**Библиографический список**

1. Орлов, С. Технологии разработки программного обеспечения / С. Орлов. СПб. : Питер, 2002.

2. Поздняков, Д. А. Компонентная программная архитектура мультиверсионных систем обработки информации и управления : автореф. дис. … канд. техн. наук / Д. А. Поздняков. Красноярск, 2006.

3. Липаев, В. В. Проектирование программных средств : учеб. пособие / В. В. Липаев. М. : Высш. шк., 1990.

4. Avizienis, A. On the Implementation of N-Version Programming for Software Fault Tolerance During Execution / A. Avizienis, L. Chen // Proc. of the IEEE COMPSAC'77. Chicago, USA, 1977. P. 149–155.

5. Randell, B. The Evolution of the Recovery Block Concept / B. Randell, J. Xu // Software Fault Tolerance / ed. M. R. Lyu. New York : John Wiley&Sons, 1995. P. 1–21.

6. Predictably Dependable Computing Systems / ed. B. Randell [et al.]. Berlin : Springer-Verlag, 1995.

7. The Methodology of N-Version Programming / A. Avizienis // Software Fault Tolerance / ed. M. R. Lyu. New York : John Wiley & Sons, 1995. P. 22–46.

8. Scott, R. K. Fault-Tolerant Software Reliability Modeling / R. K. Scott, J. W. Gault, D. F. McAllister // IEEE Transactions on Software Engineering. 1987. Vol. SE-13, № 5. P. 582–592.

9. Письман, Д. М. GERT-сетевой анализ временных характеристик работы узлов распределенных систем обработки информации : автореф. дис. … канд. техн. наук / Д. М. Письман. Красноярск, 2006.

10. Филлипс, Д. Методы анализа сетей / Д. Филлипс, А. Гарсиа-Диас. М. : Мир, 1984.

P. V. Kovalev, A. N. Laykov, S. N. Gritsenko

## THE RELIABILITY RESEARCH OF N-VERSION SOFTWARE USING METHODS OF NETWORK ANALYSIS

*The technique of the reliability research of N-version software is offered. It allows using algorithms and methods of network analysis for the reliability research of software, developed by N-version approach using.*

*Keywords: reliability, N-version, networks, estimation methods, network analysis.*

УДК 004.421.4

К. В. Бадмаева

## АЛГОРИТМ ОЦЕНКИ РЕЛЕВАНТНОСТИ ПРЕДСТАВЛЕНИЙ ДЛЯ МАТЕРИАЛИЗАЦИИ В СПЕЦИАЛИЗИРОВАННОМ ХРАНИЛИЩЕ ДАННЫХ*

*Рассмотрено проектирование специализированных хранилищ данных при отсутствии статистической информации о работе базы данных. Предложен алгоритм оценки релевантности представлений на основе данных предметной области, который позволяет принимать решение о включении в схему хранилища агрегированных данных. Алгоритм предназначен для уменьшения субъективности проектировщиков при создании эффективной первоначальной модели данных хранилища.*

*Ключевые слова: хранилище данных, материализация представлений.*

Одной из наиболее важных задач проектирования информационных хранилищ данных является задача выбора представлений для материализации. Материализованными представлениями называются собранные в базе данных (БД) сводные таблицы, полученные выбором и агрегированием данных из других таблиц. Существующие алгоритмы выбора представлений основываются на статистике, собранной по полям таблиц хранилища сервером БД в процессе эксплуатации системы [1–4]. Поэтому разработка методов, способствующих повышению производительности хранилища на ранних стадиях проектирования, когда статистика еще отсутствует, актуальна и востребованна.

При проектировании хранилищ необходимо решить проблему выбора представлений, которые наиболее целесообразно включить в схему данных при наличии ограничений на системные ресурсы. В общем случае задача выбора относится к классу NP-трудных [5], и для ее решения применяются эвристические методы.

