Поиск частых наборов

Кузьмичева Ольга, КЭ-403

Задание

1. Разработайте программу, которая выполняет поиск частых наборов объектов в заданном наборе данных с помощью алгоритма Apriori (или одной из его модификаций). Список результирующих наборов должен содержать как наборы, так и значение поддержки для каждого набора. Параметрами программы являются набор, порог поддержки и способ упорядочивания результирующего списка наборов (по убыванию значения поддержки или лексикографическое).

2. Проведите эксперименты на двух наборах из различных предметных областей. Наборы данных должны существенно отличаться друг от друга по количеству транзакций и/или типичной длине транзакции (количеству объектов). Например, наборы retail (сведения о покупках в супермаркете: см. html, скачать gzip-архив) и accidents (сведения о ДТП: см. детальное описание в формате PDF, см. html, скачать gzip-архив). В экспериментах варьируйте пороговое значение поддержки (например: 1%, 3%, 5%, 10%, 15%).

3. Выполните визуализацию результатов экспериментов в виде следующих диаграмм:

- сравнение быстродействия на фиксированном наборе данных при изменяемом пороге поддержки;
- количество частых наборов объектов различной длины на фиксированном наборе данных при изменяемом пороге поддержки.

4. Подготовьте отчет о выполнении задания и загрузите отчет в формате PDF в систему. Отчет должен представлять собой связный и структурированный документ со следующими разделами:

- формулировка задания;
- гиперссылка на каталог репозитория с исходными текстами, наборами данных и др. сопутствующими материалами;
- рисунки с результатами визуализации;
- пояснения, раскрывающие смысл полученных результатов.

В соответствии с заданием была разработана программа, выполняющая поиск частых наборов объектов с помощью алгоритма Apriori. В репозитории размещено пять файлов: apriori1.ipynb, apriori2.ipynb – реализация поиска частых наборов на разных датасетах, dataset1.txt – датасет, используемый в apriori1.ipynb, Online_Retail.xls – датасет, используемый в apriori2.ipynb, results.xls – результаты экспериментов и графики.

Предобработка данных датасетов, код алгоритма, а также результаты каждого эксперимента представлены в google colab файлах, размещенных в репозитории.

Эксперимент 1

В результате поиска наборов в первом датасете (наборы retail - сведения о покупках в супермаркете), были получены графики, представленные на рисунках 1 и 2. Значения порога поддержки взяты следующие: 1%, 3%, 5%, 10%, 15%.

На рисунке 1 можно заметить, что время выполнения алгоритма уменьшается при увеличении порога поддержки. Другими словами, быстродействие на фиксированном наборе данных при увеличении порога поддержки увеличивается. Это можно объяснить тем, что при большем числе параметра поддержки, составляется меньше наборов.

