САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

На правах рукописи

Михайлова Елена Георгиевна

Индексирование во временных базах данных

05.13.11 – Математическое и программное обеспечение
вычислительных машин, комплексов, систем и сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Санкт-Петербург 1997

Работа выполнена в Санкт-Петербургском государственном университете.

Научный руководитель: доктор физико-математических наук,
Новиков Борис Асенович

Официальные оппоненты:

- профессор, доктор технических наук **В.О. Сафонов**
- кандидат физико-математических наук **В.Н. Пономаренко**

Ведущая организация — Институт проблем информатики Российской академии наук

Защита диссертации состоится «____» _____ 199 г. в ____ час. ____ мин. на заседании диссертационного совета К 063.57.54 по присуждению ученой степени кандидата физико-математических наук в Санкт-Петербургском государственном университете по адресу: 198904 Санкт-Петербург, Старый Петергоф, Библиотечная пл. 2.

С диссертацией можно ознакомиться в библиотеке Санкт-Петербургского университета по адресу:
199034, Санкт-Петербург, Университетская набережная, 7/9.

Автореферат разослан «____» _____ 199 г.

Ученый секретарь
диссертационного совета **Б.К. Мартыненко**

Общая характеристика работы

Актуальность темы.

Автоматизированные информационные системы являются одной из важнейших областей применения вычислительной техники с момента ее появления. Принято считать, что хранимые данные в некотором смысле моделируют реальные процессы. Одним из наиболее важных аспектов моделируемой реальности является время, в котором происходят реальные процессы и события. При этом возникает задача моделирования и представления времени в системах баз данных.

В связи с этим в последнее время интенсивно развиваются концепции временных баз данных.

При выборе метода физической организации данных основными критериями являются использование памяти и скорость доступа. При тенденции к снижению стоимости внешних устройств хранения на первый план выходит скорость доступа к данным. Исходя из этого, разработка структур хранения временных данных, обеспечивающих быстрый доступ к данным, представляется весьма актуальной.

В диссертации рассматривалась задача:

- построение методов доступа для временных баз данных, эффективных для типичных запросов к данным;
- анализ предлагаемых методов;
- сравнение предложенных методов с другими известными методами доступа для временных данных.

Цель работы. Целью диссертационной работы является построение эффективных методов доступа для временных баз данных, анализ предлагаемых методов, сравнение предлагаемых методов с другими известными методами.

Методы исследования. При решении поставленных задач использовались методологические подходы, характерные для работ по моделированию временных баз данных и анализу эффективности методов доступа.

Научная новизна. Научную новизну составляют предложенные в диссертационной работе метод индексирования временной реляционной базы данных, учитывающий характерные особенности распределения запросов во временных базах данных; методы доступа к временным данным в объектно-ориентированной среде хранения; разработка модели для получения сравнительных оценок методов индексирования во временных базах данных.

Практическая ценность. Практическая ценность работы определяется возможностью использования ее результатов в системах управления базами данных, ориентированных на хранение временных данных. Результаты работы были также использованы при разработке программного прототипа, созданного в рамках системы "Синтез", ориентированной на создание интероперабельной среды неоднородных информационных ресурсов.

Апробация работы. Основные результаты диссертации докладывались на международном симпозиуме "Advances in Databases and

Information Systems" (Санкт-Петербург, 2-5 сентября 1997 г.), в международной школе по компьютерным технологиям, проводимой ACM (Албена, Болгария, 22-29 сентября 1997 г.), а также на семинарах лаборатории исследования операций. По теме диссертации опубликованы работы [1]-[2].

Структура и объем работы. Диссертация состоит из 7 глав, в том числе введения и заключения. Объем диссертации составляет 78 машинописных листов, Список литературы содержит 64 наименования.

Содержание работы

Первая глава диссертации представляет собой введение. В ней указаны основные направления развития систем управления базами данных в настоящее время.

В большинстве существующих в настоящее время систем баз данных хранится "актуальное состояние" информации (то есть хранимые данные отражают модель реальности, как она существует в каждый текущий момент времени). Существует, однако, ряд приложений баз данных, в которых необходимо моделировать и учитывать время более сложным образом. Например, в некоторых случаях может быть необходимо при любой модификации данных сохранять информацию о предшествующих состояниях. К ним относятся финансовые транзакции, множественные исторические данные, инженерный дизайн, медицинские записи и т.д. В связи с этим в последнее время интенсивно

развиваются концепции временных баз данных. В данной работе *временными базами данных* называются БД, содержащие информацию о своих предшествующих состояниях и о состояниях предметной области в различные моменты времени.

