A Quantitative Study of Two Matrix Clustering Algorithms

Александр Слесарев
СПбГУ 2-й курс

Вячеслав Галактионов, Никита Бобров, Георгий Чернышев
СПбГУ, JetBrains Research

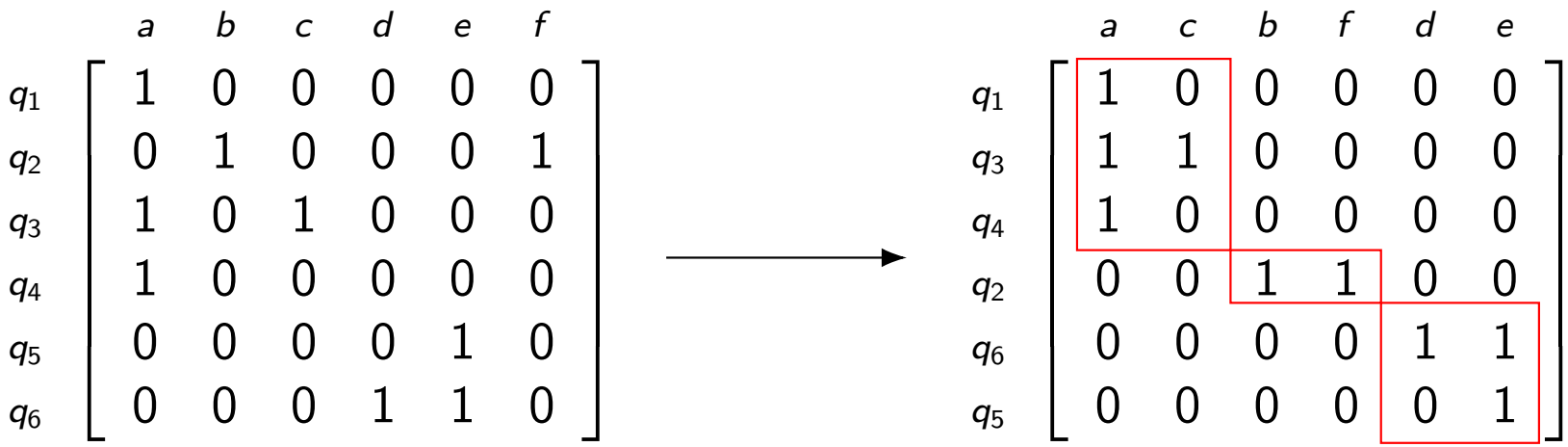
SEIM 2019
13 апреля 2019

Типы фрагментирования

1. Горизонтальное фрагментирование
2. Гибридное фрагментирование
3. Вертикальное фрагментирование
   - стоимостное
   - эвристическое
     - методы матричной кластеризации
     - графовый подход
     - data mining

Матрица запросов

q1: SELECT a FROM T WHERE a $>$ 10;
q2: SELECT b, f FROM T;
q3: SELECT a, c FROM T WHERE a = c;
q4: SELECT a FROM T WHERE a $<$ 10;
q5: SELECT e FROM T;
q6: SELECT d, e FROM T WHERE d + e $>$ 0;



Перечень статей

Описания алгоритмов:

- C. Cheng “Algorithms for vertical partitioning in database physical design”. **1993**
- C.-H. Cheng “A branch and bound clustering algorithm”. **1995**
- C.-H. Cheng and J. Motwani “An examination of cluster identification-based algorithms for vertical partitions”. **2009**
- C.-H. Cheng, K.-F. Wong, and K.-H. Woo “An improved branch-and-bound clustering approach for data partitioning”. **2011**

Предыдущие работы:

- V. Galaktionov, G. Chernishev, B. Novikov, and D. Grigoriev “Matrix clustering algorithms for vertical partitioning problem: an initial performance study”. **2016**
- V. Galaktionov “Parallelization of matrix clustering algorithms”. **2016**
- V. Galaktionov, G. Chernishev, K. Smirnov, B. Novikov, and D. A.Grigoriev “A study of several matrix-clustering vertical partitioning algorithms in a disk-based environment”. **2017**

