



Вычислительные машины и программное обеспечение

УДК 004.651.4

А.М. Бородин, С.Г. Мирвода, С.В. Поршнеv

АНАЛИЗ МНОГОМЕРНЫХ ДАННЫХ ВЫСОКОЙ РАЗМЕРНОСТИ: ПРОБЛЕМЫ ДОСТУПА К ДАННЫМ И ВОЗМОЖНЫЙ ПОДХОД К ИХ РЕШЕНИЮ

A.M. Borodin, S.G. Mirvoda, S.V. Porshnev

HIGH DIMENSIONAL DATA ANALISYS: DATA ACCESS PROBLEMS AND POSSIBLE SOLUTIONS

Обсуждена проблема доступа к многомерным данным размерности порядка 10^4 . Описана система, в ходе создания которой была обнаружена проблема; описаны технические решения, применявшиеся при решении схожих проблем. Также изучены причины, по которым они не подходят для решения описанной проблемы. Указаны особенности существующих технических решений, позволяющие предположить перспективность использования фрагментов этих технологий в обсуждаемой задаче.

МНОГОМЕРНЫЕ ДАННЫЕ; OLAP-СИСТЕМЫ; ИНДЕКСИРОВАНИЕ; ДОСТУП К ДАННЫМ; ИЕРАРХИИ ТИПОВ ПАМЯТИ.

This paper discuss multidimensional access methods applied to 10^4 -dimensional data. Authors describe informational system, development of which encountered this problem. Than this problem is stated, decomposed and detailed. Existing techniques are referenced, theirs advantages and pitfalls are stated in context of the problem.

MULTIDIMENSIONAL DATA; OLAP SYSTEMS; INDEXING; DATA ACCESS; MEMORY HIERARCHY.

Многомерные структуры данных [1, 2] используются в системах обработки информации (СОИ), применяемых в различных областях человеческой деятельности, например, в бизнес-анализе [3], астрономии [4], геоинформатике [5], биоинформатике [6] и др. Анализ опыта их эксплуатации показывает, что одним из главных факторов, определяющих скорость работы СОИ, является скорость доступа к многомерным данным. В этой связи актуальна задача разработки быстрых алгоритмов доступа к многомерным данным. Следует отметить, что с момента начала разработки (80-е гг. XX века) для рассматриваемой

задачи в ее наиболее общей постановке алгоритмов прошло более 30 лет [2], в настоящее время достигнуты определенные успехи в данной области, однако универсальных автоматизированных алгоритмов, легко адаптирующихся для любой конкретной многомерной структуры данных, пока не создано. Как следствие, при разработке СОИ оказывается, что не удается использовать известные технологии доступа к многомерным данным без существенной модернизации либо создания новых алгоритмов, адаптированных к особенностям собираемой и обрабатываемой информации на конкретном производстве.

С подобной ситуацией мы столкнулись при реализации комплексного проекта по созданию высокотехнологичного производства с участием российского высшего учебного заведения по теме «Разработка автоматизированной системы слежения, контроля, моделирования, анализа и оптимизации полного цикла выпуска металлургической продукции на основе создания и интеграции математических моделей технологических, логистических и бизнес-процессов предприятия (АС ВМП)» (Шифр 2012-218-03-167). АС ВМП, состоящая из двух взаимодействующих модулей: автоматизированной информационной системы сбора и анализа данных производства (АИС САД) и автоматизированной информационной системы моделирования технологических, логистических и организационных (бизнес) процессов предприятия (АИС МОД).

Анализ исходной информации показал, что особенностью структуры данных, собираемых и обрабатываемых разрабатываемой АИС САД, является большое число разнородных измерений (список измеряемых технологических параметров, в т. ч. и в режиме реального времени, включает в себя более семи тысяч наименований), по каждому из которых потенциально может быть произведена выборка. При этом данные, характеризующие технологические процессы металлургического производства, являются высоко интегрированными и не могут быть разделены между различными программными модулями.

