

На правах рукописи

Ухаров Андрей Олегович

**Метод приближенной обработки запросов
в системах оперативного анализа данных**

**Специальность 05.13.17 – Теоретические основы информатики
(технические науки)**

**Автореферат диссертации на соискание ученой степени
кандидата технических наук**

Москва - 2011

Работа выполнена в Московском государственном техническом университете им. Н. Э. Баумана.

Научный руководитель: доктор технических наук,
профессор Григорьев Юрий Александрович

Официальные оппоненты: доктор технических наук,
профессор Сеницын Игорь Николаевич

доктор технических наук,
доцент Романов Виктор Петрович

Ведущая организация: Федеральное государственное унитарное
предприятие «Научно-исследовательский
институт «Восход»

Защита диссертации состоится «26» мая 2011 года в 14:30 на заседании диссертационного совета Д 212.141.10 в Московском государственном техническом университете имени Н.Э. Баумана по адресу: 107005, г. Москва, 2-я Бауманская ул., д.5.

С диссертацией можно ознакомиться в библиотеке МГТУ им. Н.Э. Баумана.

Ваши отзывы в 2-х экземплярах, заверенные печатью, просим выслать по указанному адресу.

Автореферат разослан «___» _____ 2011 г.

Ученый секретарь

Диссертационного совета Д 212.141.10

к.т.н., доцент

С.Р. Иванов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Приближенная обработка запросов возникла в системах оперативного анализа данных (OLAP-системах) как один из методов анализа больших объемов числовой информации. Этот метод широко применяется в таких областях как инвестирование, экономическое прогнозирование, маркетинг, медицинские исследования и т.п. Он получил распространение благодаря специфике аналитических систем, а именно их исследовательской природе. В частности в таких системах принципиально значимым является зависимость между данными, тенденция поведения и порядок исследуемой величины, а не ее абсолютная точность. В то же время в системах оперативного анализа требуется обеспечить высокую производительность обработки запросов при больших объемах данных.

Для решения таких задач и применяется приближенная обработка запросов, которая подразумевает формирование некоторого сжатого представления исходного набора данных и получение приближенных значений на его основе.

В настоящее время распространены три основных подхода к приближенной обработке запросов: методы выборки, методы гистограмм и методы вейвлет-преобразования. Однако существующие подходы обладают рядом недостатков, что сдерживает их применение на практике. Методы выборки и гистограмм ориентированы на получение агрегатных значений с большим количеством элементов. Вычисление же единичных значений или сумм с небольшим числом слагаемых сопровождается значительной погрешностью. Более того с ростом размерности данных (а в современных OLAP-системах по оценкам аналитиков количество измерений в кубах достигает 10-15) увеличивается стоимость построения и хранения приближенного представления, обеспечивающего приемлемую погрешность. В этой связи наиболее перспективным представляется подход на основе вейвлет-преобразования, позволяющий вычислять как единичные, так и суммарные значения на основе многомерных данных. Однако существующие методы накладывают ряд ограничений на структуру данных, не позволяют производить обновление значений без полного пересчета всего набора и не предоставляют точных методов оценки погрешностей.

Таким образом, задача разработки математических методов и инструментального средства приближенной обработки запросов, позволяющих вычислять единичные и суммарные значения на основе многомерных данных произвольной структуры с возможностью обновления и оценки погрешности, является актуальной задачей. Использование таких методов расширит возможности применения приближенной обработки для анализа данных.

Цель работы. Целью данной работы является разработка математических методов и инструментального средства для приближенной обработки запросов на основе вейвлет-преобразования.

В работе решаются следующие **задачи**:

1. Разработка метода вейвлет-преобразования Хаара z -мерного набора данных с произвольной длиной измерений.
2. Разработка метода восстановления исходного элемента и суммы элементов из сжатого хранилища данных с произвольной длиной измерения.
3. Разработка метода обновления сжатого хранилища данных без полного пересчета вейвлет-коэффициентов.
4. Разработка метода оценки погрешностей восстановления исходного элемента данных и суммы элементов.
5. Разработка инструментального средства приближенной обработки запросов.
6. Разработка OLAP-системы «Надзор за заболеваемостью» и анализ результатов использования предложенных методов.

Объект исследования. Объектом исследования настоящей работы являются системы оперативного анализа данных (OLAP-системы).

Предмет исследования. Предметом исследования являются процессы приближенной обработки запросов на основе вейвлет-преобразования в системах оперативного анализа данных.

Научная новизна. В работе получены следующие новые научные результаты:

1. Разработан метод нестандартного вейвлет-преобразования Хаара, учитывающий произвольную длину измерений, отличную от 2^n , и позволяющий уменьшить объем хранимых вейвлет-коэффициентов до объема исходных значений. Доказана лемма о замещении усредненных вейвлет-коэффициентов на разных уровнях детализации вейвлет-преобразования, что позволило сократить количество операций чтения/записи.
2. Разработан метод восстановления исходного элемента данных и суммы элементов на основе приближенного вейвлет-представления с произвольной длиной измерения, учитывающий взаимную компенсацию коэффициентов и позволяющий вычислять несохраненные вейвлет-коэффициенты.
3. Доказаны леммы и теорема о независимости расчета обновленных значений в разработанных алгоритмах дискретного нестандартного вейвлет-преобразования Хаара и на их основе предложен метод обновления имеющейся вейвлет-декомпозиции без пересчета ранее вычисленных вейвлет-коэффициентов.
4. Получено выражение для случайной величины ошибки восстановления исходного элемента, а также суммарного значения, доказана сходимость функции распределения этой ошибки к нормальному закону при увеличении степени сжатия декомпозиции, исследована скорость сходимости, что позволяет оценивать доверительные интервалы ошибок восстановления.

Методы исследования. Исследования проводились на основе комплексного использования теории вейвлетов, теории вероятности, теории алгоритмов.

