Приближенная обработка запросов с использованием вейвлет преобразований в OLAP системах.

Ухаров А.О.
oukharov@gmail.com

Московский Государственный Технический Университет им Н.Э. Бумана
Информационные технологии для эпидемиологии - ЭпиИТ

Московская секция ACM SIGMOD
Март 2017 г.

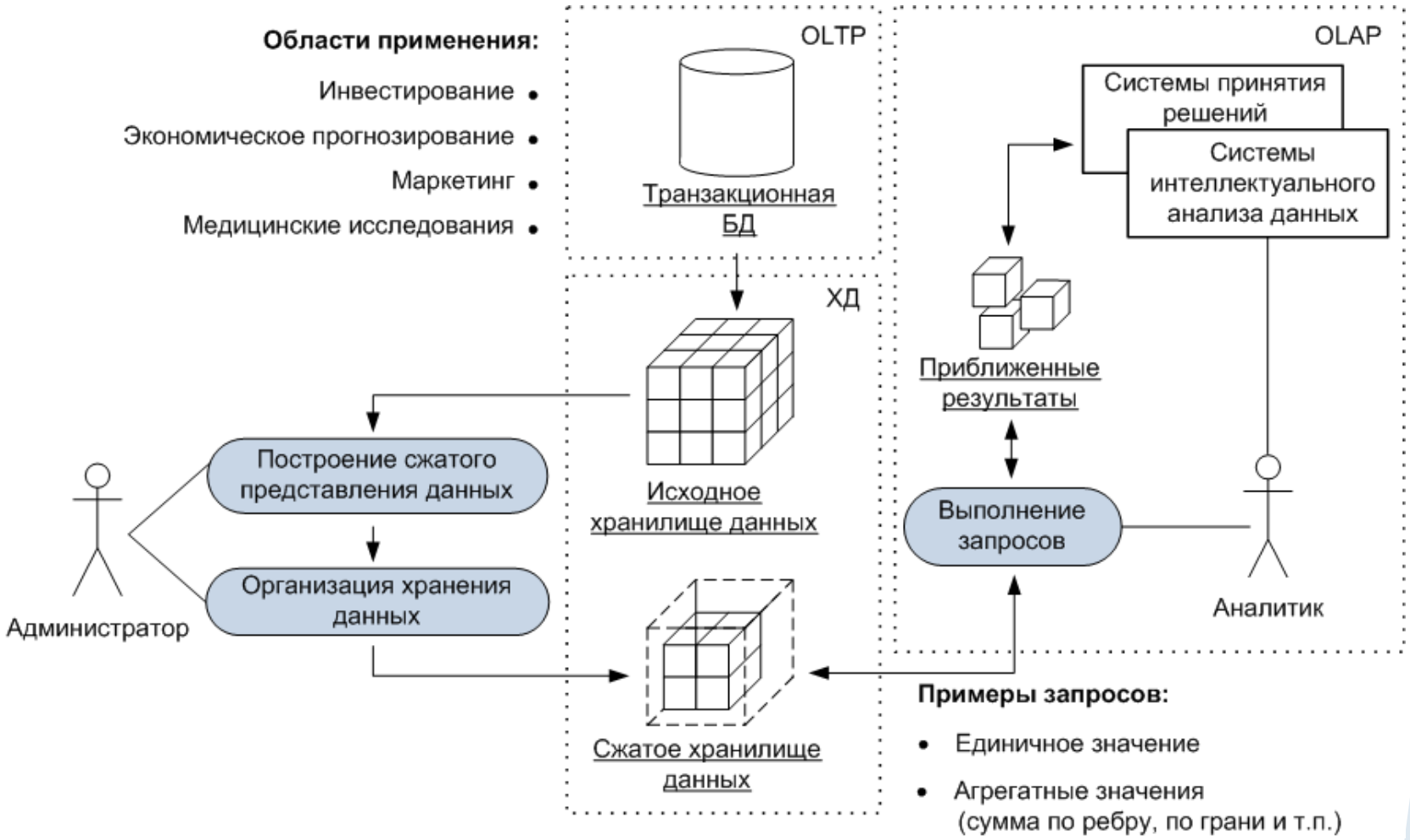
Содержание

- Приближенная обработка данных
- Принципиальные подходы к приближенной обработке данных в OLAP
  - Выборка
  - Гистограммы
  - Вейвлет-преобразования
- Проблемы вейвлет-преобразования многомерных данных
  - Массивы данных произвольной структуры
  - Сложность оценки точности приближенных вычислений
- Вейвлет-преобразование Хаара для многомерного хранилища данных
- Метод прямого и обратного вейвлет-преобразования Хаара для хранилища произвольной структуры
- Оценка погрешности приближенных вычислений
- Апробирование предложенного метода для анализа санитарно-эпидемиологической обстановки

Приближенная обработка. Зачем?

- Сжатие данных с потерями
- Исследовательская природа запросов, подразумевающая выявление зависимостей между данными
- Принципиально значимым является тенденция поведения и порядок исследуемой величины, а не ее абсолютная точность
- Необходимость обеспечения гибкости анализа на больших объемах данных при высоких требованиях к производительности системы

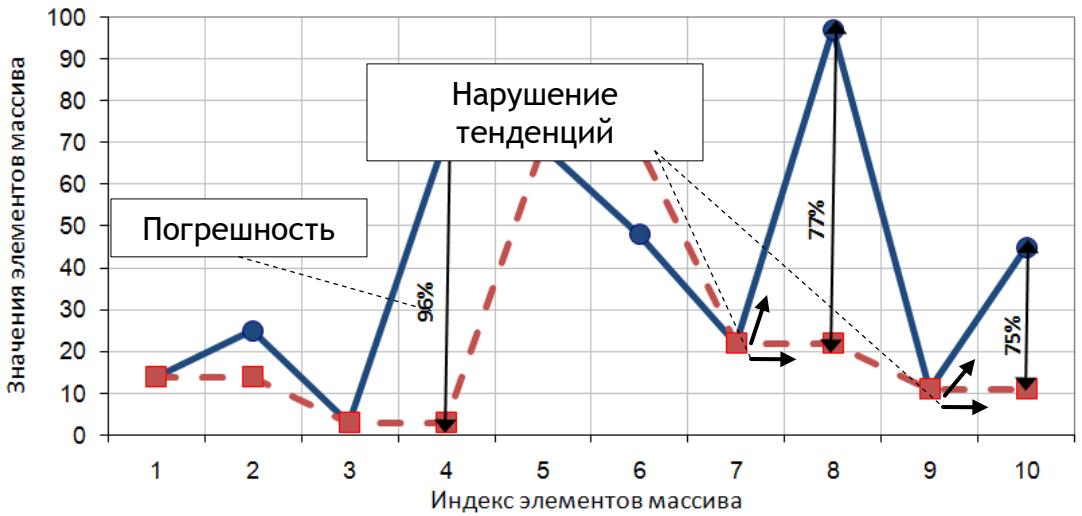
Приближенная обработка запросов в системах оперативного анализа данных

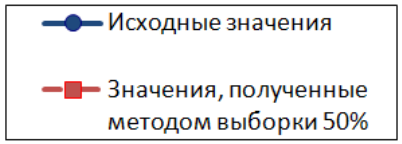


Приближенная обработка запросов методом выборки

Получение некоторого подмножества исходных значений, которое как можно ближе характеризует исходное полное множество значений.

- Низкая точность при вычислении единичных и суммарных значений с малым количеством элементов из-за отсутствии компенсации отбрасываемых значений.





Исходные данные	50% выборка	Приближенное представление данных
14	14	14
25		14
3	3	3
71		3
69	69	69
48		69
22	22	22
97		22
11	11	11
45		11

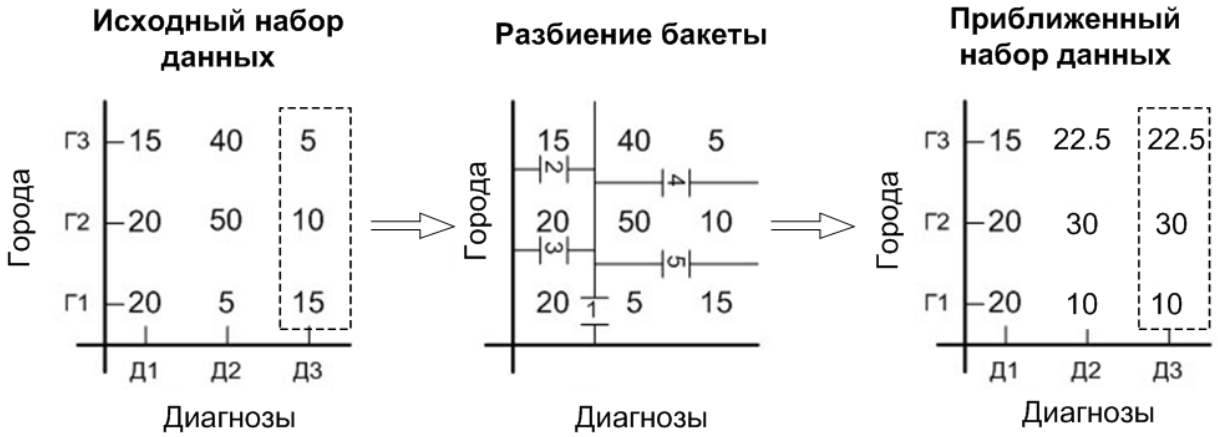
- Снижение точности и нарушение тенденций поведения величин в произвольных срезах с ростом размерности данных.

- Трудность прогнозирования погрешности вычислений в приближенном представлении данных при заданной степени сжатия и наоборот.

