Математика и математическое моделирование. 2021. № 1. С. 27–42.

DOI: **10.24108/mathm.0121.0000244**

Математика и Математическое моделирование

Сетевое научное издание

http://mathmelpub.ru ISSN 2412-5911

УДК 004.6+57.087.1

Анализ эффективности методов полиномиальной степени сложности при декомпозиции OLAP-кубов многомерных данных

Ахрем А. А.[1], Носов А. П.[1,*], Рахманкулов В. З.[1]

[1]Федеральный исследовательский центр «Информатика и управление» РАН, Москва, Россия

[*] nosov@isa.ru

В работе исследуются проблемы редукции (декомпозиции) моделей многомерных данных в виде гиперкубовых OLAP-структур. Рассматривается случай, когда структура данных определяется решеткой, разбивающей гиперкуб на нечетное количество подкубов, и декомпозиция гиперкуба осуществляется на этом множестве подкубовых структур. Установлена точная верхняя граница увеличения вычислительной производительности методов анализа OLAP-данных на подкубах, определяющая эффективность декомпозиционного подхода по сравнению с анализом OLAP-данных на полном нередуцированном гиперкубе. Проведено сравнение эффективности декомпозиции гиперкуба на два подкуба на множествах, состоящих из четного и нечетного числа подкубовых структур и показано, что при большом дроблении данных для методов полиномиальной степени сложности эффективность декомпозиции практически не зависит от этого фактора и растет с ростом степени сложности применяемых методов.

Ключевые слова: OLAP-система, декомпозиция, вычислительная производительность, гиперкуб OLAP-данных, полиномиальная сложность

Представлена в редакцию: 21.12.2020.

Введение

Обработка и анализ сверхбольших массивов информации гиперкубов аналитических OLAP-систем относятся к классу BigData и требуют больших затрат машинного времени, с увеличением объемов данных падает производительность вычислений кубовых структур, и эта проблема с распространением цифровых технологий практически на все сферы человеческой деятельности становится как никогда актуальной. Напомним, что указанные системы предназначены для анализа и обобщения детальных данных, накапливаемых в базах и хранилищах данных, при анализе бизнес-процессов и поддержке принятия жизненно важных решений в промышленности, финансовой сфере, торговле, медицине и других областях, использующих информационные технологии [1, 2, 3, 4].

Один из подходов, позволяющий решать проблему снижения производительности вычислений, связан с декомпозицией больших кубов данных на подкубы с меньшими объемами. Задачи декомпозиции структур многомерных данных с целью ускорения вычислений и увеличения производительности рассматриваются во многих работах отечественых и зарубежных авторов [5, 6, 7, 8, 9]. При решении задачи декомпозиции необходимо исследовать влияние выбранного способа декомпозиции на вычислительную сложность задач. OLAP обработка данных не допускает изменения размерности пространства, однако, поскольку вычислительная сложность OLAP-данных определяется не размерностью гиперкуба, а размером решетки подкубов, которая задает на нем структуру данных, становится возможной декомпозиция на подкубовых структрах.

В работах авторов [10, 11, 12] по математическим методам анализа многомерных данных аналитических OLAP-систем рассматриваются задачи декомпозиции куба на меньшие по размерности подкубы для снижения времени вычисления полной решётки при динамическом изменении данных в кубе, когда агрегирование критериев уже определено решёткой куба. Для методов анализа многомерных данных были введены классы вычислительной сложности, определяющие вычислительную производительность этих методов при обработке данных гиперкуба, и исследовались способы редукции моделей многомерных данных, способствующих уменьшению вычислительной сложности решения задач с большими и сверхбольшими исходными данными.

В работах [10, 11] исследуются математические методы декомпозиции (редукции) больших гиперкубов многомерных данных аналитических OLAP-систем на подкубовые компоненты. Показана возможность уменьшения вычислительной сложности методов решения задач при декомпозиции данных по сравнению с применением этих методов к анализу больших массивов информации, накапливаемых непосредственно в гиперкубах многомерных OLAP-данных, и установлены критерии уменьшения или увеличения вычислительной производительности при применении методов на подкубовых компонентах (редукционные методы) по сравнению с применением этих методов на гиперкубе (нередукционные или традиционные методы) в зависимости от классов той или иной степеней сложности рассматриваемых методов. Заметим, что полученные в этих работах критерии уменьшения или увеличения вычислительной сложности методов решения задач анализа OLAP-гиперкубов многомерных данных устанавливают лишь характер изменения этих величин при декомпозиции, но не дают их количественной оценки.

В работе [12] было проведено количественное сравнение эффективности применения редукционных методов анализа OLAP-данных по сравнению с традиционными методами в классе полиномиальной степени сложности для случая, когда решетка исходного гиперкуба данных содержит четное число подкубов.

В настоящей статье, являющейся продолжением работы [12], получена точная количественная оценка уменьшения вычислительной сложности редукционных методов анализа

OLAP-кубов по сравнению с нередукционными методами в ситуации, когда данные методы имеют полиномиальную степень сложности, а решетка исходного гиперкуба данных состоит из нечетного числа подкубов.