Один из возможных подходов к решению рассматриваемой проблемы, подтвержденный экспериментами, проведенными на этапе эскизного проектирования АС ВМП с тестовыми данными, основан на использовании самого быстродействующего на сегодняшний день серверного оборудования и программного обеспечения. Однако его практическая реализация приведет, как очевидно, к резкому удорожанию стоимости проекта. Кроме того, возможно, что в процессе продолжительной эксплуатации системы и накопления больших объемов данных в хранилище данных (ХД) выбранных вычислительных мощностей окажется недостаточно для обеспечения требуемой

скорости выполнения запросов к ХД. Описанная ситуация подтверждает актуальность задачи разработки быстрых алгоритмов доступа к разнородным многомерным данным высокой размерности (данные с общим числом измерений 10^4 и более).

В статье описывается общая постановка проблемы доступа к многомерным данным высокой размерности, изложены результаты анализа возможности использования известных методов доступа к многомерным данным, в т. ч. битовых карт, пространственного хеширования, метода соединения многомерных индексов, а также предложен подход, обеспечивающий быстрый доступ к многомерным данным большой размерности.

Проблемы доступа к многомерным данным высокой размерности

Каждая аналитическая система предназначена для выполнения тех или иных расчетов на соответствующих данных. В процессе выполнения расчетов потребляются вычислительные ресурсы, которые используются для выполнения двух условно независимых друг от друга задач: доступа к данным, хранящимся в ХД, и оперирования извлеченными данными.

Проиллюстрируем данный факт следующими простыми примерами.

Пример 1. В Microsoft Excel имеется таблица, содержащая колонки, названные А, В и С, соответственно. Колонка С содержит значения выражения, вычисляемого в соответствии с формулой $C = A + B$. Очевидно, что данную операцию можно представить в виде абстрактного синтаксического дерева, состоящего из одного узла, соответствующего операции сложения (оперирование данными), а процесс получения значений из конкретных ячеек А и В рассматривать как доступ к данным.

Пример 2. Построение на экране монитора трехмерной сцены. Здесь оперирование данными (в основном вычисление матричных произведений) является наиболее затратным с точки зрения использования вычислительных ресурсов. В то же время процесс извлечения данных требует значительно меньшего количества вычислитель-

ных ресурсов, чем оперирование данными, поскольку объем исходных данных, используемых в вычислениях, относительно невелик.

Пример 3. Поиск пути в графе знакомств большой социальной сети, связывающих двух и/или более пользователей. Здесь расчеты выполняются с минимальным количеством арифметических и условных операций, однако из-за большого объема исходных данных и, как правило, использования распределенного способа хранения этих данных, требуются значительные вычислительные ресурсы для доступа к этим данным.

Исторически сложилось так, что решение задач специализированного доступа к многомерным данным неразрывно связано с иерархией типов памяти. Отметим, что термин «иерархия» используется в т. ч. и по историческим причинам. Здесь, по сути, речь идет о способе упорядочивания типов памяти компьютера: от быстрой, ограниченной в размере и дорогой памяти (регистры процессора, кэш, ОЗУ), к медленной и дешевой памяти (дисковые системы, системы резервирования, облачные хранилища). Основная идея большинства методов доступа к многомерным данным основана на уменьшении количества обращений к медленной памяти за счет эффективного размещения данных в быстрой памяти.

Особенность доступа к многомерным данным состоит в том, что в большинстве случаев производится поиск данных (запросы), затрагивающий одновременно несколько различных свойств данных (различных измерений). Отметим, что здесь список свойств данных, которые могут участвовать в поисковом запросе, известен заранее. Однако их число, как правило, оказывается столь большим, что построение всех одномерных поисковых структур по всем комбинациям измерений оказывается практически невозможным. Например, запрос (который должен быть эффективным с точки зрения затрачиваемых ресурсов) к многомерной структуре данных, содержащей антропометрическую информацию по большому количеству человек, в котором требуется найти людей, имеющих в за-

данных интервалах длину носа, рост, дату рождения. Здесь интервалы, задающие требуемые значения длины носа, роста и даты рождения, — это критерии поиска, а собственно информационные поля, содержащие значения длины носа людей, занесенных в базу данных (БД), их рост и дата рождения — активные измерения поискового запроса. При этом в БД может быть множество различных измерений, но многомерная структура данных должна быть такой, чтобы выполняемые к ней запросы были одинаково эффективны при использовании в них любой возможной комбинации активных измерений из множества всех измерений.

