Поиск минимальных нечастых и максимальных частых наборов в частично упорядоченных данных

Драгунов Никита Аркадьевич
Дюкова Елена Всеволодовна

МГУ имени М.В. Ломоносова
ВЦ имени Дородницына ФИЦ ИУ РАН

ММРО - 2019

Постановка задачи в случае бинарных данных

- Дано конечное множество V. Подмножество $X \subset V$ называется набором
- Дана база данных D, содержащая некоторые $X \subset V$ (транзакции)
- Частота набора $\nu(X)$ — доля транзакций в D, содержащих X
- Если $\nu(X) \geq s$, то набор X — s-частый, иначе — s-нечастый
- Если $\nu(X) < s$ и $\nexists X' \subset X : \nu(X') < s$, то набор X — минимальный нечастый
- Если $\nu(X) \geq s$ и $\nexists X' \supset X : \nu(X') \geq s$, то набор X — максимальный частый
- Требуется найти X_{max} — множество всех максимальных частых наборов и Y_{min} — множество всех минимальных нечастых наборов

Постановка задачи в случае частично упорядоченных данных

- K.M. Elbassioni (2014) рассмотрен случай небинарных данных, элементы которых принимают значения из частично упорядоченных множеств
- Пусть $\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$ – декартово произведение частично упорядоченных множеств
- На множестве $\mathcal{P}$ задано отношение частичного порядка $\preceq$ следующим образом:$x \preceq y$ в $\mathcal{P}$, где $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \Leftrightarrow x_1 \preceq y_1$ в $\mathcal{P}_1, \ldots, x_n \preceq y_n$ в $\mathcal{P}_n$
- Дана база данных $\mathcal{D}(\mathcal{P})$. Частота набора $\nu(X)$ — доля транзакций в $\mathcal{D}(\mathcal{P})$, следующих за X
- В данной постановке случай бинарных данных трактуется следующим образом $\mathcal{P}_i = \{0, 1\}, 0 \preceq 1, 0 \neq 1$

Актуальность поиска X_{max} и Y_{min}

- Важное приложение поиска частых и нечастых наборов — поиск ассоциативных правил в данных
- Поиск ассоциативных правил осуществляется в два этапа. Сначала решается задача поиска s-частых наборов. Затем из найденных на первом этапе s-частых наборов формируются ассоциативные правила, при этом фактически решается задача поиска t-нечастых наборов, где $t \neq s$ определяется специальным образом
- Множества минимальных нечастых наборов Y_{min} и максимальных частых наборов X_{max} позволяют компактно хранить информацию о всех нечастых и частых наборах
- В основном изучены вопросы поиска Y_{min} и X_{max} в бинарных данных

Задача дуализации над произведением частичных порядков

- Дано множество $R \subset \mathcal{P}$
- $R^- = \{x \in \mathcal{P} | \exists a \in R, x \preceq a\}$. Множество $I(R^-)$, состоящее из всех минимальных элементов множества $\mathcal{P} \setminus R^-$, называется минимальным независимым от R
- $R^+ = \{x \in \mathcal{P} | \exists a \in R, a \preceq x\}$. Множество $I(R^+)$, состоящее из всех максимальных элементов множества $\mathcal{P} \setminus R^+$, называется максимальным независимым от R
- Каждая из задач поиска $I(R^-)$ и $I(R^+)$ называется задачей дуализации над произведением частичных порядков

Совместное перечисление множеств X_{max} и Y_{min}

- Идея метода предложена K.M. Elbassioni и автором экспериментально не изучена
- Подход реализован в настоящей работе и получил название «совместное» перечисление множеств X_{max} и Y_{min}
- Строятся последовательности: $X_1 \subset X_2 \subset \cdots \subset X_{max}$, $Y_1 \subset Y_2 \subset \cdots \subset Y_{min}$
- Шаг 1. $X_1 = \{x\}$, $Y_1 = \{y\}$
- Шаг $i+1$, $i \geq 1$. Дуализация либо $I(X_i^-)$, либо $I(Y_i^+)$ и формирование X_{i+1}, Y_{i+1}

Последовательное перечисление множеств X_{max} и Y_{min}

- Метод достаточно очевиден и основан на свойстве двойственности: $I(X^{-}_{max}) = Y_{min}$, $I(Y^{+}_{min}) = X_{max}$
- Этап 1. Поиск всех максимальных частых наборов X_{max} стандартным алгоритмом (например, Apriori)
- Этап 2. Поиск всех минимальных нечастых наборов Y_{min} путем дуализации найденного на первом этапе множества X_{max}