1. Критерии уменьшения вычислительной сложности редукционных методов решения задач анализа OLAP-гиперкубов данных

В настоящем разделе статьи будем считать, что задан гиперкуб K_nH_m с решеткой L, состоящей из n подкубов. Допустим также, что на этой решетке L решаются задачи анализа многомерных OLAP-данных из некоторого непустого множества $C(K_nH_m)$. Предположим, что структура данных из $C(K_nH_m)$ допускают декомпозицию (редукцию) гиперкуба K_nH_m на k ($2 \le k \le n$) непересекающихся подкубовых структур $L_1, \ldots, L_k$, в каждую из которых входят $n_1, \ldots, n_k$ подкубов исходной решетки L таким образом, что $n_1 + \ldots + n_k = n$.

Пусть вычислительная сложность f метода решения задачи анализа данных из множества $C(K_nH_m)$ принадлежит классу полиномиальной вычислительной степени сложности [12]:

– $F^p(n) \triangleq \{f(p,n) = n^p\colon p \in \mathbb{N}\}$, где $\mathbb{N}$ — множество натуральных чисел.

Подкласс класса $F^p(n)$ при $p \ge 2$ далее будем обозначать $F_2^p(n)$.

Напомним [12], что классы вычислительной степени сложности определяют значения вычислительной сложности $f = f(\cdot, n)$ нередукционных (недекомпозиционных) методов на всей структуре данных гиперкуба, а вычислительная сложность f^R редукционных (декомпозиционных) методов составляется из вычислительных сложностей метода на подкубовых структурах $f_i = f(\cdot, n_i)$, при этом полагаем

$$f^R(\cdot, n) = \sum_{i=1}^{k} f(\cdot, n_i), \tag{1}$$

и критерий эффективности методов определяется путем сравнения вычислительных сложностей редукционного и нередукционного методов из одного класса, и эффективнее считается тот из методов, для которого вычислительная сложность меньше.

В работах авторов [11, 12] были доказаны следующие критерии уменьшения вычислительной сложности (увеличения вычислительной производительности) редукционных методов анализа OLAP-кубов многомерных данных из класса полиномиальной степени сложности.

Теорема 1. Для методов полиномиальной степени сложности при $p \ge 2$ справедливо следующее соотношение:

$$f(p,n) > f(p,n_1) + \ldots + f(p,n_k). \tag{2}$$

При $p = 1$ ни один из методов не имеет преимуществ над другими, так как $f^R = f$ ввиду простого равенства $n^1 = n_1^1 + \ldots + n_k^1$, справедливого при любом $k \ge 2$ в силу предположения о характере разбиения на подкубовые структуры.

Отметим, что теорема 1 устанавливает необходимые и достаточные условия уменьшения вычислительной сложности редукционных методов при решении задач анализа OLAP-гиперкубов из подкласса $F_2^p(n)$ полиномиальной степени сложности, который, в силу тривиальности случая $p = 1$, содержит все возможные случаи, представляющие какой-либо интерес для исследований. Заметим также, что полученные в теореме 1 критерии уменьшения вычислительной сложности методов решения задач анализа OLAP-гиперкубов многомерных данных не дают количественной оценки значения величины уменьшения вычислительной сложности редукционных методов.

Далее для подкласса $F_2^p(n)$ редукционных методов полиномиальной степеней сложности даются точные количественные оценки величины уменьшения вычислительной сложности.

2. Точная верхняя граница уменьшения вычислительной сложности редукционных методов полиномиальной степени сложности при декомпозиции OLAP-гиперкуба данных на два подкуба

В настоящем разделе статьи рассмотрим случай, когда нередукционный метод f и редукционный метод f^R декомпозиции исходного гиперкуба K_nH_m с решеткой L на два подкуба $K_{n_1}H_m$ и $K_{n_2}H_m$, состоящих соответственно из n_1 и n_2 подкубов решетки L принадлежат подклассу полиномиальной сложности $F_2^p(n)$. Напомним [12], что относительной эффективностью редукционного метода f^R из класса сложности F по отношению к нередукционному методу f из того же класса сложности F называется отношение их вычислительных сложностей $E = f/f^R$, и это отношение зависит от способа разбиения гиперкуба на подкубовые структуры. Отметим, что в той же работе [12] рассмотрена ситуация, при которой решетка L имеет четное число подкубов, и был получен следующий результат.

Теорема 2. Пусть число n является четным: $n = 2r$. Тогда точная верхняя граница относительной эффективности

$$\overline{E}_{F^p}(f, f^R) = \max_{\{n_1,n_2\}\in\mathcal{L}_2} E = \frac{f}{\min\limits_{\{n_1,n_2\}\in\mathcal{L}_2} f^R_{\mathcal{L}_2}\big(\{n_1, n_2\}\big)} \tag{3}$$

редукционного метода f^R из класса полиномиальной степени сложности по отношению к нередукционному методу f того же класса сложности достигается на множестве подкубовых структур $\mathcal{L}_2 \triangleq \{L_1, L_2\}$ разбиений гиперкуба K_nH_m на два подкуба $K_{n_1}H_m$ и $K_{n_2}H_m$, состоящих соответственно из n_1 и n_2 ($n_1 + n_2 = n$) подкубов исходной решетки L при $n_1 = n_2 = r$, причем

$$\overline{E} = 2^{p-1}. \tag{4}$$

Отметим, что теорема 2 устанавливает точную верхнюю границу уменьшения вычислительной сложности редукционных методов анализа OLAP-гиперкуба данных по отношению к нередукционным методам в классе полиномиальной степени сложности при декомпозиции исходного гиперкуба K_nH_m на два подкуба $K_{n_1}H_m$ и $K_{n_2}H_m$. Заметим также, что

в силу (4) максимально достигаемая эффективность редукционных методов по отношению к нередукционным в случае, когда решетка L гиперкуба K_nH_m имеет четное число подкубов, не зависит от размерности решетки, а зависит лишь от степени полиномиальной сложности метода, и чем сложнее методы, тем большую эффективность дает применение редукционных методов при такой декомпозиции данных исходного OLAP-гиперкуба K_nH_m.

