УДК 519.681.3

МОДИФИКАЦИЯ АЛГОРИТМА БАЛАША ДЛЯ РЕШЕНИЯ ЗАДАЧИ НОРМАЛИЗАЦИИ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ

И.В. Клименко

ФГБОУ ВПО «Ростовский государственный университет путей сообщения», г. Ростов-на-Дону

Рецензент д-р техн. наук, профессор М.А. Бутакова

Ключевые слова и фразы: аксиомы вывода функциональных зависимостей; алгоритм Балаша; задача о наименьшем покрытии; максимальные независимые транзитивные множества; нормализация отношений реляционной модели; нормальные формы; первичный ключ; функциональные зависимости.

Аннотация: Решена актуальная проблема модификации известных алгоритмов для решения задачи оптимального покрытия с учетом специфики реляционной модели. В частности, для решения задачи о наименьшем покрытии модифицирован известный алгоритм Балаша. Вычислительный эксперимент показал существенное преимущество модифицированного алгоритма, выраженное в ускорении процедуры зондирования ветвей.

Для формализации процесса нормализации отношений реляционной модели возможно использовать методы целочисленного программирования и, в частности, задачи о наименьшем покрытии.

В этом направлении актуальной видится задача модификации известных алгоритмов для решения задачи оптимального покрытия с учетом специфики реляционной модели. Указанные модификации должны быть направлены на сокращение количества шагов, так как задача наименьшего покрытия (**ЗНП**) относится к классу *NP*-полных [1].

В качестве исходных данных для решения задачи покрытия исследуемой предметной области (**ПО**) максимальными транзитивно независимыми множествами (**МТНМ**) служит матрица $C = [t_{ij}]$, элементы которой равны 0 или 1, а вектор-столбец является МТНМ.

В матричной форме, когда строки матрицы C, состоящей из нулей и единиц, покрываются столбцами, задача покрытия ПО максимальными

Клименко Игорь Валерьевич – кандидат технических наук, доцент кафедры «Вычислительная техника и автоматизированные системы управления» e-mail: klimigor67@mail.ru, ФГБОУ ВПО «Ростовский государственный университет путей сообщения», г. Ростов-на-Дону.

ТНМ может быть сформулирована как задача линейного (целочисленного) программирования минимизировать

$$z = \sum_{j=1}^{N} c_j \xi_j , \tag{1}$$

при ограничениях

$$\sum_{j=1}^{N} t_{ij} \xi_j \geq 1, \quad i = 1, 2, \ldots, m, \tag{2}$$

где $c_j \geq 0$, $\xi_j = \begin{cases} 1, \text{если } S_j \in \Im', \\ 0, \text{если } S_j \notin \Im'. \end{cases}$

Для окончательного сведения рассматриваемой задачи к ЗНП необходимо решить вопрос о правилах присвоения стоимостей каждому МТНМ из семейства $\Im$.

Очевидно, что чем больше вершин x_i входит в максимальное свободное от конфликтов множество S_j, тем больше строк матрицы $[t_{ij}]$ покроет S_j. Следовательно, чем больше элементов входит в каждое из МТНМ, покрывающих множество X, тем больше вероятность того, что меньшее количество МТНМ потребуется для покрытия X. В свою очередь, если каждому МТНМ поставлено в соответствие отношение базы данных (**БД**), то для уменьшения общего количества отношений в проекте БД нужно стремиться использовать МТНМ с бóльшим количеством атрибутов.

В соответствии с приведенными аргументами предлагается следующее правило присвоения стоимостей МТНМ S_j.

Правило 1. Если S_j – МТНМ атрибутов предметной области, а q_j – мощность этого множества, то для решения задачи оптимизации проекта БД множеству S_j необходимо присвоить стоимость c_j

$$c_j = \frac{1}{q_j}. \tag{3}$$

Общий алгоритм нормализации

Исходные данные: ($m \times n$)-матрица C.

Шаг 1.

Каждому максимальному свободному от конфликтов множеству S_j из множества $\Im$ присвоить стоимость c_j по правилу (3).

Шаг 2.

Получить матрицу C', выполнив возможные упрощения над матрицей C.