Число возможных фрагментов

Число Белла – число всех неупорядоченных разбиений n-элементного множества

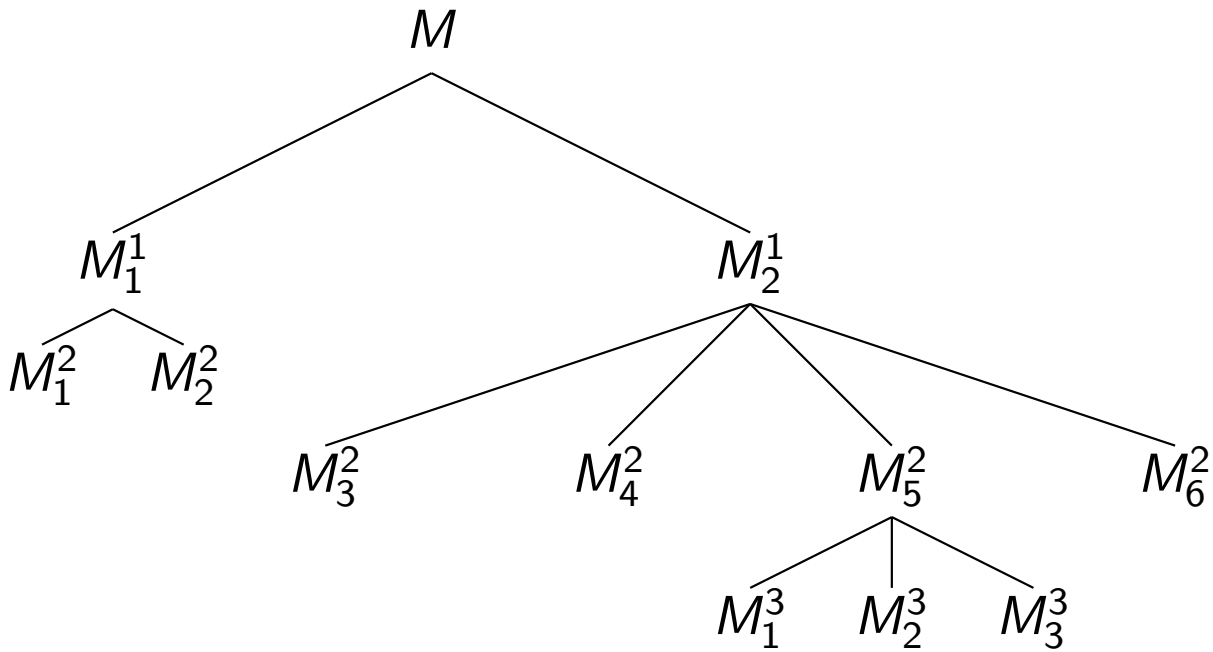
При больших n выполняется $B(n) \approx n^n$, например, $B(30) \approx 10^{23}$

n	$B(n)$
1	1
2	1
3	2
4	5
5	15
6	52
7	203
8	877
9	4140
10	21147
11	115975

Поиск решения – 1

M – матрица запросов

M_i^j – i-й узел на j-ом уровне фрагментирования



Поиск решения – 2

R - множество индексов транзакций M

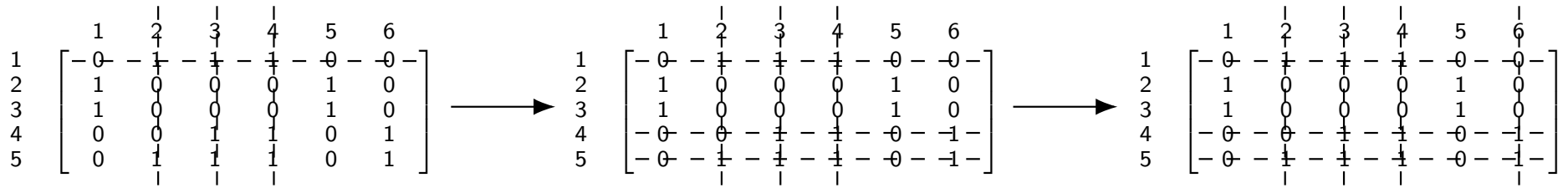
C - множество индексов атрибутов M

$$cohesion(M) = \frac{|\{a_{ij} = 1, i \in R \wedge j \in C\}|}{|R| * |C|}$$

Набор кластеров - решение, если каждый кластер S удовлетворяет условиям:

- $cohesion(S) < threshold$
- В S отсутствуют нулевые строки или столбцы

Cluster identification



$$
\begin{array}{c} \\ 1 \\ 4 \\ 5 \\ 2 \\ 3 \end{array}
\begin{array}{c}
\begin{array}{cccccc} 2 & 3 & 4 & 6 & 1 & 5 \end{array} \\
\left[\begin{array}{cccccc}
1 & 1 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1
\end{array}\right]
\end{array}
$$

Метод ветвей и границ 2009 – 1

нижняя граница: Z_L – число единиц, удаленных из матрицы транзакции в ходе ветвления

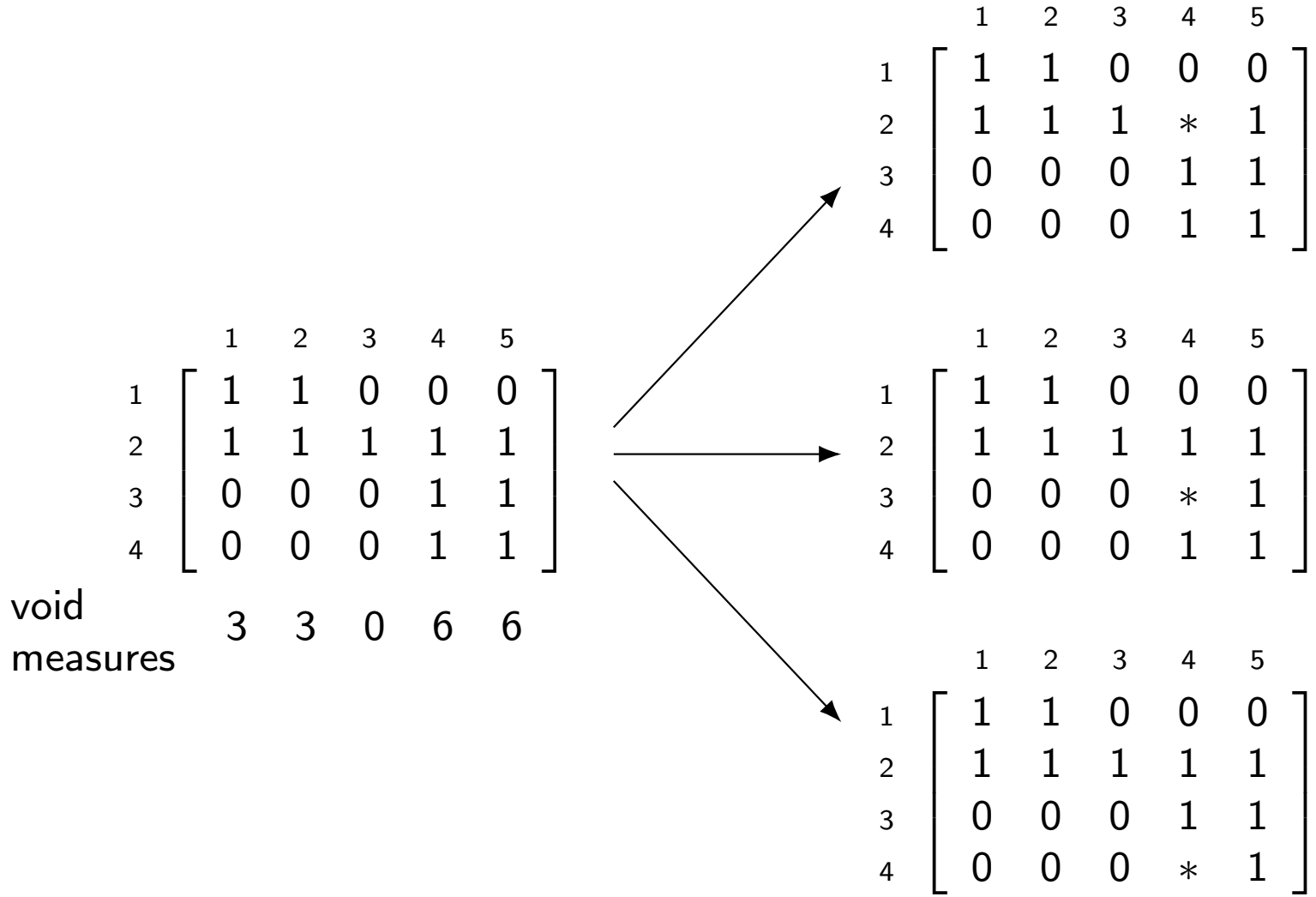
верхняя граница: Z_U – минимальный Z_L среди найденных решений