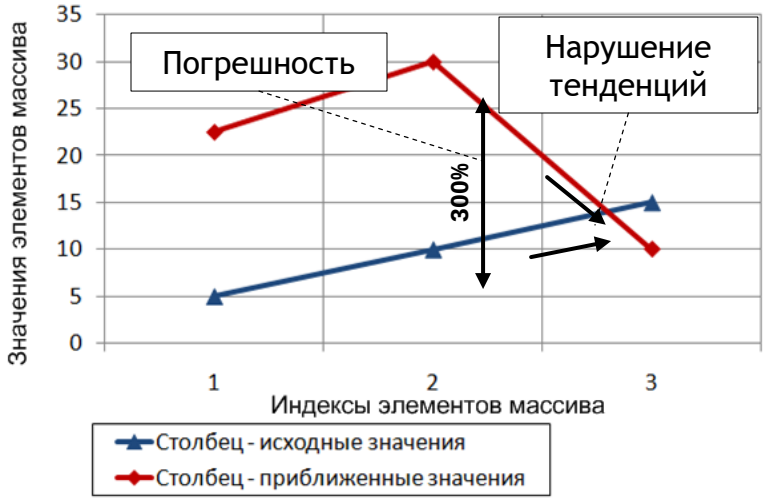
Приближенная обработка запросов методом гистограмм

Использование гистограммы исходного набора данных, т.е. разбиение исходного множества значений на группы (бакеты) на основе некоторых выбранных характеристик и замена каждого полученного бакета аппроксимированными значениями (например суммой значений элементов и их количеством).

- Потеря тенденций поведения величин внутри каждого бакета.



- Трудность прогнозирования погрешности вычислений в приближенном представлении данных при заданной степени сжатия и наоборот.



- Увеличение стоимости построения гистограмм и снижение точности вычислений с ростом размерности данных вследствие сложности разделения многомерного набора данных на непересекающиеся подмножества.

- Ориентированность на агрегатные значения с большим количеством элементов

Приближенная обработка запросов методом вейвлет-преобразования Хаара - 1

Прямое дискретное нестандартное вейвлет-преобразование Хаара (ДНВПХ)

Исходный набор данных (V)

3	2	4	3	5
2	7	5	1	3
1	9	1	2	2
0	3	5	7	9
	0	1	2	3

Нестандартное вейвлет-преобразование Хаара

3	4	-1	-1
6	2	1	-1
5	2	4	0
4	8	-1	-1

1.5	-1	1	0
-0.5	3	-2.5	-0.5
0.5	-0.5	0	-1
4.25	0.25	1.5	-0.5

1.5	-1	1	0
-0.5	3	-2.5	-0.5
4.5	3	0	-1
4.5	5	1.5	-0.5

1.5	-1	1	0
-0.5	3	-2.5	-0.5
3.75	0.75	0	-1
4.75	-0.25	1.5	-0.5

Нормализованная декомпозиция

3	-2	2	0
-1	6	-5	-1
2	-2	0	-2
17	1	3	-1

Сжатие 56% (9 из 16)

Среднеквадратическая погрешность

$$\xi = \sum_{i=1}^{N^*} c_i^2$$

Сжатая нормализованная декомпозиция (C)

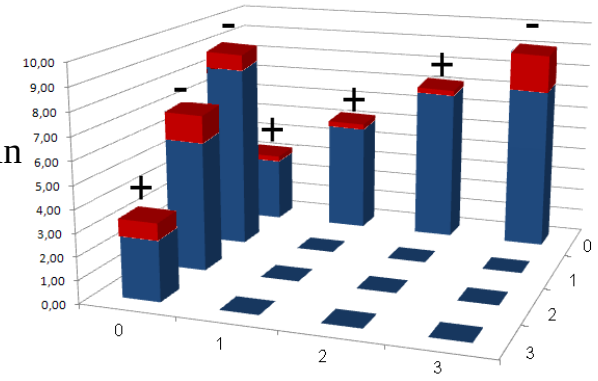
3	**0**	**0**	0	3
0	6	-5	**0**	2
0	**0**	0	**0**	1
17	**0**	3	**0**	0
	0	1	2	3

3	2,75	2,75	4,25	4,25
2	5,75	5,75	4,25	4,25
1	8,25	0,25	1,25	1,25
0	3,25	5,25	7,25	7,25
	0	1	2	3

Восстановленный набор данных (V')

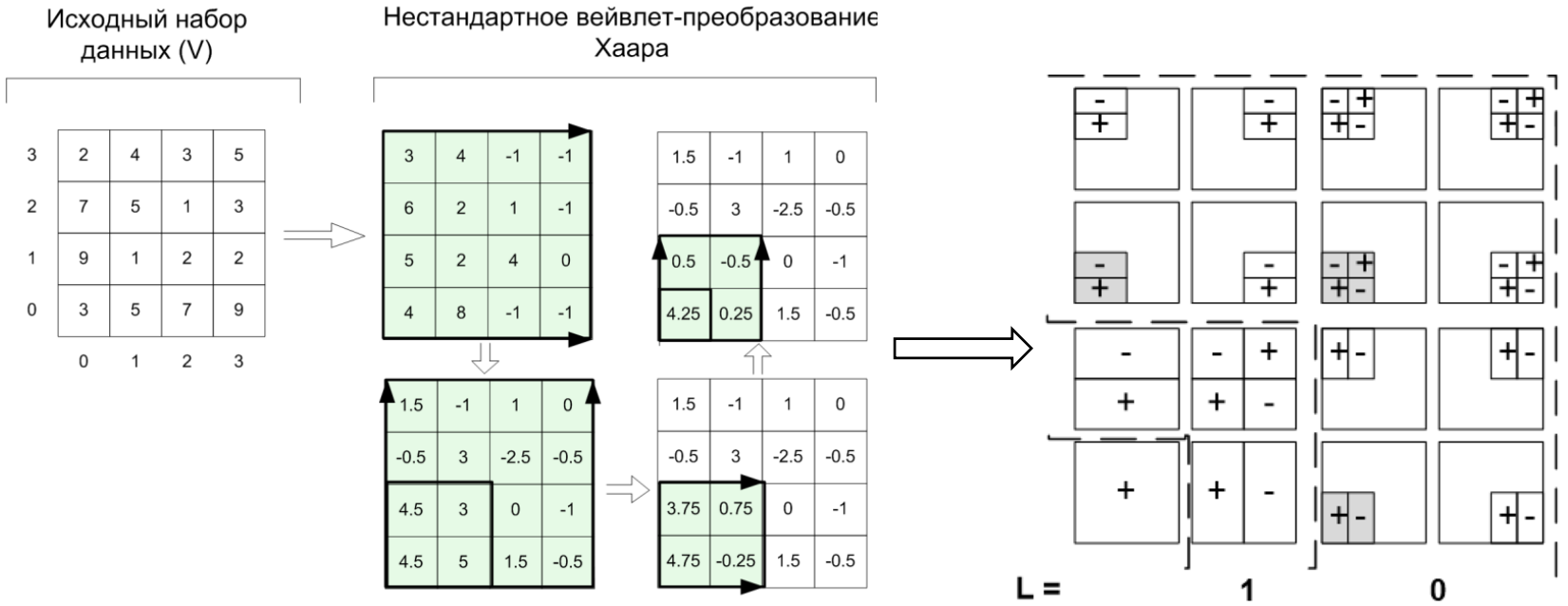
Δ

$$\sum \Delta_i^2 \to \min$$



Приближенная обработка запросов методом вейвлет-преобразования Хаара -2

Обратное дискретное нестандартное вейвлет-преобразование Хаара (ДНВПХ)



Области действия коэффициентов

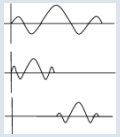
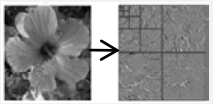
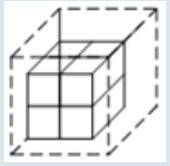
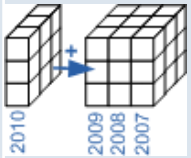
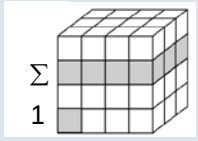
V(0;0)=	4.25	+0.5	-0.5	+0.25	-0.5	-2.5	+1.5	= 3
V(0;1)=	4.25	+0.5	-0.5	+0.25	-0.5	-(-2.5)	-(1.5)	= 5
V(1;0)=	4.25	+0.5	-0.5	+0.25	-(-0.5)	-(-2.5)	+1.5	= 9
V(1;1)=	4.25	+0.5	-0.5	+0.25	-(-0.5)	-2.5	-(1.5)	= 1
Σ =	4.25x4	0.5x4	-0.5x4	+0.25x4	0	0	0	= 18

Взаимно уничтожающиеся коэффициенты

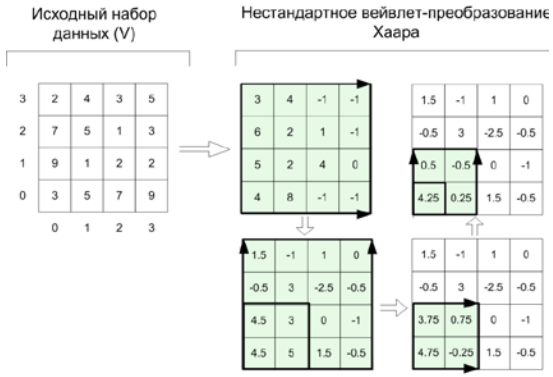
Приближенная обработка запросов методом вейвлет-преобразования Хаара -3

- Возможность вычисления единичных и суммарных значений
- Возможность обработки многомерных данных без значительного увеличения затрат
- Вычисление агрегированных значений не требует восстановления всех единичных значений
- Простой способ минимизации среднеквадратической погрешности восстановления исходного набора данных

Особенности применения вейвлет-преобразования Хаара

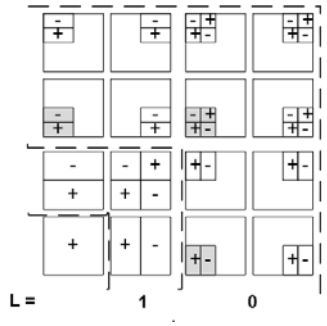
	Обновление данных	Объем данных	Сжатие данных	Восстановление данных	Погрешность
Сигнал 	Не требуется.	- Одномерный массив данных. - Ограничение на количество элементов 2^n не приводит к резкому увеличению объема.	Задание коэффициента сжатия.	Исходный сигнал	Супремальная оценка априорной погрешности
Изображение 	Не требуется.	- Двумерный массив данных. - Ограничение на количество элементов 2^n в измерениях не приводит к резкому увеличению объема.	Задание коэффициента сжатия.	Исходное изображение	Оценка апостериорной погрешности
Хранилище данных 	Требуется пересчет сжатых данных. 	- z-мерный массив данных. - Ограничение на количество элементов 2^n в измерениях ведет к резкому увеличению объема данных. $\prod_{i=1}^{z} m_i \ll 2^{nz}$ m_i – длина i-ого измерения	- Задание прогнозируемой погрешности. - Задание коэффициента сжатия.	- Исходный элемент. - Сумма исходных элементов. 	Оценка априорной погрешности

Проблемы приближенной обработка запросов методом вейвлет-преобразования Хаара



- Сложность обработки измерений произвольной длины, отличных от $2^n, n \in N$. Необходимость хранения метаданных, что приводит к увеличению объема приближенного представления.