Автором данной статьи предлагается оценивать релевантность, т. е. важность с точки зрения пользователей хранилища, представлений на основе исследования предметной области, определения целей, задач и наборов методов анализа данных еще до того, как в базе данных будет накоплена достаточная статистика. Определение ре-

* Работа выполнена при поддержке гранта Президента РФ для ведущих научных школ № НШ-3431.2008.9.

левантности материализованных представлений производится с помощью разработанного автором алгоритма, основанного на системном анализе целей и задач, которые решаются информационным хранилищем данных. Применение этого алгоритма позволяет принимать решения о включении в схему БД агрегированных данных, наиболее выгодных для использования системных ресурсов, на ранних этапах разработки хранилища.

**Постановка задачи оценки представлений.** Для решения задачи выбора представлений для материализации при исследовании предметной области требуется определить набор расчетных методик $M = \{M_1, ..., M_N\}$, предназначенных для анализа данных. Методики включают правила вычисления отдельных показателей или семантически связанных групп показателей.

Поставим каждой методике в соответствие числовое значение $S(M_i)$, отражающее оценку ее семантической важности для данной предметной области. Это можно выполнить, используя методы экспертных оценок [6].

Частоту использования данных в хранилище $F(M_i)$ можно оценить по временному периоду обновления результатов методик, связанному с поступлением новых данных. Для этого необходимо определить наибольший период обновления результатов методик $T$, характерный для данной предметной области, и рассчитать частоту применения результатов каждой методики за этот период, получаемую методами экспертных оценок [6].

Для $M = \{M_i\}$, $i = \overline{1, N}$, где $M$ – множество методик предметной области; $N$ – количество методик, семантическая важность каждой из этих методик определяется как $S(M_i) \in [0; \sigma]$, где $\sigma$ – некоторое максимальное значение, соответствующее наибольшей важности, а частота применения каждой методики – как $F(M_i) \in \left(0; \frac{1}{T}\right]$, где $T$ – наибольший период времени обновления результатов методик (пополнения данных).

Выполнив анализ и систематизацию экспертных оценок для расчетных методик предметной области, представления для материализации можно оценить на основе алгоритма релевантности. Входными данными этого алгоритма является множество методик предметной области с рассчитанной семантической важностью и частотой использования:

$$M = \{M_i\}, i = \overline{1, N}; \quad S(M_i) \in [0; \sigma]; \quad F(M_i) \in \left(0; \frac{1}{T}\right].$$

Множество методик разбивается на классы по частоте использования. Количество классов $K$ для разбиения методик выбирается на основе исследования предметной области или с помощью экспертных оценок [6]. Выходные данные алгоритма – это множество представлений $M^{CV}$, упорядоченных по релевантности.

В алгоритме определения релевантности представлений используются следующие основные обозначения:

– $M = \{M_i\}$ – множество методик предметной области, $i = \overline{1, N}$;

– $N$ – количество методик предметной области;

– $S(M_i)$ – семантическая важность $i$-й методики;

– $\sigma$ – максимальное значение семантической важности;

– $F(M_i)$ – частота применения $i$-й методики;

– $T$ – наибольший период времени обновления результатов методик (пополнения данных);

– $K$ – количество классов разбиения методик;

– $P = \{P_1, \ldots, P_K\}$ – множество классов частоты использования методик;

– $s$ – количество уровней семантической важности методик;

– $S = \{1, ..., s\}$ – множество уровней семантической важности;

– $R_M = \{R_1^{\,1}, R_1^{\,2}, ..., R_1^{\,s}, \ldots, R_K^{\,1}, R_K^{\,2}, ..., R_K^{\,s}\}$ – множество классов методик, упорядоченное по релевантности на основе частоты использования и семантической важности;

– $R_M^i = \{R_i^{\,1}, R_i^{\,2}, ..., R_i^{\,s}\}$ – множество $i$-го класса частот методик, упорядоченное по релевантности на основе семантической важности;

– $A = \{A_1, ..., A_K\}$ – множество атрибутов классов методик;

– $A_i$ – множество атрибутов, используемых в методиках $R_M^i$;

– $a \in A$ – атрибут расчетной методики;

– $M^A$ – множество классов атрибутов, построенное по заданному алгоритму;

– $M^{CV}$ – множество кандидатов представлений на основе $M^A$.