(i) Если $\exists\, x_i \in X$, такое, что $x_i \in S_k$ и $x_i \notin S_j \;\forall\, j \neq k$, то S_k должно присутствовать во всех решениях и задачу можно свести к «меньшей», положив $X = X - \{x_i\}$ и $\Im = \Im - \{S_k\}$. Включить все такие S_k в $\Im_1$.

(ii) Пусть $V_i = \{j | x_i \in S_j\}$; тогда если $\exists\ p, q \in \{1, \ldots, M\}$ такие, что $V_p \subseteq V_q$, то x_q можно удалить из X, поскольку любое множество, которое покрывает x_p, должно также покрывать x_q, то есть x_p доминирует над x_q.

(iii) Если для некоторого семейства множеств $\overline{\mathfrak{I}} \subset \mathfrak{I}$ справедливы соотношения

$$\bigcup_{S_j \in \overline{\mathfrak{I}}} S_j \supseteq S_k \quad \text{и} \quad \sum_{S_j \in \overline{\mathfrak{I}}} c_j \le c_k,$$

для любых $S_k \in \mathfrak{I} - \overline{\mathfrak{I}}$, то S_k может быть вычеркнуто из $\mathfrak{I}$, так как $\bigcup_{S_j \in \overline{\mathfrak{I}}} S_j$ доминирует над S_k.

Шаг 3.

Переформулировать задачу в неприводимой форме, соответствующей виду матрицы C'.

Шаг 4.

Решить задачу линейного программирования модифицированным алгоритмом Балаша с фильтром. Результатом решения задачи является некоторое подсемейство $\mathfrak{I}_2$ максимальных ТНМ.

Шаг 5.

Определить множество: $\mathfrak{I}' = \mathfrak{I}_1 \cup \mathfrak{I}_2$.

Семейство максимальных ТНМ $\mathfrak{I}'$ является оптимальным покрытием исследуемой ПО.

Шаг 6.

Остановка.

Модификация алгоритма Балаша. Задача оптимизации, сформулированная на шаге 3 общего алгоритма нормализации, относится к классу задач целочисленного программирования с булевыми переменными. Решение задач с булевыми переменными можно производить методом ветвей и границ, как с обычными целочисленными переменными, для которых заданы граничные условия $0 \le x_j \le 1$. Однако применение такого метода в ряде случаев оказывается нецелесообразным. Существуют специальные методы частичного перебора для решения таких задач, относящиеся к методам возврата, наиболее совершенным из которых признан алгоритм Балаша с фильтром [2]. Недостатком данного алгоритма является то, что процедура перебора предполагает исследование всех 2^n возможных наборов значений булевых переменных.

Анализ алгоритма Балаша с целью эффективного применения к задаче оптимального покрытия ПО максимальными ТНМ позволил сделать вывод о возможности сокращения количества исследуемых наборов (зондируемых ветвей).

Пусть m – количество покрываемых строк. Тогда в оптимальном решении не может быть меньше r единиц:

$$R = \inf |\mathfrak{R}|, \tag{4}$$

где $\mathfrak{R} \subseteq \mathfrak{I}$ такое, что $\sum_{j \in \mathfrak{R}} q_j \ge m$.

Определенная таким образом величина r представляет собой оценку нижней границы количества ненулевых переменных в оптимальном решении. Применение выведенной нижней границы позволяет исключить из рассмотрения наборы переменных, в которых количество единиц меньше величины r.

Учитывая специфику решаемой задачи, для сокращения количества исследуемых наборов целесообразно использовать новую процедуру поиска начального допустимого решения и соответствующего ему фильтра (см. шаг 1 описания модифицированного алгоритма).

Проведенные исследования позволили получить вывод о существовании зависимости между конкретным видом постановки задачи и продолжительностью вычислений [1–3]. Порядок, в котором рассматриваются переменные и ограничения, в ряде случаев оказывает существенное влияние на эффективность алгоритма. В частности, для рассматриваемой задачи предпочтительным является упорядочение ограничений по мере убывания их «жесткости», а переменные следует ранжировать в соответствии с порядком увеличения коэффициентов целевой функции. Вычислительный эксперимент показал, что соблюдение этих условий ускоряет процедуру зондирования частичных решений по сравнению с исходным алгоритмом Балаша с фильтром.

