УДК 004.627

© 2008 г. **Ю.А. Григорьев,** д-р техн. наук,
А.О. Ухаров
(Московский государственный технический университет им. Н.Э. Баумана),
А.Д. Плутенко, д-р техн. наук
(Амурский государственный университет, Благовещенск)

ИСПОЛЬЗОВАНИЕ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ ДЛЯ ПРИБЛИЖЕННОЙ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ МНОГОМЕРНЫХ ДАННЫХ

В статье рассматривается многомерная модель данных с указанием проблем, возникающих при аналитической обработке больших объемов данных. Предлагается методика приближенной обработки данных на основе вейвлет-преобразования, позволяющая работать с многомерными данными, обеспечивая хорошую степень сжатия и небольшую погрешность.

Введение

Одной из важнейших задач информационных систем, используемых на различных предприятиях, является аналитическая обработка данных. Означенный компонент в последнее время становится неотъемлемой частью систем автоматизации бизнеса. Это обусловлено спецификой современного производства, которое требует учета постоянно изменяющихся потребностей рынка. Своевременное выявление тенденций, обнаружение факторов, влияющих на эффективность производственного процесса, а также прогнозирование спроса позволяют эффективно организовывать работу предприятия.

Однако подобные задачи весьма нетривиальны. Невозможно заранее предугадать направление анализа, а следовательно, оптимизировать этот процесс с точки зрения автоматизации. Таким образом, возникает необходимость обрабатывать произвольные аналитические запросы к информационной системе, оперируя при этом всем накопленным объемом данных, что отрицательно сказывается на производительности, так как увеличивается время чтения/записи данных с носителей.

В этой связи представляется эффективным применение механизмов сжатия данных с потерями, что приводит к появлению понятия "приближенной" обработки данных. Очевидно, что вследствие исследовательской природы аналитических систем во многих случаях абсолютная точность не является первостепенной задачей обработки данных [1, 5]. Здесь в качестве ключевых параметров выступают гибкость анализа и скорость обработки запросов. Особенно справедливо это

на первых этапах анализа, когда важно выявить основные тенденции: например, получить закономерности между сферами деятельности заказчиков и услугами, предлагаемыми предприятием. Применение детального анализа для всех возможных вариантов нецелесообразно.

С другой стороны, использование приближенной обработки данных и сжатия с потерями может быть единственным решениям при жестком ограничении технических либо временных ресурсов. Наконец, в аналитических системах, требующих в качестве результатов анализа численные агрегированные значения, абсолютная точность этих значений, очевидно, не является принципиально значимой. Здесь важна тенденция их поведения.

Вейвлет-преобразование, представляющее собой математическое средство иерархической декомпозиции функций, совсем недавно было предложено в качестве метода приближенной обработки многомерных данных. Однако полученные результаты позволяют судить об эффективности такого подхода.

Многомерная модель данных

Системы аналитической обработки данных основаны на многомерном представлении информации. Здесь имеется в виду не многомерность визуализации, а многомерное представление при описании структур данных и поддержка многомерности в языках манипулирования данными.

Многомерное представление оперирует следующими основными понятиями:

1. Измерение – атрибут описания, – например, тип продукции.

2. Показатель – скаляр либо вектор исходных или агрегированных данных, – например, себестоимость.

3. N-мерный куб – совокупность связанных измерений и показателей.

Таким образом, многомерное представление есть представление данных в виде многомерного массива (гиперкуба), индексированного значениями измерений и содержащего в качестве элементов исходные либо агрегированные данные (показатели). На рис. 1 представлен пример куба. В целях наглядности показаны только три измерения, в действительности же их может быть больше. Здесь измерениями могут быть печатные работы, заказчик, размер изделия, цветность, способ печати, срочность, подразделения издательства и др. Структура измерений обычно является иерархией [4]. Например, в измерении «Время заказа» элемент «год» включает квартал, который, в свою очередь, включает месяц и неделю.