**Пошаговое описание алгоритма оценки релевантности представлений.** Для построения упорядоченного по релевантности множества представлений выполняются следующие шаги.

*Шаг 1 – разбить множество расчетных методик на классы по частоте их использования.*

На этом шаге определяются правила разбиения методик на классы по частоте их использования. Если $K$ – количество классов разбиения, на которое разбивается интервал частоты применения методик $\left(0; \frac{1}{T}\right] = (0; P_1] \cup (P_1; P_2] \cup \ldots \cup \left(P_{K-1}; \frac{1}{T}\right]$, каждый из которых определяет отдельный класс, то

$$M_i \in P_1, \text{ если } F(M_i) \in (0; P_1],$$

$$M_i \in P_j, \text{ если } F(M_i) \in (P_{j-1}; P_j],$$

$$M_i \in P_K, \text{ если } F(M_i) \in \left(P_{K-1}; \frac{1}{T}\right],$$

где $P = \{P_1, \ldots, P_j, \ldots, P_K\}$ – множество классов частот использования методик, $i = \overline{1, N}$, $j = \overline{2, K-1}$.

*Шаг 2 – сформировать упорядоченное по релевантности множество методик.*

На этом шаге формируются классы методик по частоте и семантической важности. Входные данные: $M = \{M_i\}$, $S(M_i) \in [0; \sigma]$, $F(M_i) \in \left(0; \frac{1}{T}\right]$, $i = \overline{1, N}$, $K$ – количество классов разбиения. Требуется выделить на интервале $[0; \sigma]$ $s$ уровней:

$$[0; \sigma] = \left[\begin{array}{l} 0;\ \text{level}1) \cup [\text{level}1; \text{level}2) \cup \\ \cup \ldots \cup [\text{level}(s-1); \sigma \end{array}\right].$$

Количество уровней $s$ и границы их интервалов определяются экспертами на этапе анализа предметной области. В первую очередь необходимо учитывать элементы, имеющие наибольшую оценку семантической важности.

Правила разбиения, которые определяют классы методик по уровням семантической важности и классам частот использования, задаются следующим образом:

$$M_i \in R_j^{\,1}, \text{ если } F(M_i) \in P_j \text{ и } S(M_i) \in [level(s-1); \sigma]$$

$M_i \in R_j^{\,2}$, если $F(M_i) \in P_j$ и
$S(M_i) \in \left[\text{level}(s-2); \text{level}(s-1)\right)$,
$M_i \in R_j^{\,s}$, если $F(M_i) \in P_j$ и
$S(M_i) \in [0; \text{level}1)$ $j = \overline{1, K}$.

Выходные данные: $R_M = \{R_1^{\,1}, R_1^{\,2}, ..., R_1^{\,s}, \ldots, R_K^{\,1}, R_K^{\,2}, ..., R_K^{\,s}\}$ – множество классов методик, упорядоченных по уменьшению релевантности и частоте использования. Пошаговое построение $R_M$ с промежуточным выделением его элементов позволит выполнять перерасчет множества релевантных представлений за счет изменения или исключения из рассмотрения отдельных уровней семантической важности.

*Шаг 3 – построить множество классов атрибутов методик.*

На этом шаге для определения множеств атрибутов, участвующих в формировании методик, элементы множества $R_M$ группируются по частоте использования: $R_M^1 = \{R_1^{\,1}, R_1^{\,2}, ..., R_1^{\,s}\}$, …, $R_M^s = \{R_K^{\,1}, R_K^{\,2}, ..., R_K^{\,s}\}$, всего $K$ групп. Одни и те же атрибуты могут входить в различные методики. Для определения множества классов релевантных атрибутов $M^A$ предлагается использовать алгоритм исключения (рис. 1).

