

# Выбор представлений для материализации в специализированном хранилище

К.В. Бадмаева

**Аннотация.** Хранилище данных рассматривается как объект адаптации. Предложен эвристический алгоритм выбора представлений для материализации в специализированном хранилище. Основным преимуществом алгоритма является возможность его применения для формирования начальной модели данных хранилища и для ее дальнейшего развития при отсутствии накопленной статистической информации о работе хранилища. Обоснована практическая значимость алгоритма на примере медико-демографических данных.

**Ключевые слова:** Хранилище данных, адаптация, материализация представлений.

## Введение

Хранилище данных определяют как предметно-ориентированный, интегрированный, поддерживающий хронологию, неизменяемый набор данных, организованный с целью поддержки управления [1]. Под предметной ориентированностью понимается наличие данных, характеризующих конкретную проблемную область. Интегрированность подразумевает объединение исходных данных из различных источников, при этом данные проверяются, очищаются и приводятся к единому виду. Данные в хранилище всегда связаны с определенным периодом времени, что хронологически упорядочивает информацию. Неизменяемость характеризуется тем, что данные, помещенные в хранилище, в дальнейшем не корректируются и не изменяются.

Технология хранилищ данных решает проблемы организации больших объемов информации, поступающей из разнородных источников. Использование хранилищ данных обеспечивает эффективное применение методов оперативной аналитической обработки для задач поддержки принятия решений в разных прикладных областях. Хранилища данных, содержащие информацию, предназначенную для исследования процес-

сов одной или нескольких задач предметной области, называются специализированными хранилищами данных. Специализированные хранилища данных могут являться подмножествами корпоративного хранилища данных или, в процессе своего развития, служить основой для более общего хранилища данных.

Хранилища данных являются развивающейся средой. В процессе эксплуатации хранилища возникает необходимость учета различных изменяющихся условий и факторов, влияющих на эффективность его функционирования. Отсутствие полной информации о состоянии этих дестабилизирующих факторов приводит к необходимости применения адаптации для обеспечения и поддержания необходимой производительности. Кроме того, наличие адаптации позволяет ослабить требования к процессу проектирования хранилища и тем самым упростить и удешевить этот трудоемкий и дорогой процесс.

## Применение адаптации в хранилище данных

Хранилище данных – это сложная система, обладающая рядом характерных, специфических черт. *Существенная зависимость от сре-*

ды. Хранилище данных как объект адаптации связано со средой, которую можно представить в виде потока решаемых задач. Поскольку главной целью хранилища является наилучшее удовлетворение запросов пользователей, то оно всегда должно находиться в строгом соответствии с требованиями среды. Адаптацию хранилища данных к изменяющимся условиям среды можно разделить на три уровня.

*Аппаратный уровень* предполагает адаптацию параметров и структуры технических средств (процессоры, модули основной памяти, устройства внешней памяти, средства телекоммуникаций и т.п.) и физических соединений между ними.

*Программный уровень* призван улучшать функционирование и взаимосвязь программного обеспечения и связан с выбранными СУБД и инструментальными средствами для разработки хранилища данных.

*Алгоритмический уровень* адаптации хранилища данных связан с адаптацией алгоритмов обработки информации с целью учета специфики решаемых в системе задач.

После определения наиболее полно отвечающих заданным целям структуры и параметров хранилища данных, можно добиться дополнительной минимизации времени отклика за счет формирования модели хранилища данных и использования различных стратегий повышения производительности обработки данных. Основные идеи сводятся к следующим способам [2]: использование многомерной модели данных, выделение активных и пассивных данных, секционирование таблиц для уменьшения их размера, параллельная обработка запросов и другие. Наряду с перечисленными стратегиями для сокращения времени получения информации конечным пользователем в хранилищах данных применяются материализованные представления [3-11].

### **Задача выбора представлений для материализации**

Материализация представлений является одной из локальных задач оптимизации, решаемых в процессе выполнения общей задачи адаптации хранилища данных. На практике не удается материализовать все возможные пред-

ставления из-за ограничений дискового пространства и необходимости поддержки представлений в актуальном состоянии. Поэтому задача состоит в выборе из множества представлений такой комбинации, которая бы минимизировала стоимость выполнения запросов и стоимость обслуживания при ограничении дискового пространства. Задача выбора представлений для материализации является одной из основных задач физического проектирования хранилищ данных [3-11].