- (инициализация) дерево состоит из корневого узла с матрицей транзакций в нем, $Z_U = \infty$
- (ветвление) находим в матрице текущего узла подматрицу $S : cohesion(S) < threshold$
- (выбор решения) у каждого нового узла обновляем Z_L; убираем узел из рассмотрения, если:
  1. $Z_L \geq Z_U$
  2. какая-то из полученных подматриц содержит пустые строки/столбцы
- (завершение фрагментирования) в случае, если не осталось узлов для выбора решения, возвращаем текущее решение, иначе начинаем ветвление

Метод ветвей и границ 2009 – 2

- $Q_j = \{j' \neq j : a_{ij'} = 1 \wedge \exists i : a_{ij} = 1\}$
- $C_j = Q_j \cup \{j\}$
- $R_j = \{i : a_{ij} = 1 \wedge \exists j' \in C_j : a_{ij'} = 1\}$
- $void\_measure(attribute) = |\{a_{ij'} = 0,\ i \in R_j \wedge j' \in C_j\}|$

Метод ветвей и границ 2009 – 3



Метод ветвей и границ 2009 – 4

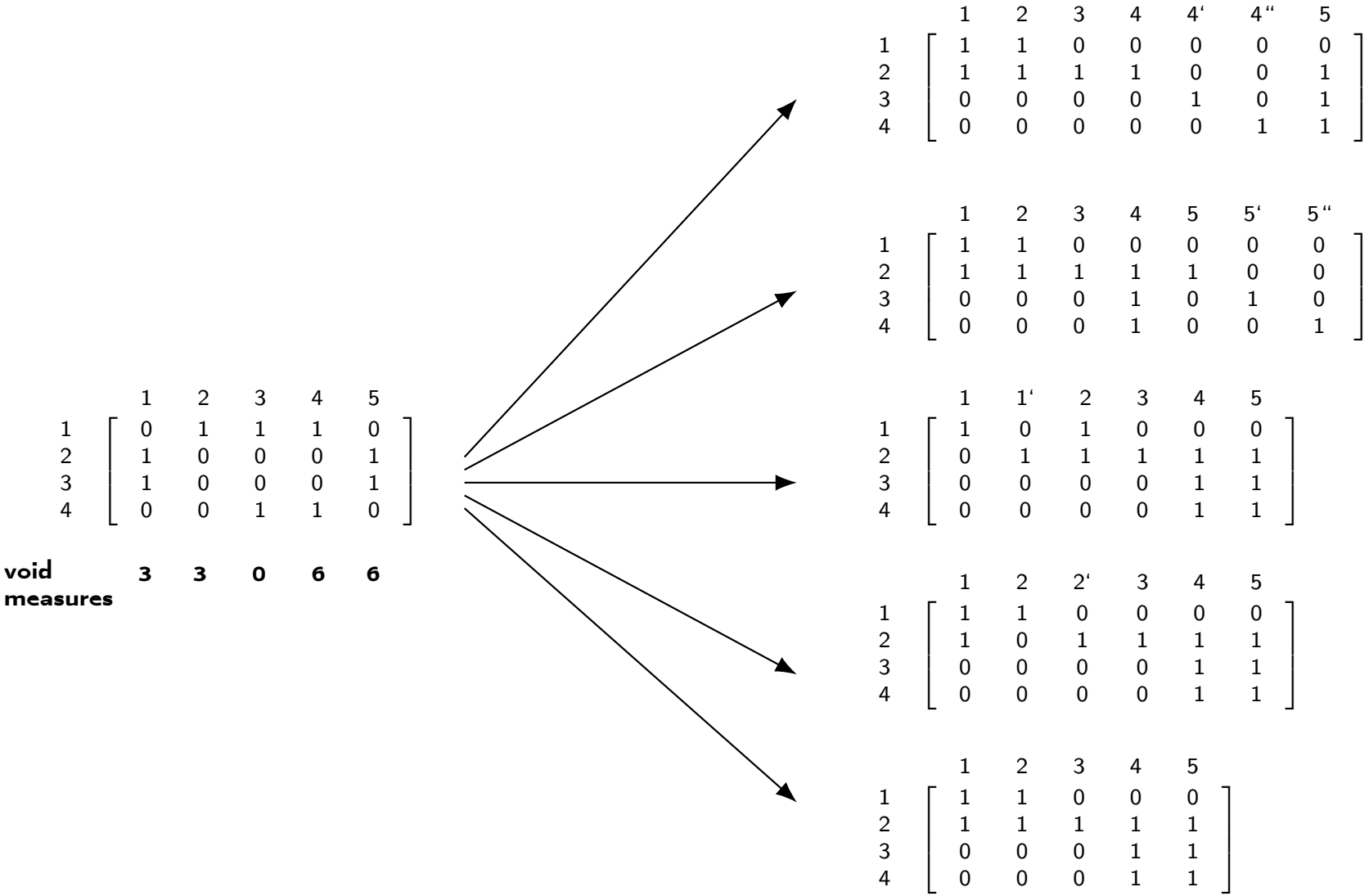
Распределение межкластерных элементов по матрицам в решении:

- (separate) составить отдельный фрагмент из межкластерных эелементов
- (nearest) добавить межкластерный столбец в тот фрагмент, где есть какая-то его часть
- (replicate) добавить в каждый фрагмент все необходимые межкластерные столбцы

Метод ветвей и границ 2011 – 1

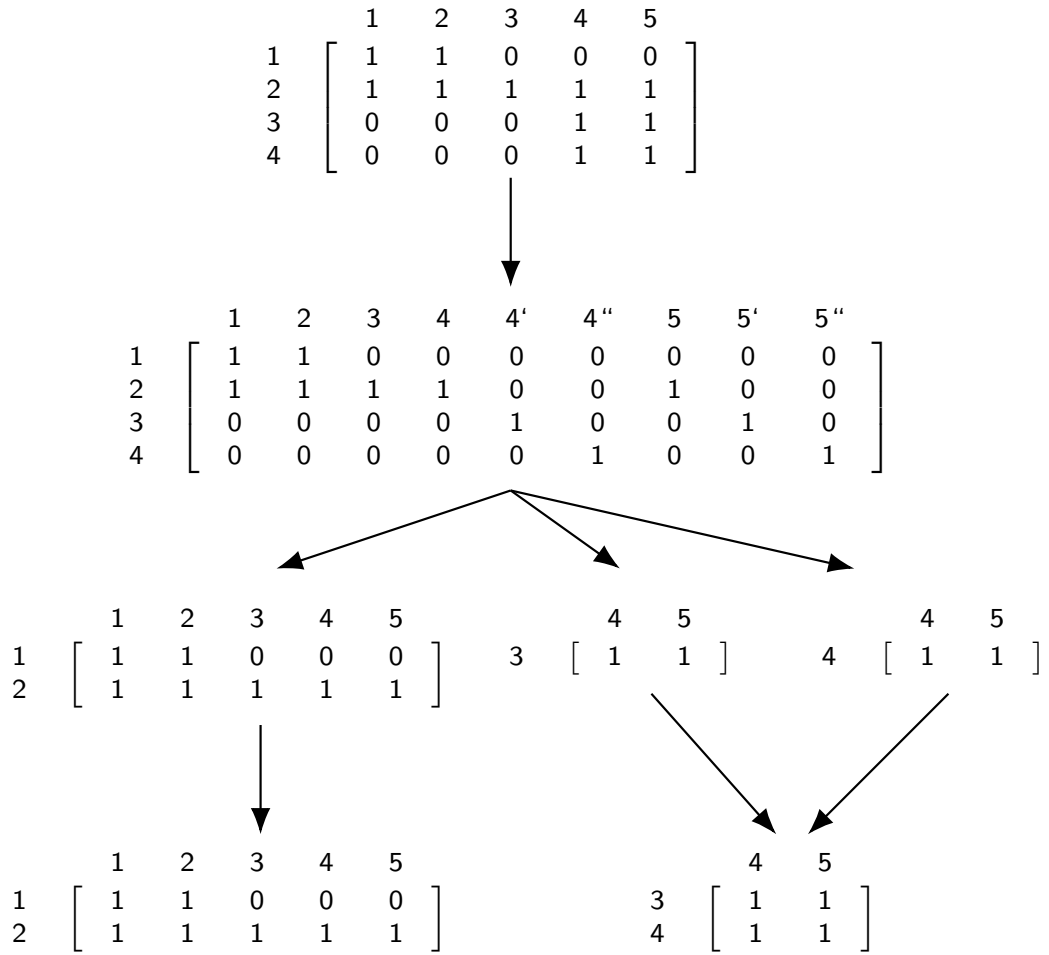
нижняя граница: Z_L – глубина узла в дереве