Этот алгоритм заключается в том, что элементы множества атрибутов, участвующих в формировании наиболее важных методик, вычисляются по формуле

$$M_1^A = A_1,\ M_j^A = A_j \setminus \bigcup_l A_l,$$

где $l = \overline{1, (j-1)}$, $j = \overline{2, K}$; $A_j$, $A_l$ – множества атрибутов методик $R_M^j$ и $R_M^l$ соответственно. Количество элементов множества $M^A$ совпадает с количеством классов методик $K$, заданным экспертным путем: $M^A = \{M_i^A\}$, где $i = \overline{1, K}$.

*Шаг 4 – упорядочить по релевантности множество представлений*.

На этом шаге частота и релевантность представлений определяется входящими в них атрибутами. На основе множества $M^A$ формируется множество кандидатов представлений $M^{CV} = \{M_i^{CV}\}$, где $M_i^{CV}$ – множество всевозможных сочетаний элементов $M_i^A \cup M_{i-1}^A$.

Полученное множество $M^{CV}$ кандидатов представлений, упорядоченное по релевантности, может использоваться для повышения эффективности модели хранилища данных.

**Пример применения алгоритма оценки релевантности представлений.** Пусть при исследовании предметной области определены множество методик $M = \{M_i\}$, $i = \overline{1, 15}$, семантическая важность каждой методики $S(M_i) \in [0; \sigma]$ и частота использования $F(M_i) \in \left(0; \frac{1}{T}\right]$, где $y = 10$; $T = 12$ месяцев. Количество уровней важности $s = 2$, граница уровней находится в $y / 2$, количество классов частот использования методик $K = 3$. Требуется разбить $M$ на классы по частоте и семантической важности (рис. 2).

Для выделения семантически важных элементов разобьем интервал $[0; \sigma]$ на две части (по значению уровней важности $s$): $[0; \sigma] = [0; \sigma/2) \cup [\sigma/2; \sigma]$. Методики, значение семантической важности которых попадает в область $[\sigma/2; \sigma]$ относятся к первому уровню, в область $[0; \sigma/2)$ – ко второму.

Выполним разбиение отрезка $\left(0; \frac{1}{T}\right]$ на классы. Например, методики, используемые ежеквартально и ежемесячно, относятся к классу $P_1$, используемые примерно два раза в год, – к классу $P_2$, используемые реже чем два раза в год, – к классу $P_3$.