Выбор представлений для материализации производится в условиях априорной неопределенности, особенно в процессе проектирования и развития хранилища данных. Решение о выборе конкретных представлений для материализации основывается на частотах выполнения запросов, использующих эти представления. Частоты определяются в процессе работы пользователя с информацией хранилища данных. Для решения задачи необходимо определить оценки частот выполнения запросов в хранилище данных, сформировать модель, позволяющую определить эффективность материализации представлений в рамках ограничений на дисковое пространство и трудозатраты, связанные с обслуживанием материализованных представлений.

Входными данными задачи выбора представлений для материализации являются: множество исходных отношений  $R = \{r_1, \dots, r_n\}$ , где  $n$  – количество отношений; множество запросов  $Q = \{q_1, \dots, q_m\}$  над  $R$ , где  $m$  – общее число запросов над множеством отношений; частота запросов  $f_{q_1}, \dots, f_{q_m}$ , частота обновления исходных данных  $f_{r_1}, \dots, f_{r_n}$ ; множество представлений  $MV = \{v_1, \dots, v_k\}$  над  $R$ , которые первоначально материализованы в хранилище,  $k$  – число представлений в хранилище. На этапе проектирования множество  $MV$  может быть пустым.

В качестве системы ограничений рассматриваются ограничения на время и ресурсы системы, требуемые для хранения и поддержания материализованных представлений в актуальном состоянии и минимизация времени ответа на запросы пользователя.

Выбор представлений в пространстве решений осуществляется на основе функций стои-

мости, которые формализуют сравнение между стоимостью запросов и стоимостью обслуживания представлений. Для формирования функций стоимости требуется статистическая информация о хранимых данных и их источниках. В общем случае стоимость материализации множества представлений  $MV$  складывается из стоимости обслуживания и стоимости выполнения запросов:

$$Cost(MV) = c_1CQ + c_2CM, \quad (1)$$

где  $c_1, c_2$  – весовые коэффициенты, которые определяет проектировщик хранилища данных;  $CQ$  – стоимость выполнения запросов на множестве материализованных представлений  $MV$ ;  $CM$  – стоимость обслуживания материализованных представлений. Задача заключается в формировании такого множества  $MV$ , которое бы минимизировало общую стоимость материализации представлений.

Областью определения функции общей стоимости материализации представлений является  $O(Cost(MV)) = [0; N]$ , где  $N$  – количество всех возможных представлений. Верхнюю границу области значений  $D(Cost(MV))$  можно обозначить суммой максимальных значений стоимостей выполнения запросов и обслуживания материализованных представлений:  $D(Cost(MV)) = [0; c_1CQ_{max} + c_2CM_{max}]$ , где  $CQ_{max}$  – наибольшая стоимость выполнения запросов на множестве материализованных представлений  $MV$ ,  $CM_{max}$  – наибольшая стоимость обслуживания материализованных представлений.

Для каждого сформированного некоторым образом множества  $MV$  требуется определить  $Cost(MV)$  в случае выполнения множества запросов  $Q$ . Решение этой задачи осуществляется разными способами. Для перезаписи запросов на множестве материализованных представлений используются нотации многомерной решетки [3-5] или мультиграфа [8, 9] и различные алгоритмы поиска решений по ним.

Сегодня существует множество алгоритмов решения задачи выбора представлений для материализации [3-10]. Кроме того, современные аналитические программные комплексы, предназначенные для разработки и поддержки хранилищ данных, имеют встроенные алгоритмы оптимизации структуры хранилища [14]. Однако основ-

ные вопросы о том, как настроить платформу хранилища данных на специфику решаемой задачи и как организовать модель данных при отсутствии накопленной статистической информации о работе хранилища, решает проектировщик. Поэтому актуальной задачей является разработка методов и алгоритмов, которые позволили бы эффективно сочетать автоматизированные средства с ручной настройкой, максимально полно используя информацию, полученную от специалистов предметной области.

### Формирование релевантных множеств представлений

Решение задачи выбора представлений на практике требует сокращения пространства поиска, размерности задачи, обобщения или исключения некоторых ограничений. Например, в [4] не учитывается стоимость обслуживания материализованных представлений. В [5] предлагается использовать не все множество запросов над исходными отношениями  $Q$ , а только часть, интересующую пользователя. Особенно актуальным выделение множества релевантных запросов является в специализированных хранилищах данных. В данной работе предлагается формирование релевантных множеств представлений на основе информации о предметной области и алгоритм последующего выбора из них представлений для материализации в хранилище.