Даны критерии выбора методов физической организации данных: использование памяти и скорость доступа. В контексте временных баз проанализированы наиболее типичные запросы к данным:

- запросы к актуальным данным, т.е. фактам, отражающим актуальное состояние предметной области;
- запросы к полному содержимому базы данных, т.е. фактам, имевшим место когда-либо;
- запросы, относящиеся к определенным моментам времени или временным интервалам.

Вторая глава содержит анализ основных направлений исследований в области временных баз данных. Все направления можно разделить на две большие группы: концептуальный уровень и физический уровень представления временных баз данных. В этой главе упомянуты и кратко охарактеризованы известные статьи, посвященные концептуальным методам представления временных данных. К этому разделу относятся работы, посвященные семантике временных данных, выявлению специфических требований к временным данным, предъявляемых приложениями, временному моделированию в контексте реляционных и объектно-ориентированных баз данных, моделированию

временных систем хранения для баз данных реального времени, проблеме эволюции схем.

Далее дан обзор наиболее интересных методов доступа к временным данным. В основном, это многомерные древовидные структуры: *S*-, *MDM*- и *TSB*-деревья. Ряд работ ориентирован на устройства хранения с однократной записью. Актуальность таких методов в настоящее время невелика, поскольку емкость обычных дисков растет очень быстро, и одноразовые диски практически вытеснены другими технологиями. Наиболее интересным методом доступа для временных баз данных представляются деревья с расщеплением по времени - *TSB*-деревья.

Также в этой главе перечислены основные задачи временного моделирования в контексте объектно-ориентированных баз данных. Последний раздел второй главы посвящен индексированию в объектно-ориентированных системах баз данных в контексте временного моделирования. Эта область является слабо изученной и не имеет пока практического применения в существующих системах управления базами данных.

Третья глава представляет метод индексирования для временной реляционной базы данных. Эффективность методов организации баз данных в среде хранения зависит от операций, производимых над базой данных приложения. Этот метод построен в предположении, что наиболее важными являются актуальные данные базы данных, и в первую очередь ставилась задача обеспечить наиболее быстрый доступ к акту-

альным данным. Описываемый метод является гибридом TSB-дерева и хеширования.

Предполагается, что каждая запись базы данных однозначно идентифицируется ключом, и, кроме этого, снабжена временной меткой, соответствующей времени помещения записи в базу данных. Запись актуальна до тех пор, пока в базу не будет занесена новая запись с таким же ключом и новой временной меткой.

Метод хранения основывается на использовании двух уровней хеширования. Сперва файл делится на m групп, для адресации которых применяют функцию $H(K)$. На следующем уровне для каждого блока подбирается отдельная функция хеширования h_i которая перемешивает данные по блокам внутри группы. При переполнении группы одной из стратегий расщепления является выделение архивных данные в цепочку архивных блоков.

Для ускорения поиска в архивной базе данных создаются дополнительные индексы для временных срезов, относящихся к некоторым моментам в прошлом. Моменты, на которые создаются такие индексы, определяются в зависимости от ограничений на среднюю длину цепочки вершин архива.

Сделаны оценки данного метода и сравнение его с другой структурой хранения для временной реляционной базы данных - TSB-деревом. Зависимость скорости доступа от временной метки данных приведены в следующей таблице.

Время	TSB-дерево	Описываемый метод
t_{now}	3	1
$t_{now} - \delta t$	3	1
$t_{now} - 2 * \delta t$	3.4	4
$t_{now} - 3 * \delta t$	3.6	4

Здесь t_{now} - актуальное время.

Чтобы оценить среднюю скорость доступа к данным были сделаны предположения, что измерения начали производиться в момент времени t_0 - время первоначального заполнения базы данных. За время δt база пополнялась записями (на 5000 записей). Средняя скорость доступа к данным в предложенной структуре и при использовании TSB-дерева приводятся ниже.

Время	TSB-дерево	Описываемый метод
t_0	3	2.8
$t_0 + \delta t$	3.1	3.2
$t_0 + 2 * \delta t$	3.2	3.5
$t_0 + 3 * \delta t$	3.4	3.9

В целом можно оценить предложенный метод следующим образом. Доступ к актуальным данным при предлагаемой организации хранения осуществляется за одно обращение к диску (при условии, что таблица оглавления помещается в оперативной памяти). Среднее время доступа к архивным данным линейно зависит от средней длины цепочек переполнения.