Практическая ценность полученных результатов. В настоящей работе для практического использования предложенных методов приближенной обработки запросов на основе вейвлет-преобразования и оценки погрешности восстановления значений разработано инструментальное средство. Оно позволяет сформировать приближенное представление многомерного набора данных (OLAP-куба) с произвольной длиной измерения по заданным критериям: допустимая прогнозируемая погрешность вычисления единичного значения, допустимая прогнозируемая погрешность вычисления суммы значений с учетом ее мерности, объем приближенного представления. На основе приближенного представления с помощью инструментального средства можно восстанавливать единичные и суммарные значения. Таким образом, разработанное инструментальное средство предоставляет возможность сформировать приближенное представление данных с прогнозируемой погрешностью и требуемым объемом, и осуществлять аналитическую обработку над ним, получая единичные либо суммарные значения. Приближенное представление при необходимости может быть обновлено. При этом не накладывается ограничение на длину измерений.

Внедрение результатов исследований. Разработанные методы и инструментальное средство были использованы для создания компонента аналитического модуля системы «Надзор за заболеваемостью». Создание компонента приближенной обработки позволило расширить список факторов, участвующих в анализе и выявить новые направления для последующей детальной обработки. При этом не потребовалось увеличения существующих вычислительных мощностей. Высчитываемая прогнозируемая погрешность позволила оценить достоверность результатов и уменьшить ресурсные и временные затраты на детальные исследования тех факторов, которые не представляют интереса для решаемых задач.

Публикации по теме. По материалам настоящей работы опубликовано 6 печатных работ, из них 4 – в научных журналах, рекомендуемых ВАК.

Апробация работы. Материалы работы были изложены автором на семинарах кафедры ИУ-5 МГТУ им. Н.Э. Баумана в 2008 и 2009 годах.

Объем работы. Диссертационная работа содержит 187 страниц, 52 рисунка и 16 таблиц, 1 страницу копий актов о внедрении, список литературы из 141 наименования.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность проблемы. Формулируются цели и задачи исследований, приводится перечень основных результатов, выносимых на защиту, а так же излагается краткое содержание глав диссертации.

В **первой главе** “Критический анализ существующих методов приближенной обработки запросов в OLAP-системах” приведено описание систем аналитической обработки данных, выделены их особенности. В рамках OLAP-систем рассмотрена концепция приближенной обработки запросов,

указаны ее основные этапы. Подробно проанализированы существующие подходы к приближенной обработке запросов и выделен ряд существенных недостатков. На основе проведенного анализа сформулирована концепция и требования к новым методам приближенной обработки запросов.

Для систем аналитической обработки данных приведено описание модели данных (многомерные кубы) и рассмотрены основные операции (формирование срезов, вращение, укрупнение и детализация). Выделены ключевые принципы OLAP-систем: 1) многомерное представление данных, 2) гибкость анализа, т.е. возможность построения произвольных неформализованных запросов, 3) устойчивая производительность анализа, т.е. производительность при выполнении произвольного запроса, 4) глубина анализа, т.е. поддержка необходимого числа измерений, 5) поддержка вычисления суммарных значений.

Показано, что использование OLAP-систем при больших объемах многомерных данных ведет к снижению производительности. Анализ особенностей аналитической обработки данных показал, что в силу исследовательской природы в ряде случаев ключевыми факторами являются гибкость анализа, выявление тенденций и скорость обработки запросов, а не абсолютная точность результата. В этом случае возможно применить методы приближенной обработки запросов, подразумевающие формирование сжатого представления данных и получение значений на его основе с некоторой погрешностью.

При анализе существующих методов приближенной обработки данных в аналитических системах было выявлено, что они обладают рядом недостатков, ограничивающих их применение на практике:

1. Методы выборки не обеспечивают приемлемую точность при вычислении единичных значений либо сумм с небольшим числом слагаемых, ограничивая тем самым гибкость анализа, а с ростом размерности данных значительно усложняется задача получения репрезентативной выборки.

2. Методы гистограмм ориентированы на работу с суммарными значениями, вычисление единичных значений приводит к возникновению значительных погрешностей, а размерность данных увеличивает стоимость построения и поддержания гистограмм.

В результате анализа было выявлено, что методы приближенной обработки данных на основе вейвлет-преобразования являются наиболее перспективными с точки зрения обработки наборов данных с множеством измерений и вычисления как единичных, так и суммарных значений. Был проведен детальный анализ существующих методов, основанных на вейвлет-преобразовании. Анализ показал, что они обладают рядом существенных недостатков:

1. Недостаточно разработанные методы оценки погрешности восстановления значений: верхняя граница общей относительной погрешности не позволяет адекватно оценить качество приближенного ответа; не оцени-

вается погрешность вычисления суммы значений; ряд оценок подразумевает использование динамического программирования.

2. Необходимость полного пересчета ранее полученного приближенного представления при обновлении данных, что помимо ресурсных затрат, приводит к смещению ранее рассчитанных приближенных значений и вновь добавленных точных значений, затрудняя тем самым оценку погрешности в обновленном наборе.

3. Сложность обработки измерений произвольной длины. Существующие методы исходят из предположения, что длина каждого из измерений равна 2^n , значительно ограничивая возможность их применения на практике, либо предполагают хранение метаданных, увеличивающих объем сжатого представления.

В связи с недостатками существующих математических методов возникла необходимость разработать следующее:

1. Математические методы приближенной обработки запросов на основе вейвлет-преобразования, позволяющие учитывать измерения произвольной длины и проводить обновление данных без полного пересчета ранее полученной декомпозиции.

2. Математические методы оценки погрешности восстановления единичного значения и суммы значений.

3. Инструментальное средство, реализующее предлагаемые методы приближенной обработки запросов и оценку качества приближения.

Во **второй главе** “Методы дискретного нестандартного вейвлет-преобразования Хаара для произвольной длины измерения исходного массива данных с возможностью обновления декомпозиции, оценки распределения погрешности восстановления исходных элементов” предложен новый метод получения приближенного представления данных с измерениями произвольной длины. Разработаны алгоритмы вычисления исходного элемента и суммы элементов из такого представления, предложен метод оценки погрешности восстановления элементов. Разработан новый метод обновления приближенного представления без полного пересчета значений.

