Поиск неприводимых ассоциативных правил в частично упорядоченных данных

И.Е. Генрихов[1], Е.В. Дюкова[2]

[1]*ООО «Мобайл парк ИТ», Панфилова 21/1, Химки, Россия, 141407*
[2]*ФИЦ «Информатика и управление» РАН, Вавилова 44/2, Москва, Россия, 119333*

Аннотация
Рассматривается задача поиска ассоциативных правил специального вида в данных с элементами из декартового произведения конечных частично упорядоченных множеств. Для сокращения временных затрат при анализе небинарных данных используются модификация классического бинарного FP-дерева и параллельные вычисления на основе технологии CUDA. Приводятся результаты тестирования последовательного и параллельного алгоритмов.

Ключевые слова
Декартовое произведение частичных порядков, база данных, частый элемент, ассоциативное правило, FP-дерево, параллельные вычисления, CUDA

1. Введение

Задача поиска ассоциативных правил в данных является одной из центральных задач интеллектуального анализа информации и имеет важное прикладное значение. Поиск ассоциативных правил обычно осуществляется на основе нахождения в данных часто встречающихся элементов (частых событий). Ассоциативное правило (АП) устанавливает зависимость между двумя частыми событиями, согласно которой одно частое событие X с некоторой «достоверностью» влечёт другое частое событие Y . При этом элементы X и Y порождаются одним общим частым элементом, обозначаемым далее (X, Y) . Вопросы поиска АП наиболее изучены в случае бинарных данных [1].

Один из известных способов нахождения частых элементов в бинарных данных основан на построении FP-дерева [2]. В случае небинарных данных, как правило, осуществляется бинаризация исходных данных по некоторому числовому набору порогов и задача сводится к построению бинарного FP-дерева. Результат существенно зависит от выбора набора порогов. Однако перебор по всем возможным вариантам бинаризации требует больших временных затрат. В [3] для сокращения времени анализа небинарных данных, в том числе с элементами из декартового произведения конечных частично упорядоченных множеств, предложена модель порогового FP-дерева (TFP-дерева), которая является модификацией бинарного FP-дерева. Частые элементы и АП, найденные с использованием TFP-дерева, названы пороговыми. В [4] на основе TFP-дерева и технологии CUDA разработаны эффективные параллельные алгоритмы поиска максимальных пороговых частых элементов.

Наиболее информативными считаются те АП, которые порождаются максимальными частыми элементами (X, Y) с «минимальной» посылкой X . Такие правила называются неприводимыми, и задача их поиска особенно важна в случае больших данных. В настоящей работе разработаны и исследованы два алгоритма поиска неприводимых пороговых АП на основе TFP-дерева: последовательный алгоритм TFP-tree и его параллельная версия на основе технологии CUDA (алгоритм DPTFP-tree).

2. Основные результаты

Каждый шаг алгоритма TFP-tree состоит из трех этапов. На первом этапе на основе анализа базы данных D строится множество так называемых значимых наборов порогов H_D и для

каждого $H \in H_D$ ищутся максимальные пороговые частые элементы. На втором этапе для каждого найденного максимального порогового частого элемента строится условное FP-дерево. На третьем этапе по построенному условному FP-дереву ищутся неприводимые пороговые АП. В параллельном алгоритме DPTFP-tree множество H_D разбивается на непересекающиеся подмножества, каждое из которых подаётся на отдельный вычислительный блок графического процессора (GPU) для поиска искомых правил. В этом алгоритме для ускорения вычислений активно применяется динамический параллелизм.

На рис. 1 приведено время поиска множества максимальных пороговых частых элементов $|F_D|$ алгоритмами TFP-tree и DPTFP-tree (это самый трудоёмкий этап их работы). Указано среднее время поиска при обработке 20 случайных небинарных баз данных размера $m \times n$ с элементами из декартового произведения конечных цепей.

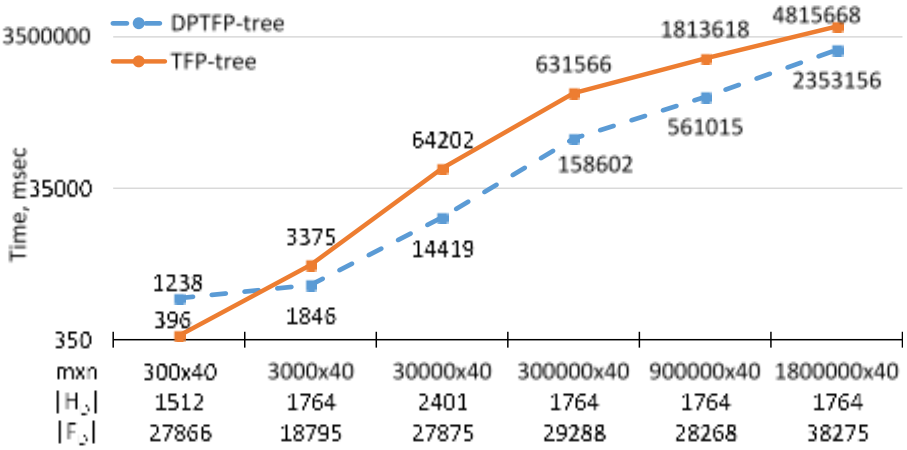

Рисунок 1: Зависимость времени поиска множества максимальных пороговых s -частых элементов от числа транзакций m при числе атрибутов $n = 40$, $s = 0.3$

3. Заключение

Рассмотрена актуальная задача логического анализа данных – задача поиска неприводимых ассоциативных правил в произведении частичных порядков. Для её решения на основе использования TFP-дерева (порогового FP-дерева) и технологии CUDA разработаны последовательный и параллельный алгоритмы. Приведены результаты тестирования построенных алгоритмов на модельных данных, свидетельствующие об эффективности предлагаемого решения поставленной задачи.

4. Благодарности

Работа выполнена при частичной финансовой поддержке РФФИ (проект № 19-01-00430-а).

5. Литература

[1] Agrawal, R. Mining association rules between sets of items in large databases / R. Agrawal, T. Imielinski, A. Swami // Proc. of the 1993 ACM SIGMOD Inter. Conf. on management of data. – 1993. – P. 207-216. DOI: 10.1145/170036.170072.

[2] Han, J. Mining Frequent Patterns without Candidate Generation / J. Han, H. Pei, Y. Yin // Proc. Conf. on the Management of Data (SIGMOD'00, Dallas, TX). – 2000. – P. 1-12.

[3] Genrikhov, I. Finding Frequent Elements for a Product of Partial Orders and Association Rules / I. Genrikhov, E. Djukova // Inter. Conf. on Information Technology and Nanotechnology (ITNT). IEEE Publisher. – 2020. – P. 1-5. DOI: 10.1109/ITNT49337. 2020.9253275.

[4] Генрихов, И.Е. О поиске частых элементов в небинарных данных на основе технологии CUDA / И.Е. Генрихов, Е.В. Дюкова // Тезисы докладов межд. конф. ИОИ-13. Россия, Москва. – 2020. – С. 59-63.