В четвертой главе предложена структура хранения для временной объектно-ориентированной базы данных.

Предполагается, что в базе данных хранятся объекты, которые могут принадлежать различным классам. Объекты могут иметь подтипы и супертипы, т.е. может иметь место наследование классов.

Предполагается, что каждый объект базы данных однозначно идентифицируется объектным идентификатором, и, кроме этого, снабжен временной меткой, соответствующей времени помещения записи в базу данных. Данные никогда не удаляются - все актуальные и предшествующие версии состояний различных классов хранятся в базе данных.

Объекты могут содержать или ссылаться на вложенные значения атрибутов. Задача этого метода - организация поисковой структуры, которая позволяла бы находить объект данного класса, содержащий заданное значение атрибута в определенное время. При этом искомый объект может принадлежать любому классу в иерархии классов.

Метод основан на известном методе вложенных наследуемых индексов для объектов с вложенными атрибутами. В предлагаемом методе индекс строится по вложенным значениям атрибута.

Нелистовые индексные вершины содержат следующие записи:

$< value, adr >$,

где $value$ - значение атрибута, а adr - следующий индексный блок, содержащий записи с данным значением атрибута.

По листовым вершинам можно определить, какой объект в какое время содержал или ссылался на данное значение атрибута. Так как

объекты могут принадлежать к различным классам, листовые вершины содержат следующие записи:

$< class, list >$,

где *class* указывает на класс объектов, указатели которых перечислены далее, а *list* - список следующего вида:

$< value, oid, timestamp >$.

Здесь *value* - значение атрибута, *oid* - идентификаторы объектов, которые содержат или ссылаются на данное значение атрибута, а *timestamp* - время, когда этот факт имел место в моделируемой реальности.

При переполнении блоков данных можно производить расщепление двух видов: добавлять новый блок листовой данных, добавляя указатель на него в индексный блок, или выделять архивные данные в цепочку переполнения. Выбор стратегии расщепления зависит от характера данных, составляющих переполненный блок. Для ускорения поиска в архивной базе данных возможно создание дополнительных индексов для временных срезов, относящихся к некоторым моментам в прошлом. Моменты, на которые создаются такие индексы, определяются в зависимости от ограничений на среднюю длину цепочки вершин архива.

Предложенная структура дает кратчайший путь от значения атрибута до объекта верхнего уровня.

Оценки поиска при использовании предлагаемого метода индексирования приведены в следующей таблице.

Временная метка объекта	Скорость доступа
Актуальное время	$log_k G$
Архивное время	$log_k G + l/2$

Здесь k - количество ключей в индексной вершине, G - количество блоков нижнего уровня, а l - средняя длина цепочки архивных данных.

В пятой главе намечена структура хранения, которая является гибридом методов, предложенных во третьей и четвертой главах. Эта структура ориентирована на объектно-ориентированную временную базу данных. Доступ к актуальным данным осуществляется за одно обращение к диску, если таблица оглавления содержится в оперативной памяти. Описание операций и основные оценки этого метода совпадают с описанным выше методом индексирования для реляционной базы данных, и поэтому этот метод не рассматривается в диссертации подробно, а лишь намечена его схема.

В шестой главе описан программный прототип, в котором реализованы предложенные методы. Прототип разработан в рамках системы хранения "СИНТЕЗ". В первом разделе шестой главы кратко описана архитектура системы хранения. Второй раздел главы описывает программную реализацию предложенных методов. Приводятся экспериментальные оценки методов и делается сравнение этих оценок с теоретическими.

Глава седьмая является заключительной. В ней перечислены результаты данной работы.

Основные результаты

Основными результатами диссертации являются:

1. Исследование методов и структур хранения временных данных.
2. Разработка модели для получения сравнительных оценок методов индексирования во временных базах данных.
3. Сравнительный анализ методов индексирования временных баз данных.
4. Разработка метода индексирования для временной реляционной базы данных.
5. Анализ основных задач временного моделирования в контексте ООБД.
6. Разработка метода индексирования для временной ООБД.
7. Программная реализация предложенных методов.
8. Сравнение экспериментальных оценок с теоретическими.

Работы автора по теме диссертации

[1] Е.Г. Михайлова. Структура хранения для временных баз данных. "Программирование", N.6, 1997, стр.83-90.

[2] E. Michailova. Indices for Temporal Object Databases. Proceedings of the First East-European Symposium on Advances in Databases and Information Systems, St.-Petersburg, ADBIS-97, pages 92-94, September 2-5, 1997.