Известно, что вейвлет-преобразование Хаара представляет собой математический метод иерархической декомпозиции функций. Исходный массив данных можно представить в виде кусочно-постоянной функции. Множество функций представляют собой пространство V^j , где j -количество интервалов, на которых функция постоянна. Базис для пространства V^j может быть задан масштабирующими характеристическими функциями, каждая из которых отличается от нуля на некотором отрезке.

$$\varphi_1^j(x) = \varphi(2^j \cdot x - i), i = \overline{0..2^j - 1}; \varphi(x) = \begin{cases} 1, 0 \leq x < 1 \\ 0, \text{иначе} \end{cases} \quad (1)$$

Можно определить еще одно пространство W^j как ортодополнение для V^j в пространстве V^{j+1} . Совокупность множества линейно-независимых функций ψ_i^j из пространства W^j называется множеством вейвлетов Хаара.

$$\psi_i^j(x) = \psi(2^j \cdot x - i), i = \overline{0..2^j - 1}; \psi(x) = \begin{cases} 1, 0 \leq x < \frac{1}{2} \\ -1, \frac{1}{2} \leq x < 1 \\ 0, \text{иначе} \end{cases} \quad (2)$$

Таким образом, вейвлеты из W^j являются инструментом для представления той части функций в V^{j+1} , которую невозможно представить в V^j .

Например, набор данных из четырех значений может быть представлен как $F(x) = c_0^2 \phi_0^2(x) + c_1^2 \phi_1^2(x) + c_2^2 \phi_2^2(x) + c_3^2 \phi_3^2(x)$, где коэффициенты c_i^2 соответствуют значениям исходного массива. Переписывая это же выражение в базисе V^1 и W^1 , на следующем уровне детализации имеем $F(x) = c_0^1 \phi_0^1(x) + c_1^1 \phi_1^1(x) + d_0^1 \psi_0^1(x) + d_1^1 \psi_1^1(x)$, где $c_0^1 = \frac{c_0^2 + c_1^2}{2}$; $c_1^1 = \frac{c_2^2 + c_3^2}{2}$ - усредненные коэффициенты, а $d_0^1 = \frac{c_0^2 - c_1^2}{2}$; $d_1^1 = \frac{c_2^2 - c_3^2}{2}$ - уточняющие коэффициенты.

Наконец, переходя к базису V^0, W^0, W^1 , на последнем уровне детализации получаем $F(x) = c_0^0 \phi_0^0(x) + d_0^0 \psi_0^0(x) + d_1^1 \psi_1^1(x) + d_1^1 \psi_1^1(x)$. Таким образом, через представление исходной функции в различных базисах можно описать последовательное усреднение и выявление уточняющих коэффициентов.

Дискретное нестандартное вейвлет-преобразование Хаара (ДНВПХ) для многомерного набора данных заключается в последовательном применении операций усреднения и уточнения по каждому из измерений на всех уровнях детализации. Как видно, существующий метод подразумевает, что длина измерения равна 2^n .

Особенность вейвлет-преобразования состоит в том, что отбрасывание из вейвлет-декомпозиции наименьших по модулю нормализованных значений вейвлет-коэффициентов приводит к минимальной средней квадратичной ошибке при восстановлении исходных значений.

Следует подчеркнуть, что это свойство, позволяющее сохранить тенденции изменения данных, определяет важное преимущество метода приближенной обработки данных на основе вейвлет-преобразования по сравнению с другими подходами.

В диссертационной работе предложен метод нестандартного вейвлет-преобразования, на основе ДНВПХ Чакрабартти, позволяющий работать с измерениями произвольной длины, сокращая при этом количество операций чтения/записи. В процессе разработки данного метода была доказана лемма о замещении значений.

Лемма 1. При ДНВПХ Чакрабартти усредненные значения A_i^j уровня детализации j заменяются усредненными и уточняющими значениями A_i^{j-1}, Q_i^{j-1} уровня $j-1, j = 1..Lev_{\max}$, где $Lev_{\max}$ - максимальный уровень детализации. ■

Определение 1. Пусть задано измерение d_i произвольной длины $\|d_i\| = m$. Измерение d_i^{Add} будем называть дополненным, если оно эквивалентно d_i , но при этом длина которого считается равной $2^n : 2^n \geq m$. Несуществующие значения в интервале $(m, 2^n]$ будем называть фиктивными значениями. ■

Определение 2. Измерение d_i на уровне детализации j на некотором шаге k является неактивным, если на этом шаге $r_i^j > \|d_i\| - 1$, т.е. рассчитанное на шаге k значение правой координаты попадает в область фиктивных значений. ■

Суть метода заключается в дополнении измерений исходного набора фиктивными значениями, чтобы длины измерений стали равны между собой и в то же время равны 2^n . При осуществлении операций усреднения и уточнения фиктивные значения считаются равными нулю, а уточняющие коэффициенты, полученные на основе фиктивных значений, не сохраняются в результирующей декомпозиции. В итоге количество вейвлет-коэффициентов остается равным количеству исходных значений, т.е. равно $m \leq 2^n$. Таким образом, введение фиктивных значений не влияет на размер результирующей декомпозиции. Получение вейвлет-декомпозиции на основе предложенного метода для одномерного набора данных представлено на рис.1.