- Недостаточно разработанные методы оценки погрешности приближенной обработки: отсутствие оценки погрешности при вычислении агрегатных значений, сложность вычисления относительных погрешностей.



- Сложность обновления вейвлет-декомпозиции. Требуется полный пересчет декомпозиции. Невозможность оценки погрешности при повторном сжатии.

$$\sum_i (v_i - \tilde{v}_i)^2 = \sum_j c_j^2 = 2^{z-n}\Omega^2$$

Решаемые задачи

- Разработка метода вейвлет-преобразования Хаара n-мерного набора данных с произвольной длиной измерений.

- Разработка методов восстановления исходного элемента и суммы элементов из сжатого хранилища данных.

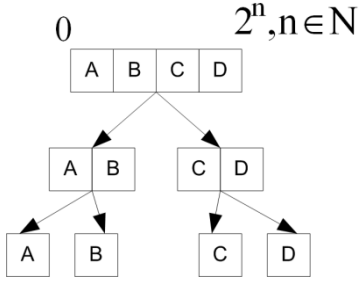
- Разработка методов оценки погрешностей восстановления исходного элемента хранилища данных и суммы таких элементов.

- Апробирование предложенного метода и оценок погрешности на реальных данных

ДНВПХ по алгоритму Чакрабарти*

Недостатки существующих методов

- Требование к длине измерения быть кратной $2^n, n \in N$
- Необходимость хранения метаданных

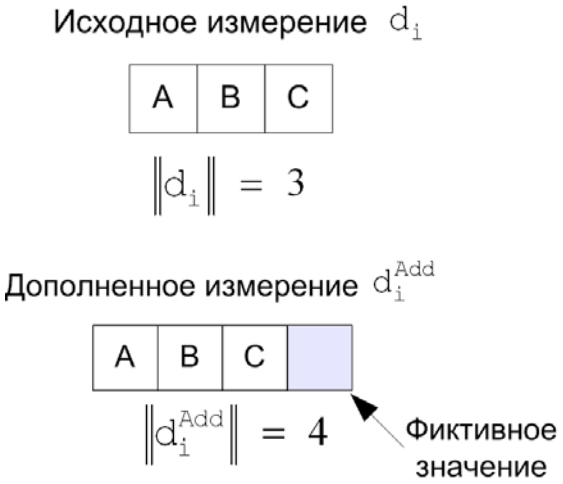


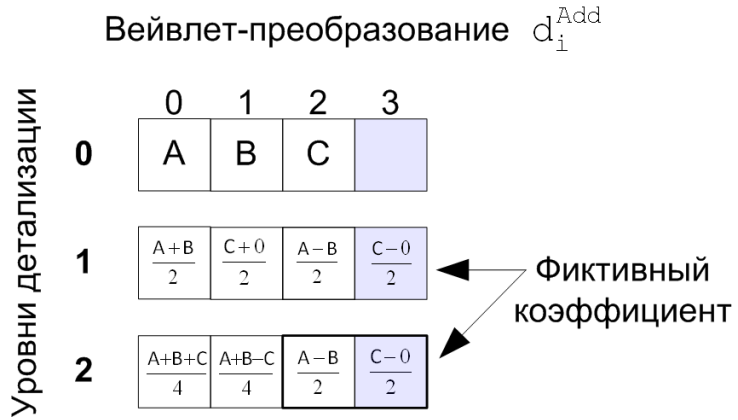
При ДНВПХ Чакрабарти усредненные значения A_i^j *уровня детализации j заменяются усредненными и уточняющими значениями* A_i^{j-1}, Q_i^{j-1} *уровня j-1.*

*

- Chakrabarti, K., Garofalakis, M., Rastogi, R. 2001. Approximate query processing using wavelets. The VLDB Journal — The International Journal on Very Large Data Bases.- Vol. 10(2).- P. 199-223.
- Garofalakis, M. 2006. Wavelet-Based Approximation Techniques in Database Systems. IEEE Signal Processing Magazine.- Vol. 23(6).- P. 54-58.

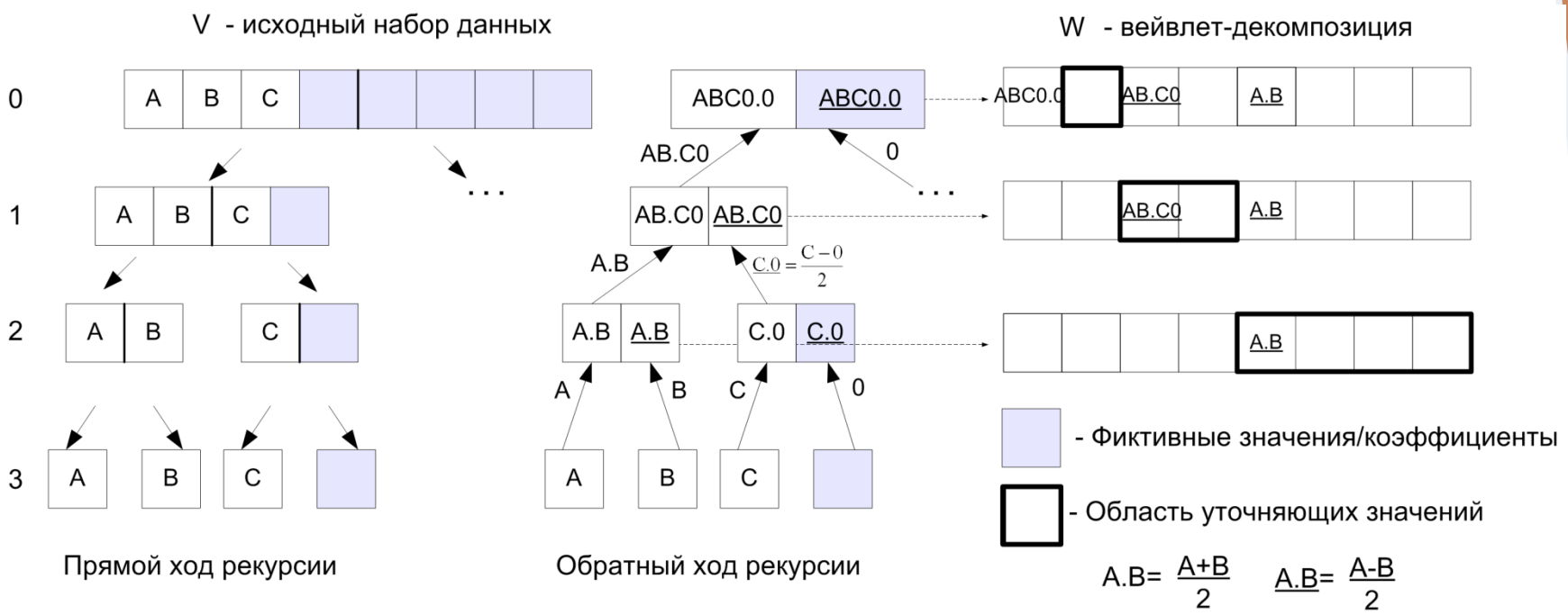
Принцип работы дискретного нестандартного вейвлет-преобразования Хаара (ДНВПХ) с произвольной длиной измерения





На основе регулярной структуры получаемого набора данных предложенный алгоритм при нахождении фиктивного (т.е. отсутствующего) значения (массива значений) в исходном наборе данных принимает его равным нулю при расчетах, однако не сохраняет связанный с этим значением (массивом значений) вейвлет-коэффициент в декомпозиции.

Прямое ДНВПХ с произвольной длиной измерения



- Количество вейвлет-коэффициентов в полученной декомпозиции равно количеству исходных значений.

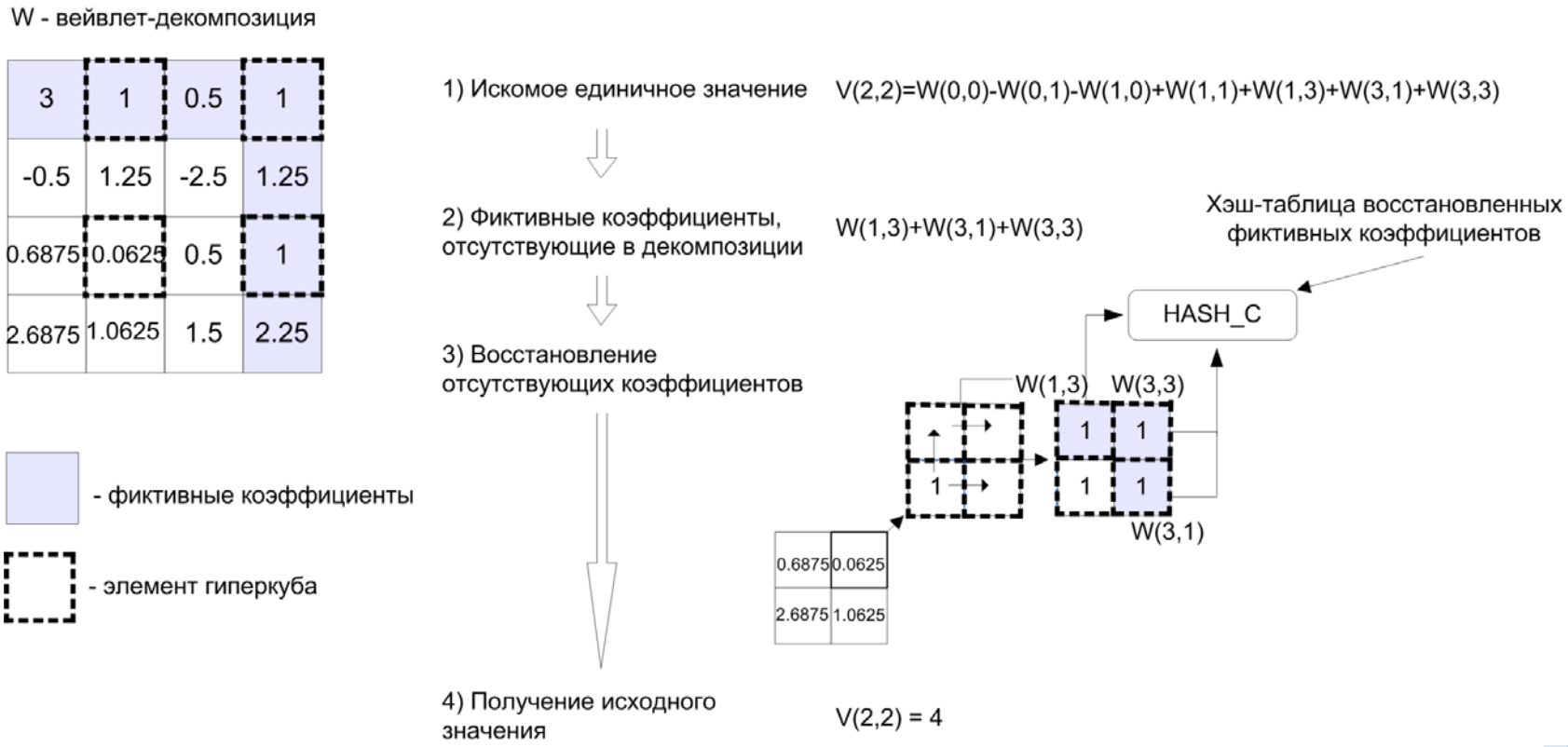
- При вычислении вейвлет-декомпозиции сохраняются только те коэффициенты, которые не будут перезаписаны в дальнейшем. Если набор данных содержит z измерений , длиной 2^n, то выигрыш в количестве операций чтения/записи составит $\frac{1}{2^z} \cdot 100\ \%$.