Последовательно-совместное перечисление множеств X_{max} и Y_{min}

- Метод предложен в настоящей работе и является синтезом последовательного и совместного подходов
- Работает итеративно
- Строится одна из последовательностей: $X_1 \subset X_2 \subset \cdots \subset X_{max}$ или $Y_1 \subset Y_2 \subset \cdots \subset Y_{min}$
- Пусть для определенности строится $X_1 \subset X_2 \subset \cdots \subset X_{max}$
- Шаг 1. $X_0 = \varnothing$, $X_1 = \{x\}$
- Шаг $i+1$, $i \geq 1$. Дуализация $I\left((X_i \setminus X_{i-1})^-\right)$ и формирование X_{i+1}
- $Y_{min} = I(X_{max}^-)$ получается путем дуализации X_{max}

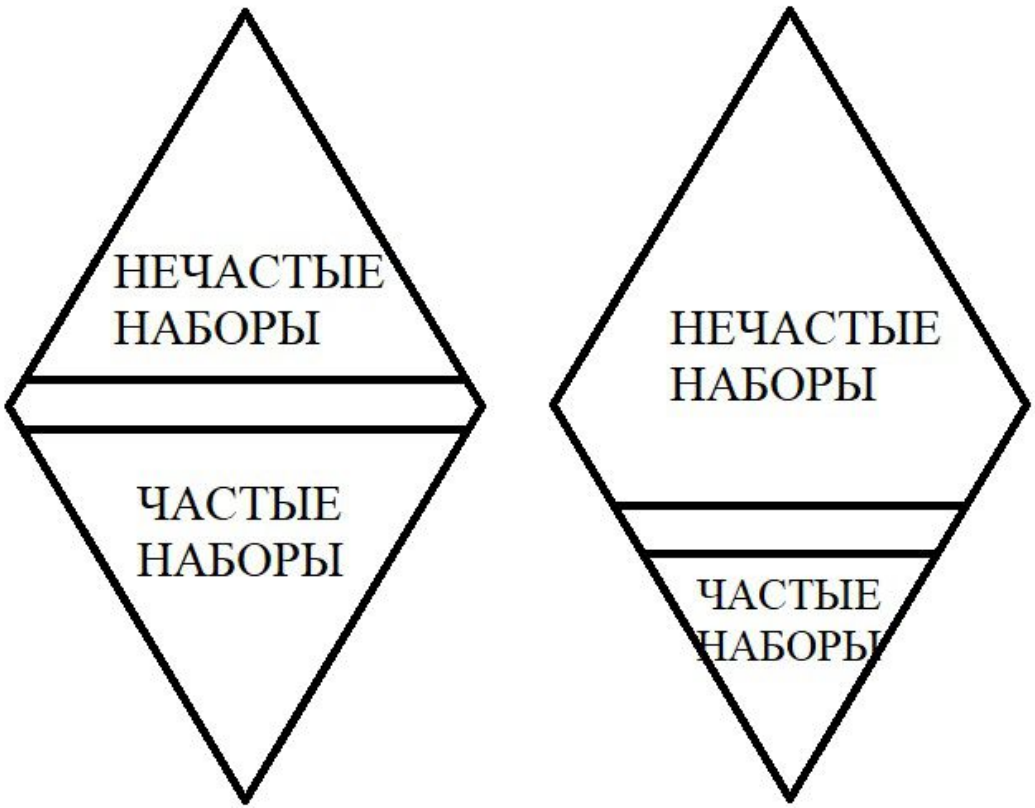
Обоснование корректности совместного и последовательно-совместного подхода

Утв. 1. *Если $X \subset X_{max}$, $y \in I(X^-)$ — нечастый набор, то y — минимальный нечастый набор.*

Утв. 2. *Пусть $X \subseteq X_{max}$, $Y \subseteq Y_{min}$. Тогда $I(X^-) = Y$ в том и только в том случае, когда $X = X_{max}$ и $Y = Y_{min}$.*

Экспериментальное сравнение (случай цепей)