Специализированные хранилища данных предназначены для решения одной или нескольких задач предметной области и включают заданный ряд расчетных методик, необходимых для проведения анализа  $M = \{M_i\}$ ,  $i = \overline{1, q_M}$ . Предложенный в работе алгоритм определения релевантных представлений основан на понятии «расчетная методика». Под расчетной методикой понимается совокупность атрибутов таблиц хранилища и правил расчета важного для анализа показателя:

$$M_i = \langle A_i, Pr_i \rangle, \quad (2)$$

где  $A_i \subseteq A$ ,  $i = \overline{1, s}$ ,  $A_i$  – множество атрибутов, участвующих в расчете  $M_i$  методики,  $A$  – множество атрибутов таблиц измерений и фактов хранилища данных,  $Pr_i$  – правила, по которым выполняется расчет результата по атрибутам.

Например, расчет методики половозрастной структуры смертности  $M_l$ , включающей в себя атрибуты  $A_l = \{\text{дата рождения, пол, тип поселения, дата смерти, причина смерти, число случаев смерти}\}$ , требует выполнения следующих операций:  $Pr_l = \{pr_1, pr_2, pr_3\}$ , где  $pr_1$  – определение возрастной группы по централизованным справочникам хранилища,  $pr_2$  – расчет случаев смерти от отдельных причин по заданным атрибутам,  $pr_3$  – расчет структуры смертности в процентах.

На этапе проектирования, с помощью расчетных методик, можно выделить наиболее релевантные данные, чтобы обеспечить к ним более высокую скорость доступа. Для этого требуется для каждой методики определить частоту ее использования за период  $T$ , семантическую важность, трудоемкость вычисления и выполнить классификацию по заданным критериям.

Период  $T$  определяется разработчиками хранилища совместно с экспертами предметной области и представляет собой максимальный период обновления данных в хранилище. Затем эксперты определяют, насколько часто они используют ту или иную методику за период  $T$ , и присваивают каждой из них некоторую семантическую важность. Трудоемкость вычисления методик оценивают разработчики хранилища данных.

Далее требуется разбить на классы множество расчетных методик предметной области по заданной семантической важности, частоте использования и трудоемкости вычисления. Границами классов при этом являются уровни семантической важности, уровни трудоемкости вычисления и временные периоды, которые можно выделить в периоде  $T$ . Диапазоны каждого из временных периодов, уровней важности и трудоемкости определяются разработчиками хранилища совместно с экспертами предметной области. Далее на основе методов классификации [12, 13] формируется множество классов расчетных методик, упорядоченное по уменьшению релевантности и частоты использования  $R^M = \{R_i^M\}$ ,  $i = \overline{1, K}$ .

Элементы множества атрибутов  $M^A$ , участвующие в формировании наиболее важных расчетных методик, могут быть определены по формуле  $M_1^A = A_1$ ,  $M_j^A = A_j \setminus \bigcup_{i=1}^{j-1} A_i$ , где

$l = \overline{1, (j-1)}$ ,  $j = \overline{2, K}$ ,  $A_j, A_l$  – множества атрибутов методик  $R_j^M$  и  $R_l^M$  соответственно. Релевантность представлений определяется входящими в них атрибутами. На основе множества  $M^A$  формируется множество представлений  $M^{CV} = \{M_i^{CV}\}$ , где  $M_i^{CV}$  – множество всевозможных сочетаний элементов  $M_i^A \cup M_{i-1}^A$  [10].

Использование упорядоченного по релевантности множества представлений  $M^{CV}$  при выборе представлений для материализации позволит учесть требования пользователей, сократить пространство поиска и получить решение без накопленной информации о работе хранилища данных.

### Пример применения алгоритма

Пусть при исследовании предметной области определено множество методик  $M = \{M_i\}$ ,  $i = \overline{1, 15}$ , определены семантическая важность каждой методики  $S(M_i) \in [0; \sigma]$  и частота использования  $F(M_i) = (0; 1/T]$ , где  $\sigma = 10$ , а  $T = 12$  месяцев. Количество уровней важности:  $s = 2$ , граница уровней находится в  $\sigma/2$ , количество классов частот использования методик:  $K = 3$ .

Разбиение методик на классы по семантической важности и частоте использования выполняется по заданным правилам. Для выделения семантически важных элементов интервал  $[0; \sigma]$  разбивается на 2 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.