С точки зрения аналитика каждая ячейка куба есть мера, отражающая влияние параметров, находящихся на его осях. Количество измерений куба определяется составом анализируемых параметров. Тогда анализ данных будет сводиться к формированию срезов куба по тем или иным измерениям. Например, с помощью хранилища данных (см. рис. 1) могут быть решены следующие задачи оперативного ситуационного анализа:

1) определение профиля заказчика конкретного типа продукции (распределение тиражей конкретного типа продукции по заказчикам за все месяцы),

2) предсказание изменений ситуации на рынке (распределение тиражей

конкретного типа продукции по месяцам для определенного заказчика),

3) корреляционный анализ данных хранилища (зависимость объемов тиражей для разных заказчиков) и др.

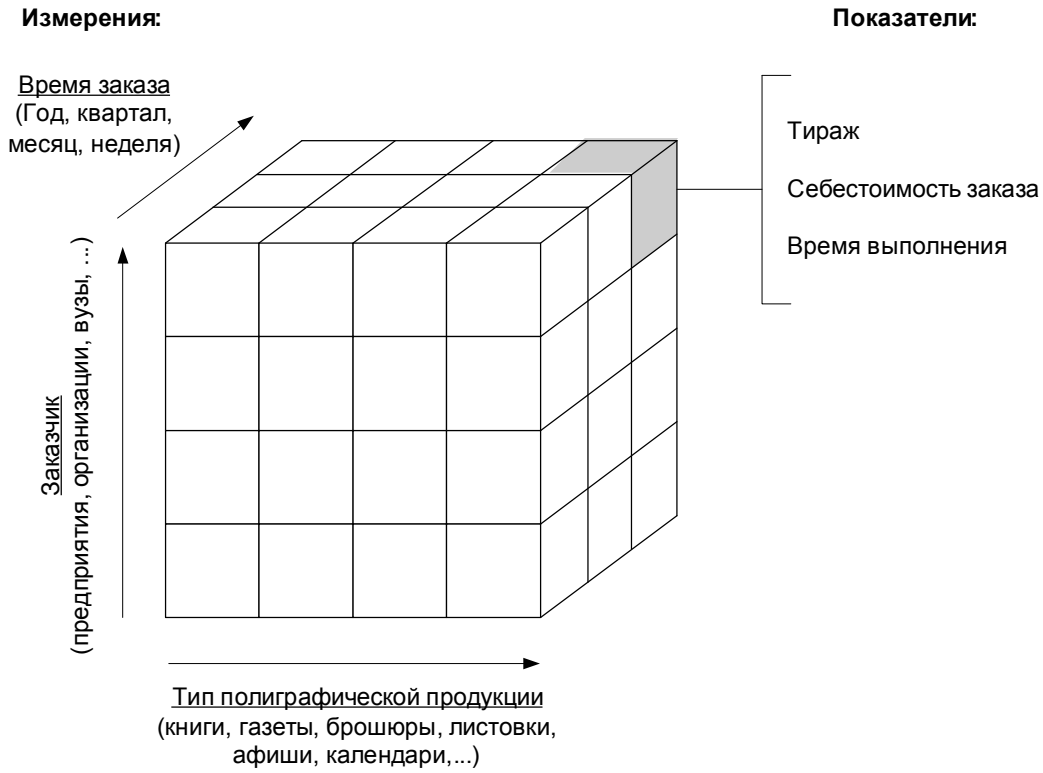

Рис. 1. Пример 3-мерного куба.

При рассмотренной структуре данных такие операции не потребуют больших вычислительных ресурсов. Однако этот способ хранения ведет к существенному увеличению объема данных. Например, если количество измерений равно 10, а каждое измерение включает 5 элементов, то число ячеек куба будет равно 5^{10}. При этом многие ячейки куба (значения показателя) будут нулевыми (например, не все потребители заказывают одну и ту же продукцию и т.д.). Обычно плотность данных в кубах очень мала (менее 10%). Однако в отличие от реляционной базы данных в многомерной модели не удается избавиться от нулевых значений, они включаются в результирующий набор [3].

