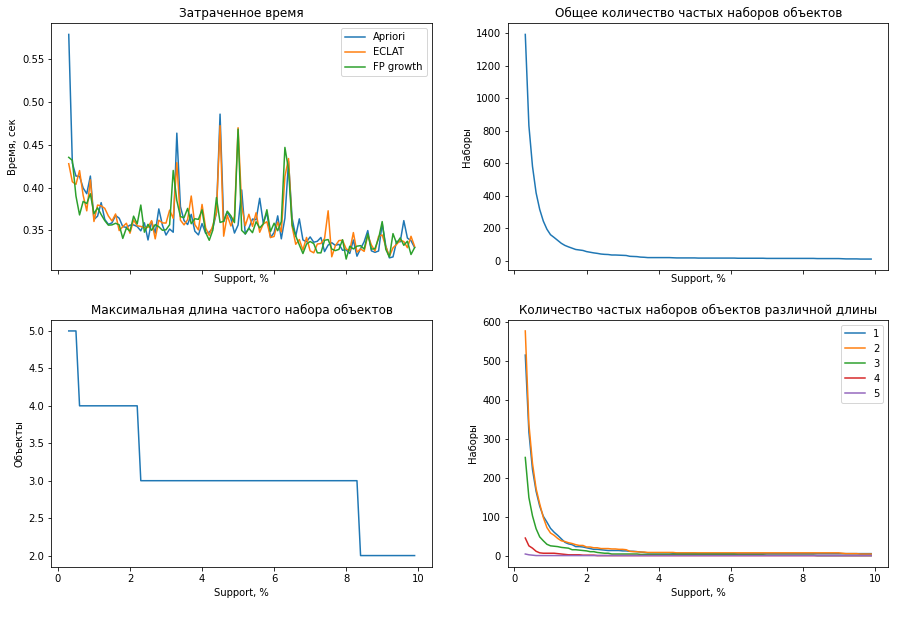


Рис. 1. Результаты для датасета retail

Для визуализации были использованы пороговые значения поддержки от 0.3 до 10 с шагом 0.1.

Для первого набора данных максимальная длина частого набора 7 и столько же частых наборов объектов различной длины В остальных наборах данных максимальная длина меньше 4 и 5 и количество частных наборов объектов различной длины также меньше.

Стоит заметить, что общее количество частых наборов объектов всегда уменьшается при увеличении значения минимальной поддержки независимо от набора данных, поэтому вторая диаграмма на всех рисунках имеет одинаковый вид.

Все алгоритмы работают быстро, но если обратить внимание на результаты для первого и второго набора данных, то Apriori по сравнению с FP-frowth и ECLAT работает медленнее.