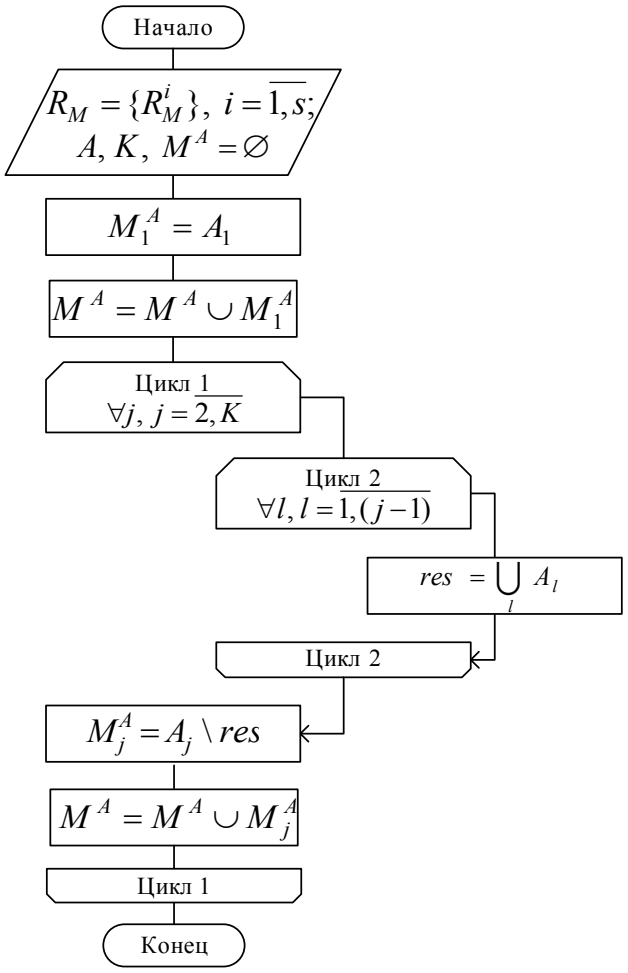


Рис. 1. Блок-схема алгоритма исключения

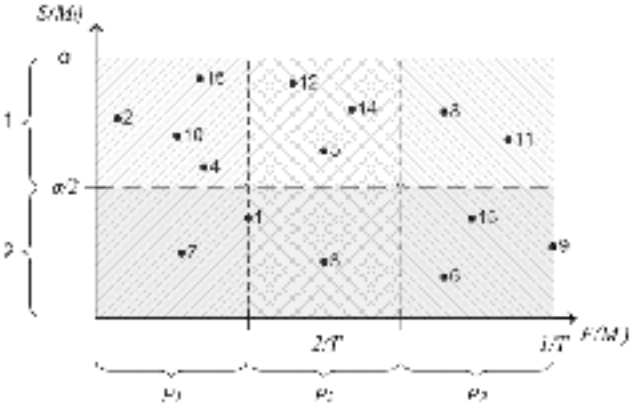

Рис. 2. Разбиение методик на классы по семантической важности (ось абсцисс) и частоте использования (ось ординат) (методики обозначены точками с указанием порядкового номера)

Правила формирования множества релевантных элементов для данного примера будут следующими:

$M_i \in R_j^{\,1}$, если $F(M_i) \in P_j$ и $S(M_i) \in [\sigma/2;\sigma]$,

$M_i \in R_j^{\,2}$, если $F(M_i) \in P_j$ и $S(M_i) \in [0;\sigma/2)$; $j = \overline{1,K}$.

Сформируем упорядоченное по релевантности множество классов:

$$R_M = \{R_1^{\,1}, R_1^{\,2}, R_2^{\,1}, R_2^{\,2}, R_3^{\,1}, R_3^{\,2}\} =$$
$$= \begin{Bmatrix}(M_2, M_4, M_{10}, M_{15}); (M_1, M_7); (M_5, M_{12}, M_{14});\\ (M_3); (M_8, M_{11}); (M_6, M_9, M_{13})\end{Bmatrix}.$$

Для определения множеств атрибутов, участвующих в формировании методик, сгруппируем элементы множества $R_M$ по частоте использования: $R_M^1 = \{R_1^1, R_1^2\}$, $R_M^2 = \{R_2^1, R_2^2\}$, $R_M^3 = \{R_3^1, R_3^2\}$.

Сформируем множество релевантных атрибутов по алгоритму исключения (рис. 3).

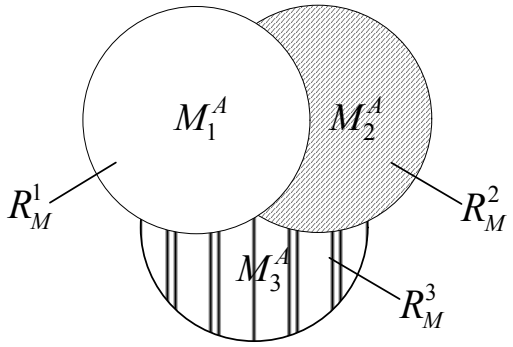


Рис. 3. Пример алгоритма исключения

Определим множество классов $M^A$, включающее множества атрибутов, упорядоченных по убыванию релевантности. Упорядоченные множества содержат атрибуты множеств методик, исключая уже учтенные атрибуты, и число полученных множеств совпадает с количеством классов методик: $M_1^A = A_1$, $M_2^A = A_1 \setminus A_2$, $M_3^A = A_3 \setminus (A_1 \cup A_2)$. Итак, множество $M^A = \{M_1^A, M_2^A, M_3^A\}$ для формирования кандидатов представлений $M^{CV}$ построено.