Исследуем теперь ситуацию, когда декомпозиция гиперкуба для редукционного метода f^R из подкласса полиномиальной степени сложности $F_2^p(n)$ осуществляется на множестве подкубовых структур $\mathcal{L}_2 \triangleq \{L_1, L_2\}$ разбиений гиперкуба K_nH_m на два подкуба $K_{n_1}H_m$ и $K_{n_2}H_m$, состоящих соответственно из n_1 и n_2 ($n_1 + n_2 = n$) подкубов исходной решетки L при нечетном n: $n = 2r + 1$, r — натуральное число.

В этом случае для верхней границы относительной эффективности (3) справедлива следующая оценка.

Теорема 3. Пусть число n является нечетным: $n = 2r + 1$. Тогда точная верхняя граница относительной эффективности редукционного метода f^R из подкласса полиномиальной степени сложности $F_2^p(n)$ по отношению к нередукционному методу f того же подкласса сложности $F_2^p(n)$ достигается на $\mathcal{L}_2$ при $n_1 = r$, $n_2 = r + 1$ вне зависимости от r, и вычисляется по формуле

$$\overline{E} = 2^{p-1}\left[\frac{2n^p}{(n-1)^p + (n+1)^p}\right]. \tag{5}$$

Доказательство. Относительная эффективность для любой пары $\{n_1, n_2\} \in \mathcal{L}_2$ определяется по формуле

$$E = \frac{n^p}{n_1^p + n_2^p}. \tag{6}$$

Пусть для определенности $n_2 > n_1$. Тогда для любой такой пары $\{n_1, n_2\} \in \mathcal{L}_2$ имеем, что $1 \le n_1 \le r$ и $r + 1 \le n_2 \le 2r$.

1. Рассмотрим вначале случай, когда степень p равна двум ($p = 2$).

Положим $c = n_1/n_2$. Тогда для любой пары $\{n_1, n_2\} \in \mathcal{L}_2$ имеем, что $0 < c < 1$, и также,

$$(c+1)^2 = \left(\frac{n_1}{n_2} + 1\right)^2 = \left(\frac{n_1 + n_2}{n_2}\right)^2 = \frac{n^2}{n_2^2}. \tag{7}$$

$$c^2 + 1 = \left(\frac{n_1}{n_2}\right)^2 + 1 = \frac{n_1^2 + n_2^2}{n_2^2}. \tag{8}$$

Для величины E в силу (6)–(8) в данном случае справедливы соотношения

$$E = \frac{n^2}{n_1^2 + n_2^2} = \frac{(c+1)^2}{c^2+1} = \frac{c^2 + 2c + 1}{c^2+1} = 1 + \frac{2c}{c^2+1}. \tag{9}$$

Так как

$$\frac{2c}{c^2+1} - 1 = \frac{2c - c^2 - 1}{c^2+1} = -\frac{(c-1)^2}{c^2+1} < 0,$$

то при любом $0 < c < 1$ всегда $\dfrac{2c}{c^2+1} < 1$, и учитывая (9), получаем, что

$$E < 2, \tag{10}$$

причем E принимает максимальное значение в (10) в случае, когда $c \to 1$, так как функция $E = E(c)$ монотонно возрастает при $c \in (0,1)$, поскольку ее производная $\dfrac{dE}{dc} = \dfrac{2-2c^2}{(c^2+1)^2}$ строго положительна на этом интервале.

Теперь найдем разбиение $\mathcal{L}_2$ на пары $\{n_1, n_2\}$, при котором достигается максимальная относительная эффективность. Так как $n_1 = n - n_2$, имеем $c = n/n_2 - 1$. Поскольку $n = 2r+1$ и n_2 может принимать только значения $n_2 = r+k$, $k = 1, 2, \ldots, r$, в силу предположения $n_2 > n_1$, получаем для последовательных значений c_k при $k = 1, 2, \ldots, r$

$$c_k = \frac{2r+1}{r+k} - 1 = \frac{2r+1-r-k}{r+k} = \frac{r+1-k}{r+k}, \tag{11}$$

откуда видим, что c_k принимает максимальное значение, равное $r/(r+1)$ при $k = 1$, так как с ростом k числитель в (11) уменьшается, а знаменатель растет, и, таким образом, значение дроби (11) лишь уменьшается. Как уже было отмечено в ходе доказательства, относительная эффективность E увеличивается с ростом c, поэтому достигает максимального значения при максимально возможном значении c_k, т. е. при $k = 1$ на разбиении, когда $n_2 = r+1$, и соответственно $n_1 = r$.