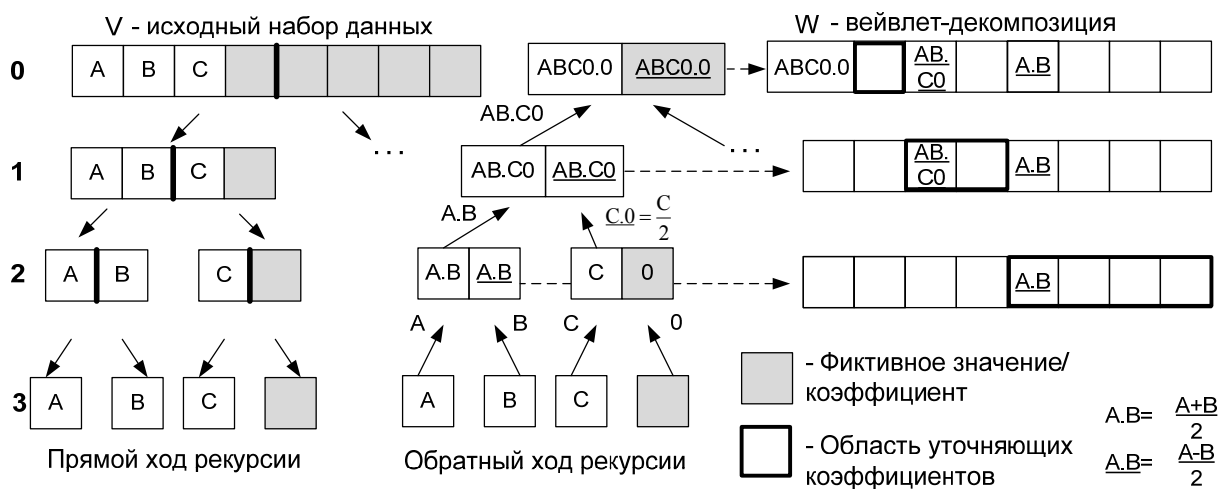


Рис. 1. Вейвлет-преобразование с произвольной длиной измерения

Исходный набор из трех элементов был дополнен пятью фиктивными значениями, так что его длина стала равной 2^n . Можно было дополнить измерение всего лишь одним фиктивным значением, однако в дальнейшем потребуются дополнять измерения заведомо большим числом фиктивных значений. При обратном ходе рекурсивного алгоритма вычисления декомпозиции фиктивные значения считаются нулями, а полученные на их основе фиктивные коэффициенты, например $\underline{C.0}$, не сохраняются в результирующем наборе.

Для полученной декомпозиции неприменимы существующие методы восстановления единичного и суммарного значений, т.к. значения фиктивных коэффициентов не были сохранены в декомпозиции. В диссертационной работе разработан метод обратной вейвлет-декомпозиции, обеспечивающий безошибочное восстановление всех единичных и суммарных значений исходного массива, позволяющий рассчитывать несохраненные фиктивные коэффициенты на любом уровне детализации и учитывающий взаимно компенсирующиеся коэффициенты.

На рис. 2 представлено вычисление суммы трех значений исходного набора. На каждом уровне детализации (от 3 до 0) в соответствии с областью действия определяются коэффициенты, участвующие в восстановлении слагаемых суммы. Значения отсутствующих коэффициентов (второй и шестой коэффициенты) вычисляются исходя из их расположения в декомпозиции и знания того, что исходные фиктивные значения трактовались нулями. На уровнях 1 и 0 учитывается взаимная компенсация коэффициентов.

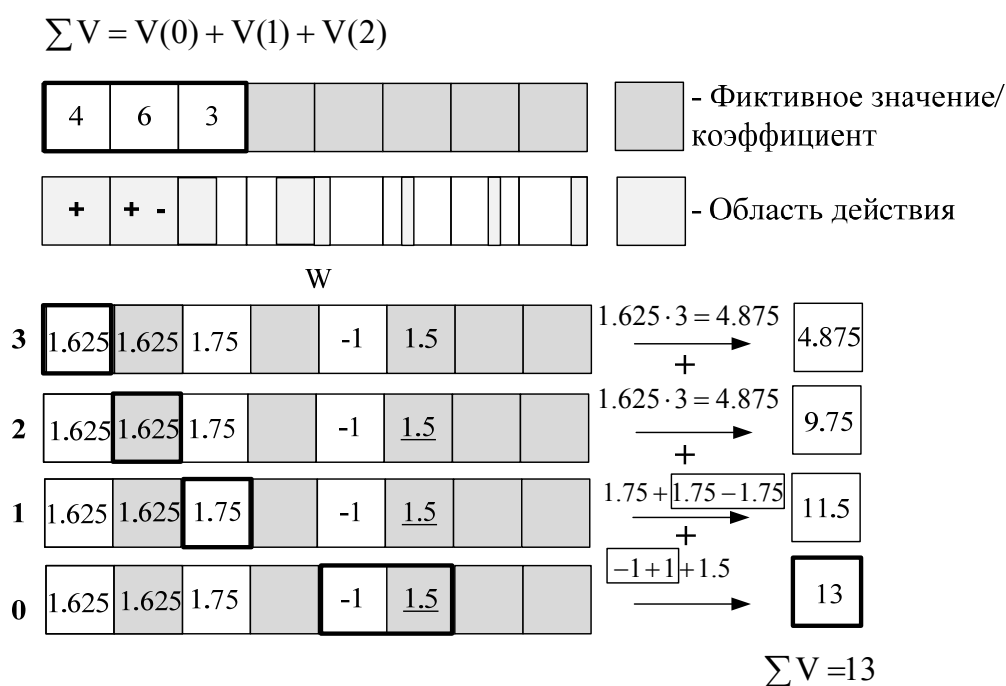


Рис. 2. Обратное вейвлет-преобразование для суммы значений

В диссертационной работе предложен метод обновления вейвлет-декомпозиции без пересчета ранее полученных значений, основанный на использовании фиктивных значений. Идея обновления декомпозиции заключается в том, чтобы дополнить каждое измерение намеренно большим числом фиктивных значений, а в дальнейшем, при обновлении, заменить часть фиктивных значений на вновь добавленные, не затрагивая при этом ранее рассчитанные коэффициенты. Данный метод базируется на двух леммах и теореме, доказанных в диссертационной работе.