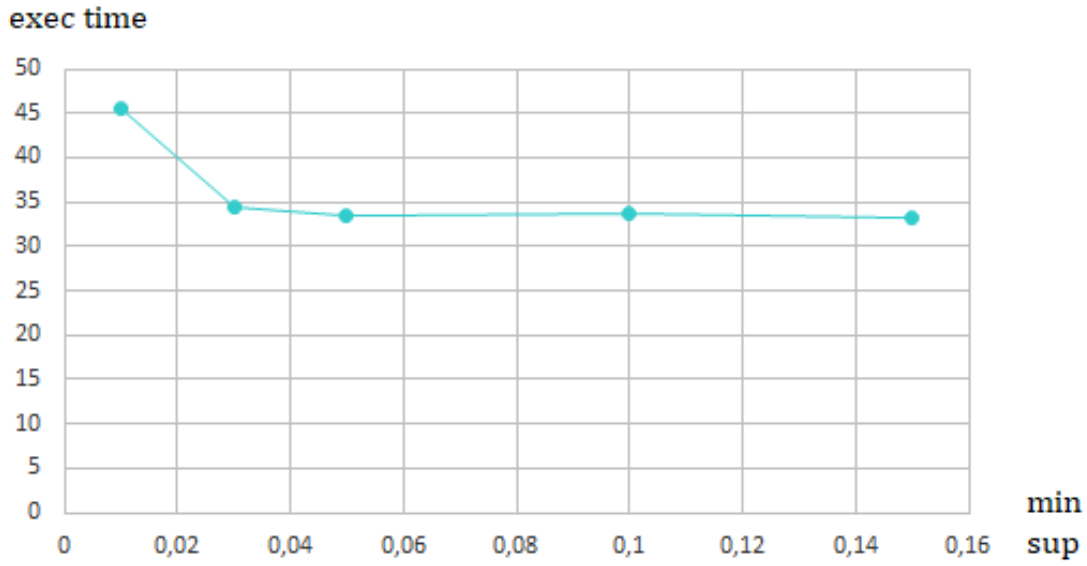

Рисунок 1 – Быстродействие при изменяемом пороге поддержки

На рисунке 2 можно заметить, что количество частых наборов объектов различной длины на фиксированном наборе данных при увеличении порога поддержки уменьшается.

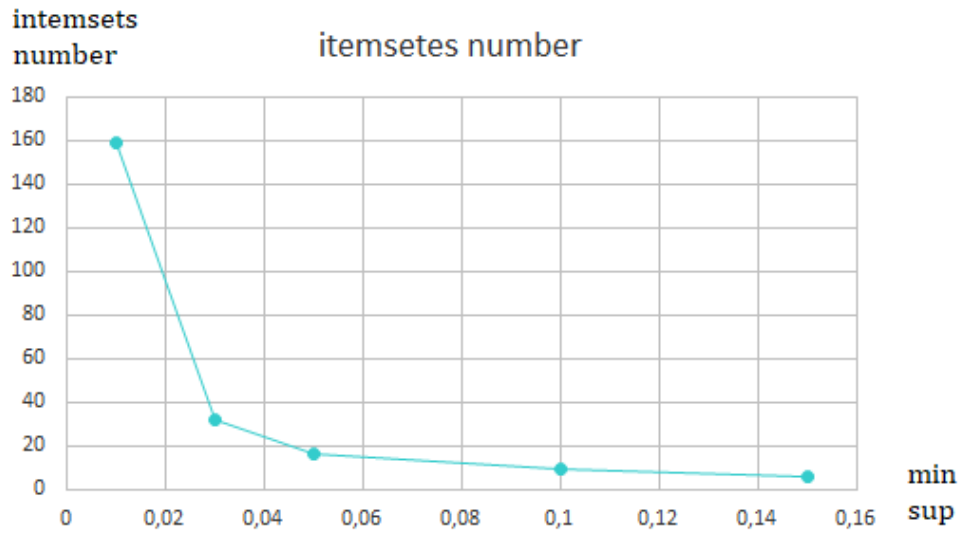

Рисунок 2 – Количество частых наборов при изменяемом пороге поддержки

Эксперимент 2

В результате поиска наборов во втором датасете (наборы online retail - сведения о покупках в супермаркете во Франции), были получены графики, представленные на рисунках 3 и 4. В связи со сложностью датасета, не удалось выполнить эксперимент на тех же значениях порога поддержки, что и в предыдущем эксперименте: при значениях ниже 17% происходило переполнение оперативной памяти. Были взяты следующие значения порога поддержки: 17%, 25%, 35%, 45%, 60%.

Несмотря на другие значения порога поддержки и отличный от первого эксперимента датасет, на рисунке 3 можно заменить, что график быстродействия ведет себя так же, как в первом эксперименте: быстродействие на фиксированном наборе данных при увеличении порога поддержки увеличивается.