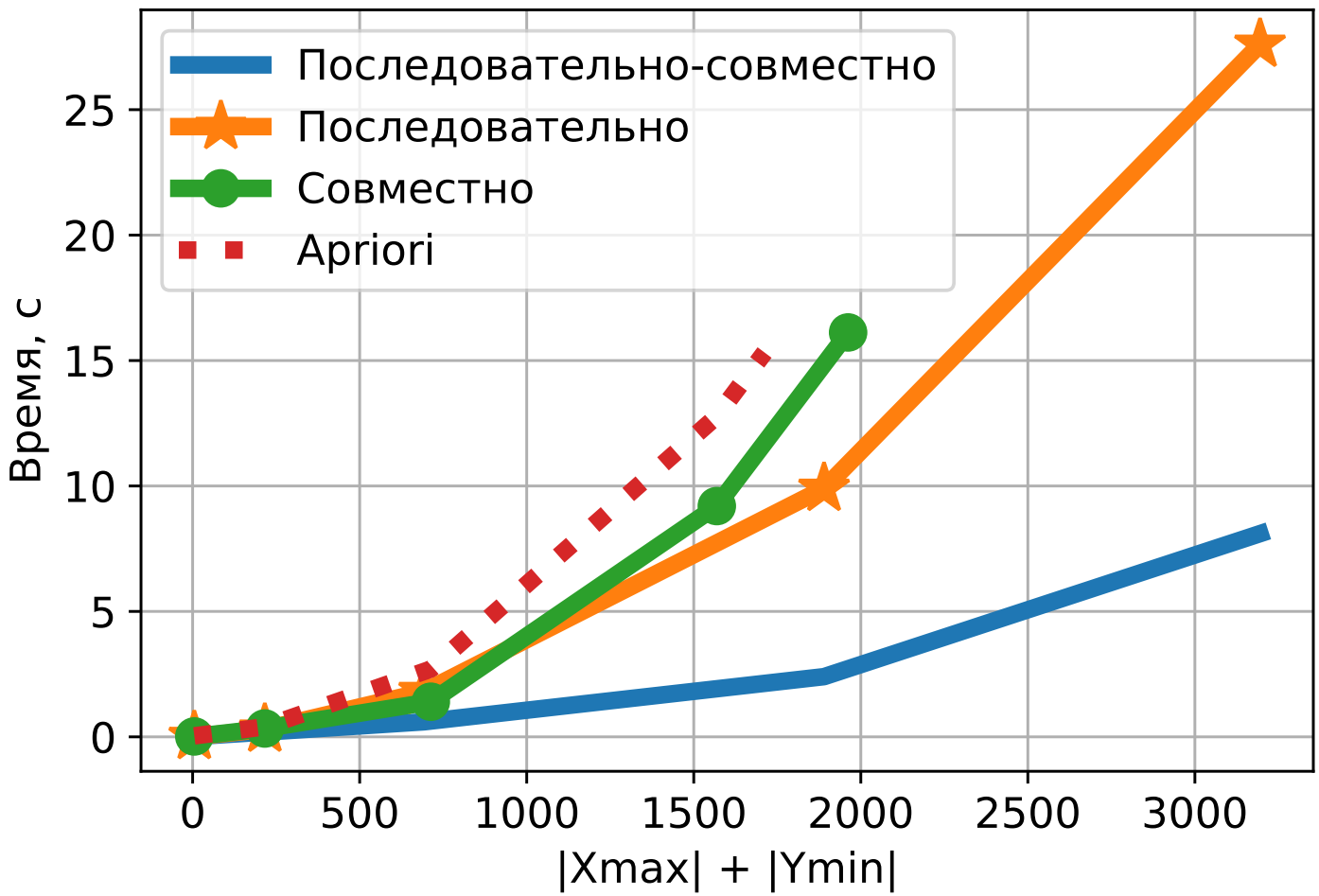
- Случай 1: мощность множества частых наборов примерно равна мощности множества нечастых наборов
- Случай 2: мощность множества частых наборов существенно меньше (больше) мощности множества нечастых наборов