- Преобразование осуществляется за один проход по набору данных.

Обратное ДНВПХ с произвольной длиной измерения - 1

Существующие методы восстановления элемента исходных данных из вейвлет-декомпозиции не применимы, т.к. часть вейвлет-коэффициентов, полученная на основе фиктивных значений, не была сохранена в декомпозиции.

Восстановление единичного значения исходного набора данных



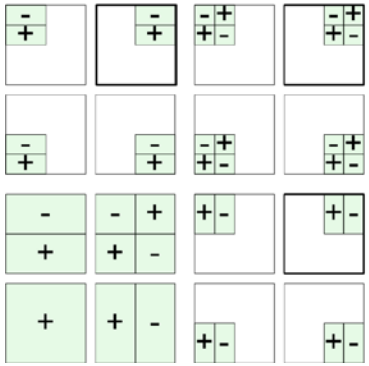
Обратное ДНВПХ с произвольной длиной измерения - 2

Восстановление суммарного значения исходного набора данных

W- Вейвлет-декомпозиция

3	1	0.5	1	3
-0.5	1.25	-2.5	1.25	2
0.6875	0.0625	0.5	1	1
2.6875	1.0625	1.5	2.25	0
0	1	2	3	

Области действия вейвлет-коэффициентов



1) Искомое суммарное значение

Sum=V(0,0)+V(0,1)+V(1,0)+V(1,1)

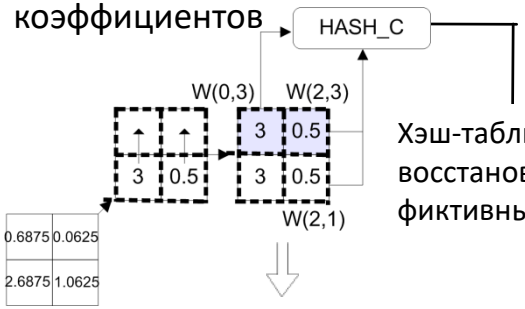
2) Учет компенсируемых значений

V(0;0)=	2.6875	+0.6875	+0.0625	+1.0625	-0.5	-2.5	+1.5	= 3
V(0;1)=	2.6875	+0.6875	+0.0625	+1.0625	+0.5	+2.5	+1.5	= 9
V(1;0)=	2.6875	+0.6875	+0.0625	+1.0625	-0.5	+2.5	-1.5	= 5
V(1;1)=	2.6875	+0.6875	+0.0625	+1.0625	+0.5	-2.5	-1.5	= 1
Σ =	2.6875x4	0.6875x4	0.0625x4	1.0625x4	0	0	0	= 18

Взаимно-компенсирующиеся коэффициенты 12 из 28

3) Фиктивные коэффициенты, отсутствующие в декомпозиции W(0,3)+W(2,3)+...

4) Восстановление отсутствующих коэффициентов



Хэш-таблица восстановленных фиктивных коэффициентов

5) Получение искомой суммы Sum = 18

- Сохраняется взаимная компенсация коэффициентов при расчете агрегированных значений, что позволяет уменьшить количество операций чтения и суммирования.

Погрешность восстановления исходных элементов хранилища данных - 1

Недостатки существующих методов

- Оценка только верхней границы относительной ошибки по всему объему данных.
- Отсутствие оценки погрешности вычисления суммы элементов.

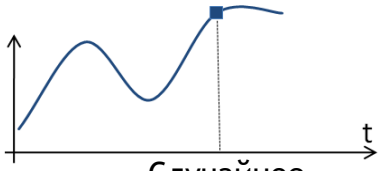
Вероятностное пространство

$V(v_i)(i=1..2^n, n \in N)$ - исходный набор данных.

$\tilde{V}(\tilde{v}_i)(i=1..2^n, n \in N)$ - восстановленный набор данных, полученный после обнуления некоторых вейвлет - коэффициентов $c_j, j=1..N^*, c_j \neq 0$ в соответствующей декомпозиции W.

$v_i = \tilde{v}_i + \Delta v_i$ - элемент исходного набора данных.

Процесс изменения состояния куба



Случайное состояние куба

Погрешность восстановления исходных элементов хранилища данных - 2

Пусть конкретные $\{c_j \mid j=1..N^*, c_j \neq 0\}$ случайно равномерно распределены по кубу вейвлет-декомпозиции. Тогда конкретная декомпозиция W, с коэффициентами $\{c_j \mid j=1..N^*, c_j \neq 0\}$, которые находятся в определенных позициях и которые впоследствии будут обнулены, является единичной выборкой из соответствующей генеральной совокупности.

$$\Delta v = \sum_j \xi_j, j=1..N^*$$

ошибка восстановления некоторого исходного v, где ξ_j - случайная величина, соответствующая использованию вейвлет - коэффициента c_j при восстановлении v.

Предполагаем, что ξ_j независимы.

Распределение ξ_j

Вероятность	p_L^+	p_L^-	$1-p_L^+-p_L^-$
Значение	$k_L \cdot c_j$	$-k_L \cdot c_j$	0

p_L^+, p_L^-- вероятности участия коэффициента c_j со знаком +/- при восстановлении элемента v на уровне детализации L.

k_L - коэффициент нормирования c_j на уровне L.

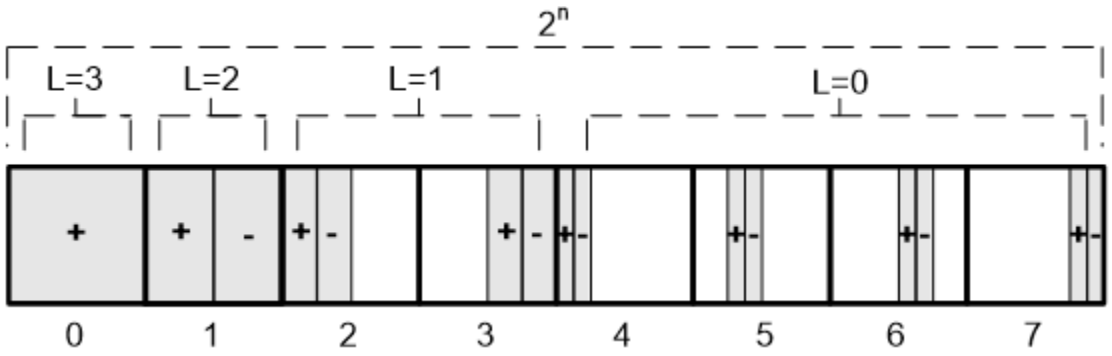
Погрешность восстановления исходных элементов хранилища данных - 3

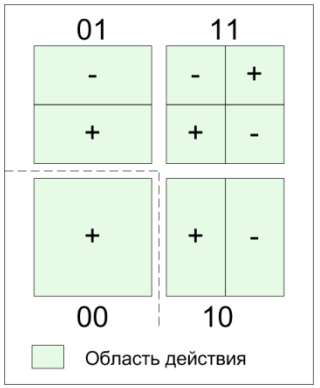
На основе регулярной структуры расположения областей действия найдем вероятности P_L^+ и P_L^-

$P_L = \frac{2^z - 1}{2^{z(L+1)}}, L \in [n-1, 0]$ - вероятность попадания коэффициента на уровень детализации L.

$P_{LS} = 2^{z(L+1-n)}, L \in [n-1, 0]$ - вероятность попадания коэффициента в область действия на уровне L.

$P_L^+ = P_L^- = \frac{(2^z - 1)}{2^{zn+1}}$ - вероятности участия коэффициента ‘c’ в восстановлении ‘v’ с положительным и отрицательным знаком, где z-количество измерений, а длина измерений 2^n.





Погрешность восстановления исходных элементов хранилища данных - 4

На основе центральной предельной теоремы Ляпунова доказали:

Распределение $\Delta v = \sum_j \xi_j, j = 1..N^*$ стремится к нормальному при $N^* \to \infty$.

Скорость сходимости распределения $\Delta v = \sum_j \xi_j, j = 1..N^*$ к нормальному распределению

$\sup_x \left| P\left(\frac{\Delta v}{B} < x \right) - \Phi(x) \right| \le A' \cdot \frac{1}{\sqrt{N^*}}$, где $\Phi(x)$ - нормальное распределение с параметрами (0,1), $B^2 = \sum_{i=1}^{N^*} D(\xi_i)$

A – константа.