Приведем примеры информационных систем, в которых реализуется обработка многомерной информации.

Пример 4. Геоинформационные системы. Данный класс СОИ, несмотря на то, что они традиционно оперируют с 3–4 измерениями (три пространственные координаты и время), также относят к классу информационных систем, в которых реализуется обработка многомерной информации. Это обусловлено высокой вариативностью используемых в них запросов, например, «поиск дома в заданном прямоугольнике», «поиск ближайшей АЗС к точке нахождения автомобиля» или «поиск оптимального маршрута с тремя и более пересадками» и т. д.

Пример 5. Системы бизнес-анализа. СОИ, относящиеся к данному классу систем, оперируют данными с несколькими десятками измерений поиска [7]. Например, запрос системы бизнес-анализа может выглядеть следующим образом: рассчитать сумму дохода от всех операций по определенной группе тарифов, в определенный интервал дат, в определенном районе города, для абонентов с подключенным цифровым телевидением и широкополосным доступом в Интернет.

На сегодняшний день создано множество различных структур данных, часть из которых описана в [8]. Большинство современных структур данных являются подмножеством т. н. *обобщенного сбалансированного поискового дерева* (GiST [9]), создаваемого

в результате разделения всего пространства данных на иерархически вложенные группы, которые являются узлами этого дерева. При этом группировка объектов в пространстве индексируемых данных осуществляется так, чтобы минимизировать число групп, затрагиваемых наиболее вероятными запросами к данным.

Выбор той или иной структуры доступа к данным зависит от структуры данных и структуры запроса. Для количественного обоснования выбора структуры сбалансированного дерева используют т. н. *модели эффективности поискового запроса*. Например, в [10] для оценки эффективности пространственных индексов, используемых для выполнения запросов к многомерным данным бизнес-анализа, предложено использовать следующую модель:

$$DA = 1 + \sum_{x=1}^{\lceil \log_f N \rceil} \frac{N}{F^x} \prod_{j=1}^D p(W(x, |s_j|), |q_j|),$$

здесь DA (disk access) – количество рассмотренных узлов древовидного индекса данных; N – количество индексируемых записей; F (fanout) – фактор ветвления древовидного индекса; $p(s, q)$ – вероятность пересечения диапазонов s и q ; $W(x, s)$ – средний диапазон данных по одному измерению, сгруппированный x раз; D – размерность индексируемых данных; s – средний диапазон, занимаемый одним элементом исходных данных,

$$p(|q|, |s|) = \frac{|s| + |q|}{(1 - |q|)(1 - |s|)},$$

$$W(0, |s|) = |s|,$$

$$W(x + 1, |s|) = \frac{\sqrt[x]{F} - 1}{\sqrt[x]{N(x)}} + W(x, |s|).$$

Кроме того, в [10] показано, что в общем случае при оперировании данными бизнес-анализа с количеством измерений порядка 10^1 , сложность расчета запроса с фиксированным результатом (количество данных, отвечающих критериям запроса) – отношение DA к размеру результирующей выборки имеет сложность $O(N \log N)$. Однако при больших D (например, порядка 10^4) $\sqrt[x]{F} \approx 1$ и $W(x+1, |s|) \approx W(x, |s|)$, это, в свою

очередь, означает, что межуровневая группировка данных оказывается неэффективной вследствие уменьшения диапазона рассматриваемых узлов дерева при спуске вниз по дереву индекса. В результате сложность расчета запроса с фиксированным результатом оказывается $O(N)$.

