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 системного программирования

ОТЧЁТ ПО ЗАДАНИЮ №1

по дисциплине «Интеллектуальный анализ больших данных»

Тема: Поиск частых наборов

Выполнил
студент группы КЭ-120
Глизница Максим Николаевич
E-mail: letadllo@mail.ru

Челябинск 2021

1. Задание

Задание 1. Поиск частых наборов

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.

2. Краткие сведения о наборах данных

Использованные наборы данных:

Groceries dataset (https://www.kaggle.com/heeraldedhia/groceries-dataset). Содержит данные о покупках продуктов в формате «покупатель-дата-продукт». Каждая запись содержит один продукт, поэтому для получения списка транзакций требуется предобработка данных. Всего содержит 14963 транзакции, средняя длина транзакции: 2.54.

Dataset for Apriori Algorithm - Frequent Itemsets (https://www.kaggle.com/akalyasubramanian/dataset-for-apriori-algorithm-frequent-itemsets). Также содержит данные о покупках продуктов, но уже сгруппированные в транзакции. Всего содержит 7501 транзакций, средняя длина – 3.91.

MyAnimeList Dataset (https://www.kaggle.com/azathoth42/myanimelist?select=AnimeList.csv) – содержит информацию о различных аниме, взятую с сайта

myanimelist.net, из которой был использован список жанров. Таким образом, каждое аниме было рассмотрено как транзакция, а ассортимент товаров составили жанры, использующиеся на сайте (такие как Action, Adventure и т.д.). Всего содержит 14414 транзакции, средняя длина – 2.91.

3. Краткие сведения о средствах реализации

Для реализации методов была использована библиотека PyFIM, автор Christian Borgelt (https://borgelt.net/pyfim.html). Библиотека содержит используемые в задании алгоритмы Apriori, ECLAT и FP-growth, а также некоторые другие.

Репозиторий по дисциплине: https://github.com/Airpllane/DAAlgorithms.

Каталог для задания: 1. Itemsets.

4. Частые наборы

В ходе анализа первого набора данных о продуктах были обнаружены следующие наборы с поддержкой около 1%:

- Йогурт + Молоко;
- Газированная вода + Молоко;
- Хлеб + Овощи;
- Хлеб + Молоко;
- Овощи + Молоко.

Одним из наиболее популярных наборов из 4 элементов оказался набор:

- Колбаса + Йогурт + Хлеб + Молоко (поддержка 0.03%).

В целом, первый набор о продуктах содержит достаточно малое количество частых наборов данных, причиной чего может быть низкая средняя длина транзакции. Выявленные наборы данных в основном содержат просто популярные продукты, которые не имеют логической связи друг с другом.

В ходе анализа второго набора данных о продуктах были обнаружены следующие наборы с поддержкой около 5%:

- Шоколад + Минеральная вода;
- Яйца + Минеральная вода;
- Спагетти + Минеральная вода.

Самым популярным набором из 4 элементов оказался набор:

- Молоко + Шоколад + Спагетти + Минеральная вода (поддержка 0.5%).

Из этих наборов можно сделать вывод, что минеральная вода в этом магазине является популярным напитком, который часто покупается в комплекте с едой.

Были также обнаружены наборы, состоящие из продуктов, которые могут использоваться вместе для приготовления еды:

- Фарш + спагетти (поддержка 3.9%);
- Яйца + спагетти (поддержка 3.7%).

В ходе анализа набора данных о жанрах аниме были обнаружены наборы с поддержкой около 7%:

- Sci-Fi + Action;
- Fantasy + Adventure;
- Adventure + Action.

Можно увидеть, что с научно-фантастическим сеттингом чаще сочетается аниме жанра Action, в то время как к сеттингу фэнтэзи больше подходит жанр Adventure.

Более интересные результаты можно увидеть, если просмотреть наборы по 3 элемента, которые включают в себя такие, как:

- Mecha + Sci-Fi + Action (поддержка 3.4%)
- School + Romance + Comedy (поддержка 2.1%)
- Fantasy + Adventure + Comedy (поддержка 3%)

Жанры, находящиеся в этом списке, часто используются в сочетании при производстве аниме, что говорит о том, что их легко совмещать в пределах одной истории.

5. Визуализация

Для визуализации были использованы пороговые значения поддержки от 0.5 до 10 с шагом 0.1. Результаты визуализации для первого набора данных приведены на рис. 1.

На рисунке можно увидеть, что общее количество наборов, максимальная длина набора и количество наборов определённых длин падают с увеличением порога поддержки. Это ожидаемо, так как чем выше установлен порог поддержки, тем меньше наборов могут его пройти. Временные затраты при этом меняются в очень узких пределах и почти не зависят от порога поддержки.