⇓

Оценка погрешности единичного значения

$$M_1(\Delta v) = 0 \qquad D_1(\Delta v) = \frac{2^{zn} - 1}{2^{zn}} \cdot \frac{1}{2^{zn}} \cdot \sum_{i=1}^{N^*} c_i^2$$

С помощью полученных выражений можно оценить зависимость погрешности восстановления исходного элемента данных (доверительный интервал ошибки) от степени сжатия данных N*

$$(-r\sigma, +r\sigma), \ \sigma = \sqrt{D(\Delta v)}$$

Интересно, что принимая во внимание $\sum_i (v_i - \tilde{v}_i)^2 = \sum_j c_j^2 = 2^{z \cdot n} \Omega^2$ при больших Значениях z, n и N*дисперсия единичного значения равна квадрату средней квадратической ошибки восстановления всех элементов

Погрешность восстановления исходных элементов хранилища данных - 5

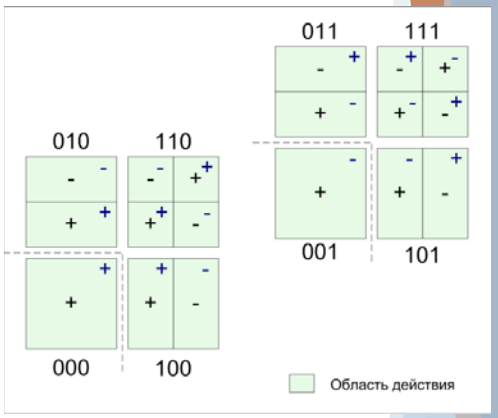
Оценка погрешности восстановления суммы исходных элементов хранилища данных

$P'_L = \frac{2^z - 1}{2^{z(L+1)}}, L \in [n-1, 0]$ - вероятность попадания коэффициента на уровень детализации L.

$P'_{LS} = 2^{(t-z)(n-L-1)}, L \in [n-1, 0], t \in [1, z]$ - вероятность попадания коэффициента в область действия для суммы значений.

$P'_{1NC} = \frac{2^{z-t} - 1}{2^z - 1}, t \in [1, z]$ - вероятность, что данный коэффициент не будет компенсирован.

$P'^{+}_L = P'^{-}_L = 2^{t(n-L-1)-(zn+1)} \cdot (2^{z-t} - 1)$ - вероятности участия коэффициента с в восстановлении суммы значений Sum(v) размерностью t с положительным и отрицательным знаком, где z-количество измерений, а длина измерений 2^n.



$$D_\Sigma(\Delta v) = \frac{2^{(z-t)n} - 1}{2^{(z-t)n}} \cdot \frac{1}{2^{(z-t)n}} \cdot \sum_{i=1}^{N^*} c_i^2 \qquad M_\Sigma(\Delta v) = 0$$

Оценка погрешности восстановления исходных элементов хранилища данных для запросов общего вида

Запрос на подмножестве ячеек X от 1 до M представим в виде функции от многих переменных $f(v_1, \ldots, v_M)$.

$\Delta f = \sum_{i=1}^{M} \frac{\partial f}{\partial v_i} \Delta v_i$ - погрешность выполнения запроса. Пусть $\{\Delta v_i\}$ независимые случайные величины.

$$D(\Delta f) = D_1(\Delta v) \sum_{i=1}^{M} \left(\frac{\partial f(\tilde{v}_i)}{\partial v_i} \right)^2 \qquad M(\Delta f) = 0$$

Сумма M значений: $D(\Delta f) = M \cdot D_1(\Delta v)$

Среднее M значений: $D(\Delta f) = \frac{D_1(\Delta v)}{M}$

Пример расчета распределения погрешности восстановления исходных элементов хранилища данных

Получение сжатого представления данных

Отброшено 9 из 16 коэффициентов.
Коэффициент сжатие 56%

Вычисление распределения погрешностей

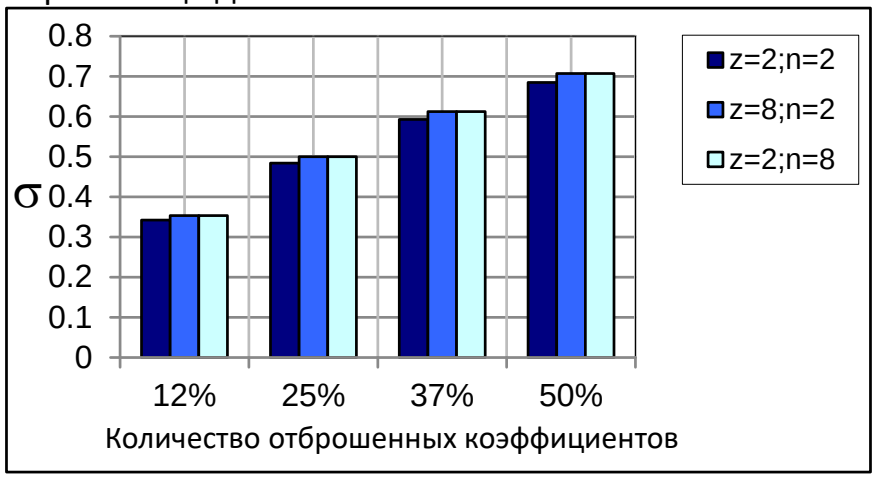
	Дисперсия	Доверительный интервал ошибки		
		σ	2σ	3σ
Элемент данных	$\frac{2^{2n}-1}{2^{4n}} \cdot \sum_{i=1}^{N^*} c_i^2$	1.19	2.37	3.56
Агрегат по ребру $(t=1)$	$\frac{2^{n}-1}{2^{2n}} \cdot \sum_{i=1}^{N^*} c_i^2$	2.12	4.24	6.37
Агрегат по всему набору $(t=2)$	$0 \cdot \sum_{i=1}^{N^*} c_i^2$	-	-	-

Анализ результатов

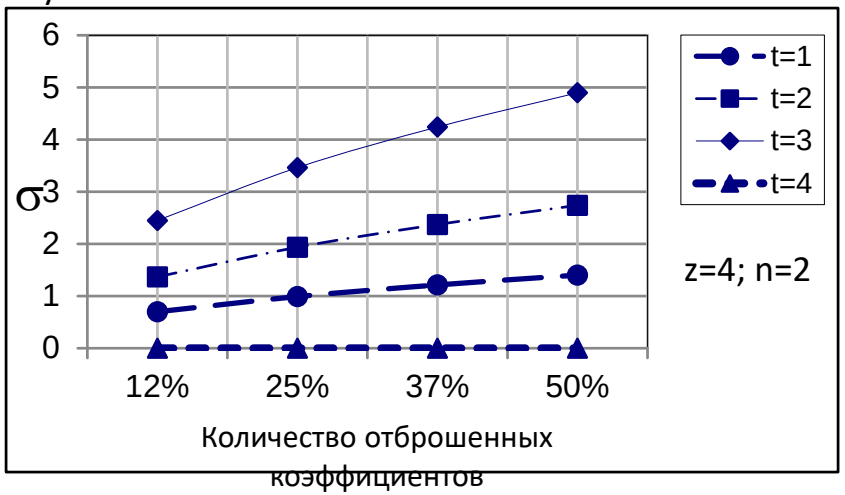
- Максимальная ошибка вычисления составляет 3.25 (ΔV_{22}), т.е. попадает в интервал 3σ. Средняя ошибка восстановления элемента данных, равная 1, попадает в интервал σ.
- Дисперсия суммы по всему набору данных равняется 0. При расчете агрегата по ребру максимальная ошибка вычисления равна 4, т.е. попадает в интервал 2σ. При этом средняя ошибка восстановления агрегатного значения по ребру, равная 2, также попадает в интервал σ. Имеет место сохранение тренда суммарных значений по строкам и столбцам.

Исследование погрешностей восстановления исходного элемента и суммы исходных элементов хранилища данных

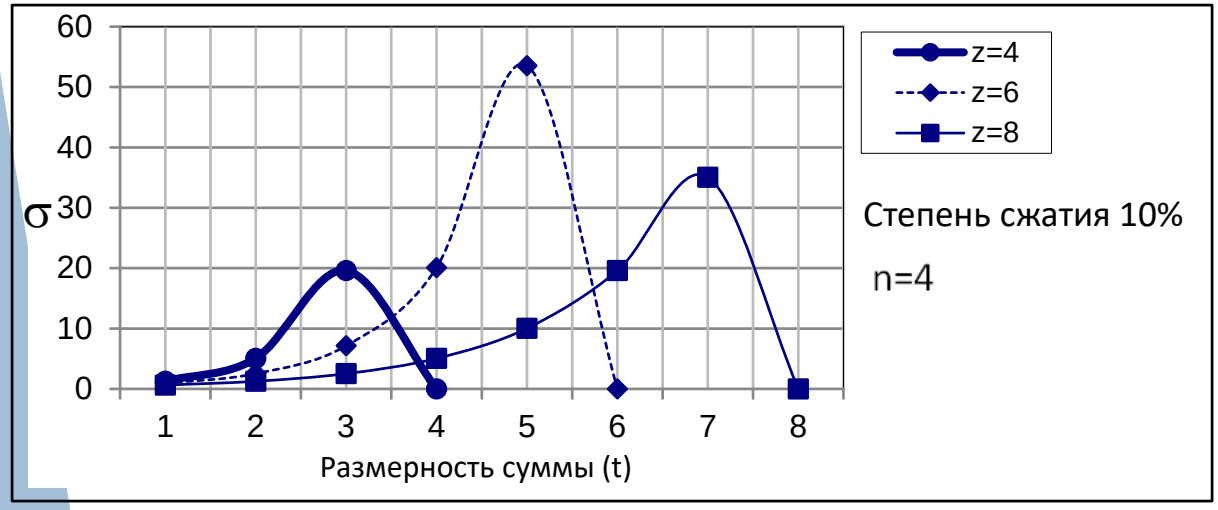
1. Погрешность восстановления исходного элемента данных для различных конфигураций хранилища данных и степеней сжатия



2. Погрешность восстановления суммы исходных элементов для различных степеней сжатия и размерности суммы t



3. Погрешность восстановления суммы исходных элементов данных при одной и той же степени сжатия для различных конфигураций хранилищ данных и размерности суммы t



z - количество измерений.

