МЕТОДЫ ФИЗИЧЕСКОЙ ОРГАНИЗАЦИИ ДАННЫХ
ХЕШИРОВАНИЕ

Размещаемые записи идентифицируются с помощью ключа — поля фиксированной длины, располагаемого в каждой записи в одной и той же позиции. Ключ д.б. уникальным.

Хеширование — один из методов, позволяющих по первичному ключу записи получить ее физический адрес.

Ключ записи преобразуется в квазислучайное число, которое используется для определения местоположения записи. Число может указывать на адрес по которому расположена запись или на область в которой расположена группа записей. Область, в которой расположена группа записей называется участком записей или пакетом записей.

При первоначальной загрузке, адрес, по которому д.б. размещена запись определяется следующим образом:

- Ключ записи преобразуется в квазислучайное число от 0 до N, где N — количество пакетов записей;
- Полученное число преобразуется в физический адрес пакета;
- Если в пакете есть свободное место, то запись располагается в нем;
- Если нет, размещается в области переполнения.

Факторами, влияющими на эффективность размещения являются:

- Размер пакета;
- Плотность заполнения (отношение количества записей в пакете к максимальной вместимости пакета);
- Алгоритм преобразования ключа в адрес;
- Организация области переполнения.

Алгоритм преобразования ключа в физический адрес пакета выполняется в 3 этапа:

1. Ключ преобразуется в цифровое представление;
2. Цифровое представление ключа преобразуется в совокупность произвольно-распределенных чисел по возможности равномерно в диапазоне допустимых адресов. Ключ преобразуется в адрес;
3. Адрес умножается на константу для размещения адресов в основной памяти, т.е. фактическое масштабирование адреса.

АЛГОРИТМЫ

1. Метод деления: ключ делится на простое число (или число не имеющее малых делителей) близкое по значению к числу пакетов N. Остаток от деления — относительный адрес пакета.
2. Метод средних квадратов: ключ возводится в квадрат, выбираются центральные цифры. Умножаются на константу.
3. Метод сдвига разрядов: числовое значение ключа делится на две части, младшая часть складывается со старшей. Описанный процесс продолжается до тех пор, пока количество цифр результата не окажется равным количеству цифр числа пакетов N. Результат — относительный адрес пакета.
4. Метод складывания: число разбивается на 3 части, центральная содержит столько же цифр, сколько цифр в числе пакетов N, первая и третья заворачиваются и складываются.
5. Метод преобразования системы счисления: преобразуется основание системы счисления. Число представляется в новой системе счисления. Последние цифры числа — относительный адрес пакета.
6. Метод анализа отдельных разрядов ключа: из числа вычеркиваются те разряды, которые имеют распределение сильно отличающееся от равномерного.

ОРГАНИЗАЦИЯ ОБЛАСТИ ПЕРЕПОЛНЕНИЯ

Цепочки участков переполнения

Для связи основой области памяти с областью переполнения используются цепочки адресов. Т.е. пакет в основной области хранит адрес пакета (записи) в области переполнения.

Метод распределенной области переполнения

Пакеты переполнения размещены через определенные интервалы среди первичных пакетов. Если первичный пакет переполнен, направленная в него запись направляется в ближайший пакет переполнения, следующий за данным первичным пакетом. Достоинства — пакеты переполнения располагаются в непосредственной близости к первичным пакетам — отпадает необходимость в частом перемещении головок чтения-записи дискового устройства.

Алгоритмы перемешивания (хеширования) осуществляют преобразование ключа к номеру первичного пакета. Однако алгоритм, преобразующий номер первичного пакета в его машинный адрес, должен учитывать наличие пакетов переполнения между первичными пакетами.

both

ASM?

Если каждый 10-й пакет — пакет переполнения, то адрес текущего N-го пакета записей будет определяться по формуле:

Адрес пакета = B_0 + B(N+[N/9]) , где

B_0 — адрес первого байта;
B — размер пакета в байтах;
N — номер пакета, полученный на выходе алгоритма хеширования

Метод открытой адресации (Петерсона)
Происходит рассеивание записей переполнения в основной области.
Если пакет переполнен, то запись помещается в следующий за ним пакет.
Данный метод уменьшает время поиска записей т.к. обычно смежные пакеты расположены друг за другом на расстоянии, не превышающем длины дорожки. Однако если размер пакетов невелик, то это приводит к необходимости последовательного перебора пакетов в поисках свободного места. При размерах пакета более 10 записей этот метод эффективен при периодической реорганизации БД. Позволяет увеличивать плотность заполнения.