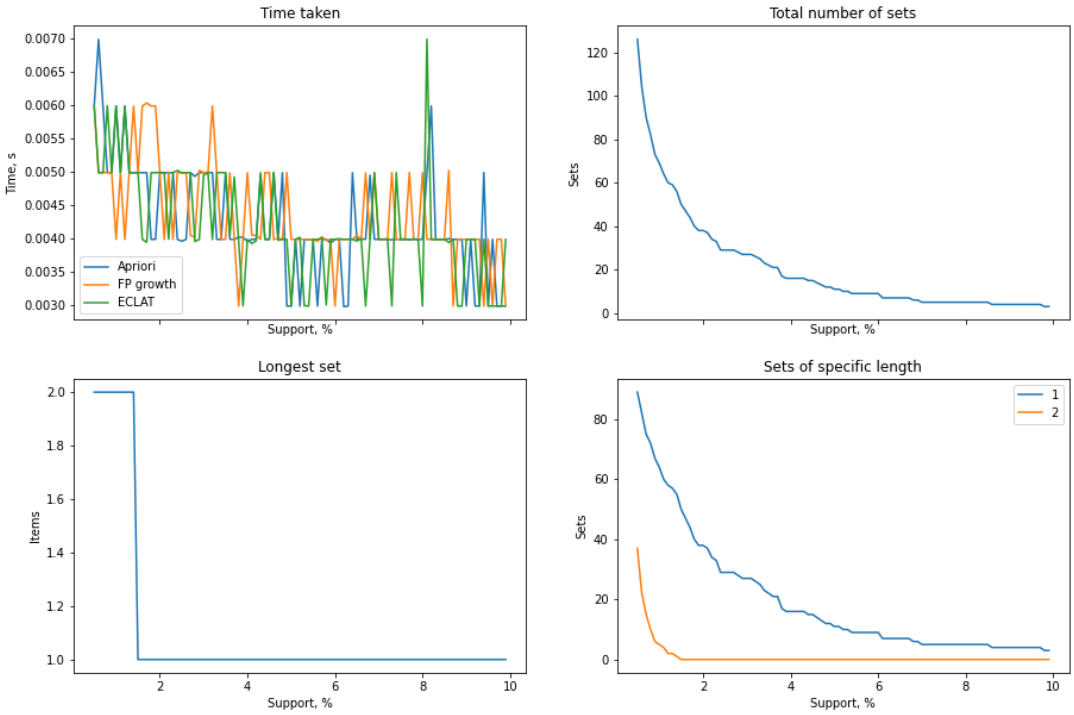

Рис. 1. Результаты визуализации для первого набора данных

Результаты визуализации для второго набора данных приведены на рис. 2.

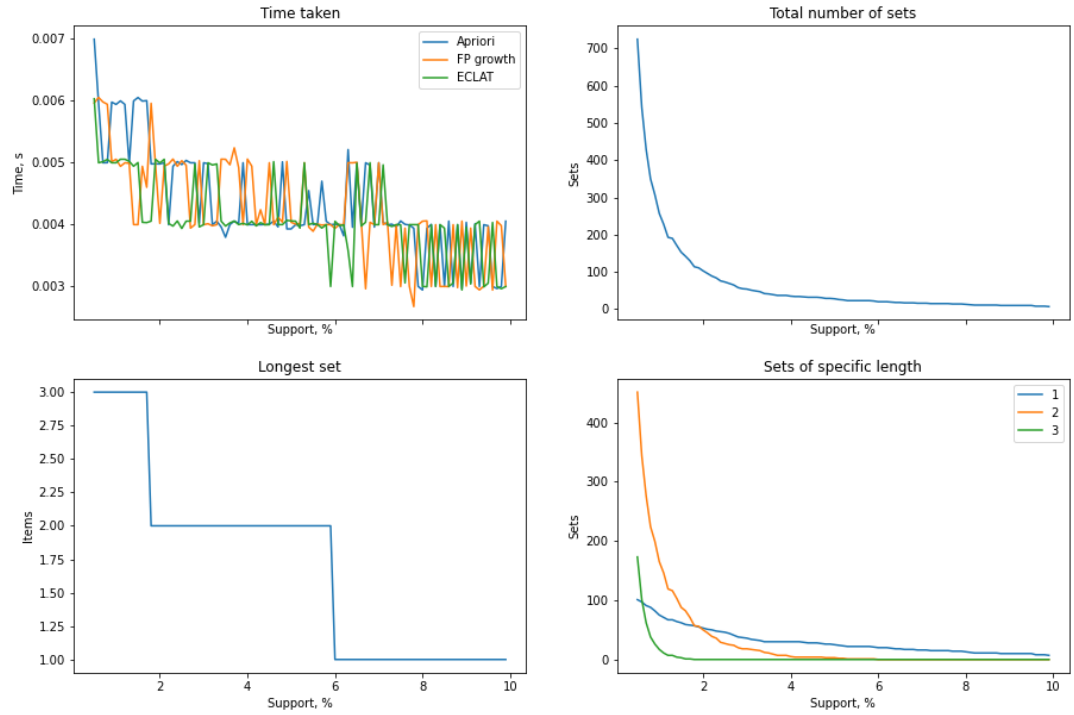

Рис. 2. Результаты визуализации для второго набора данных

Характер зависимости измеряемых значений от порога поддержки не меняется, но можно увидеть, что во втором наборе данных обнаруживается значительно больше частых наборов. В нём также находятся наборы длины 3, которые не удавалось найти в первом наборе при данном пороге поддержки.