Проблемы индексирования многомерных данных высокой размерности при помощи битовой карты

Метод bitmap-индексирования [11] (метод битовой карты) по сравнению с методом сбалансированных деревьев позволяет реализовать более простую с вычислительной точки зрения технологию организации быстрого поиска данных по многомерным условиям. В основу данной технологии положены идеи, используемые в двоичной вычислительной технике. В методе bitmap-индексирования каждому из возможных значений измерения запроса ставится в соответствие битовая карта, в которую в зависимости от соответствия либо несоответствия значения полей каждой из строк БД значению полей запроса заносятся единицы или нули. При этом последовательность заполнения bitmap-карты повторяет последовательность участвующих в запросе информационных полей БД. Для поиска строк, удовлетворяющих условиям запроса с конъюнкцией, осуществляется битовое умножение карт. После этого оставшиеся в полученном произведении единицы соответствуют номерам строк, удовлетворяющих условию многомерного запроса. Описанная технология доказала свою эффективность и применяется в большинстве современных СУБД.

Данная технология позволяет эффективно определять идентифицирующую информацию строк данных. Однако в ней остается не до конца решенным вопрос доступа к данным, т. к. после нахождения строк, удовлетворяющих условиям запроса, далее по идентифицирующей информации еще необходимо получить собственно значения атрибутов найденных строк (lookup-операция). В терминологии, используемой в СУБД, это означает, что bitmap является некластерным индексом. Для физического

извлечения многомерных данных, например, в Microsoft Analysis Services, судя по комментариям одного из авторов данного метода, размещенным в его блоге [12], используется упорядочивание по кодам Гильберта. Однако в [13] описаны примеры, в которых данный метод оказывается малоэффективным при работе с данными большой размерности.

Проблему, возникающую при использовании lookup-операций для расчета запроса к многомерным данным, можно проиллюстрировать следующим примером. Предположим, что у нас имеется N строк с D измерениями, каждое из которых принимает в среднем M возможных значений. Данную информацию необходимо разместить в блоках данных по F строк так, что для каждого запроса с одним условием (общее количество возможных вариантов – $D \cdot M$), использовалось минимально возможное количество блоков информации DA , к которым придется обратиться при получении доступа к данным, удовлетворяющим критериям запроса. Однако когда $D \gg F$, минимизировать DA невозможно, если среди D измерений нет взаимозависимых. Другими словами, группировка строк данных по одному признаку, приводит к значительным отклонениям по другому признаку.

Здесь необходимо отметить, что в большинстве известных методов пространственного индексирования в значительной степени используется зависимость между измерениями. При этом пространственные индексы могут использовать отдельные локальные зависимости, обнаруживаемые в определенных подмножествах данных, которые, однако, считаются линейными. С нашей точки зрения представляется перспективным в методе bitmap-индексов при доступе к данным в высокой размерности использовать и нелинейные зависимости (например, между группами измерений).

Метод пространственного хеширования

Метод пространственно-чувствительного хеширования (Locality-sensitive hashing – LSH [6]) необходимо упомянуть в связи с тем, что он успешно применяется в

различных областях информационных технологий [14], работающих с многомерными данными большой размерности. Основная идея метода состоит в том, чтобы близким точкам в пространстве генерировать хэш-код, совпадающий с высокой вероятностью.

Применять метод LSH для доступа к данным можно тремя основными способами:

1. Генерировать для запроса и строки данных хэш-код таким образом, чтобы одному запросу соответствовал один хэш-код. Все записи, удовлетворяющие этому запросу, соответственно, должны иметь такой же хэш-код.

2. Генерировать хэш-код так, чтобы одному запросу соответствовало множество (в большинстве случаев – многомерный диапазон (параллелотоп)) хэш-кодов. Соответственно, записи, удовлетворяющие данному запросу, должны иметь один из возможных хэш-кодов – точку многомерного пространства, принадлежащую параллелотопу.