Найдем максимальное значение E при таком разбиении, это можно сделать непосредственно из формулы (6) при $p = 2$. Имея ввиду, что $r = (n-1)/2$ находим

$$E = \frac{n^2}{n_1^2+n_2^2} = \frac{n^2}{r^2+(r+1)^2} = \frac{n^2}{\dfrac{(n-1)^2}{2^2} + \left(\dfrac{n-1}{2}+1\right)^2} =$$

$$= \frac{n^2}{\dfrac{(n-1)^2}{2^2} + \dfrac{(n+1)^2}{2^2}} = 2^2\left(\frac{n^2}{(n-1)^2+(n+1)^2}\right) = 2\left(\frac{2n^2}{(n-1)^2+(n+1)^2}\right), \tag{12}$$

и получаем, что (12) определяет точную верхнюю достижимую границу $\overline{E}$, которая зависит от n.

З а м е ч а н и е 1. Условие $E < 2$, доказанное ранее в (10), влечет необходимость

$$\frac{2n^2}{(n-1)^2+(n+1)^2} < 1.$$

Проверим, что это так. Действительно,

$$(n-1)^2 + (n+1)^2 = n^2 - 2n + 1 + n^2 + 2n + 1 = 2n^2 + 2,$$

а $\dfrac{2n^2}{2n^2+2} < 1$ при любом натуральном n, поэтому, согласно (12) действительно имеем $E < 2$, что согласуется с (10).

2. Рассмотрим теперь общий случай при произвольных натуральных числах $p \geq 3$. Идеи доказательства в этом случае в основном остаются те же, что и при $p = 2$. Выберем также $c = n_1/n_2$. Тогда

$$(c+1)^p = \left(\frac{n_1}{n_2}+1\right)^p = \left(\frac{n_1+n_2}{n_2}\right)^p = \frac{n^p}{n_2^p} \tag{13}$$

и

$$c^p + 1 = \left(\frac{n_1}{n_2}\right)^p + 1 = \frac{n_1^p + n_2^p}{n_2^p}. \tag{14}$$

Для величины E^{-1} в силу (6), (13), (14) имеет место

$$E^{-1} = \frac{n_1^p + n_2^p}{n^p} = \frac{c^p + 1}{(c+1)^p} = \left(\frac{c}{c+1}\right)^p + \frac{1}{(c+1)^p} = \left(1 - \frac{1}{c+1}\right)^p + \frac{1}{(c+1)^p}. \tag{15}$$

Положим $u = 1/(c+1)$, тогда, учитывая (15), получаем, что

$$E^{-1} = (1-u)^p + u^p, \tag{16}$$

при $1/2 < u < 1$ (так как $0 < c < 1$).

Используя (16), найдем первую производную:

$$\frac{dE^{-1}}{du} = -p(1-u)^{p-1} + pu^{p-1}, \tag{17}$$

которая строго положительна, так как $1 - u < u$ при всех $u \in (1/2,\ 1)$. В самом деле, обратное предположение $1 - u \geq u$ влечет, что $1 \geq 2u$ и $u \leq 1/2$. Поэтому

$$-p(1-u)^{p-1} + pu^{p-1} = -\left(\frac{1-u}{u}\right)^{p-1} + 1 > 0.$$

Следовательно, функция $E^{-1}(u)$ монотонно возрастает на этом интервале, т.е. для любых $u_1 < u_2$ из интервала $(1/2,\ 1)$ выполняется $E^{-1}(u_1) < E^{-1}(u_2)$, а значит, и функция $E(u)$, как обратная функция, монотонно убывает, т.е. для любых $u_1 < u_2$ из интервала $(1/2,\ 1)$ выполняется $E(u_1) > E(u_2)$.

Поскольку u изменяется от $1/2$ до 1 на интервале $(1/2,\ 1)$, когда c меняется соответственно от 1 до 0 на интервале $(0,\ 1)$, то при обратном изменении c от 0 до 1 на интервале $(0,\ 1)$ функция $E(c)$ уже монотонно возрастает, и для любых $c_1 < c_2$ из интервала $(0,\ 1)$ выполняется $E(c_1) < E(c_2)$.

Теперь, чтобы найти разбиение $\mathcal{L}_2$ на пары $\{n_1, n_2\}$, при котором функция $E(c)$ принимает наибольшее значение, воспользуемся конструкцией (11), которая не зависит от p, где уже показано, что при $n_2 = r + k$, $k = 1, 2, \ldots, r$, для последовательных значений c_k при $k = 1, 2, \ldots, r$ выполняется $c_1 > c_k$ для любого $k > 1$. Следовательно, $E(c_k) < E(c_1)$ при $k = 2, \ldots, r$ и максимальная эффективность достигается на разбиении, когда $n_2 = r + 1$, и, соответственно, $n_1 = r$.

Найдем максимальное значение E при таком разбиении. Непосредственно из формулы (15) имеем

$$E = \frac{(c+1)^p}{c^p + 1}, \tag{18}$$

и, подставляя значение $c_1 = r/(r+1)$ в (18), находим

$$E = \frac{(2r+1)^p}{r^p + (r+1)^p}. \tag{19}$$

Переходя к $n = 2r + 1$, откуда $r = (n-1)/2$, окончательно получаем

$$E = \frac{(2r+1)^p}{r^p + (r+1)^p} = \frac{n^p}{r^p + (r+1)^p} = \frac{n^p}{\dfrac{(n-1)^p}{2^p} + \left(\dfrac{n-1}{2} + 1\right)^p} =$$

$$= \frac{n^p}{\dfrac{(n-1)^p}{2^p} + \dfrac{(n+1)^p}{2^p}} = 2^p \left(\frac{n^p}{(n-1)^p + (n+1)^p}\right) = 2^{p-1}\left(\frac{2n^p}{(n-1)^p + (n+1)^p}\right), \quad (20)$$

т.е. точная верхняя граница $\overline{E}$ действительно вычисляется по формуле (5) и достигается на разбиении $\mathcal{L}_2$ при $n_1 = r$ и $n_2 = r + 1$. Теорема 3 полностью доказана.

