УДК 001.891.573

Е. Ю. Алехова

ИССЛЕДОВАНИЕ КЛАСТЕРИЗУЕМОСТИ СТРОК БОЛЬШИХ РАЗРЕЖЕННЫХ МАТРИЦ НАД GF(2)

Рассмотрены алгоритмы кластеризации больших объемов данных, которые, как правило, или очень ресурсоемки, или имеют эвристические этапы, или и то, и другое одновременно. Перед применением этих алгоритмов следует оценить по каким-либо известным характеристикам содержание во взятом объеме данных, интересных для практики кластеров.

E-mail: 0allena0@gmail.com

Ключевые слова: *решето числового поля, параллельное матрично-векторное умножение, разбиение разреженных матриц, кластеризация*

В рамках данной работы автор проанализировал специфический вид данных – строки матриц, образующиеся при факторизации методами решета числового поля [1, 2]. Исследуемые матрицы имеют следующие отличительные характеристики: разреженные матрицы большого размера; каждая строка в них является случайным сильно разреженным двоичным вектором, причем веса векторов варьируются достаточно слабо; размер матрицы может достигать 109 столбцов/строк, при этом средний вес строк для такой матрицы относительно невелик, порядка 140 ненулей. Одинаковые строки в матрице отсекаются специальной процедурой предварительной фильтрации [3], поэтому все строки в матрице различны. Метрика близости для строк следует из задачи минимизации коммуникаций при параллельном матрично-векторном умножении [1, 4] и является количеством совпадающих столбцов, содержащих ненулевые элементы. В качестве целевых вычислительных комплексов представляют интерес комплексы с 512–32 768 узлами. Исходя из этого была сформулирована следующая постановка задачи: какова вероятность p, что в матрице размера $l \times n$, каждая строка которой является случайным двоичным вектором с весом m, найдется не менее k строк, таких, что все их единицы содержатся в s столбцах.

В работе были рассмотрены следующие диапазоны указанных выше параметров:

$$n = 2^{30}; \tag{1}$$

$$2^7 < m < 2^8; \tag{2}$$

$$2^{15} < k < 2^{21}; \tag{3}$$

$$s \le mk, \ s \le 2^{29}; \tag{4}$$

$$k \le l \le n. \tag{5}$$

Построение оценки. Оценим вероятность p. Общее число допустимых в качестве строк векторов составляет $\binom{n}{m}$. Число строк, ненули которых укладываются в заданные s стобцов – $\binom{s}{m}$; число вариантов выбора из них k строк – $\binom{\binom{n}{m}}{k}$. Рассмотрим выбор остальных $l - k$ строк. Если мы считаем количество матриц, содержащих ровно k удовлетворяющих условию строк, то в $l - k$ строк не должна попасть ни одна, удовлетворяющая условию, и их надо выбирать из $\binom{n}{m} - \binom{s}{m}$ возможных строк. Таким образом, для выбора $l - k$ строк существует $\binom{\binom{n}{m} - \binom{s}{m}}{l-k}$ вариантов. Отметим, что количество матриц, в которых условию удовлетворяют ровно k строк, заведомо меньше количества матриц, в которых условию удовлетворяют не менее k строк.

При этом, количество матриц, содержащих не менее k строк, удовлетворяющих условию, можно оценить сверху следующим образом: если при выборе $l - k$ строк не отсекать все строки, удовлетворяющие условию, а отсечь лишь k уже выбранных строк, то все матрицы, содержащие не менее k нужных строк будут подсчитаны, причем многократно. Общее число матриц без ограничений на расположение ненулевых элементов составляет $\binom{\binom{n}{m}}{l}$.

Исходя из этих соображений, оценим сверху и снизу вероятность наличия среди l строк не менее k строк, все единицы которых содержатся в заданных s столбцах:

$$\frac{\left(\binom{n}{m}-\binom{s}{m}\right)\binom{s}{m}}{\binom{n}{m}} < p < \frac{\left(\binom{n}{m}-k\right)\binom{s}{m}}{\binom{n}{m}}. \quad (6)$$

Нижняя и верхняя оценки отличаются одним сомножителем, который равен для нижней и для верхней оценок. Оценим отношение

$$r = \frac{\binom{\binom{n}{m}-\binom{s}{m}}{l-k}}{\binom{\binom{n}{m}-k}{l-k}}. \quad (7)$$