Причина такой ситуации понятна. В случае одномерного массива можно отбросить те значения измерения, для которых ячейки являются нулевыми. При наличии нескольких измерений этого уже нельзя сделать, если хотя бы одна ячейка с данным значением координаты оказывается ненулевой. Это связано с тем, что соответствующее значение применяется при индексировании.

Для уменьшения объема данных используются различные методы физического сжатия. Но часто их применение является малоэффективным. Большой объем данных многомерного куба представляет серьезную проблему, сдерживающую применение систем аналитической обработки данных на практике. Актуальной становится задача сжатия данных на логическом уровне.

Одним из решений может стать применение дискретного вейвлет-преобразования к многомерным наборам данных. Вейвлеты долго и успешно используются в области сжатия сигналов и изображений [6, 7]. Основная идея применения этого метода для приближенной обработки данных заключается в преобразовании исходных данных в набор вейвлет-коэффициентов, где абсолютная величина и местоположение коэффициента определяют степень его влияния при восстановлении данных. Таким образом, можно отбросить те коэффициенты, значения которых близки к нулю. При этом не возникает большой ошибки при восстановлении данных и обеспечивается заданная степень сжатия.

Дискретное вейвлет-преобразование

Применительно к многомерным данным наиболее удобно использовать вейвлеты Хаара благодаря их свойству ортогональности [2, 6]. Формально вейвлет-преобразование может быть описано следующим образом. Исходный массив данных можно представить как кусочно-постоянную функцию. Так, набор из одного значения можно представить функцией, непрерывной на всем интервале [0,1), набор из двух значений – функцией, непрерывной на интервалах [0,1/2) и [1/2,1) и т.д. Каждое множество таких функций представляет собой пространство V^j, где j – количество интервалов, на которых функция постоянна. Очевидно, что $V^0 \subset V^1 \subset \ldots \subset V^n$.

Базис для пространства V^j может быть задан так называемыми масштабирующими характеристическими функциями:

$$f_i^j(x) = f(2^j x - i),$$

где $i = 0, 1, \ldots, 2^j - 1$.

Каждая такая функция характеризуется областью действия, т.е. той областью, где она не равна нулю. Область действия функции определяется следующим образом:

$$f(x) = \begin{pmatrix} 1,\ 0 \le x \le 1 \\ 0,\ \textit{иначе} \end{pmatrix}$$

Теперь определим еще одно пространство W^j как ортодополнение для V^j в пространстве V^{j+1}. Другими словами, W^j – это пространство всех функций в V^{j+1}, ортогональных всем функциям в V^j при определенном скалярном произведении. Совокупность множества линейно независимых функций y_i^j из пространства W^j называется множеством вейвлетов Хаара. Таким образом, можно считать вейвлеты из W^j инструментом для преставления той части функций в V^{j+1}, которую невозможно представить в V^j.

Функция y_i^j определяется следующим образом:

$y_i^j(x) = y(2^j x - i)$, где i = 0,1,… $2^j - 1$,

причем

$$y\ (x) = \begin{pmatrix} 1,\ 0 \le x < 1/2 \\ -1,\ 1/2 \le x < 1 \\ 0,\ \textit{иначе} \end{pmatrix}.$$

Например, одномерный массив {3, 1, 5, 7} может быть преобразован следующим образом.

Шаг 1. Запишем выражение в V^2:

$$\mathrm{F(x)} = c_0^2 f_0^2(x) + c_1^2 f_1^2(x) + c_2^2 f_2^2(x) + c_3^2 f_3^2(x),$$

где c_i^j коэффициенты, соответствующие исходным значениям массива.