3. Генерировать хэш-код так, чтобы одному запросу соответствовал один хэш-код. При этом большинству записей, удовлетворяющих условию данного многомерного запроса, будет присвоен данный хэш-код.

Первый из описанных способов доступа к данным, по сути, является одномерным методом, поскольку для разных комбинаций измерений потребуются разные хэш-функции.

Второй способ при увеличении числа измерений подвержен «комбинаторному взрыву» объема параллелотопа. Как следствие, их производительность оказывается ниже производительности других известных методов доступа к многомерным данным.

Третий способ не гарантирует получение правильного результата выполнения запроса.

Метод соединения многомерных индексов

Одной из наиболее известных инженерных практик, распространенных в реляционных базах данных при решении данной задачи, является построение отдельного индекса по каждому измерению. При вы-



Структура кластерного индекса однотабличного хранилища многомерных данных

полнении многомерного запроса РСУБД выполняет соединение (join) результатов фильтрации отдельных условий соответствующими им индексами.

Мы предлагаем распространить данный подход на случай многомерных запросов и использовать единый кластерный индекс, структура которого представлена на рисунке. Здесь первая колонка (вершина пирамиды) содержит идентификатор измерения (#). Далее, за идентификатором в индексе размещается колонка индексируемого значения и за ней – строки данных.

К очевидным достоинствам предложенной структуры кластерного индекса однотабличного хранилища многомерных данных следует отнести простоту его построения и надежность (в части снижения вероятности логических ошибок в алгоритмах расчета многомерного запроса). Также важно отметить, что эффективность этого индекса оказывается независимой от количества измерений в индексируемых данных. В то же время можно ожидать, что данный показатель будет достаточно сильно зависеть от количества измерений, представленных в условиях запроса. Действительно, системе доступа к данным потребуется выполнить расчет всех выборок по отдельным условиям запроса, каждая из которых может оказаться значительно масштабнее выборки, соответствующей конъюнкции условий запроса.

Результаты проведенного анализа существующих подходов к организации доступа к данным для расчета многомерных запросов в системах с количеством независимых измерений данных порядка 10^4 показали, что ни один из них не является в полной мере универсальным инструментом решения рассматриваемой проблемы. В этой ситуации существует объективная необходимость в разработке новых подходов к ее решению. При этом представляется перспективным подход, основанный на использовании гибридной технологии, интегрирующей одну или несколько описанных выше методов доступа к многомерным данным. Отметим, что сегодня подобные технологии уже доказали свою эффективность в задачах доступа к многомерным данным с количеством измерений порядка 10^0-10^2 , например, т. н. BR-tree дерево [15], в узлы индекса которого внедрен битовый фильтр Блума [16].

Достоинство битовых карт, которое можно использовать в гибридном методе, – отсутствие явной зависимости производительности расчета запроса от количества измерений индексируемых данных. Поэтому битовые карты используются при решении задачи индексирования 10^4 -мерных данных. Кроме того, при использовании пространственных индексов оказывается возможным использовать локальные нелинейные зависимости между измерениями. Мы ожидаем, что сочетание этих технологий позволит создать систему анализа многомерных данных высокой размерности, в которой время получения результата расчета будет измеряться не десятками минут, а десятками миллисекунд.

Работа выполнена в рамках договора № 02.G25.31.0055 (проект 2012-218-03-167).

СПИСОК ЛИТЕРАТУРЫ

1. Gaede V., Günther O. Multidimensional Access Methods // ACM Computer Surveys. – 1998. – Vol. 30. – № 2. – P. 170–231.
2. Greene D. An implementation and performance analysis of spatial data access methods // In Proc. of the 5 IEEE Internat. Conf. on Data Engineering. – 1989. – P. 606–615.
3. Бородин А.М., Поршнев С.В., Сидоров М.А. Использование пространственных ин-

- дексов для обработки аналитических запросов и агрегирования многомерных данных в ИАС // Известия Томского политехн. ун-та. – 2008. – № 5. – С. 64–86.
4. Frialis M. Data Management and Mining in Astrophysical Databases // PhD thesis. – Univ. of Udine, Italy. – 2005.
5. Chang K.T. Introduction to Geographical Information Systems. – New York: McGraw Hill,

2008. — 184 p.