Результаты визуализации для третьего набора данных приведены на рис. 3.

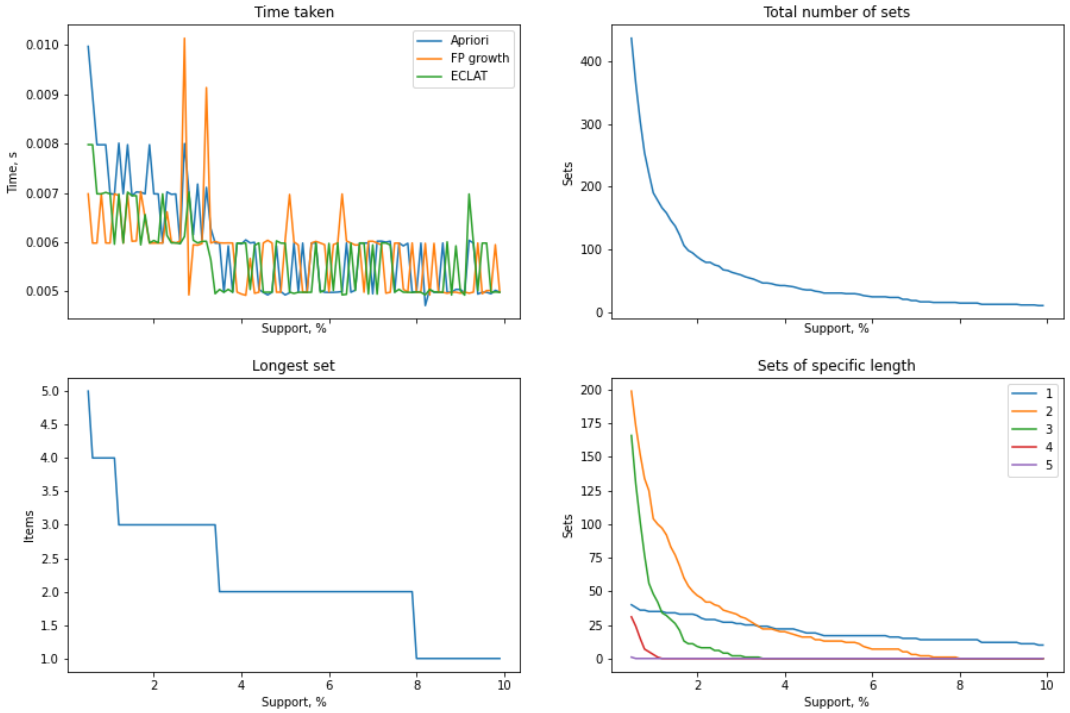

Рис. 3. Результаты визуализации для третьего набора данных

Можно увидеть, что в третьем наборе данных меньше общее количество обнаруженных частых наборов, но заметно больше средняя длина набора. Это объясняется тем, что множество элементов транзакций в третьем наборе данных значительно уже.

Ни в одном из трёх наборов не удалось получить подходящих данных для сравнения времени выполнения алгоритмов (во всех полученных случаях время выполнения было практически одинаковым). Поэтому алгоритмы были запущены на третьем наборе ещё раз с другими пороговыми значениями поддержки: от 0.01 до 0.1, с шагом 0.01. Результат этой визуализации приведён на рис. 4.

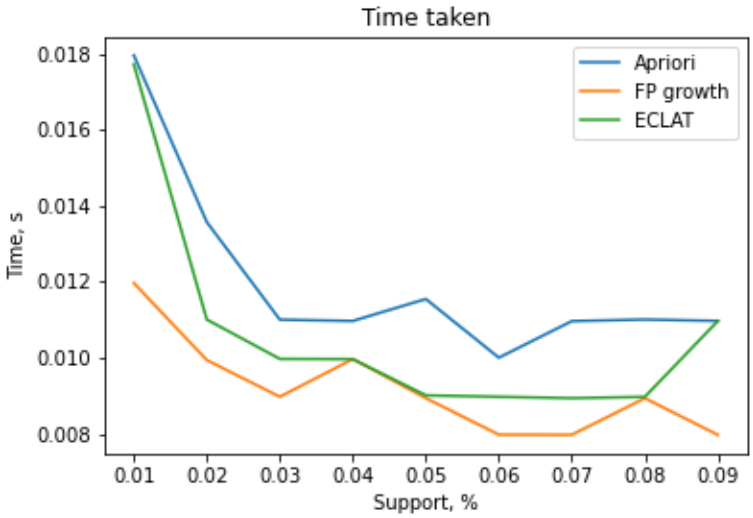

Рис. 4. Временные затраты алгоритмов при низких порогах поддержки

Можно увидеть, что несмотря на то, что все использованные алгоритмы работают очень быстро, FP-growth имеет некоторое преимущество по скорости, в то время как Apriori несколько отстаёт. Алгоритм ECLAT показывает скорость медленнее FP-growth, но быстрее Apriori.