Для этого представим числитель и знаменатель в следующем виде:

$$\binom{\binom{n}{m}-k}{l-k} = \frac{\prod_{i=k}^{l-1}\left[\binom{n}{m}-i\right]}{(l-k)!} = \frac{\binom{n}{m}^{l-k}}{(l-k)!}\prod_{i=k}^{l-1}\left[1-\frac{i}{\binom{n}{m}}\right]; \quad (8)$$

$$\binom{\binom{n}{m}-\binom{s}{m}}{l-k} = \frac{\prod_{i=k}^{l-1}\left[\binom{n}{m}-\binom{s}{m}-i\right]}{(l-k)!} = \frac{\binom{n}{m}^{l-k}}{(l-k)!}\prod_{i=k}^{l-1}\left[1-\frac{i}{\binom{n}{m}}-\frac{\binom{s}{m}}{\binom{n}{m}}\right]. \quad (9)$$

Выпишем нижние оценки для выражений (8) и (9):

$$\prod_{i=k}^{l-1}\left[1-\frac{i}{\binom{n}{m}}\right] > 1-\frac{(l-k)(l-k-1)}{2\binom{n}{m}}; \quad (10)$$

$$\prod_{i=k}^{l-1}\left[1-\frac{i}{\binom{n}{m}}-\frac{\binom{s}{m}}{\binom{n}{m}}\right]>1-\frac{(l-k)(l-k-1)}{2\binom{n}{m}}-\frac{(l-k)\binom{s}{m}}{2\binom{n}{m}}. \qquad (11)$$

Оценим отношение

$$\frac{\binom{s}{m}}{\binom{n}{m}}<\left[\frac{s}{n}\right]^m<(1/2)^{127}=2^{-127}. \qquad (12)$$

Учитывая, что $0{,}5(l-k-1)<\binom{s}{m}$, получаем, что и выражение (9), и (8) отличаются от (1) не более, чем на $2(l-k)\cdot 2^{-127}<2^{-96}$.

Таким образом, при интересных на практике значениях параметров матриц, вероятность наличия среди l строк не менее k строк, все единицы которых содержатся в заданных s столбцах, можно вычислить по следующей формуле:

$$p\approx\frac{\binom{\binom{n}{m}-k}{l-k}\binom{\binom{s}{m}}{k}}{\binom{\binom{n}{m}}{l}}. \qquad (13)$$

Отметим,что верность этого утверждения зависит от соотношения размера матрицы n, веса строки m и количества вычислительных узлов Np, $kN_p = n$. Оценка сложности заведомо применима, если $2m <= N_p$ и $n >> m$.

Представляет интерес анализ функции (13), однако не всегда представляется возможным, вычислить ее в нужных точках. Для упрощения вычисления функции были выполнены некоторые преобразования, описываемые далее.

Представим знаменатель выражения (13) в следующем виде:

$$\binom{\binom{n}{m}}{l}=\frac{\prod_{i=0}^{l-1}\left[\binom{n}{m}-i\right]}{l!} \qquad (14)$$

и выразим отношение первого сомножителя числителя к знаменателю, воспользовавшись формулой (8):

$$\frac{\prod_{i=k}^{l-1}\left[\binom{n}{m}-i\right]}{\frac{l!}{\prod_{i=0}^{l-1}\left[\binom{n}{m}-i\right]}{l!}}=\frac{l!}{(l-k)!\prod_{i=0}^{k-1}\left[\binom{n}{m}-i\right]}=\frac{l!}{(l-k)!\binom{n}{m}^k\prod_{i=1}^{k-1}\left[1-\frac{i}{\binom{n}{m}}\right]}. \quad (15)$$

Оценим

$$\prod_{i=1}^{k-1}\left[1-\frac{i}{\binom{n}{m}}\right]\approx 1-\frac{\sum_{i-1}^{k-1} i}{\binom{n}{m}}=1-\frac{k(k-1)}{2\binom{n}{m}}\approx 1.$$

Используя эту оценку, можно упростить выражение (15):

$$\frac{l!}{(l-k)!\binom{n}{m}^k\prod_{i=1}^{k-1}\left[1-\frac{i}{\binom{n}{m}}\right]}\approx\frac{l!}{(l-k)!\binom{n}{m}^k}. \quad (16)$$

Рассмотрим оставшийся сомножитель из числителя выражения (13):

$$\binom{\binom{s}{m}}{k}=\frac{\prod_{i=0}^{k-1}\left[\binom{s}{m}-i\right]}{k!}=\frac{\binom{s}{m}^k}{k!}\prod_{i=k}^{k-1}\left[1-\frac{i}{\binom{n}{m}}\right]. \quad (17)$$

Поскольку разница между m и s менее значительна, чем между m и n, с выражением (17) нельзя поступить так же, как с (15).

