УДК 681.3

А.В. Бакулев, М.А. Бакулева

ПРИМЕНЕНИЕ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ ДЛЯ АНАЛИЗА ДАННЫХ ХРАНИЛИЩА

Рассматривается применение методов кратномасштабного анализа на основе вейвлет–преобразований в задачах оперативной аналитической обработки (OLAP) данных в хранилище. Приведены разработанные алгоритмы, позволяющие увеличить производительность аналитического процесса.

В настоящее время современные вычислительные системы и компьютерные сети позволяют накапливать большие массивы данных. По мере увеличения объемов и сложности хранимых данных и по мере их интеграции растет потребность в средствах аналитической обработки, а условия беспрерывно растущей конкуренции требуют, чтобы эта обработка была оперативной (время ответа не менее 5 секунд [1]).

Основу информационно-поисковых систем производственных предприятий, муниципальных структур и коммерческих образований составляют многочисленные БД, построенные на базе реляционной модели. Основной идеей реляционной модели является нормализация с целью экономии ресурсов памяти [2]. Сложные по структуре и многообразию связей реляционные БД не отвечают требованиям производительности аналитических приложений и соответственно не могут выполнять функции информационной поддержки процедуры принятия решений, так как современные требования оперативности менеджмента не сопоставимы с производительностью реляционных СУБД.

Поэтому в современных информационных системах наиболее востребованы денормализованные БД – хранилища данных (ХД). ХД является информационной основой оперативно-аналитических систем поддержки принятия решений (СППР) и, следовательно, от него зависит эффективность этих систем. В хранилище происходит интеграция данных из различных оперативных источников (ИОД). Затем выполняется предобработка данных (*очистка*, «*выравнивание*» [3], предварительный подсчет числовых показателей – *агрегирование*). После этого данные готовы для использования аналитическими приложениями.

Анализ информации является необходимым условием при планировании деятельности предприятия, принятии решений, поиске путей повышения производительности. Поэтому нахождение в большой базе данных структур, тенденций, аномалий и релевантной информации является одной из новых, наиболее впечатляющих задач СППР.

Для их решения предлагается привлечение методов кратномасштабного анализа на основе теории вейвлет-преобразований [4]. Классическая форма кратномасштабного анализа преобразует временной ряд в иерархическую структуру посредством вейвлет–преобразований. Иерархичность представления во многом упрощает анализ исследуемого процесса [5]. Верхние уровни иерархии позволяют обобщить смысл происходящего, выявить общую тенденцию, обозначить полярность результатов (положительные или отрицательные); если результаты не оправдывают ожидания, то рассматриваются нижние уровни иерархии с целью выявления того момента, с которого начался спад, и причин, его вызвавших. Базис вейвлетов позволяет осуществлять переход между уровнями иерархии кратномасштабного представления [6].

Из множества существующих базисов вейвлет-разложения базис Хаара является наиболее простым и быстродействующим [6]. Рассмотрим вейвлет-разложение данных в базисе Хаара [7]. Пусть временной ряд *W(t)* отображает показатели бизнес-процесса за исследуемый период, тогда кратномасштабное представление ряда получается путем попарного усреднения исходных значений, которое даст вдвое меньшее количество новых значений, над ними, в свою очередь, снова производится операция попарного усреднения, и так далее, вплоть до получения единственного коэффициента, представляющего общее среднее значение.

Пусть мощность данного ряда $|W(t)| = n$, тогда количество уровней иерархии p вычисляется по формуле $p = \log_2 n$. Кратномасштабное представление имеет вид (см. рисунок 1).

Одним из важнейших аналитических показателей бизнес-процесса является *тренд*, т.е. выраженная тенденция движения бизнеса (постоянный спад или подъем). Тренд показывает динамику развития бизнеса вне зависимости от периодических колебаний. Очевидно, что возрастающий тренд позитивных факторов (прибыли) либо убывающий тренд негативных (расходов) является показателем успешной деятельности предприятия.