Лемма 2. Пусть задан z -мерный набор данных, в котором длины всех измерений равны 2^n . Количество реальных значений в каждом измерении

равно k , такое что $k < 2^n$ и $m = \lceil \log_2 k \rceil$. Тогда при ДНВПХ с произвольной длиной измерения уточняющие вейвлет-коэффициенты будут добавлены в декомпозицию только на уровнях $\text{Lev}_{\max} - t, t = 1..m$, а общее среднее на уровне 0. ■

Лемма 3. При ДНВПХ с произвольной длиной измерения z -мерного набора данных, в котором длины всех измерений равны, левая и правая части набора данных по каждому измерению на каждом уровне детализации $0.. \text{Lev}_{\max} - 1$ рассчитываются независимо друг от друга. ■

Теорема 1. Пусть задан z -мерный набор данных V с дополненными измерениями длиной 2^n , где координаты фиктивных значений лежат в интервале $[2^m..2^n - 1], m < n$. Пусть для этого набора получена декомпозиция W на основе ДНВПХ с произвольной длиной измерения. Если добавить в исходный набор данных значения с координатами $[2^m..2^{m+k} - 1], m + k \leq n$, то при вычислении декомпозиции W' обновленного набора данных V' на уровне детализации $n-m-1$ декомпозицией левой части V' будет являться декомпозиция W , а усредненным значением на данном уровне - $W(\{0,..0\})$. ■

В диссертационной работе предложен новый метод вычисления погрешности приближенной обработки данных. Он позволяет оценить распределения ошибки восстановления как исходного элемента данных, так и суммы элементов. Обозначим исходный набор данных как $V(v_i)(i = 1..2^n)$, а восстановленный как $\tilde{V}(\tilde{v}_i)(i = 1..2^n)$. Будем считать, что восстановленный набор был получен после обнуления некоторых нормализованных вейвлет-коэффициентов $c_j, j = 1..N^*, c_j \neq 0$ в соответствующей декомпозиции W . Таким образом, единичное исходное значение может быть представлено как $v_i = \tilde{v}_i + \Delta v_i$, где Δv_i - ошибка восстановления.

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

Пусть v - есть некоторое исходное значение. Выразим ошибку при восстановлении v следующим образом:

$$\Delta v = \sum_j \xi_j, j = 1..N^* \quad (3)$$

В выражении (3) ξ_j - есть случайная величина соответствующая использованию вейвлет-коэффициента c_j , который был обнулен и который должен был быть использован при восстановлении v . Величины ξ_j были признаны независимыми. В дальнейшем индексы будем опускать. В работе

было показано, что ξ_j имеет распределение на уровне $L \in [n-1..0]$, представленное в табл. 1. P_L^+, P_L^- - вероятности вхождения коэффициента 'с' с положительным и отрицательным знаком на уровне L , а k_L - денормализующий коэффициент.

Таблица 1.

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

Далее была решена задача нахождения вероятностей P_L^+, P_L^- . Было показано, что для того, чтобы некоторый коэффициент 'с' принял участие в восстановлении единичного исходного значения v на уровне детализации L необходимо выполнение 2-ух условий: попадание коэффициента на уровень L (вероятность P_L) и попадание этого коэффициента в область действия S (вероятность P_{LS}), соответствующую исходному значению. Для указанных вероятностей на основе структуры распределения областей действия вейвлет-коэффициентов были найдены выражения (4), где z – количество измерений.

$$P_L = \frac{2^z - 1}{2^{z(L+1)}}; P_{LS} = 2^{z(L+1-n)}, L \in [n-1, 0] \quad (4)$$

Из (4) были получены окончательные выражения для P_L^+, P_L^- :

$$P_L^+ = P_L^- = \frac{(2^z - 1)}{2^{zn+1}} \quad (5)$$

Из выражения (3) имеем, что Δv представляет собой сумму независимых случайных величин ξ_j , которые имеют распределение, представленное в табл. 1. С использованием центральной предельной теоремы Ляпунова в диссертационной работе было доказано, что при достаточно большом количестве отброшенных (обнуленных) коэффициентов (N^*) распределение $\Delta v = \sum_j \xi_j, j=1..N^*$ близко к нормальному.

Была исследована скорость сходимости распределения Δv к нормальному закону, которая может быть выражена следующим образом:

$$\sup_x \left| P\left(\frac{\Delta v}{B} < x\right) - \Phi(x) \right| \leq A' \cdot \frac{1}{\sqrt{N^*}}, \text{ где } A' - \text{некоторая константа.}$$

Для распределения погрешности восстановления элемента исходных данных Δv были вычислены выражения математического ожидания и дисперсии.

$$M_1(\Delta v) = 0; D_1(\Delta v) = \frac{2^{zn} - 1}{2^{zn}} \cdot \frac{1}{2^{zn}} \cdot \sum_{i=1}^{N^*} c_i^2 \quad (6)$$

Аналогичные рассуждения были проведены для случая вычисления суммы элементов. Вероятность использования некоторого коэффициента 'с' на уровне детализации L для вычисления некоторой суммы $\sum v$ рассчитывается с помощью следующих величин: P_L'' - вероятности попадания на уровень L (аналогично формуле (4)), P_{LS}' - вероятности попадания в область действия для суммы значений $\sum v$, P_{LNC}' - вероятности, что данный коэффициент не будет компенсирован. При этом соответствующие события являются независимыми. В работе были получены следующие выражения, где z – количество измерений, а t – мерность суммы:

$$P_{LS}' = 2^{(t-z)(n-L-1)}; P_{INC}' = \frac{2^{z-t} - 1}{2^z - 1}, t \in [1, z], L \in [n-1, 0] \quad (7)$$

Из (4) и (7) получаем окончательные выражения для P_L^+, P_L^- :

$$P_L^+ = P_L^- = \frac{2^{t(n-L)} \cdot (2^{z-t} - 1)}{2^{zn+t+1}} \quad (8)$$

Получены выражения для математического ожидания и дисперсии ошибки восстановления суммы элементов, которые представлены ниже:

$$M_{\sum}(\Delta V) = 0; D_{\sum}(\Delta V) = \frac{2^{(z-t)n} - 1}{2^{(z-t)n}} \cdot \frac{1}{2^{(z-t)n}} \cdot \sum_{i=1}^{N^*} c_i^2 \quad (9)$$

С помощью выражений (6) и (9) можно оценить доверительные интервалы ошибки восстановления как исходного элемента данных, так и суммы элементов.