Учитывая вышесказанное, модифицированный алгоритм Балаша с фильтром можно описать следующим образом.

Исходные данные: ($m\times n$)-матрица C.

Шаг 1.

Найти начальное допустимое решение, используя следующую процедуру.

Процедура поиска начального допустимого решения

1. Найти МТНМ S_j' с минимальной стоимостью c_j'.

2. Переменной ξ_j', соответствующей МТНМ S_j', присвоить значение 1.

3. Вычеркнуть из матрицы столбец, соответствующей МТНМ S_j', и покрытые им строки.

(i) Если вычеркнуты все строки, то перейти к п. 5.

(ii) В противном случае перейти к п. 4.

4. Для невычеркнутых МТНМ пересчитать локальные стоимости:

$$c_{\text{лок}\,j} := c_{\text{лок}\,j} + \frac{1}{q_{\text{лок}\,j}}.$$

Перейти к п. 1.

5. Переменным ξ_j, соответствующим невычеркнутым МТНМ, присвоить значение 0. Начальное допустимое решение найдено.

Шаг 2.

Значение целевой функции z^* при удовлетворении шага 1 принять в качестве дополнительного фильтрующего ограничения (фильтра).

Шаг 3.

Найти нижнюю границу r в соответствии с (4).

Шаг 4.

Начиная с найденного на шаге 1 набора, методом перебора «вниз» определять количество единиц в исследуемых наборах.

Шаг 5.

(i) Если количество единиц в исследуемом наборе меньше r, то отбросить этот набор и продолжить перебор.

(ii) По окончании перебора перейти к шагу 7.

(iii) В противном случае перейти к шагу 6.

Шаг 6.

(i) Если значение целевой функции на исследуемом наборе больше фильтра z^* или не выполняется хотя бы одно из ограничений (2), то отбросить этот набор и продолжить перебор.

(ii) Если значение целевой функции z' на исследуемом наборе меньше фильтра z^* и выполняются все ограничения (2), то $z^* = z'$ и продолжить перебор.

(iii) По окончании перебора перейти к шагу 7.

Шаг 7.

Набор ξ, на котором достигается z^*, является оптимальным. Решение найдено.

Остановка.

Модифицированный метод Балаша с фильтром целесообразно рассмотреть на примере [4].

Пример. Решить задачу минимального покрытия для исходных данных, заданных следующей матрицей C

$$C = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0{,}5 & 0{,}5 & 0{,}33 & 0{,}33 & 0{,}5 \end{pmatrix}$$

Соответствующая задача линейного программирования имеет вид:

$$\left.\begin{array}{ll} z = 0{,}5\xi_1 + 0{,}5\xi_2 + 0{,}33\xi_3 + 0{,}33\xi_4 + 0{,}5\xi_5 \to \min; & (0) \\ \xi_1 + \xi_2 + \xi_4 \geq 1; & (1) \\ \xi_3 + \xi_4 \geq 1; & (2) \\ \xi_2 + \xi_3 \geq 1; & (3) \\ \xi_4 + \xi_5 \geq 1; & (4) \\ \xi_3 + \xi_4 + \xi_5 \geq 1; & (5) \\ \xi[0;1]. & \end{array}\right\} \quad (5)$$

Процесс поиска начального допустимого решения показан на рисунке, а в таблице даны результаты поиска оптимального решения задачи (5).

	S_1	S_2	S_3	S_4	S_5
x_1	1	1		1	
x_2			1	1	
x_3		1	1		
x_4	1				1
x_5			1	1	1
c_j	0,5	0,5	0,33	0,33	0,5