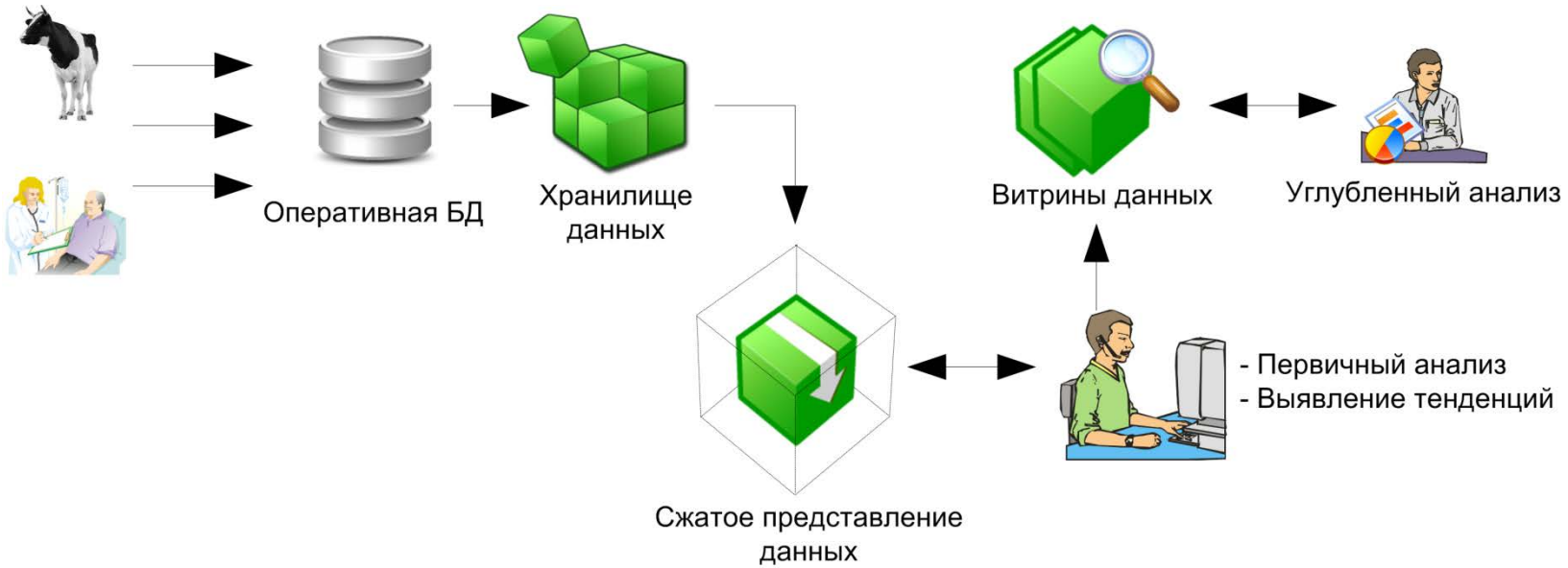
2^n - длина измерения.

t - размерность суммы значений.

Для удобства принимаем $c_i = 1$

OLAP модуль Системы надзора за эпидемиологической обстановкой - СНЭО

Предметная область Национальная система сбора, передачи, хранения и анализа информации о заболеваниях человека и животных, а так же связанных с ними данных об образцах и лабораторных тестах



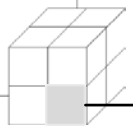
Примеры задач аналитического модуля

- Исследование случаев заболеваемости.
- Поиск факторов, коррелирующих с появлением заболевания.
- Исследование способов распространения заболеваний.
- Прогнозирование возможных вспышек заболеваний.
- Исследование факторов, способствующих передачи зоонозных заболеваний.
- Анализ качества постановки диагнозов.
- Оценка существующих методик выявления заболеваний.
- Уточнение критериев карт эпидемического расследования случаев.

Примеры результатов использования OLAP модуля СНЭО - 1

Анализ частоты проявления заданного набора клинических признаков для различных типов туляремии

Тип туляремии
Воспаление лимфатических узлов
Температура > 38
Кровавый понос
Уплотнения на коже
Кровяная слюна
Подкожная эмфизема
Головная боль
Увеличение селезенки
Лихорадка
Увеличение печени
Септический шок
Гнойные нарывы
Кашель
Квартал регистрации случая



Количество случаев заболевания

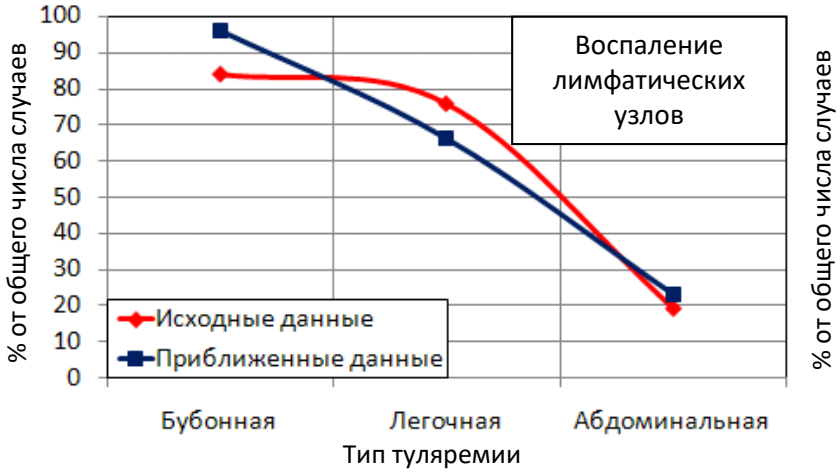
Пример исследования

Какова зависимость между воспалением лимфатических узлов и типом туляремии?

Коэффициент сжатия - 60%

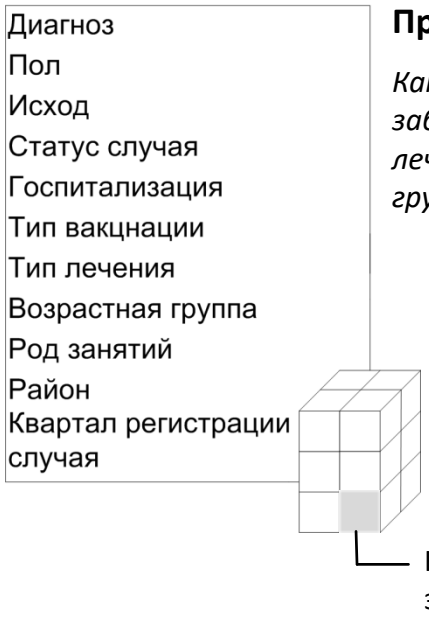
Сокращение времени выполнения запроса - 54 %

Средняя относительная погрешность - 15.5%



Примеры результатов использования OLAP модуля СНЭО - 2

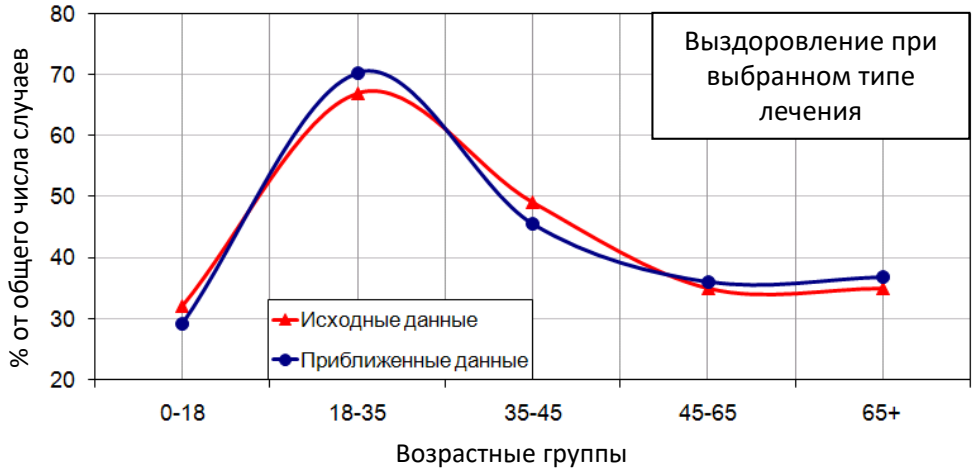
Анализ заболеваемости и эффективности лечения



Пример исследования

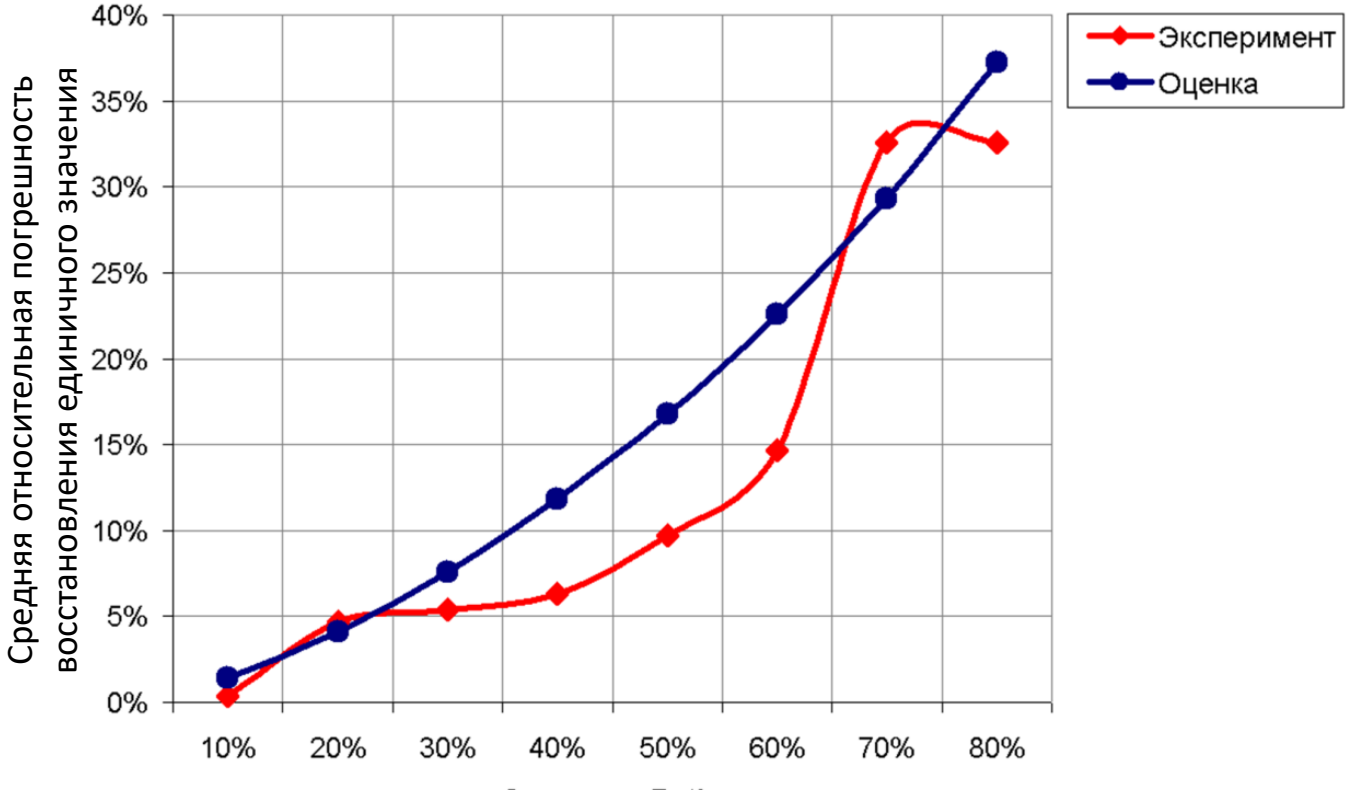
Какова зависимость исхода заболевания от выбранного типа лечения в различных возрастных группах по стране?