Погрешность выполнения запроса общего вида можно оценить следующим образом. Пусть X - подмножество ячеек от 1 до M гиперкуба Γ , $M \ll \Gamma$. Поставим в соответствие i -ой ячейке этого подмножества переменную v_i . Представим запрос на множестве X в виде функции от многих переменных: $f(v_1, \dots, v_M)$. Пусть f - дифференцируемая функция. Ошибку выполнения запроса можно представить в виде дифференциала функции f : $\Delta f = \sum_{i=1}^M \frac{\partial f}{\partial v_i} \Delta v_i$, где Δv_i - достаточно малая ошибка восстановления единичного значения. Пусть $\{\Delta v_i\}$ - независимые случайные величины. Тогда получим:

$$M(\Delta f) = M_1(\Delta v) \sum_{i=1}^M \frac{\partial f(\tilde{v}_i)}{\partial v_i} = 0; D(\Delta f) = D_1(\Delta v) \sum_{i=1}^M \left(\frac{\partial f(\tilde{v}_i)}{\partial v_i} \right)^2 \quad (10)$$

где $\tilde{v}_i$ - восстановленное значение i -ой ячейки из множества X .

Например, для суммы M значений из (10) получим $D(\Delta f) = M \cdot D_1(\Delta v)$,
а для среднего - $D(\Delta f) = \frac{D_1(\Delta v)}{M}$.

В третьей главе “Разработка инструментального средства приближенной обработки запросов на основе ДНВПХ с произвольной длиной измерения” разработан проект инструментального средства (ИС) приближенной обработки запросов на основе методологии объектно-ориентированного анализа и проектирования с использованием унифицированного языка моделирования (UML - Unified Modeling Language). Приводится концептуальное, функциональное и структурное описание системы, описание архитектуры ИС, а также формализованное описание схемы БД и программного обеспечения. Для описания прикладной части ИС была разработана объектная модель, представленная в виде диаграммы классов.

Основной задачей, поставленной при разработке ИС, была возможность использования уже существующих механизмов организации хранения, управления и доступа к многомерным данным. В работе было предложено использовать многомерные БД OLAP-систем для размещения приближенного представления и осуществления запросов к нему. Таким образом, вейвлет-декомпозиция исходного набора данных представляется в виде OLAP-куба, содержащего вместо исходных значений вейвлет-коэффициенты. Доступ к вейвлет-коэффициентам осуществляется средствами языка MDX (Multidimensional Expressions), позволяющего оперировать многомерными данными.

В основу проектируемого ИС был положен модульный принцип, в результате чего были выделены два модуля и ряд библиотек, а также разработана схема их взаимодействия. На рис.3 приведена упрощенная структурная схема ИС.

Административный модуль позволяет управлять параметрами вейвлет-декомпозиции. С помощью библиотеки вейвлет-декомпозиции осуществляется вейвлет-преобразование исходного набора данных, а также его последующие обновления. Библиотека сжатия вейвлет-декомпозиции используется как для расчета погрешности восстановления значений при заданной степени сжатия, так и для расчета степени сжатия при заданной погрешности. С ее помощью осуществляется обнуление вейвлет-коэффициентов в полученной декомпозиции.

Пользовательский модуль предоставляет возможность сформировать запрос к сжатому представлению данных (OLAP-кубу) и отобразить полученный приближенный результат. Для восстановления исходных значений из вейвлет-декомпозиции используется библиотека приближенных вычислений: на основе пользовательского MDX запроса определяется набор коэффициентов, необходимых для восстановления запрашиваемых значений, затем формируется множество MDX запросов, возвращающих нужные вейвлет-коэффициенты, и вычисляется искомое значение.