Наиболее часто для формализации тренда временного ряда *x(t)* используется метод наименьших квадратов [8]. Функция тренда принимается линейной, вида $r = a * t + b$, $b = \bar{x}$,

где $\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$, $a = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(t - \frac{t(t+1)}{2})}{\sum_{t=1}^{n}(t - \frac{t(t+1)}{2})^2}$.

На основе кратномасштабного представления разработан алгоритм, превосходящий этот метод по быстродействию, что особенно актуально для современных аналитических систем.

Алгоритм выделения тренда. Пусть в исходном ряде *W(t)* необходимо выявить тренд на интервале [*x, y*].

Шаг 1. Длина интервала исходного диапазона: $h = y - x + 1$.

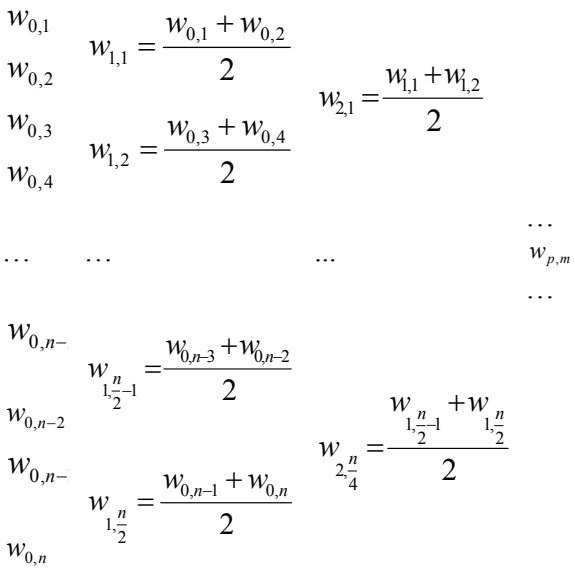


Рисунок 1 – Кратномасштабное представление ряда, где $w_{p,m}$ — элемент с номером *m* на уровне разложения *p*

Шаг 2. Базовый уровень вейвлет–разложения *l*, масштаб которого будет соответствовать размеру исходного диапазона: $l = \lfloor \log_2 h \rfloor$.

Шаг 3. $n_x = \left[\frac{x-1}{2^l}\right] + 1$ — индекс элемента W_{l,n_x} вейвлет-разложения на уровне *l*.

Шаг 4. $n_y = \left[\frac{y-1}{2^l}\right] + 1$ — индекс элемента W_{l,n_y} вейвлет-разложения на уровне *l*.

Шаг 5. уравнение тренда $r = a * t + b$, где $a = \frac{W_{l,n_y} - W_{l,n_x}}{y - x}$, $b = W_{l,n_x} - \left(\frac{W_{l,n_y} - W_{l,n_x}}{y-x}\right) * x$.

Другим не менее важным аналитическим показателем бизнес-процесса является его *периодичность*, т.е. повторяемость через определенные промежутки времени. Наличие информации о периодических составляющих бизнес-процесса и глубине их колебаний позволяет грамотно планировать деятельность предприятия на основании данных предыдущих периодов, а также эффективно распределять резервы в течение периода в зависимости от его фазы (спад, подъем, текущий пик или упадок).

Наличие или отсутствие периодичности в заданном диапазоне эффективно выявляется в кратномасштабном представлении. Индикатором периодичности является равенство коэффициентов разложения. В практических задачах абсолютное равенство этих коэффициентов встречается крайне редко, поэтому при сравнении допускается приближенное равенство с некоторой погрешностью ε.

Алгоритм поиска периодичности.

Алгоритм использует исходный ряд значений данных $W_{0,i}$, $i \in [0, n]$ и его однократное

вейвлет-разложение: $W_{j,i}, j \in [1, p]$, $p = \lfloor \log_2 n \rfloor$.