Подставим в выражение (13) полученные соотношения:

$$p\approx\frac{\binom{\binom{n}{m}-k}{l-k}\binom{\binom{s}{m}}{k}}{\binom{\binom{n}{m}}{l}}=\frac{l!}{(l-k)!\binom{n}{m}^k}\frac{\binom{s}{m}^k}{k!}\prod_{i=k}^{k-1}\left[1-\frac{i}{\binom{n}{m}}\right]=$$

$$= \binom{l}{k} \left[\frac{\binom{s}{m}}{\binom{n}{m}} \right]^k \prod_{i=k}^{k-1} \left[1 - \frac{i}{\binom{n}{m}} \right]. \qquad (18)$$

Представим биномиальные коэффициенты в более удобном для преобразований виде:

$$\binom{n}{m} = \frac{n!}{m!(n-m)!} = \frac{\prod_{i=0}^{m-1}[n-i]}{m!} = \frac{n^m}{m!} \prod_{i=1}^{m-1} \left[1 - \frac{i}{n} \right]; \qquad (19)$$

$$\binom{l}{k} = \frac{l^k}{k!} \prod_{i=1}^{k-1} \left[1 - \frac{i}{l} \right]; \qquad (20)$$

$$\binom{s}{m} = \frac{s^m}{m!} \prod_{i=1}^{m-1} \left[1 - \frac{i}{s} \right]. \qquad (21)$$

Таким образом, получим следующую оценку для выражения (13):

$$p \approx \frac{l^k}{k!} \prod_{i=1}^{k-1} \left[1 - \frac{i}{l} \right] \left[\frac{s^m \prod_{i=1}^{m-1} \left[1 - \frac{i}{s} \right]}{n^m \prod_{i=1}^{m-1} \left[1 - \frac{i}{n} \right]} \right]^k \prod_{i=1}^{k-1} \left[1 - \frac{i}{\binom{s}{m}} \right]. \qquad (22)$$

Применим к произведениям из этого выражения преобразование

$$\prod_{i=1}^{z-1} \left[1 - \frac{i}{N} \right] \approx \exp \frac{-z(z-1)}{2N}$$

и получим

$$\prod_{i=1}^{k-1} \left[1 - \frac{i}{l} \right] \approx \exp \frac{-k(k-1)}{2l};$$

$$\prod_{i=1}^{m-1} \left[1 - \frac{i}{s} \right] \approx \exp \frac{-m(m-1)}{2s};$$

$$\prod_{i=1}^{m-1} \left[1 - \frac{i}{n} \right] \approx \exp \frac{-m(m-1)}{2n};$$

$$\prod_{i=1}^{k-1}\left[1-\frac{i}{\binom{s}{m}}\right]\approx\exp\frac{-k(k-1)}{2\binom{s}{m}}.$$

Таким образом получена оценка вероятности того, что в случайной матрице с заданными характеристиками n, l, m найдется не менее k строк, все ненули которых лежат в заданных s столбцах:

$$p\approx\frac{l^k s^{mk}}{k!n^{mk}}\exp\left[-\frac{k(k-1)}{2l}-\frac{k(k-1)}{2\binom{s}{m}}-k\frac{m(m-1)}{2s}+k\frac{m(m-1)}{2n}\right]=$$
$$=\frac{l^k s^{mk}}{k!n^{mk}}\exp\left[-\frac{k(k-1)}{2l}\left(\frac{1}{l}+\frac{1}{\binom{s}{m}}\right)-k\frac{m(m-1)}{2s}\left(\frac{1}{s}-\frac{1}{n}\right)\right]. \quad (23)$$

Используя результат оценки (23), выразим оценку интересующей нас вероятности для s произвольных, а не заданных столбцов. Вероятность того, что не найдется ни одного набора из s столбцов, удовлетворяющих нужным требованиям, равна $(1-p)^{\binom{n}{s}}$. Таким образом, вероятность того, что найдется хотя бы один такой набор

$$\tilde{p}=1-(1-p)^{\binom{n}{s}}=1-\exp\left[\binom{n}{s}\ln(1-p)\right]\approx 1-\exp\left[-p\cdot\binom{n}{s}\right]=1-\exp(-R). \quad (24)$$