$\xi_3 = 1$

	S_1	S_2	S_4	S_5
x_1	1	1	1	
x_4	1			1
$c_{\text{лок},j}$	1,0	1,5	1,33	1,5

$\xi_1 = 1$

$\xi_1 = 1, \xi_2 = 0, \xi_3 = 1, \xi_4 = 0, \xi_5 = 0$

$r = 2,\ z^* = 0{,}83$

Процесс поиска начального допустимого решения

Результаты поиска оптимального решения

Номер набора	Набор ξ					Количество единиц в наборе	Значение z'	Значения ограничений					Выполнение всех ограничений	Оптимальное решение
	ξ_3	ξ_1	ξ_2	ξ_4	ξ_5			(1)	(2)	(3)	(4)	(5)		
24	1	1	0	0	0	2	0,83	–	–	–	–	–	Да	*
23	1	0	1	1	1	4	1,66	–	–	–	–	–	Нет	
22	1	0	1	1	0	3	1,13	–	–	–	–	–	»	
21	1	0	1	0	1	3	1,33	–	–	–	–	–	»	
20	1	0	1	0	0	2	–	–	–	–	–	–	»	
19	1	0	0	1	1	3	1,13	–	–	–	–	–	»	
18	1	0	0	1	0	2	–	–	–	–	–	–	»	
17	1	0	0	0	1	2	–	–	–	–	–	–	»	
16	1	0	0	0	0	1	–	–	–	–	–	–	»	
15	0	1	1	1	1	4	1,83	–	–	–	–	–	»	
14	0	1	1	1	0	3	1,33	–	–	–	–	–	»	
13	0	1	1	0	1	3	1,5	–	–	–	–	–	»	
12	0	1	1	0	0	2	–	–	–	–	–	–	»	
11	0	1	0	1	1	3	1,33	–	–	–	–	–	»	
10	0	1	0	1	0	2	–	–	–	–	–	–	»	
9	0	1	0	0	1	2	–	–	–	–	–	–	»	
8	0	1	0	0	0	1	–	–	–	–	–	–	»	
7	0	0	1	1	1	3	1,33	–	–	–	–	–	»	
6	0	0	1	1	0	2	–	–	–	–	–	–	»	
5	0	0	1	0	1	2	–	–	–	–	–	–	»	
4	0	0	1	0	0	1	–	–	–	–	–	–	»	
3	0	0	0	1	1	2	–	–	–	–	–	–	»	
2	0	0	0	1	0	1	–	–	–	–	–	–	»	
1	0	0	0	0	1	1	–	–	–	–	–	–	»	
0	0	0	0	0	0	0	–	–	–	–	–	–	»	

Таким образом, в рассмотренном примере вместо $2^5 = 32$ наборов исследовано только 9.

Применение модифицированного алгоритма позволяет сократить количество зондируемых ветвей на величину Δ

$$\Delta = 2^{n-q} + \sum_{k=1}^{r} C_n^k, \tag{6}$$

где n – количество переменных в целевой функции; r – нижняя оценка количества единиц в оптимальном решении; q – количество базисных переменных в начальном допустимом решении.

Таким образом, для нахождения оптимального решения при применении предложенной модификации алгоритма Балаша необходимо из 2^n ветвей проверить только $K_{\text{зонд}}$

$$K_{\text{зонд}} = 2^n - \Delta = 2^{n-q}(2^q - 1) - \sum_{k=1}^{r} C_n^k. \tag{7}$$

Список литературы

1. Рейнгольд, Э. Комбинаторные алгоритмы. Теория и практика / Э. Рейнгольд, Ю. Нивергельт, Н. Део ; пер. с англ. Е.П. Липатова ; под ред. В.Б. Алексеева. – М. : Мир, 1980. – 480 с.

2. Кофман, А. Методы и модели исследования операций. Целочисленное программирование : пер. с франц. / А. Кофман, А. Анри-Лабордер. – М. : Мир, 1977. – 432 с.

3. Кофман, А. Введение в прикладную комбинаторику : пер. с англ. / А. Кофман. – М. : Наука, 1975. – 480 с.

4. Клименко, И.В. Метод формальной нормализации отношений реляционной модели / И.В. Клименко, А.В. Лозовский // Науч. мысль Кавказа. Прил. № 5. – 2004. – С. 115–119.

Modification of Balash Algorithm to Solve the Problem of Relational Databases Normalization

I.V. Klimenko

Rostov State University of Railways, Rostov-on-Don

Key words and phrases: axioms of functional dependences output; Balash algorithm; least covering problem; maximal independent transitive sets; normal forms; primary key; relational model normalization; functional dependences.

Abstract: The paper solves the topical problem of the modification of known algorithms for solving problems of optimal coverage taking into account the specifics of the relational model. In particular, to solve the least covering problem we have modified the well-known Balash algorithm. Numerical experiment shows a significant advantage of the modified algorithm, expressed in the acceleration of the flexing procedure.