Теорема 3 устанавливает точную количественную оценку уменьшения вычислительной сложности (относительную эффективность) декомпозиционного метода $f^R \in F_2^p(n)$ по сравнению с традиционными методами анализа OLAP-данных, принадлежащими тому же классу сложности, когда решетка исходного гиперкуба имеет нечетное число подкубов. Отметим также, что в силу (5) максимально достигаемая эффективность редукционных методов по отношению к нередукционным в случае, когда решетка L гиперкуба $K_n H_m$ имеет нечетное число подкубов, не зависит от размерности гиперкуба, а зависит от размерности решетки и от степени сложности метода, и чем сложнее методы и выше разбиение решеткой, тем большую эффективность дает применение редукционных методов при такой декомпозиции данных исходного OLAP-гиперкуба.

3. Сравнение точных верхних границ уменьшения вычислительной сложности для решеток с четным и нечетным количеством подкубов

Сравним точные значения верхних границ уменьшения вычислительной сложности для решеток с четным (формула (4)) и нечетным (формула (5)) количеством подкубов. Выражения (4) и (5) отличаются сомножителем, зависящим от n, который появляется в случае нечетного числа подкубов в структуре данных гиперкуба и имеет вид

$$R(n) = \frac{2n^p}{(n-1)^p + (n+1)^p}. \quad (21)$$

Для $R(n)$, определенного в (21), справедливо следующее утверждение.

Утверждение 1. *При любых натуральных $n \geq 2$ и $p \geq 2$*

$$|R(n)| < 1. \quad (22)$$

Кроме того, функция $R(n)$ строго монотонно возрастает с ростом n при фиксированном p и при любом натуральном $p \geq 2$

$$\lim_{n\to\infty} R(n) = 1. \quad (23)$$

Доказательство. Пусть $m = 2n$ и k — натуральное число ($m/2 \leq k < m$). Выберем $c = m - k/k$. При таком выборе имеем $0 < c \leq 1$, и тогда для функции

$$R^{-1} = \frac{k^p + (m-k)^p}{m^p}$$

по аналогии с (15) имеет место представление

$$R^{-1} = \frac{k^p + (m-k)^p}{m^p} = \frac{c^p + 1}{(c+1)^p} = \left(1 - \frac{1}{c+1}\right)^p + \frac{1}{(c+1)^p}. \tag{24}$$

Рассмотрим обратную к R^{-1} функцию

$$R(m,k) = \frac{m^p}{k^p + (m-k)^p}. \tag{25}$$

Для $k = m/2 + 1$ получим представление $R(n)$ в виде

$$R(n) = \frac{2n^p}{(n-1)^p + (n+1)^p} = \frac{m^p}{2^{(p-1)}\Big((m/2-1)^p + (m/2+1)^p\Big)} = \\ = 2^{1-p} \cdot \frac{m^p}{k^p + (m-k)^p} = 2^{1-p}\, R(m, (m/2+1)). \tag{26}$$

Далее используем схему доказательства теоремы 2 из [12] при четном m. Положим $u = 1/(c+1)$. Тогда, учитывая (24), получаем, что

$$R^{-1} = (1-u)^p + u^p \tag{27}$$

при $1/2 \le u < 1$ (так как $0 < c \le 1$). Используя (27), найдем первую производную:

$$\frac{dR^{-1}}{du} = -p(1-u)^{(p-1)} + pu^{(p-1)}, \tag{28}$$

которая, как нетрудно заметить, обращается в нуль в одной из точек полуинтервала $[1/2,\ 1)$, а именно

$$\frac{dR^{-1}}{du} = 0 \quad \text{при} \quad u = \frac{1}{2}. \tag{29}$$

Находим вторую производную:

$$\frac{d^2R^{-1}}{du^2} = p(p-1)(1-u)^{(p-2)} + p(p-1)u^{(p-2)}. \tag{30}$$

Видим, что при любом $p \ge 2$ она положительна:

$$\left.\frac{d^2R^{-1}}{du^2}\right|_{u=1/2} = 2p(p-1)\left(\frac{1}{2}\right)^{(p-2)} > 0. \tag{31}$$

Следовательно, в точке $u = 1/2$ функция R^{-1}, как функция от u, достигает своего строгого минимума, равного

$$\left.R^{-1}\right|_{u=1/2} = \left(\frac{1}{2}\right)^p + \left(\frac{1}{2}\right)^p = 2 \cdot 2^{-p} = 2^{-(p-1)}. \tag{32}$$

В этой же точке значение $R(m)$ максимально и

$$\overline{R}(m,k) = \max_k R(m,k) = 2^{(p-1)}. \tag{33}$$

Согласно теореме 2, максимум $\overline{R}(m,k)$ достигается при $k = m/2$, в чем можно убедиться непосредственно, так как в силу выбора u, при $u = 1/2$, в точке $1/(c+1) = 1/2$ получаем $c = 1$ и $k = m - k$, т.е. $2k = m$ или $k = m/2$.