Выразив удобную для вычислений оценку для R, используя элементарные преобразования и следующие оценки:

$$\binom{n}{s}\approx\frac{n^s\exp\frac{-s(s-1)}{2n}}{s!};\quad k!=\frac{1}{e}\left[\frac{k}{e}\right]^k;\qquad s!=\frac{1}{e}\left[\frac{s}{e}\right]^s,$$

получим:

$$R \approx \frac{n^s l^k s^{mk}}{k^k s^s n^{mk}} \exp \times$$

$$\times \left[k + s + 2 - \frac{k(k-1)}{2} \left(\frac{1}{l} + \frac{1}{\binom{s}{m}} \right) - k \frac{m(m-1)}{2} \left(\frac{1}{s} - \frac{1}{n} \right) - \frac{s(s-1)}{2n} \right] =$$

$$= \left[\frac{l}{k} \right]^k \left[\frac{n}{s} \right]^{s-mk} \exp \times \tag{25}$$

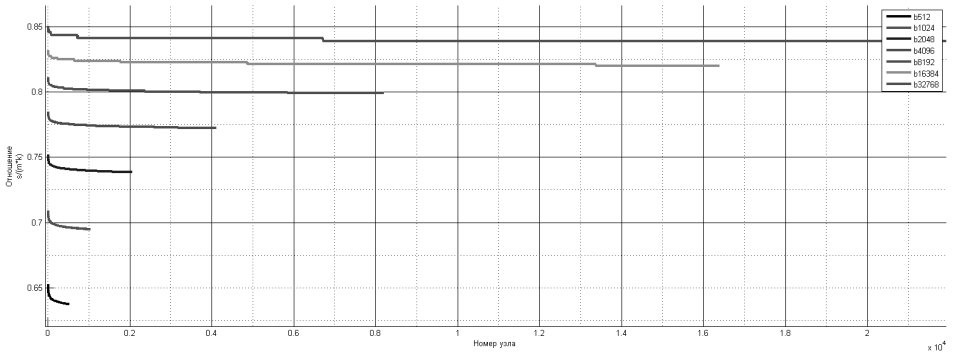
$$\times \left[2 + s \left(1 - \frac{s-1}{2n} \right) + k \left(1 - \frac{m(m-1)}{2} \left(\frac{1}{s} - \frac{1}{n} \right) - \frac{(k-1)}{2} \left(\frac{1}{l} + \frac{1}{\binom{s}{m}} \right) \right) \right].$$

Прологорифмировав, получим удобный для расчетов вид:

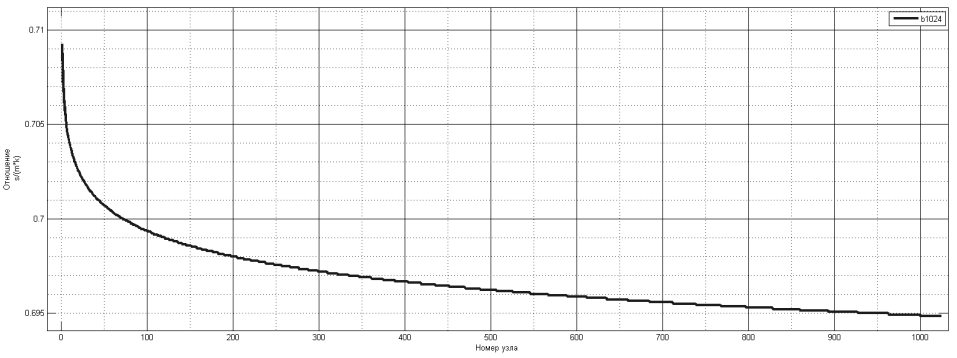
$$\ln R = k \ln \frac{l}{k} + (s - mk) \ln \frac{n}{s} + 2 + s \left(1 - \frac{s-1}{2n} \right) +$$

$$+ k \left(1 - \frac{m(m-1)}{2} \left(\frac{1}{s} - \frac{1}{n} \right) - \frac{(k-1)}{2} \left(\frac{1}{l} + \frac{1}{\binom{s}{m}} \right) \right). \tag{26}$$