верхняя граница: Z_U – минимальный Z_L среди найденных решений



Метод ветвей и границ 2011 – 2

Постобработка решения



Реализация – 1

Аппаратные средства :

- Inspiron 15 7000 Gaming (0798)
- 8GiB RAM
- Intel(R) Core(TM) i5-7300HQ CPU @ 2.50GHz
- TOSHIBA 1TB MQ02ABD1

Программное обеспечение :

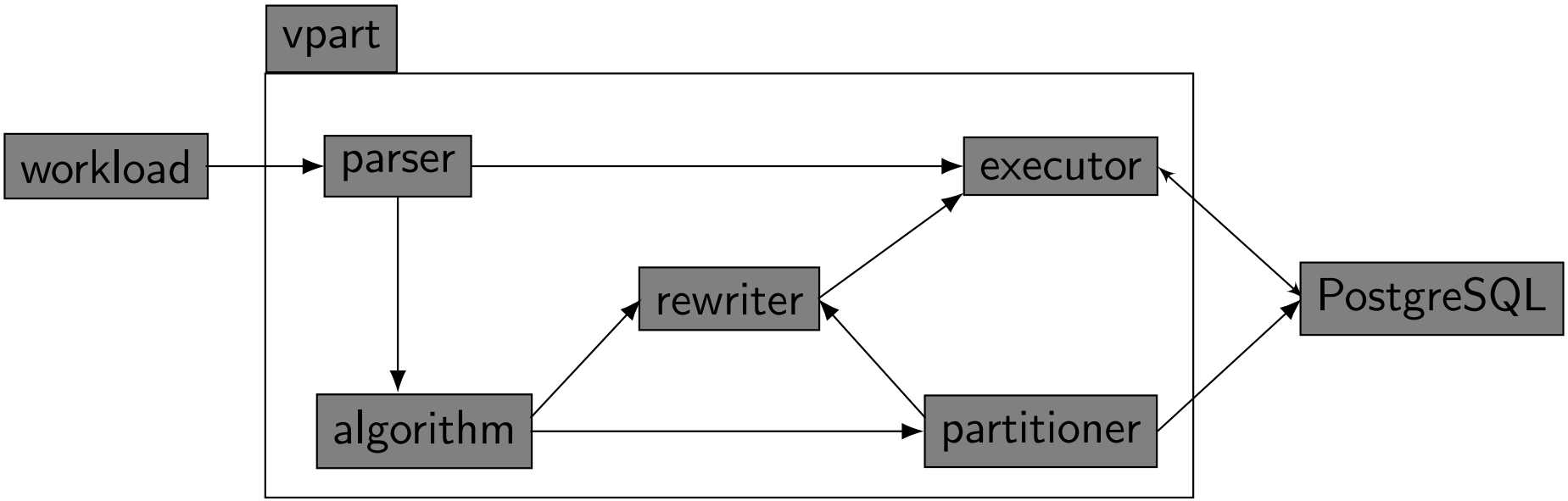
- Ubuntu 18.10
- PostgreSQL 11.1
- gcc 8.2.0

Датасет :

- TPC-H
- SDSS Star table

Реализация – 2

Структура программы



Критерии оценки алгоритмов:

- скорость кластеризации
- скорость выполнения запросов после применения алгоритма
- затраты памяти на хранение кластеров

Проверка чистоты экспериментов

- сбросить системный кеш, записав 3 в `/proc/sys/vm/drop_caches`
- запретить параллельное выполнение запросов: *setmax_parallel_workers_per_gather to* 0;

Эксперименты – 1

Эксперименты – 2

Эксперименты – 3

Эксперименты – 4

Эксперименты – 5