Вход: f — длина диапазонов исходного ряда, в которых ведется поиск периодичности; ε – погрешность вычислений периодичности.

Выход: сообщение о наличии или отсутствии периодичности, начальное значение интервалов, где наблюдаются равные гармоники, значение периода.

Шаг 1. Определение базового уровня вейвлет–разложения $l = \lfloor \log_2 f \rfloor$, масштаб которого будет соответствовать размеру исходного диапазона.

Шаг 2. Сравнение коэффициентов вейвлет-разложения масштаба l $W_{l,j}$ $(j = 1, \overline{\frac{n}{f}})$.

Если $W_{l,j} - W_{l,x} < \varepsilon$ $(x = j + 1, \overline{\frac{n}{f}})$,

то переход на шаг 3, $j := x$;

иначе вывод сообщения «на заданном уровне приближения периодичности не наблюдается».

Шаг 3. Вычисление начального значения интервала первой гармоники

$d_1 = (j-1) * 2^l + 1$;

$d_1 + f$ – длина первой гармоники.

Шаг 4. Вычисление начального значения интервала второй гармоники

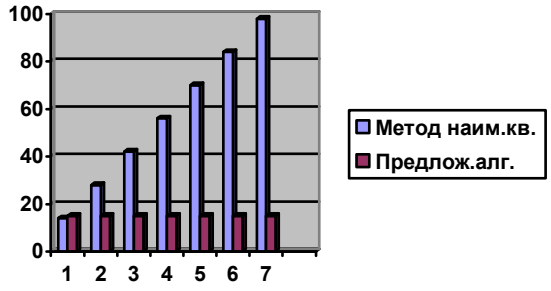
$d_2 = (x-1) * 2^l + 1$.

Шаг 5. Вычисление периода $p=(x-i)*f$.

Шаг 6. Вывод d_1, d_2, p.

Шаг 7. Конец.

Проведена оценка сложности предложенного алгоритма выявления тренда и метода наименьших квадратов от длины рассматриваемого интервала (рисунок 2).



Сложность метода наименьших квадратов–O(n). Сложность предложенного алгоритма–O(n^0).

Рисунок 2 – Сравнение алгоритмов

Таким образом, разработанные алгоритмы позволяют значительно повысить оперативность аналитических систем. Это достигается посредством наиболее адаптированного к анализу кратномасштабного представления данных хранилища [9].

Библиографический список

1. *Баргесян А. А., Куприянов М. С.*, Степаненко В. В., Холод И. И. Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004. 336 с.

2. *Дейт К.Дж.* Введение в системы баз данных. М.: «Вильямс», 1999. 848 с.

3. *Inmon, W.H.* Building The Data Warehouse" Third Edition: New York, John Wiley&Sons, 1996.

4. *Столниц Э., ДеРоуз Т., Салезин Д.* Вейвлеты в компьютерной графике. Ижевск: НИЦ «Регулярная и хаотическая динамика», 2002. 272 с.

5. *Острейковский В.А.* Теория систем: Учеб. для вузов по спец. «Автом. сист. обр. информ. и упр.». М.: Высш. школа, 1997. 240 с.

6. *Добеши И.* Десять лекций по вейвлетам. Ижевск: НИЦ «Регулярная и хаотическая динамика», 2001. 464 с.

7. *Бакулева М.А.* Применение современных математических методов для поиска информации в базах данных библиотек. Библиотековедение. Информационная деятельность: проблемы науки и практики: Материалы второй Международной научно-практической конференции. К., 2005. Ч.1. С. 165-167.

8. *Тюрин Ю.Н., Макаров А.А.* Анализ данных на компьютере. М.: ИНФРА-М, 2003. 544 с.

9. *Бакулева М.А* Применение вейвлет-преобразований для представления данных хранилища.// Вестник РГРТА. Научно-технический журнал. Выпуск 18. Рязань: РГРТА, 2006. С. 80-86.