Правила формирования множества релевантных элементов для данного примера:  $M_i \in R_j^1$ , 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  $S(M_i) \in [0; \sigma/2)$ ,  $j = \overline{1, K}$ .

По исходным данным формируется упорядоченное по релевантности множество классов:  $R_M = \{R_1^1, R_1^2, R_2^1, R_2^2, R_3^1, R_3^2\} = \{(M_2, M_4, M_{10}, M_{15}); (M_1, M_7); (M_5, M_{12}, M_{14}); (M_3); (M_8, M_{11}); (M_6, M_9, M_{13})\}$ . Для определения множеств атрибутов, участвующих в формировании методик, элементы множества  $R_M$  группируются по частоте использования:  $R_M^1 = \{R_1^1, R_1^2\}$ ,  $R_M^2 = \{R_2^1, R_2^2\}$ ,  $R_M^3 = \{R_3^1, R_3^2\}$ .

Далее определяется множество классов  $M^A$ , содержащее множества атрибутов, упорядоченных по убыванию релевантности. Упорядоченные множества содержат атрибуты множеств методик, исключая уже учтенные атрибуты; число полученных множеств совпадает с количеством классов методик:  $M_1^A=A_1$ ,  $M_2^A=A_1 \setminus A_2$ ,  $M_3^A=A_3 \setminus (A_1 \cup A_2)$ .

Построенное множество  $M^A = \{M_1^A, M_2^A, M_3^A\}$  используется для формирования кандидатов представлений  $M^{CV}$ .

### Эвристический алгоритм выбора представлений для материализации из релевантных подмножеств

В алгоритме выбора представлений для сокращения пространства поиска используются упорядоченные по релевантности подмножества представлений, сформированные на основе информации о предметной области. Алгоритм включает 24 шага и описан ниже.

Входными данными алгоритма являются: множество уже материализованных представлений в хранилище  $MV$  (может быть пустым); множество представлений кандидатов на материализацию, упорядоченное по релевантности  $M^{CV}$ ; объем дискового пространства  $D_F$ , выделенного для материализации представлений, и пространство, необходимое для материализации всех кандидатов  $D(M^{CV})$ . Другие обозначения, используемые в описании алгоритма:  $Cost(MV + V_j)$  – общая стоимость материализации при добавлении  $V_j$  представления к  $MV$ ;  $V \in M^{CV}$  – представление из множества  $M^{CV}$ ;  $D(V_j)$  – пространство, занимаемое представлением  $V_j$ . Множество превышающих выделенное дисковое пространство представлений –  $M_0^{CV}$ . В алгоритме используется показатель, позволяющий определить выгоду материализации в пространстве, занимаемом представлением  $MV$ , который предложен М.-Ч. Хунгом [6] и рассчитывается по формуле:

$$g_m(V_j) = (Cost(MV) - Cost(MV + V_j)) / D(V_j). (3)$$

### Алгоритм выбора представлений для материализации в специализированном хранилище данных

**Input:**  $MV, M^{CV}, D(M^{CV})$ .

**Output:**  $MV$ .

- 1: if  $D(M^{CV}) \leq D_F$  then
- 2:  $MV = MV \cup M^{CV}$ ;
- 3: end if
- 4: for all  $M_i^{CV} \in M^{CV}$  do
- 5: if  $D(MV_i^{CV}) < D_F$  then
- 6:  $MV = MV \cup M_i^{CV}$ ;
- 7:  $D_F = D_F - D(M_i^{CV})$ ;
- 8: else
- 9: if  $\min_{V \in M_i^{CV}} D(V) < D_F$  then
- 10: for all  $V_j \in M_i^{CV}$
- 11: if  $D_F < D(V_j)$  then
- 12:  $M_0^{CV} = M_0^{CV} \cup V_j$ ;
- 13: else
- 14:  $g_m(V_j) = (Cost(MV) - (Cost(MV + V_j)) / D(V_j)$ ;
- 15: end if
- 16: if  $g_m(V_j) = \max_{V \in M_i^{CV}} g_m(V)$  and  $g_m(V_j) > 0$  and  $Cost(MV + V_{j-1}) > Cost(MV + V_j)$  then
- 17:  $MV = MV \cup V_j$ ;
- 18:  $D_F = D_F - D(V_j)$ ;
- 19: end if
- 20: end for
- 21:  $M_i^{CV} = M_i^{CV} / V_j$ ;
- 22: end if
- 23: end if
- 24: end for