Шаг 2. Перепишем теперь это же выражение в базисе V^1 и W^1:

$$\mathrm{F(x)} = c_0^1 f_0^1(x) + c_1^1 f_1^1(x) + d_0^1 y\ {}_0^1(x) + d_1^1 y\ {}_1^1(x),$$

где $c_0^1 = \dfrac{c_0^2 + c_1^2}{2} = 2,\ c_1^1 = \dfrac{c_2^2 + c_3^2}{2} = 6,$

$$d_0^1 = \frac{c_0^2 - c_1^2}{2} = 1,,\ d_1^1 = \frac{c_2^2 - c_3^2}{2} = -1.$$

Шаг 3. Перейдем к базису V^0, W^0, W^1:

$$\mathrm{F(x)} = c_0^0 f_0^0(x) + d_0^0 y\ {}_0^0(x) + d_0^1 y\ {}_0^1(x) + d_1^1 y\ {}_1^1(x),$$

где $c_0^0 = \dfrac{c_0^1 + c_1^1}{2} = 4,\ d_0^0 = \dfrac{c_0^1 - c_1^1}{2} = -2.$

После преобразований будет получен массив {4, –2, 1, –1}. Таким образом, мы описали последовательное усреднение и выявили детальные коэффициенты через представления исходной функции в различных базисах. Другими словами, произвели вейлет-разложение по базису Хаара. Каждый из описанных шагов соответствует определенному уровню детализации. Алгоритм обратного преобразования очевиден и заключается в последовательном выполнении операций усреднения коэффициентов (+ и -), начиная с шага 3.

Вейвлет-преобразование Хаара может быть распространено на многомерный случай. Воспользуемся нестандартным вейвлет-преобразованием, представляющим собой чередование операций усреднения и выделения уточняющих коэффициентов над строками и столбцами. На первом этапе реализуется один шаг одномерного вейвлет-преобразования для каждой строки, а затем – для каждого столбца. Далее рекурсивно повторяем эту процедуру только для областей, содержащих средние значения.

На рис. 2 представлен пример нестандартного 2-мерного вейвлет-преобразования.

Для восстановления исходных данных можно воспользоваться областями действия коэффициентов, представленными на рис. 3. В ненормированном случае в областях, обозначенных знаками '+' и '-', функции имеют значение +1 и -1 соответственно и значение 0 – в необозначенных областях [8]. Подматрица Q(i,j) со-

ответствует всей матрице A и указывает, с каким знаком (плюс, минус или нулевое значение) коэффициент $W_A(i,j)$ входит в сумму при восстановлении элемента матрицы A.