6. **Andoni A., Indyk P.** Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions // *Communications of the ACM*. —2008. —Vol. 51. —№ 1. —P. 117–122.

7. **Бородин А.М., Поршнева С.В.** Сравнительный анализ возможностей и скорости обработки многомерных данных программными средствами бизнес-аналитики на основе индексирующих структур основной памяти // *Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление*. —СПб.: Изд-во Политехн. ун-та, 2010. —№ 1(93). —С. 99–102.

8. **Гулаков В.К., Трубаков А.О.** Многомерные структуры данных. —Брянск: Изд-во БГТУ, 2010. —387 с.

9. **Hellerstein J., Naughton J., Pfeffer A.** Generalized Search Trees for Database Systems // *Proc. 21st Int'l Conf. on Very Large Data Bases*. Zürich. Sept., 1995. —P. 562–573.

10. **Бородин А.М., Поршнева С.В.** Аналитические способы оценки эффективности применения пространственных индексов в OLAP-системах // *Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации.*

Управление. —СПб.: Изд-во Политехн. ун-та, 2011. — № 2(120). —С. 93–100.

11. **Johnson T.** Performance Measurements of Compressed Bitmap Indices // *Proc. of 25th Internat. Conf. on Very Large Data Bases*. Sept. 7–10, 1999. —P. 278–289.

12. **Посуманский М.** Хроника номер 9. «Мы — великие умы» [электронный ресурс] / URL: <http://web.archive.org/web/20040306084024/>, <http://www.mosha.com/XRONIKI/win-xronika9.html> (дата обращения 13.10.2013).

13. **Бородин А.М., Поршнева С.В.** Алгоритмы быстрого доступа к многомерным данным в OLAP-системах. —Saarbrücken: LAP Lambert Academic Publishing, 2012. —176 с.

14. **Rajaraman A., Ullman J.** Mining of Massive Datasets. — Stanford University, California, 2010. — Ch. 3. —326 p.

15. **Yu Hua, Xiao Bin, Wang Jianping.** BR-Tree: A Scalable Prototype for Supporting Multiple Queries of Multidimensional Data // *Computers, IEEE Transactions on*. —2009. —Vol. 58. —Iss. 12. —P. 1585–1598.

16. **Bloom B.H.** Space/time trade-offs in hash coding with allowable errors // *Communications of the ACM*. —1970. —Т. 13(7). —P. 422–426.

REFERENCES

1. **Gaede V., Günther O.** Multidimensional Access Methods / *ACM Computer Surveys*. — 1998. — Vol. 30. — № 2. — P. 170–231.

2. **Greene D.** An implementation and performance analysis of spatial data access methods / In *Proc. of the 5 IEEE Internat. Conf. on Data Engineering*. — 1989. — P. 606–615.

3. **Borodin A.M., Porshnev S.V., Sidorov M.A.** Ispol'zovanie prostranstvennykh indeksov dlia obrabotki analiticheskikh zaprosov i agregirovaniia mnogomernykh dannykh v IAS / *Izvestiia Tomskogo politekhn. un-ta*. — 2008. — № 5. — S. 64–86. (rus)

4. **Frialis M.** Data Management and Mining in Astrophysical Databases / *PhD thesis*. — Univ. of Udine, Italy. 2005.

5. **Chang K.T.** Introduction to Geographical Information Systems. — New York: McGraw Hill, 2008. — P. 184

6. **Andoni A., Indyk P.** Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions / *Communications of the ACM*. — 2008. — Vol. 51. — № 1. — P. 117–122.