Таким образом, предложенный в данной статье алгоритм оценки релевантности представлений, использующий данные предметной области, позволяет принимать решения о включении в схему хранилища агрегированных данных. Множество представлений, полученное в результате выполнения алгоритма, может использоваться для повышения эффективности модели хранилища данных на начальном этапе его разработки. Дальнейший выбор представлений для материализации предполагается выполнять с учетом ограничения дискового пространства и минимизации общей стоимость материализации, состоящей из стоимости обслуживания и времени ответа на запросы к базе данных [7].

Апробация рассмотренных методов и алгоритмов выполнена при реализации хранилища медико-демографических данных и специализированного хранилища показателей социально-экономического развития региона [8; 9].

**Библиографический список**

1. Theodoratos, D. A general framework for the view selection problem for data warehouse design and evolution / D. Theodoratos, M. Bouzeghoub // Proc. of the 3rd ACM Intern. Workshop on Data Warehousing and OLAP. New York, 2000. P. 1–8.

2. ROLAP implementations of the data cube / K. Morfonios, S. Konakas, Y. Ioannidis, N. Kotsis // ACM Comput. Survey. 2007. Vol. 39, № 4. P. 12.

3. Efficient approaches for materialized views selection in a data warehouse / M. Hung, M. Huang, D. Yang, N. Hsueh // Inform. Sci. 2007. № 177. P. 1333–1348.

4. Gupta, H. Selection of Views to Materialize in a Data Warehouse / H. Gupta, F. N. Afrati, P. G. Kolaitis // Proc. of the 6th intern. Conf. on Database Theory. London : Springer-Verlag, 1997. P. 98–112.

5. Golfarelli, M. View Materialization for Nested GPSJ Queries / M.Golfarelli, S.Rizzi // Proc. of the Intern. Workshop on Design and Management of Data Warehouses. Stockholm, Sweden, 2000. P. 6.

6. Бешелев, С. Д. Математико-статистические методы экспертных оценок / С. Д. Бешелев, Ф. Г. Гурвич. 2-е изд., перераб. и доп. М. : Статистика, 1980.

7. Бадмаева, К. В. Формирование стоимостной модели для проектирования хранилищ данных / К. В. Бадмаева // Современные информационные технологии в науке, образовании и практике : материалы VII Всерос. науч.-практ. конф. с междунар. участием / Оренбург. гос. ун-т. Оренбург, 2008. С. 332–341.

8. Исаева, О. С. Проблемы построения специализированного хранилища демографических данных / О. С. Исаева, К. В. Шалдыбина // Вестник КрасГУ. Серия «Физико-математические науки». 2006. № 1. С. 222–227.

9. Пенькова, Т. Г. Проектирование специализированного хранилища показателей социально-экономического развития региона / Т. Г. Пенькова, Л. Ф. Ноженкова, К. В. Шалдыбина // Вестник КрасГАУ. 2007. № 14. С. 122–128.

K. V. Badmaeva

## THE RELEVANCE ESTIMATION ALGORITHM REPRESENTATIONS FOR MATERIALIZATION IN SPECIALIZED DATA WAREHOUSE

*In the paper specialized data warehouses design at the absence of the statistical information about database work is considered. The relevance estimation algorithm of representation, on the basis of data domain is offered. It allows decision making about aggregated data including the data warehouse scheme. The algorithm is intended to reduce a designer's subjectivity at the development of effective initial model of data warehouse.*

*Keywords: data warehouse, representations materialization.*