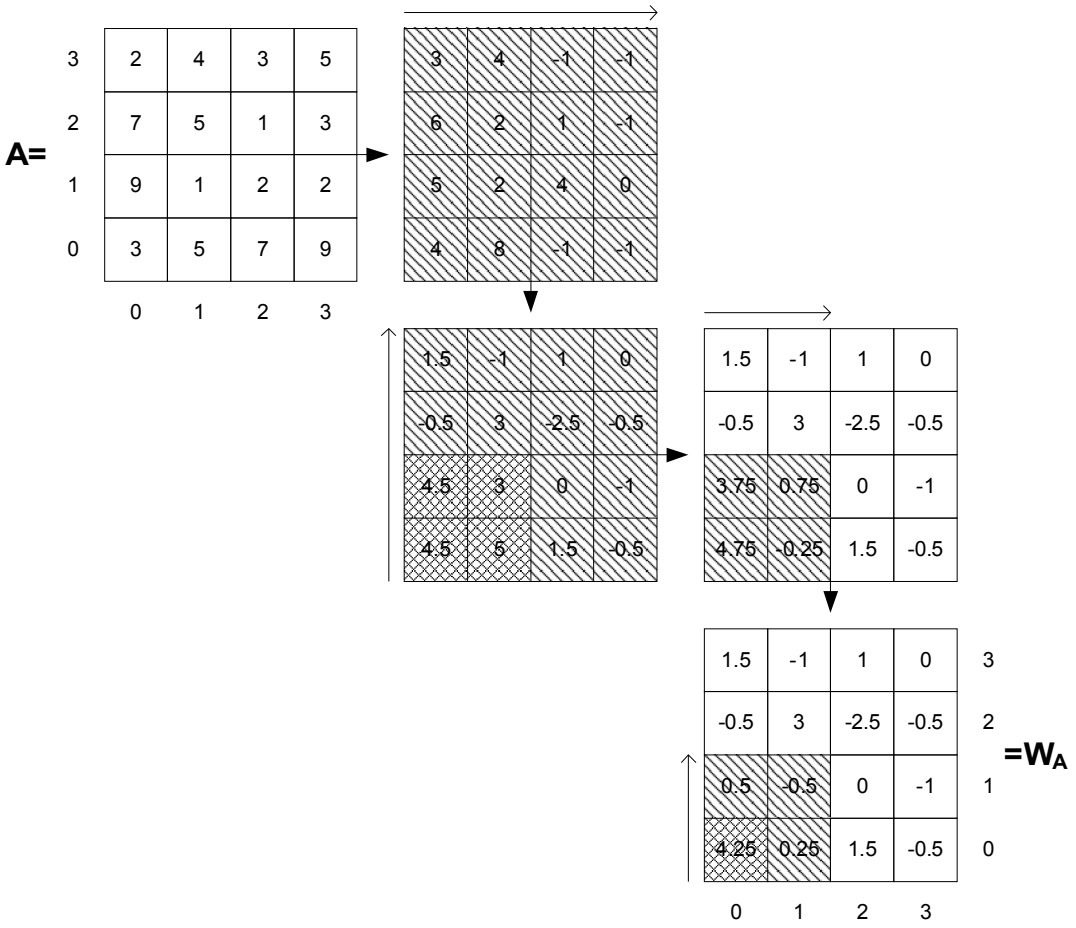

Рис. 2. Нестандартное вейвлет-преобразование.

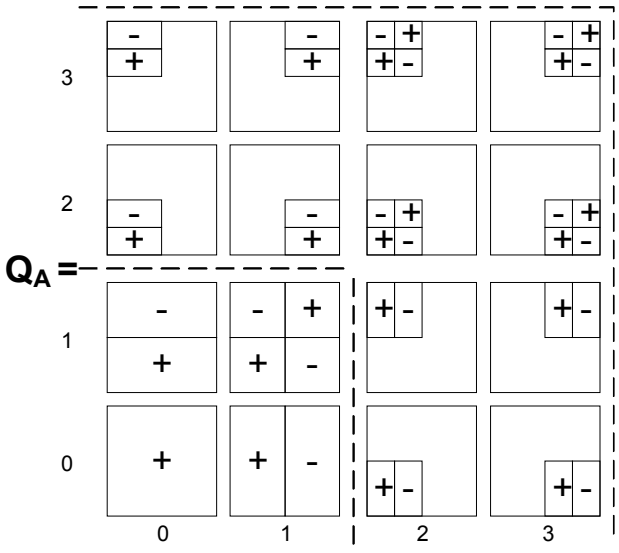

Рис. 3. Области действия вейвлет-коэффициентов.

Таким образом, принимая во внимания структуру областей действия функций, восстановим для примера значение исходного массива данных с координатами (1,1):

$A(1,1) = W_A(0,0) + W_A(1,0) + W_A(0,1) + W_A(1,1) – W_A(0,2) + W_A(2,2) –$
$– W_A(2,0) = 4.25 + 0.25 + 0.5 + (–0.5) – (–0.5) + (–2.5) – 1.5 = 1.$

Аналогично можно восстановить все остальные значения.

Приближенная обработка данных

Очевидно, что коэффициенты общих уровней (в приведенном примере это $W_A(0,0)$, $W_A(1,0)$, $W_A(0,1)$, $W_A(1,1)$) будут участвовать в восстановлении большего числа значений исходной матрицы, чем коэффициенты детальных уровней (остальные элементы матрицы W). Поэтому прежде чем осуществлять дальнейшие манипуляции с коэффициентами, их необходимо нормализовать. Нормализация будет заключаться в делении каждого коэффициента в пространстве j на $\sqrt{2^j}$. Подчеркнем, что это справедливо только для вейвлетов Хаара.