Для расчета общей стоимости материализации предлагается использовать способ, описанный в работе [11]. Перебор представлений из  $M_i^{CV}$  продолжается, пока размер наименьшего из представлений рассматриваемого подмножества не будет превышать  $D_F$ . Расчет метрик прироста выполняется только для представлений, занимаемое пространство которых меньше, чем имеющееся в наличии дисковое пространство. К множеству  $MV$  добавляются представления с максимальным и положительным значением метрики прироста  $g_m(V_j)$  и в

случае минимизации общей стоимости материализации по сравнению с предыдущей итерацией. Результат считается достигнутым, когда множество  $MV$  сформировано или когда исчерпано доступное дисковое пространство.

Основным преимуществом алгоритма является возможность его применения для формирования начальной модели данных хранилища и для ее дальнейшего развития при отсутствии накопленной статистической информации о работе хранилища.

### Экспериментальные результаты

Применение предложенного в работе алгоритма выбора представлений рассмотрено на примере группы расчетных методик половозрастной смертности. Каждая методика включает обязательные для рассмотрения измерения «пол», «возраст» и факт «число случаев смерти». Узлы решетки представляют результат сгруппированных по заданным измерениям данных таблицы фактов. Для упрощения записи введены обозначения узлов в соответствии с входящими в них измерениями:  $Y$  – год,  $R$  – район,  $T$  – тип поселения,  $P$  – причина смерти,  $M$  – место смерти, обязательные измерения (пол, возраст) объединяются под одной буквой –  $S$ .

На примере половозрастной смертности исследуется выбор представлений при условии ограничения дискового пространства и минимизации общей стоимости материализации, используя оценку занимаемого представлением дискового пространства.

В таблице представлены результирующие множества представлений для материализации, полученные с помощью предложенного алгоритма и алгоритма [6] при ограничении дискового пространства, выделенного для материализованных представлений. Наиболее трудоемкой частью выбора представлений является вычисление общей стоимости материализации, поэтому для сравнения результатов работы рассчитано количество ее вычислений и количество циклов. Под циклом понимается выполнение части алгоритма, после которой происходит изменение множества материализованных представлений и требуется перерасчет использованного дискового пространства и определение дальнейших действий. Результи-

рующие множества представлений для материализации, полученные применением алгоритмов, совпадают, но для расчета с помощью предложенного алгоритма требуется меньше вычислений общей стоимости материализации и циклов.

При уменьшении выделенного дискового пространства полученные в результате применения алгоритмов множества материализованных представлений различны (таблица). В результате выполнения алгоритма [6] представление  $YRPS$ , интересующее пользователя, не материализовано. Предложенный алгоритм позволяет найти решение, которое лучше соответствует пользовательским требованиям, что является важным для практических приложений.

Сравнение результатов работы алгоритмов

Алгоритм	Выделенное дисковое пространство, Mb	Количество вычислений $Cost(MV)$	Количество циклов	Результат
Алгоритм [6]	1600	179	14	$MV=\{YRTPMS, TS, YS, YTS, RS, PS, YRS, TPS, YPS, YRMS, YTPS, YPMS, YTPMS, YRPS\}$
Предложенный алгоритм		33	8	$MV=\{YRTPMS, YTPS, YTS, YPS, YS, TPS, TS, PS, RS, YRS, YRPS, YRMS, YPMS, YTPMS\}$
Алгоритм [6]	1400	178	13	$MV=\{YRTPMS, TS, YS, YTS, RS, PS, YRS, TPS, YPS, YRMS, YTPS, YPMS, YTPMS\}$
Предложенный алгоритм		28	7	$MV=\{YRTPMS, YTPS, YTS, YPS, YS, TPS, TS, PS, RS, YRS, YRPS, YRMS, YPMS\}$

Трудозатраты, связанные с формированием релевантных множеств кандидатов для материализации, зависят от специфики предметной области. Преимуществом предложенного подхода является использование простых и понятных для экспертов приемов выделения релевантной информации. Общение с экспертами показало, что оценить важность методик им

легче, чем сравнить атрибуты разных методик между собой.

Использование предложенного подхода обеспечивает сокращение пространства поиска и позволяет получать решения, учитывая требования пользователей даже без накопленной статистической информации о работе хранилища данных.