Коэффициент сжатия	- 60%
Сокращение времени выполнения запроса	- 52 %
Средняя относительная погрешность	- 5.8 %



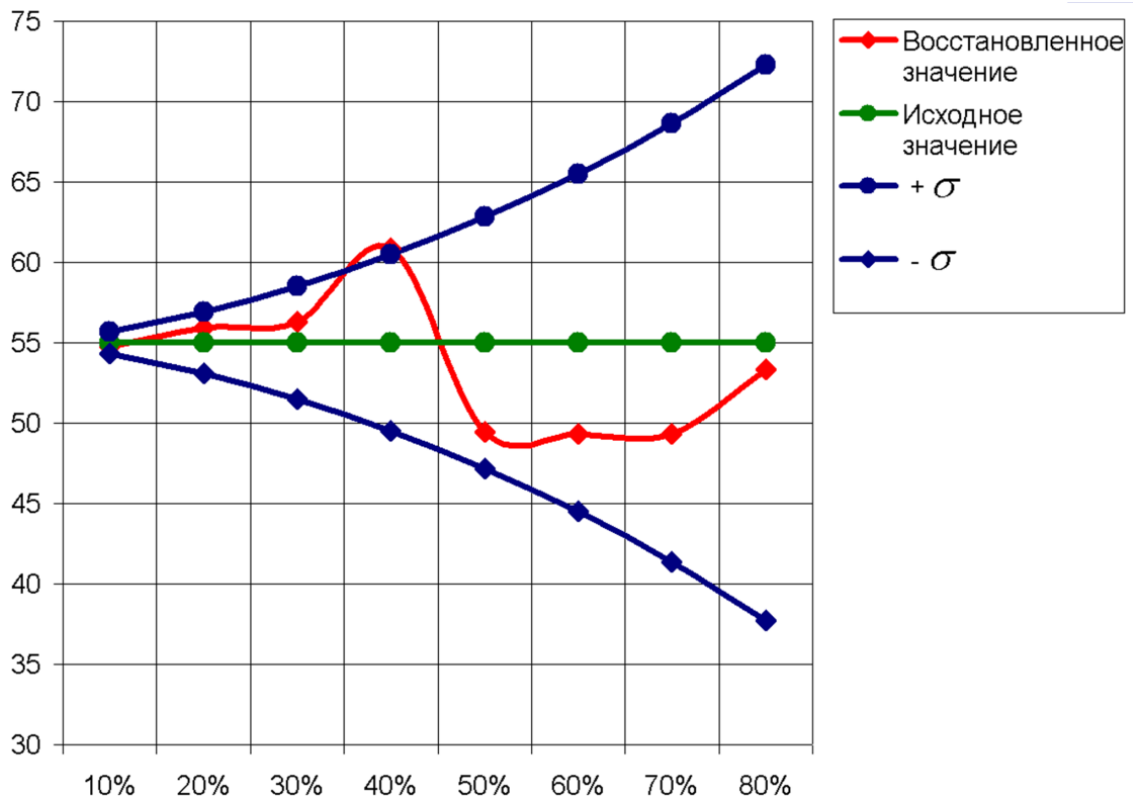
Анализ результатов использования. Погрешность восстановления исходного элемента хранилища данных - 1

Средняя относительная погрешность восстановления элемента исходных данных для различных степеней сжатия данных.



Анализ результатов использования. Погрешность восстановления исходного элемента хранилища данных - 2

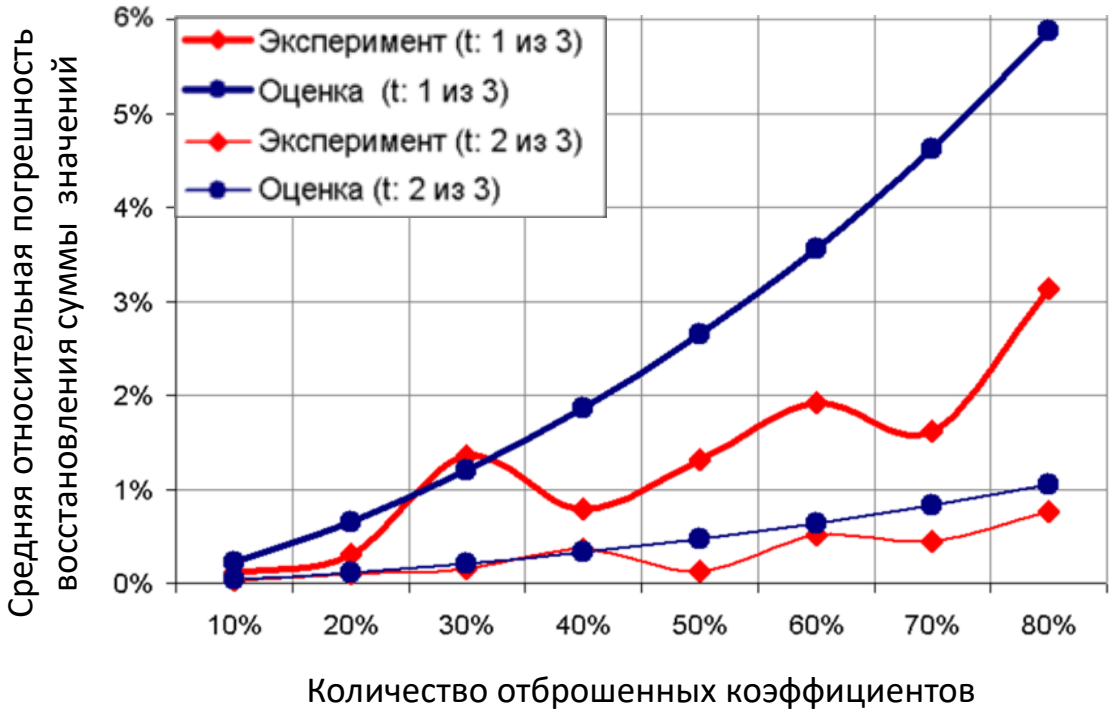
Рассчитанный доверительный интервал относительной погрешности восстановления элемента исходных данных для различных степеней сжатия хранилища данных.



Анализ результатов использования.

Погрешность восстановления суммы исходных элементов хранилища данных - 1

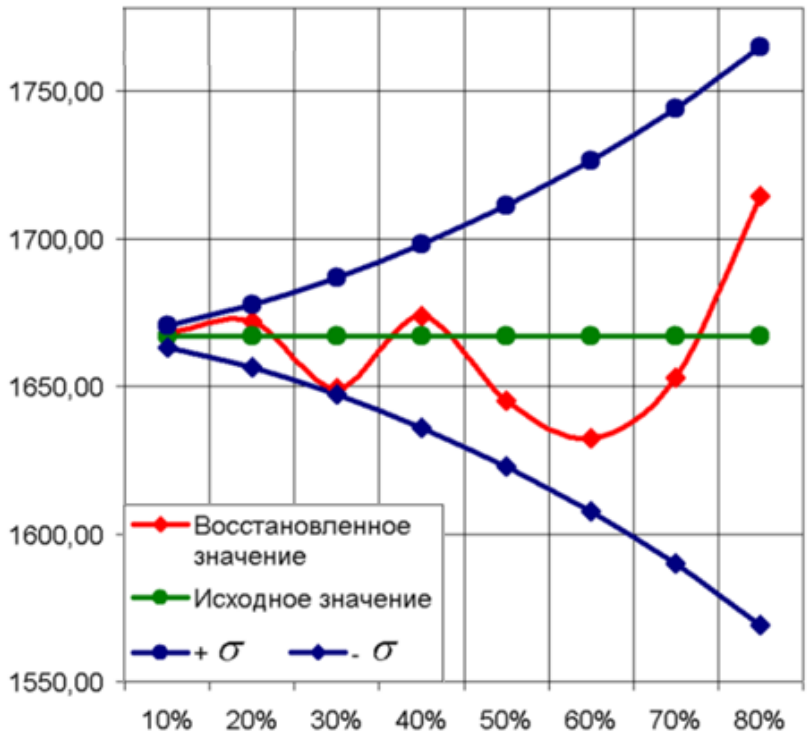
Средняя относительная погрешность восстановления суммы исходных элементов хранилища данных для различных степеней сжатия данных и размерности суммы t.



Анализ результатов использования.

Погрешность восстановления суммы исходных элементов хранилища данных - 2

Рассчитанный доверительный интервал относительной погрешности восстановления суммы исходных элементов данных для различных степеней сжатия хранилища данных.