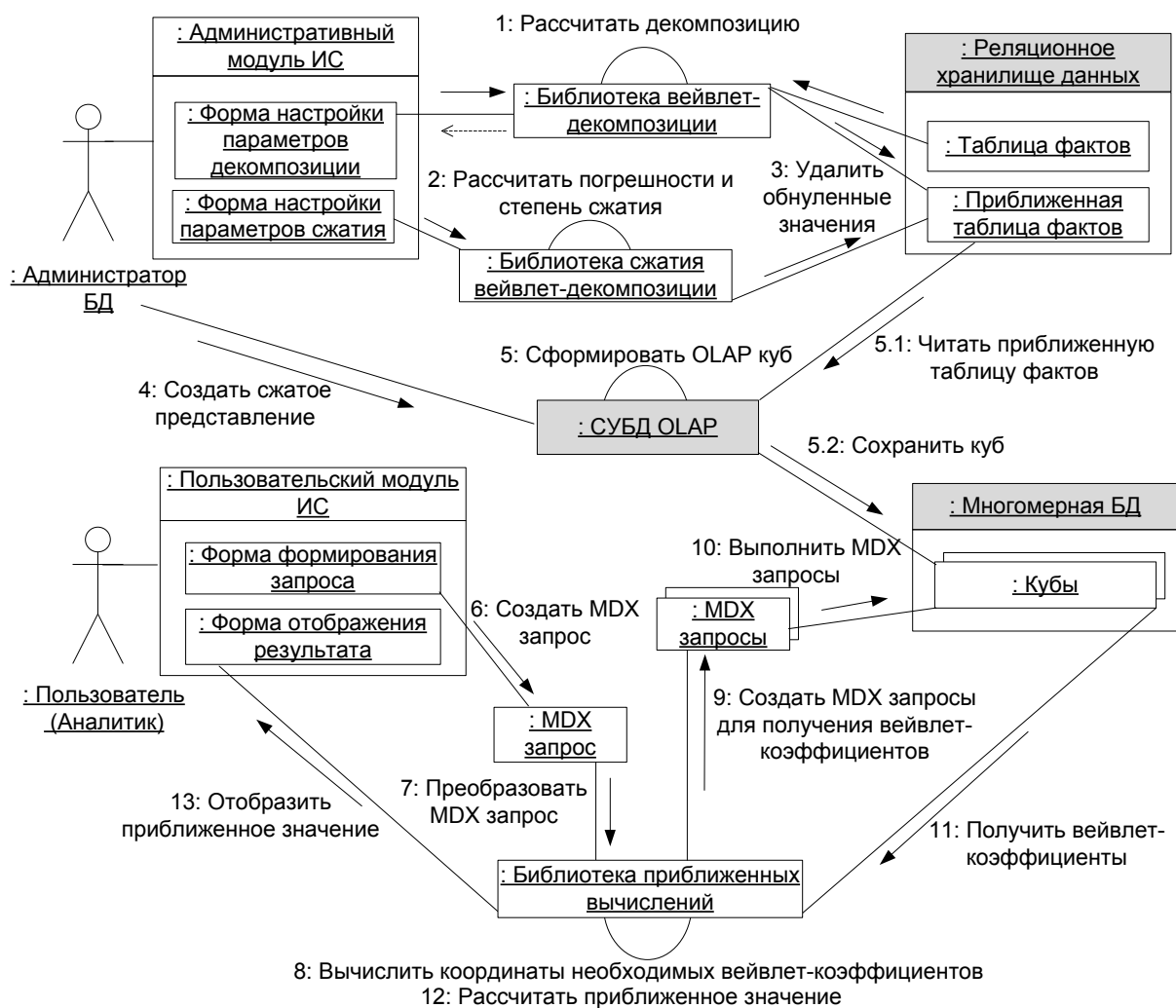


Рис. 3. Упрощенная структурная схема ИС

В четвертой главе “Использование разработанного средства для создания компонента OLAP-модуля системы “Надзор за заболеваемостью” приведено описание реализации компонента аналитического модуля системы “Надзор за заболеваемостью”, а так же исследованы результаты использования разработанного инструментального средства.

Система “Надзор за заболеваемостью” - государственная система сбора, передачи, хранения и анализа информации о заболеваниях человека и животных, а также связанных с ними данных об образцах и лабораторных тестах. Она реализована в виде распределенной БД с единым центром, в который стекаются данные, начиная с уровня районов. В год в системе регистрируется порядка 400 000 случаев заболеваний.

В задачи аналитического модуля системы, используемого на государственном и региональных уровнях, входят:

1. Поиск факторов, коррелирующих с появлением заболевания.
2. Прогнозирование возможных вспышек заболеваний.
3. Анализ качества постановки диагнозов.
4. Исследование факторов передачи зоонозных заболеваний.
5. Уточнение карт эпидемиологического расследования случаев.

Для оперативного решения подобного рода исследовательских задач было использовано разработанное инструментальное средство приближенной обработки запросов. Была выполнена формализация предметной области и выделены группы факторов, подлежащих анализу, на основе которых были реализованы приближенные многомерные представления.

Для примера на рис. 4 показано, что в результате использования разработанного ИС удалось сократить объем исходного набора данных на 60%, снизив время выполнения запросов на 52%, при этом погрешность вычисления суммы исходных элементов составила 5.8% при прогнозе 6.2%. Здесь и далее используется доверительный интервал на уровне σ .

Пример исследования *Какова зависимость количества выздоровевших от выбранного типа лечения в различных возрастных группах по стране?*

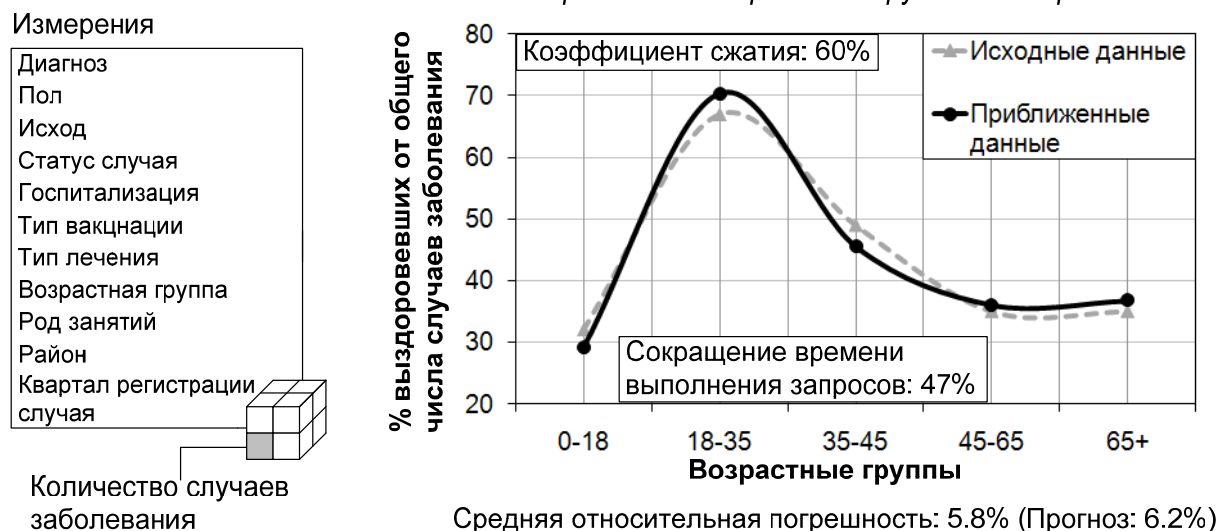
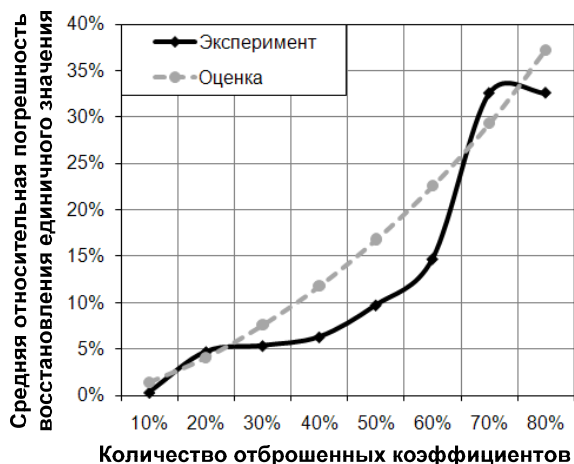