Предложенный подход выделения множеств релевантных представлений кандидатов для материализации можно использовать в сочетании с другими алгоритмами поиска представлений в подмножествах и алгоритмами вычисления общей стоимости материализации. Решение, полученное на основе предложенного алгоритма, зависит от качества исходных экспертных оценок.

## Заключение

Рассмотрено хранилище данных как объект адаптации и выделены уровни адаптации к изменяющимся условиям среды. Предложен алгоритм для решения локальной задачи адаптации – материализации представлений. В предложенном алгоритме используются релевантные множества представлений, сформированные на основе информации о предметной области. Выполнено сравнение результатов применения предложенного алгоритма выбора представлений для материализации и алгоритма Хунга [6]. Показано, что для выполнения предложенного алгоритма требуется меньшее количество вычислений общей стоимости материализации. При усилении ограничения дискового пространства предложенный алгоритм позволяет найти решение, которое лучше соответствует требованиям пользователей. Предложенный алгоритм применим для повышения скорости получения информации конечным пользователем за счет включения в модель данных материализованных представлений при создании и развитии специализированных хранилищ, в случае изменения условий эксплуатации и появлении новой информации или дополнительного дискового пространства, независимо от наличия накопленной статистики о работе хранилища данных.

**Бадмаева Ксения Владимировна.** Научный сотрудник Института вычислительного моделирования СО РАН, г. Красноярск. Окончила Красноярский государственный технический университет в 2005 году. Кандидат технических наук. Автор 21 печатной работы. Область научных интересов: хранилища данных, оперативный анализ данных, адаптация сложных систем. E-mail: ksenya@icm.krasn.ru.

## Литература

1. Inmon W. Building the Data Warehouse / W. Inmon. – John Wiley & Sons, New York, 1992.
2. Спирли Э. Корпоративные хранилища данных. Планирование, разработка, реализация / Э. Спирли. – М.: Вильямс, 2001. – 400 с.
3. Horng, J.-T. Applying evolutionary algorithms to materialized view selection in a data warehouse / J.-T. Horng, Y.-J. Chang, B.-J. Liu // *Soft Computing*, Springer-Verlag. – 2003. – No.7. – P. 574-581.
4. Harinarayan, V. Implementing data cubes efficiently / V. Harinarayan, A. Rajaraman, J.D. Ullman // *Proc. of the 1996 ACM SIGMOD Inter. Conf. on Management of Data*, Quebec. – 1996. – P. 205-216.
5. Baralis, E. Materialized views selection in a multidimensional database / E. Baralis, S. Paraboschi, E. Teniente // *Proc. of the 23rd Inter. Conf. on Very Large Data Bases*, Eds. Very Large Data Bases. Morgan Kaufmann Publishers, San Francisco, CA. – 1997. – P. 156-165.
6. Hung, M. Efficient approaches for materialized views selection in a data warehouse / M. Hung, M. Huang, D. Yang, N. Hsueh // *Information Sciences* – 2007. – No. 177. – P. 1333-1348.
7. Morfonios, K. ROLAP implementations of the data cube / K. Morfonios, S. Konakas, Y. Ioannidis, N. Kotsis // *ACM Computing Surveys*, 2007. – Vol. 39. – No. 4. – Article 12. – 53 p.
8. 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 Inter. Workshop on Data Warehousing and OLAP*, New York. – 2000. – P. 1-8.
9. M. Golfarelli and S. Rizzi. View Materialization for Nested GPSJ Queries. In *Proc. of the Intl. Workshop on Design and Management of Data Warehouses*, pages 10-19, 2000.
10. Бадмаева К.В. Алгоритм оценки релевантности представлений для материализации в специализированном хранилище данных / *Вестник Сибирского государственного аэрокосмического университета им. академика Решетнева*. Выпуск 1(22). В 2 частях. Ч.2. – 2009. – С. 60-64.
11. Бадмаева К.В. Формирование стоимостной модели для проектирования хранилищ данных / *Материалы VII всероссийской научно-практической конференции «Современные информационные технологии в науке, образовании и практике» / ОГУ. – Оренбург, 2008. – С. 332-341.*
12. Классификация и кластер / Под ред. Дж. Вен Райзина. – М.: Мир, 1980. – 390с.
13. Рубан А.И. Методы анализа данных: в 2 ч. – Красноярск: КГТУ, 1994. – 220 с
14. Рынок платформ для Хранилищ данных: результаты исследования IDC [Электронный ресурс]: статья пер. с англ. – 2008 – Режим доступа: <http://citcity.ru/20054/>.