После этого можно приступить к процессу обнуления части коэффициентов в соответствии с заданной целевой функцией. Пусть у нас есть функция, выраженная через сумму базисных функций:

$$F(x) = \sum_{i=1}^{n} c_i u_i(x).$$

Задача состоит в нахождении функции, аппроксимирующей *F(X)* с меньшим количеством коэффициентов при заданном условии. Для примера рассмотрим одну из простейших постановок задачи сжатия: упорядочивание вейвлет-коэффициентов таким образом, чтобы первые m элементов последовательности давали приближение исходной функции с минимальной средней квадратической ошибкой.

Пусть p_i – перестановка n вейвлет-коэффициентов, а функция $J(x)$ – функция, использующая коэффициенты, соответствующие первым m членам перестановки p_i, т.е.:

$$J(x) = \sum_{i=1}^{m} c_{p_i} u_{p_i},$$

тогда квадрат погрешности определяется в виде

$$\langle F(x) - J(x) \mid F(x) - J(x) \rangle = \left\langle \sum_{i=m+1}^{n} c_{p_i} u_{p_i} \mid \sum_{j=m+1}^{n} c_{p_j} u_{p_j} \right\rangle =$$

$$= \sum_{i=m+1}^{n} \sum_{j=m+1}^{n} c_{p_i} c_{p_j} \langle u_{p_j} \mid u_{p_i} \rangle = \sum_{i=m+1}^{n} (c_{p_i})^2.$$

Последний шаг связан со свойством нормирования базиса функций Хаара. Таким образом, средняя квадратическая ошибка равна сумме квадратов отброшенных коэффициентов. Следовательно, оптимальной перестановкой будет та, которая располагает коэффициенты в порядке убывания их абсолютных величин.

Для примера, осуществим вейвлет-сжатие рассмотренного ранее массива данных. На рис. 4 представлен исходный массив данных (a), его нормализованное вейвлет-представление (b), вейвлет-представление после обнуления коэффициентов, абсолютные величины которых меньше либо равны 0,5 (c), восстановленный массив данных (d).

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

a

1.06	-0.69	0.69	0
-0.35	2.12	-1.77	-0.35
0.5	-0.5	0	-0.69
4.25	0.25	1.06	-0.35

b

1.06	-0.69	0.69	0
0	2.12	-1.77	0
0	0	0	-0.69
4.25	0	1.06	0

c

2.48	3.89	4.25	5.66
6.01	4.60	2.83	4.25
7.07	1.42	2.12	2.12
3.54	4.95	6.37	6.37

d

Рис. 4. Сжатие вейвлет-коэффициентов.

Расчеты показывают, что при обнулении 50% вейвлет-коэффициентов, т.е. уменьшении плотности данных на 50%, средняя ошибка при вычислении суммы значений по каждому измерению составляет всего 10,05 %. При этом минимальная ошибка равняется 0,93 %, а максимальная – 19,77%. Полученные результаты, безусловно, не могут служить оценкой степени сжатия и точности результатов восстановления при применении вейвлет-преобразования для других, более сложных случаев. Однако они позволяют судить в целом об эффективности и целесообразности применения данного подхода в системах аналитической обработки данных на основе многомерных моделей данных.

Приведенные результаты позволяют говорить о том, что даже при существенных степенях сжатия закономерности, которым подчиняются данные, мало изменяются. Последнее положение является принципиально важным для осуществления аналитической обработки данных. Например, пусть рассмотренный на рис. 4 куб содержит по вертикальной оси тип печатной продукции (0 – газеты, 1 – книги, 2 – брошюры, 3 – календари), по горизонтальной оси – время заказа (0 – 2001 г., 1 – 2002 г., 2 – 2003 г., 3 – 2004 г.), а в качестве показателя выступает тираж. Тогда точное изменение тиража брошюр по годам на основе всей совокупности данных и его приближенное изменение при 50-процентном сжатии будут демонстрировать одинаковые тенденции (рис. 5).