Рис. 4. Пример результатов использования ИС

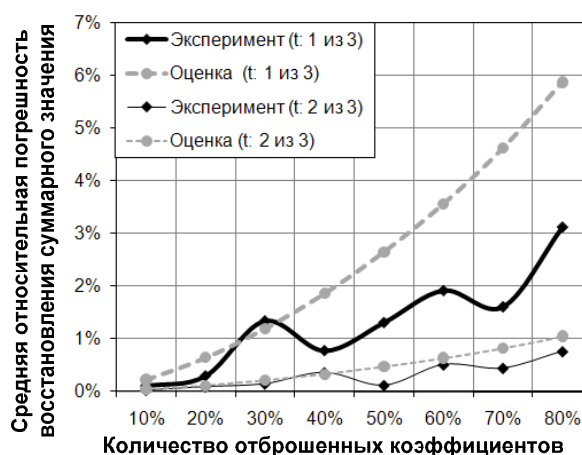
Разработанное ИС позволило провести исследование качества предложенных методов оценки погрешностей, а также проанализировать выигрыш в объеме хранилища данных и скорости выполнения запросов.

На рис. 5а представлено сравнение рассчитанной и экспериментальной средней относительной погрешности восстановления единичного значения (элемента данных) для различных степеней сжатия данных. Аналогичное исследование было проведено для суммы значений. При этом помимо объема данных, была рассмотрена зависимость погрешности восстановления суммы значений от количества слагаемых. На рис. 5б представлено сравнение рассчитанной и экспериментальной средней относительной погрешности восстановления суммы значений для различных степеней сжатия и количества слагаемых (в приведенном примере количество слагаемых определяется мерностью суммы t). Как видно, предложенные методы оценки погрешности верны при различных степенях сжатия данных и различном количестве элементов в сумме.

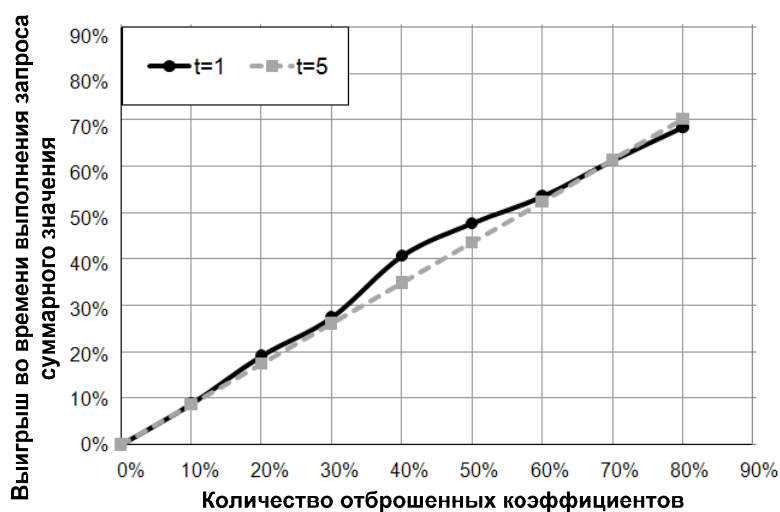
На рис.5в показан выигрыш во времени выполнения запроса вычисления суммы значений в зависимости от степени сжатия данных для различной мерности суммы t .



а



б



в

Рис. 5. Анализ результатов исследования

Анализ результатов исследования показывает, что при существенном сжатии исходного набора данных (порядка 60%) происходит значительное уменьшение времени выполнения запросов (в среднем на 50%), при этом погрешности вычисления единичного и суммарного значений остаются небольшими (15% и 5% соответственно), что позволяет изучать тенденции поведения исследуемых величин при изменении тех или иных факторов, а в целом говорит о целесообразности применения предложенного метода приближенной обработки запросов.

ЗАКЛЮЧЕНИЕ

1. Разработан метод получения приближенных представлений данных на основе вейвлет-преобразования Хаара, позволяющий обрабатывать массивы с произвольными длинами измерений и сокращающий число операций чтения/записи.
2. 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.

3. Предложен метод обновления имеющегося приближенного представления без пересчета ранее полученных значений. При этом обеспечивается возможность использовать разработанные методы оценки погрешности для обновленного набора данных.

4. 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.

5. Разработано инструментальное средство, реализующее предложенные методы приближенной обработки запросов на основе вейвлет-преобразования, которое является частью аналитического модуля системы “Надзор за заболеваемостью”.

6. Исследованы результаты использования разработанного инструментального средства и предложенных методов 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.

7. В дальнейшем планируется провести анализ и предложить методы получения приближенных представлений и оценки погрешностей при наличии заранее вычисленных агрегатов.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Григорьев Ю.А., Ухаров А.О. Обзор концепции многомерной модели данных в технологии OLAP // Проблемы построения и эксплуатации систем обработки информации и управления. – 2006. – №4. – С. 15-30.

2. Григорьев Ю.А., Ухаров А.О. Возможности приближенной аналитической обработки многомерных данных на основе вейвлет-преобразования // Проблемы полиграфии и издательского дела. – 2007. – №1. – С. 60-68.

3. Григорьев Ю.А., Ухаров А.О. Распределение ошибки при вейвлет-сжатии в хранилищах данных OLAP // Информатика и системы управления. – 2008. – №3. – С. 3-14.

4. Григорьев Ю.А., Ухаров А.О., Плутенко А.Д. Использование вейвлет-преобразования для приближенной аналитической обработки многомерных данных // Информатика и системы управления. – 2008. – №1. – С. 3-11.

5. Григорьев Ю.А., Ухаров А.О. Вейвлет-сжатие в хранилищах данных OLAP // Информатика и системы управления. – 2008. – №4. – С. 3-10.

6. Ухаров А.О. Использование вейвлет-преобразования для представления многомерных данных в системах OLAP // Интеллектуальные технологии и системы. – 2005. – №7. – С.256-271.