Таким образом, $R(m,k) < R(m,m/2) = 2^{p-1}$ при любом натуральном $m/2 < k < m$, откуда получаем $R(m,(m/2+1)) < 2^{p-1}$ и из (26) видим, что

$$R(n) = 2^{1-p}\,R(m,(m/2+1)) < 2^{1-p}\cdot 2^{p-1} = 1, \tag{34}$$

Это доказывает неравенство (22) утверждения 1.

Монотонное возрастание $R(n)$ по n является естественным следствием монотонного возрастания $R(m,(m/2+1))$ по m. Чтобы убедиться в последнем, рассмотрим вначале последовательность c_m значений c при $k = m/2+1$. Имеем

$$c_m\Big|_{k=m/2+1} = \frac{m-k}{k} = \frac{m-m/2-1}{m/2+1} = \frac{m-2}{m+2}. \tag{35}$$

Для любых натуральных $m_2 > m_1$ выполняется

$$c_{m_2} - c_{m_1} = \frac{m_2-2}{m_2+2} - \frac{m_1-2}{m_1+2} = \frac{4(m_2-m_1)}{(m_2+2)(m_1+2)} > 0, \tag{36}$$

т.е. $c_{m_2} > c_{m_1}$ при любых $m_2 > m_1$, так что последовательность c_m монотонно возрастает согласно известному определению монотонно возрастающей последовательности.

Далее,

$$\lim_{m\to\infty} c_m = \lim_{m\to\infty} \frac{m-2}{m+2} = 1, \tag{37}$$

и последовательность c_m с ростом m стремится к своему предельному значению, равному единице.

Используя те же рассуждения, что и при доказательстве теоремы 3, приходим к выводу, что $u_m = 1/(c_m+1)$ стремится строго монотонно справа налево к предельному значению $1/2$ при изменении на полуинтервале $[1/2,\ 1)$, когда c_m стремится к 1 слева направо на полуинтервале $(0,\ 1]$. Кроме того,

$$\lim_{m\to\infty} R_m^{-1} = \lim_{m\to\infty} [(1-u_m)^p + u_m^p] = \lim_{m\to\infty} \left[\left(1-\frac{1}{c_m+1}\right)^p + \left(\frac{1}{c_m+1}\right)^p\right] =$$
$$= \lim_{m\to\infty} \left[\left(\frac{c_m}{c_m+1}\right)^p + \left(\frac{1}{c_m+1}\right)^p\right] = \left(\frac{1}{2}\right)^p + \left(\frac{1}{2}\right)^p = 2^{-(p-1)}. \tag{38}$$

Соответственно,

$$\lim_{m\to\infty} R\Big(m,(m/2+1)\Big) = 2^{p-1}. \tag{39}$$

Учитывая (26), находим

$$\lim_{n\to\infty} R(n) = 2^{1-p}\cdot \lim_{m\to\infty} R\Big(m,(m/2+1)\Big) = 2^{1-p}\cdot 2^{p-1} = 1, \tag{40}$$

что и доказывает равенство (23) утверждения 1. Утверждение 1 полностью доказано.

В утверждении 1 получено следующее свойство: верхняя граница эффективности редукционных методов на решетках с разбиением гиперкуба на нечетное число подкубов отличается от верхней границы эффективности на решетках с разбиением гиперкуба на четное число подкубов. Наиболее сильно отличие в верхних границах проявляется при малых n, но с ростом n оно «сглаживается».

З а м е ч а н и е 2. Можно и непосредственно проверить, что

$$\lim_{n\to\infty} R(n) = \lim_{n\to\infty} \frac{2n^p}{(n-1)^p + (n+1)^p} = 1.$$

Приведенное же в работе доказательство утверждения 1, с одной стороны, позволяет увидеть монотонность стремления $R(n)$ к предельному значению, а с другой, раскрывает структурные особенности различий в декомпозиции на четной и нечетной решетках.

Следствием теорем 2, 3 и утверждения 1 является тот факт, что при достаточно больших n при одинаковой степени сложности эффективность редукционных методов на решетках с нечетным числом подкубов сопоставима с эффективностью на решетках с четным числом подкубов.

Следствие 1. При нечетном n точная верхняя граница относительной эффективности редукционного метода f^R из подкласса полиномиальной степени сложности $F_2^p(n)$ по отношению к нередукционному методу f того же подкласса сложности $F_2^p(n)$ при $n \to \infty$ вычисляется по формуле

$$\overline{E} = 2^{p-1} \tag{41}$$

и совпадает с верхней границей относительной эффективности редукционного метода f^R из подкласса полиномиальной степени сложности $F_2^p(n)$ по отношению к нередукционному методу f того же подкласса сложности $F_2^p(n)$ при четном n.

Отметим, что на решетках с четным числом подкубов точная верхняя граница (41) достигается на множестве $\mathcal{L}_2$ при $n_1 = n_2$, тогда как на решетках с нечетным числом подкубов эта граница недостижима, хотя значение относительной эффективности становится сколь угодно близким к границе с ростом n, и наилучшее приближение на множестве $\mathcal{L}_2$ достигается при $n_1 = n_2 - 1$.

Заключение

Исследованы проблемы редукции (декомпозиции) моделей многомерных данных в виде гиперкубовых OLAP-структур. Рассмотрен случай, когда структура данных определяется решеткой, разбивающей гиперкуб на нечетное количество подкубов, и декомпозиция гиперкуба осуществляется на два подкуба на множестве, состоящем из нечетного числа подкубовых структур, методами полиномиальной степени сложности.