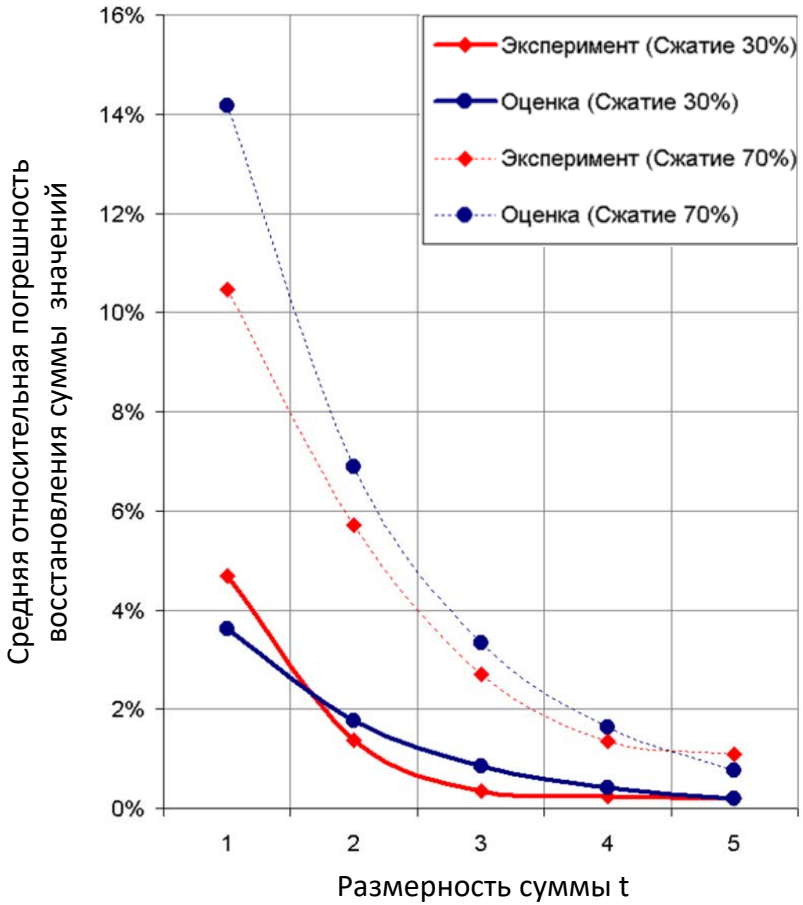


Количество отброшенных коэффициентов

Анализ результатов использования.

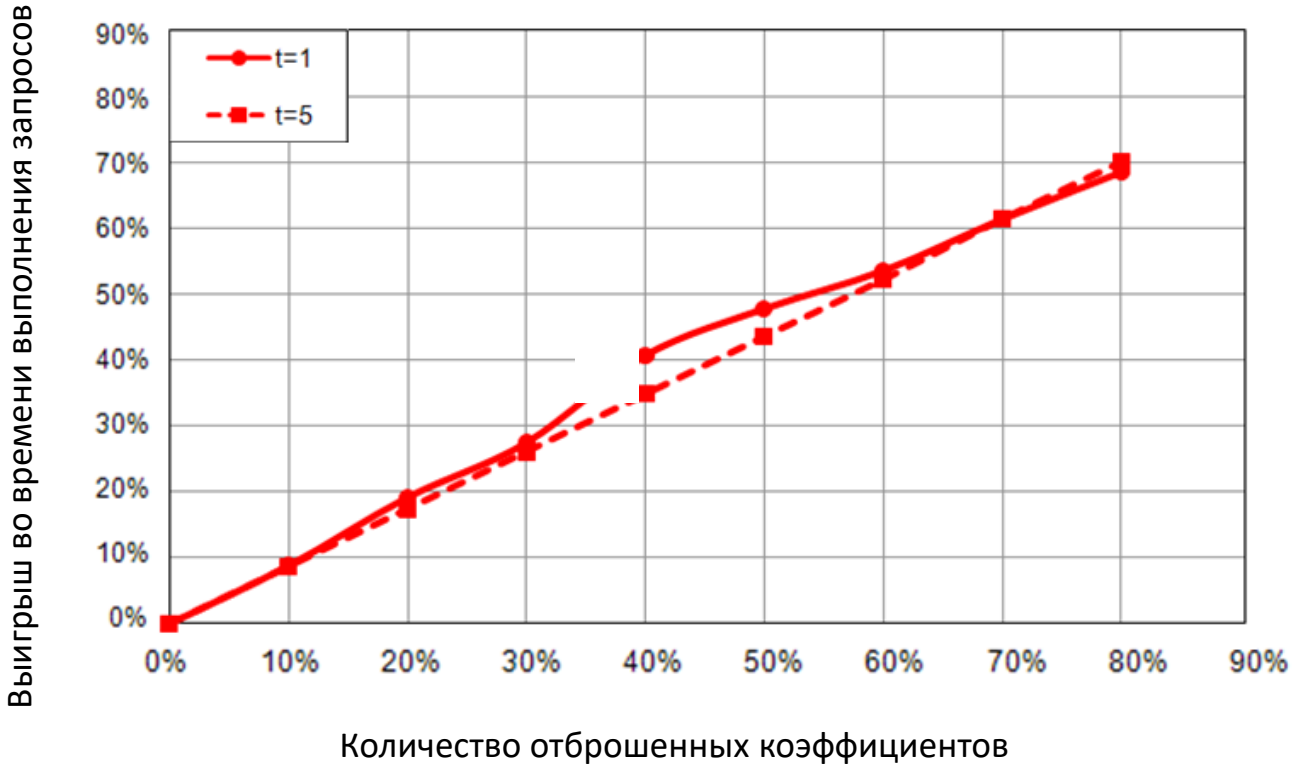
Погрешность восстановления суммы исходных элементов хранилища данных - 3

Средняя относительная погрешность восстановления суммы исходных элементов данных для различных степеней сжатия хранилища данных в зависимости от размерности суммы t.



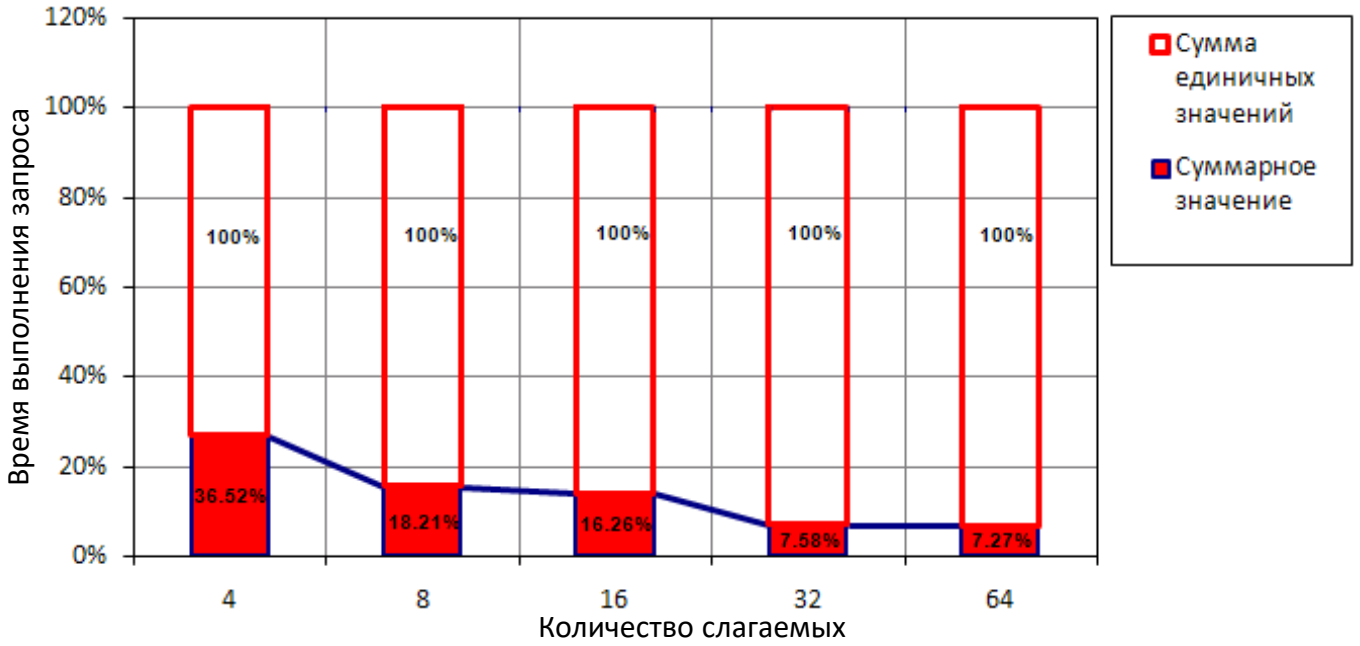
Анализ результатов использования. Время выполнения запроса и степень сжатия - 1

Выигрыш во времени выполнения запросов в зависимости от степени сжатия хранилища данных для различной размерности суммы t.



Анализ результатов использования. Время выполнения запроса и степень сжатия - 2

Время вычисления суммы исходных элементов хранилища данных по сравнению с вычислением той же суммы поэлементно в зависимости от количества слагаемых.



Выводы

- Разработан метод получения приближенных представлений данных на основе вейвлет-преобразования Хаара, позволяющий обрабатывать массивы с произвольными длинами измерений и сокращающий число операций чтения/записи.

- Разработан метод восстановления исходного элемента и суммы элементов хранилища данных на основе приближенного вейвлет-представления с произвольной длиной измерения, учитывающий взаимную компенсацию коэффициентов и позволяющий вычислять несохраненные вейвлет-коэффициенты.

- Предложен метод оценки погрешности восстановления исходного элемента и суммы элементов хранилища данных, позволяющий вычислять прогнозируемые доверительные интервалы ошибок в зависимости от степени сжатия хранилища данных.

- Исследованы результаты использования предложенного метода приближенной обработки запросов и оценки погрешностей. Сжатие данных на 60% уменьшает время выполнения запросов в среднем на 50% при погрешности восстановления исходного элемента и суммы элементов хранилища данных, равной 15% и 5% соответственно.

Расширение сотрудничества

- **Развитие темы приближенной обработки запросов**
  - Использования предложенного метода в параллельных СУБД для аналитической обработки сверхбольших массивов данных
- **Использование ИТ технологий (включая анализ больших объектом данных) в области эпидемиологического надзора**
  - Синдромный надзор
  - Надзор за окружающей средой
  - Мониторинг СМИ
- **Базы данных**
  - Технологии обработки больших данных (от OLAP к платформам Hadoop, Spark и т.д.)
  - Анализ больших графов (социальные сети, транспортные сети, сети финансовых транзакций и др.)
  - Машинное обучение на больших данных
- **Совместная подготовка статей на конференции, оформление, выступления по другим направлениям**