Случай 1 Случай 2

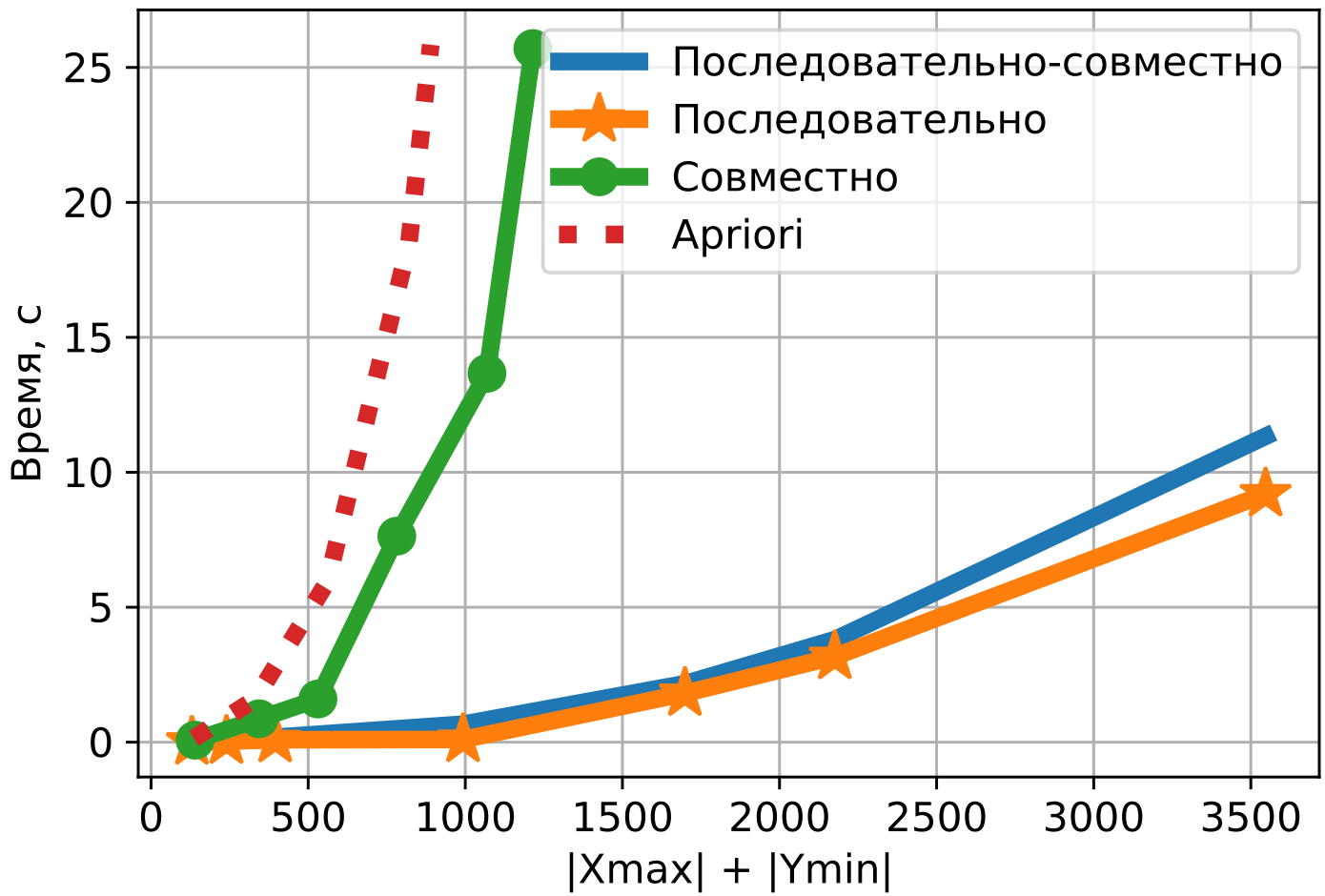
Экспериментальное сравнение (случай цепей)

Случай 1: мощность множества частых наборов примерно равна мощности множества нечастых наборов (наилучшие результаты показывает последовательно-совместный метод)



Экспериментальное сравнение (случай цепей)

Случай 2: мощность множества частых наборов существенно меньше мощности множества нечастых наборов (наилучшие результаты показывает последовательный метод)



Заключение

- Рассмотрены важнейшие задачи информационного поиска: поиск минимальных нечастых и максимальных частых наборов в данных, представленных в виде произведения частично упорядоченных множеств
- Описаны известные подходы к решению указанных задач и предложен оригинальный подход, основанный на решении задачи дуализации над произведением частичных порядков
- Проведено экспериментальное исследование рассмотренных подходов в случае произведения конечных цепей, и выявлены условия их эффективности