Установлена точная количественная оценка уменьшения вычислительной сложности (повышения относительной эффективности) декомпозиционного метода по сравнению с традиционными методами анализа OLAP-данных, принадлежащими тому же классу сложности.

Показано, что максимально возможная эффективность редукционных методов по отношению к нередукционным методам в случае, когда решетка разбивает гиперкуб на нечетное число подкубов, не зависит от размерности гиперкуба, а зависит от размерности решетки и от степени сложности метода, и чем сложнее методы и выше разбиение решеткой, тем большую эффективность дает применение редукционных методов при такой декомпозиции данных исходного OLAP-гиперкуба.

Проведено сравнение эффективности редукционных методов при декомпозиции гиперкуба на два подкуба на множествах, состоящих из четного и нечетного числа подкубовых структур. Получены свойства, которыми верхняя граница эффективности редукционных методов на решетках с разбиением гиперкуба на нечетное число подкубов отличается от верхней границы эффективности на решетках с разбиением гиперкуба на четное число подкубов. Показано, что эти отличия "сглаживаются" с увеличением размерности решетки, и что при большом дроблении данных для методов полиномиальной степени сложности на нечетных решетках максимальная эффективность редукционных методов практически достигается и сопоставима с их эффективностью на четных решетках.

Изложенные в статье результаты по редукции OLAP-гиперкубов используются в междисциплинарном проекте РФФИ, исследующим принципы и методы виртуального моделирования искусственных биологических органов на основе моделей OLAP и Data Mining.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований, грант 19-07-00686 а.

Список литературы

1. Андрейчиков А.В., Андрейчикова О.Н. Интеллектуальные информационные системы: учеб. М: Финансы и статистика, 2004. 422 с.
2. Ноженкова Л.Ф., Шайдуров В.В. OLAP-технологии оперативной информационно-аналитической поддержки организационного управления // Информационные технологии и вычислительные системы. 2010. № 2. С. 15–27.
3. Замятин А.В. Введение в интеллектуальный анализ данных: учеб. пособие. Томск: Изд-во Томского гос. ун-та, 2016. 120 с.
4. Вайнштейн Ю.В. Планирование медицинской помощи с применением аналитических OLAP-моделей // Вестник Томского гос. ун-та. Сер. «Математика. Кибернетика. Информатика». 2004. Приложение № 8470, 9(II). С. 16–22.
5. Петровский А.Б., Ройзензон Г.В. Снижение размерности признакового пространства в задачах многокритериальной классификации: стратификация кортежей // 11-я национ. конф. по искусственному интеллекту с международным участием: КИИ-2008 (г. Дубна, Россия, 29 сентября – 3 октября 2008 г.): Тр. М.: ЛЕНАНД, 2008. Т. 2. С. 262–270.

6. Петровский А.Б., Лобанов В.Н. Многокритериальный выбор в пространстве признаков большой размерности: мультимедийная технология ПАКС-М // Искусственный интеллект и принятие решений. 2014. № 3. С. 92–104.
7. Agarwal S., Agrawal R., Deshpande P.M., Gupta A., Naughton J.F., Ramakrishnan R., Sarawagi S. On the computation of multidimensional aggregates // Materialized views: techniques, implementations and applications / Ed. by A. Gupta. Camb.: MIT Press, 1999. Pp. 506–521. DOI: 10.7551/mitpress/4472.003.0030
8. Чубукова И.А. Data Mining: учеб. пособие. 2-е изд. М: Бином. Лаборатория знаний, 2008. 382 с.
9. Макаров И.М., Рахманкулов В.З., Ахрем А.А., Ровкин И.О. Исследование свойств гиперкубовых структур в OLAP-системах // Информационные технологии и вычислительные системы. 2005. № 2. С. 4–9.
10. Akhrem A.A., Rakhmankulov V.Z., Yuzhanin K.V. On the complexity of the reduction of multidimensional data models // Scientific and Technical Information Processing. 2017. Vol. 44, no. 6. Pp. 406–411. DOI: 10.3103/S0147688217060028
11. Ахрем А.А., Носов А.П., Рахманкулов В.З., Южанин К.В. Вычислительная производительность методов редукции гиперкубов многомерных данных аналитических OLAP-систем // DOI: 10.14357/20718594190403
12. Ахрем А.А., Носов А.П., Рахманкулов В.З., Южанин К.В. Анализ вычислительной сложности методов декомпозиции OLAP-гиперкубов многомерных данных // Математика и математическое моделирование. 2020. № 4. С. 52–64. DOI: 10.24108/mathm.0420.0000221

Mathematics and Mathematical Modeling, 2021, no. 1, pp. 27–42.
DOI: **10.24108/mathm.0121.0000244**

Mathematics & Mathematical Modelling
Electronic journal
http://mathmelpub.ru ISSN 2412-5911

Analysing Efficiency Methods of Polynomial Complexity Degree in Multidimensional OLAP Cube Data Decomposition

Akhrem A. A.[1], Nosov A. P.[1,*], Rakhmankulov V. Z.[1]

Federal Research Center "Informatics and Control" of RAS, Moscow, Russian Federation
[*] nosov@isa.ru

Keywords: OLAP-system, decomposition, computational performance, OLAP-data hypercube, polynomial complexity

Received: 21.12.2020.