Справочник свободных пакетов. dynamic structure?how many (in general)?
Организуется специальный справочник, который содержит сведения о том, какие пакеты еще не заполнены. Если первичный пакет заполнен, то с помощью справочника определяется свободный пакет и после занесения в него записи в первичный пакет заносится ссылка на него. Необходимость в справочнике возрастает в случае частой модификации файлов. В качестве справочника может использоваться простой список заполнения пакетов. Метод эффективен, если размер пакета более 20 записей. Оптимизация предполагает загрузку файла, когда в основной области помещаются наиболее часто используемые записи. necessary?

ЗАДАНИЕ для выполнения

Этап № 1

Произвести анализ двух методов хеширования в соответствии с вариантом задания, предложенным преподавателем. Для этого:
Разместить 1- м из двух методов, файл содержащий последовательность записей. Каждая запись состоит минимум из трех полей: ключевое поле и несколько информационных полей. Длина последовательности не менее 1 000 000 элементов. Ключевое поле символьное (6 символов, ключ уникален). Из информационных полей одно строковое, другое числовое.
При размещении место выделить под 1 200 000 записей (20% прибавляется на расширение базы и несовершенство алгоритма хеширования).
Размещение производить по ключу, изменяя количество пакетов по следующему правилу:

- от 20 до 200 с шагом 20;
- от 200 до 2000 с шагом 200;
- от 2000 до 20000 с шагом 2000;
- от 20000 до 200000 с шагом 20000.

При каждом изменении числа пакетов длина пакета должна автоматически пересчитываться. Например, если пакетов 20, то число записей в пакете 60 000, если пакетов 200 000, то в каждом по 6 записей. При каждом переразмещении оценить плотность заполнения основной области и процентное отношение записей, попавших в область переполнения к общему числу записей. Результат каждого размещения поместить в отдельный файл. На основании полученных данных построить график.
Те же действия проделать с этим же файлом, но для второго метода хеширования.
Сравнив два графика оценить эффективность каждого метода для хеширования символьных ключей.
Выдвинуть предположение об оптимальном соотношении числа пакетов и записей в пакете.

Этап № 2

1. Произвести поиск записи по ключу в хешируемом файле. Ключ записи для поиска вводиться с клавиатуры.
2. Оценить время поиска набора записей в прохешированном файле, для чего:

Создать тестовую последовательность для поиска путем добавления к исходной 200-ста несуществующих записей (последовательность сохранить в отдельном файле). Оценить время поиска всех записей новой последовательности для каждого метода хеширования, в каждом из

созданных в процессе выполнения первой лабораторной работы файлов. На основании полученных результатов построить графики. Сравнить полученное оптимальное количество пакетов для хеширования с заявленным в предыдущей работе.
Выбрать оптимальный алгоритм хеширования и наилучшее для этого метода соотношение число пакетов/размер пакета. Файл, хранящий выбранную структуру, будем в дальнейшем именовать файлом, хранящим данные, размещенные с помощью алгоритма хеширования. Прочие файлы с результатами размещения методом хеширования могут быть удалены.

Варианты заданий:

1. Методы хеширования: "метод средних квадратов", "метод складывания"
   Область переполнения в виде цепочек пакетов переполнения.
2. Методы хеширования: "метод деления", "метод преобразования системы счисления"
   Распределенная область переполнения.
3. Методы хеширования: "сдвиг разрядов", " метод деления".
   Область переполнения в виде цепочек пакетов переполнения.
4. Методы хеширования: "метод деления", "метод складывания".
   Распределенная область переполнения.
5. Методы хеширования: "метод деления", "анализ отдельных разрядов ключа".
   Область переполнения в виде цепочек пакетов переполнения.
6. Методы хеширования: "метод средних квадратов", "сдвиг разрядов"
   Организация области переполнения методом открытой адресации (Петерсона)
7. Методы хеширования: "метод преобразования системы счисления", "метод складывания".
   Организация области переполнения методом открытой адресации (Петерсона)
8. Методы хеширования: "сдвиг разрядов", "анализ отдельных разрядов ключа".
   Организация области переполнения методом открытой адресации (Петерсона)
9. Методы хеширования: "метод средних квадратов", "метод преобразования системы счисления"
   Распределенная область переполнения.