Заключение

Предложенный подход к приближенной обработке многомерных данных представляется перспективным. Вейвлет-преобразование может быть применено к массивам большой размерности без потери эффективности анализа данных. Оно обеспечивает небольшую ошибку расчетов при хорошей степени сжатия. При этом погрешность распределена по всему набору вейвлет-коэффициентов, благодаря чему сохраняются зависимости в данных. Таким образом, становится возможным производить анализ тенденций на основе сжатого набора данных.

Использование рассмотренной методики при анализе показателей функ-

ционирования предприятия позволяет эффективно оперировать большим объемом информации, обеспечивая выявление латентных зависимостей в данных.

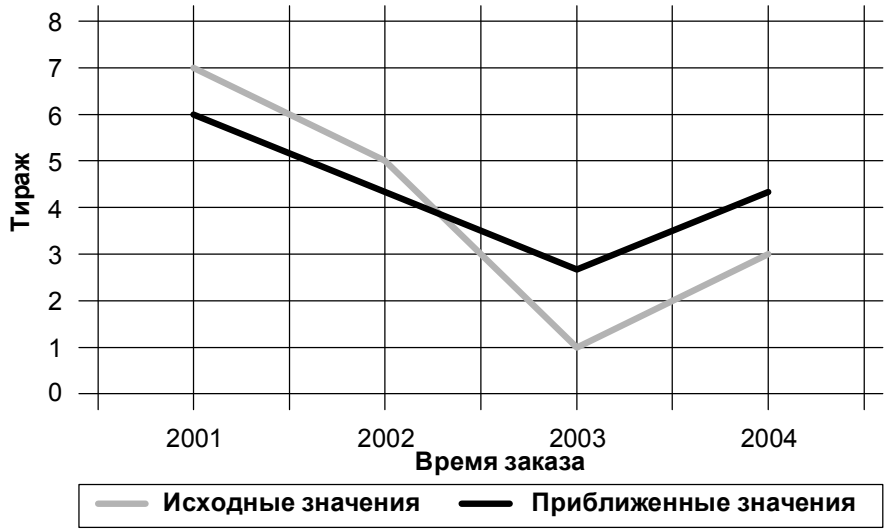

Рис. 5. Изменение тиражей брошюр по годам.

ЛИТЕРАТУРА

1. *Архипенков С.Я., Голубев Д.В., Максименко О.Б.* Хранилища данных. – М.: Диалог – МИФИ, 2002.
2. *Добеши И.* Десять лекций по вейвлетам / пер. с англ. – Ижевск: НИЦ «Регулярная и хаотическая динамика», 2001.
3. *Григорьев Ю.А., Ухаров А.О.* Обзор концепции многомерной модели данных в технологии OLAP// Проблемы построения и эксплуатации систем обработки информации и управления. Сб. статей. Вып. 4 / под ред. В.М. Черненького. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2006.
4. *Сахаров А.А.* Концепция построения и реализации информационных систем, ориентированных на анализ данных // СУБД. – 1996. – № 4. – С. 55-70.
5. *Спирли Э.* Корпоративные хранилища данных. Планирование, разработка, реализация. – Т.1. / пер. с англ. – М.: Изд. дом «Вильямс», 2001.
6. *Столниц Э., Дероуз Т., Салезин Д.* Вейвлеты в компьютерной графике. Теория и приложения / пер. с англ. – Ижевск: НИЦ «Регулярная и хаотическая динамика», 2002.
7. *Уэлстид С.* Фракталы и вейвлеты для сжатия изображений в действии / пер. с англ. – М.: Триумф, 2003.
8. *Chakrabarti K., Garofalakis M., Rastogi R., Shim K.* Approximate query processing using wavelets // The Very Large Databases. – N.Y. – 2001. – №10. – P.199-223.