The article investigates the problems of reduction (decomposition) of multidimensional data models in terms of hypercube OLAP structures. Describes the case when a data structure is defined by the array that slices and dices the hypercube into the odd number of subcubes, and this set of subcube structures becomes decomposed. Defines an exact upper bound for increasing a computational performance of methods to analyze OLAP data on subcubes, which determines the decomposition approach efficiency in comparison with the OLAP data analysis on a complete unreduced hypercube. A compared efficiency of the hypercube decomposition into two subcubes on the sets consisting of the even and odd number of subcube structures has shown that with considerable data partitioning for methods of a polynomial complexity degree the decomposition efficiency essentially is independent on this factor and rises with increasing complexity degree of methods applied.

When using the mathematical methods to study decomposition (reduction) of large hyper-cubes of multidimensional data of analytical OLAP systems into subcube components, there is a need to find conditions for minimising the computational complexity of methods to solve the problems of the OLAP hyper-cube analysis during data decomposition in comparison with using these methods for analyzing large amounts of information that is accumulated directly in the hyper-cubes of multidimensional OLAP data to establish the criteria for decreasing or increasing computational performance when applying methods on the subcube components (reduction methods) as compared to applying these methods on a hypercube (non-reduction or traditional methods), depending on one or another degree of complexity of complex methods.

The article provides an accurate quantitative estimate of decreasing computational complexity of reduction methods for analyzing OLAP cubes as compared to the non-reduction methods in the case when said methods have the polynomial complexity and the original hypercube array of data comprises the odd number of subcubes.

References

1. Andrejchikov A.V., Andrejchikova O.N. *Intellektual'nye informatsionnye sistemy* [Intelligent information systems]: a textbook. Moscow: Finansy i statistika Publ., 2004. 422 p. (in Russian).
2. Nozhenkova L.F., Shaydurov V.V. OLAP-technology of operative information-analytical support of organizational management. *Informatsionnye tekhnologii i vychislitel'nye sistemy* [Information Technologies and Computing Systems], 2010, no. 2, pp. 15–27 (in Russian).
3. Zamiatin A.V. *Vvedenie v intellektual'nyj analiz dannykh* [Introduction to data mining]: a textbook. Tomsk: Tomsk State Univ. Publ., 2016. 120 p. (in Russian).
4. Vainstein Yu.V. Planning of medical care using analytical OLAP models. *Vestnik Tomskogo gosudarstvennogo universiteta. Ser. Matematika. Kibernetika. Informatika* [Bulletin of the Tomsk State Univ. Ser. Mathematics. Cybernetics. Computer science], 2004, suppl. No. 8470, 9(II), pp. 16–22 (in Russian).
5. Petrovsky A.B., Rojzenson G.V. Snizhenie razmernosti priznakovogo prostranstva v zadachakh mnogokriterial'noj klassifikatsii: stratifikatsiia kortezhej [Reducing the dimension of the feature space in multi-criteria classification problems: stratification of tuples]. *11-ia natsionalnaia konferentsiia po iskusstvennomu intellektu s mezhdunarodnym uchastiem: KII 2008* [11th national conf. on artificial intelligence with international participation: KII-2008 (Dubna, Russia, Sept. 29th - October 3rd, 2008)]: Proc. Moscow: LENAND Publ., 2008. Vol. 2. Pp. 262–270 (in Russian).
6. Petrovsky A.B., Lobanov V.N. Multiple criteria choice in the attribute space of large dimension: multi-method technology PAKS-M. *Isskustvennyj intellekt i priniatie reshenij* [Artificial Intelligence and Decision Making], 2014, no. 3, pp. 92–104 (in Russian).
7. Agarwal S., Agrawal R., Deshpande P.M., Gupta A., Naughton J.F., Ramakrishnan R., Sarawagi S. On the computation of multidimensional aggregates. *Materialized views: techniques, implementations and applications* / Ed. by A. Gupta. Camb.: MIT Press, 1999. Pp. 506-521. DOI: 10.7551/mitpress/4472.003.0030
8. Chubukova I.A. *Data Mining*: a textbook. 2nd ed. Moscow: BINOM. Laboratoriia znanij Publ., 2008. 382 p. (in Russian).
9. Makarov I.M., Rakhmankulov V.Z., Akhrem A.A., Rovkin I.O. Investigation of properties of hypercube structures in OLAP systems. *Informatsionnye tekhnologii i vychislitel'nye sistemy* [Information Technologies and Computing Systems], 2005, no. 2, pp. 4–9 (in Russian).

10. Akhrem A.A., Rakhmankulov V.Z., Yuzhanin K.V. On the complexity of the reduction of multidimensional data models. *Scientific and Technical Information Processing*, 2017, vol. 44, no. 6, pp. 406–411. DOI: 10.3103/S0147688217060028

11. Akhrem A.A., Nosov A.P., Rakhmankulov V.Z., Yuzhanin K.V. Computational performance of hypercube reduction methods for multidimensional data of analytical OLAP systems. *Iskustvennyj intellekt i priniatie reshenij* [Artificial Intelligence and Decision Making], 2019, no. 4, pp. 23–28. DOI: 10.14357/20718594190403

12. Akhrem A.A., Nosov A.P., Rakhmankulov V.Z., Yuzhanin K.V. Computational complexity analysis of decomposition methods of OLAP hyper-cubes of multidimensional data. *Matematika i matematicheskoe modelirovanie* [Mathematics and Mathematical Modeling], 2020, no. 4, pp. 52–64. DOI: 10.24108/mathm.0420.0000221 (in Russian).