7. **Borodin A.M., Porshnev S.V.** Sravnitel'nyi analiz vozmozhnostei i skorosti obrabotki mnogomernykh dannykh programmnyimi sredstvami biznes-analitiki na osnove indeksiruiushchikh struktur osnovnoi pamiatii / *Nauchno-tehnicheskie vedomosti SPbGPU. Informatika. Telekommunikatsii.*

Управление. — St.-Petersburg: Izd-vo Politekhn. un-ta, 2010. — № 1(93). — S. 99–102. (rus)

8. **Gulakov V.K., Trubakov A.O.** Mnogomernye struktury dannykh. — Briansk: Izd-vo BGTU, 2010. —387 s. (rus)

9. **Hellerstein J., Naughton J., Pfeffer A.** Generalized Search Trees for Database Systems / *Proc. 21st Int'l Conf. on Very Large Data Bases*. — Zürich, Sept. 1995. — P. 562–573.

10. **Borodin A.M., Porshnev S.V.** Analiticheskie sposoby otsenki effektivnosti primeneniia prostranstvennykh indeksov v OLAP-sistemakh / *Nauchno-tehnicheskie vedomosti SPbGPU. Informatika. Telekommunikatsii. Управление*. — St.-Petersburg: Izd-vo Politekhn. un-ta, 2011. — № 2(120). — S. 93–100. (rus)

11. **Johnson T.** Performance Measurements of Compressed Bitmap Indices / *Proc. of 25th Internat. Conf. on Very Large Data Bases*. — Sept. 7–10, 1999. — P. 278–289.

12. **Posumanski M.** Khronika nomer 9. «Мы — velikie umy». Available <http://web.archive.org/web/20040306084024/http://www.mosha.com/XRONIKI/win-xronika9.html> (Accessed 13.10.2013).

13. **Borodin A.M., Porshnev S.V.** Algoritmy bystrogo dostupa k mnogomernym dannyim v OLAP-sistemakh. — Saarbrücken: LAP Lambert Academic

Publishing, 2012. – 176 p.

14. **Rajaraman A., Ullman J.** Mining of Massive Datasets. – Stanford University, California, 2010. – Ch. 3. – 326 p.

15. **Hua Yu, Xiao Bin, Wang Jianping.** BR-Tree: A Scalable Prototype for Supporting Multiple

Queries of Multidimensional Data / Computers, IEEE Transactions. – 2009. – Vol. 58. – Iss. 12. – P. 1585–1598.

16. **Bloom B.H.** Space/time trade-offs in hash coding with allowable errors / Communications of the ACM. – 1970. – T. 13(7). – P. 422–426.

БОРОДИН Андрей Михайлович – доцент кафедры радиоэлектроники информационных систем Уральского федерального университета, кандидат технических наук.

620002, Россия, г. Екатеринбург, ул. Мира, д. 32.

E-mail: amborodin@acm.org

BORODIN, Andrey M. Ural State Federal University.

620002, Mira Str. 32, Yekaterinburg, Russia.

E-mail: amborodin@acm.org

МИРВОДА Сергей Геннадьевич – ассистент кафедры радиоэлектроники информационных систем Уральского федерального университета.

620002, Россия, г. Екатеринбург, ул. Мира, д. 32.

E-mail: sergey@mirvoda.com

MIRVODA, Sergey G. Ural State Federal University.

620002, Mira Str. 32, Yekaterinburg, Russia.

E-mail: sergey@mirvoda.com

ПОРШНЕВ Сергей Владимирович – заведующий кафедрой радиоэлектроники информационных систем Уральского федерального университета, доктор технических наук, профессор.

620002, Россия, г. Екатеринбург, ул. Мира, д. 32.

E-mail: sergey_porshnev@mail.ru

PORSHNEV, Sergey V. Ural State Federal University.

620002, Mira Str. 32, Yekaterinburg, Russia.

E-mail: sergey_porshnev@mail.ru