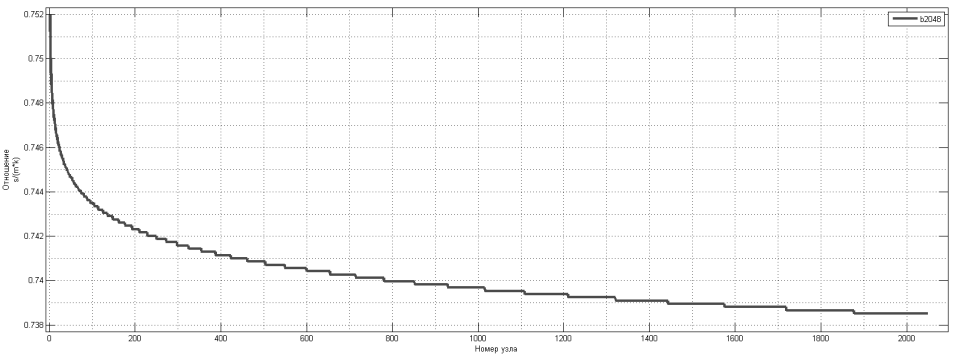
Формула (26) использовалась для расчета матрицы размером 109 с весом строки $m = 140$ и распределеним ее строк на 512, 1 024, 2 048, 4 096, 8 192, 16 384 и 32 768 узлах. Для каждого узла было посчитано, при каком количестве s столбцов с ненулевыми элементами вероятность нахождения в матрице нужного количества подходящих строк близка к 1. На рисунке представлены сводные графики для всех узлов, и для узла 1024 и для узла 2048 в увеличенном масштабе.



а



б



в

Зависимость отношения s_i/mk от номера узла:

а – для всех узлов; *б* – для узла 1024; *в* – для узла 2048

Таким образом рассмотрено итеративное параллельное матрично-векторное умножение, в котором на каждом узле обрабатывается одинаковое количество строк матрицы – k. Всего узлов N_p. На каждой итерации каждый узел вырабатывает k разрядов результата, используя s_i разрядов из результата предыдущей итерации. В худшем случае все s_i разрядов он должен получить от других узлов, в лучшем случае от других узлов нужно получить $s_i \times k$ разрядов.

Для общего объема коммуникаций Com выполняются следующие неравенства (под объемом коммуникаций подразумевается общее количество передаваемых между узлами разрядов):

$$\sum_{i=1}^{N_p}(s_i - k) \le \mathrm{Com} \le \sum_{i=1}^{N_p}(s_i). \qquad (27)$$

В худшем случае $s_i = mk, \ \forall i$ и оценки принимают вид

$$N_p(m-1)k \le \mathrm{Com} \le N_p mk. \qquad (28)$$

Приведенные в статье расчеты позволяют найти для каждого узла такое значение s_i, что вероятность того, что в матрице заданного размера найдется k подходящих для этого узла строк близка к 1. Значение s отличается для разных узлов, так как предполагается, что первый узел выбирает из всех строк матрицы, второй из всех, кроме выбранных первым, и т. д. На рисунках отображено изменение отношения s_i к mk с номером узла для разного общего количества узлов. Фактически, это соотношение указывает какую часть общего объема коммуникаций можно сэкономить за счет «оптимального» распределения строк по узлам по сравнению с наихудшим вариантом.

Расчеты показывают, что потенциальная экономия составляет от 0,4 до 0,15 от объема в наихудшем случае, и уменьшается с ростом количества узлов. Однако отметим, что из тех же расчетов следует, что разница в значениях s_i для узла, для которого моделируется ситуация выбора из всех строк матрицы, и для узла, у которого выбора, фактически, нет, – невелика и также уменьшается с ростом количества узлов.

СПИСОК ЛИТЕРАТУРЫ

1. Rob H. Bisseling. Multiplication by using sparse matrix partitioning. Society, 2009. 31(4). – P. 3128–3154.
2. Joppe W. Bos, Pierrick Gaudry, Alexander Kruppa, Peter L. Montgomery, Dag Arne Osvik, Herman Riele, Andrey Timofeev, and Paul Zimmermann. Factorization of a 768-bit RSA modulus. 2010. – P. 1–22.
3. Cavallar S. Strategies in filtering in the number field sieve. Proceedings of the 4th International Symposium on Algorithmic Number Theory, 2000. – P. 209–232.
4. Pinar A., Michael T. Heath. Improving performance of sparse matrix-vector multiplication // Proceedings of the 1999 ACM/IEEE conference on Supercomputing – Supercomputing ’99, 1999. – P. 30.

Статья поступила в редакцию 14.05.2012