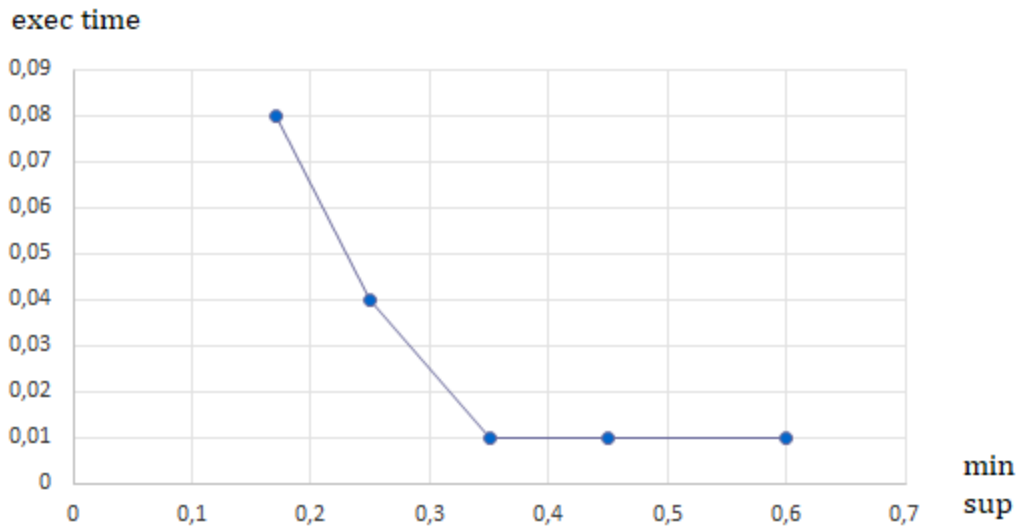

Рисунок 3 – Быстродействие при изменяемом пороге поддержки

На рисунке 3 можно заметить, что так же, как в эксперименте 1, количество частых наборов объектов различной длины на фиксированном наборе данных при увеличении порога поддержки уменьшается.

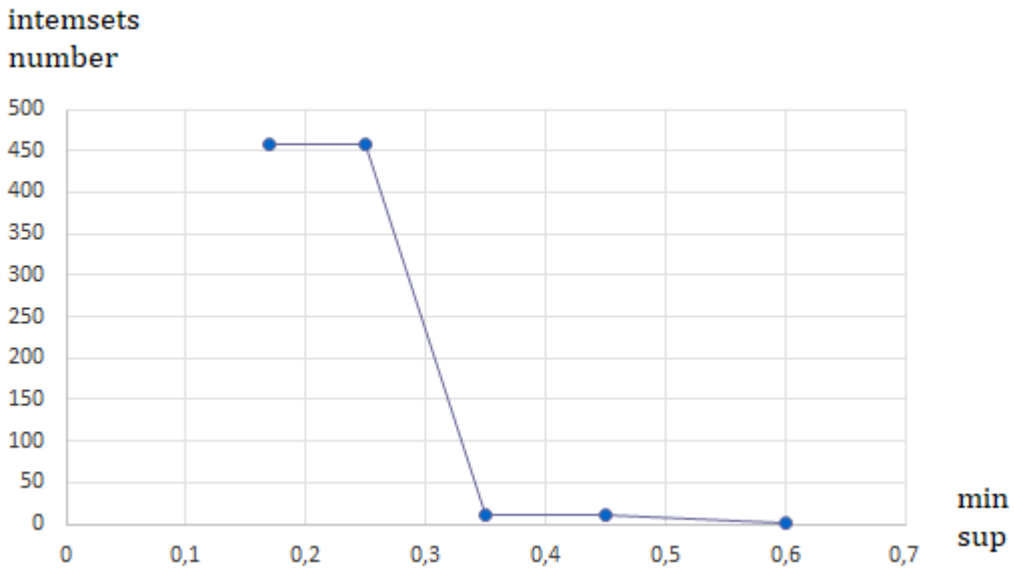

Рисунок 4 – Количество частых наборов при изменяемом пороге поддержки

Вывод

- Быстродействие на фиксированном наборе данных при увеличении порога поддержки увеличивается.
- Количество частых наборов объектов различной длины на фиксированном наборе данных при увеличении порога поддержки уменьшается.