Литература

I. Евдокимов В.П., Покрасс В.М. Методы обработки данных в научных космических экспериментах. М., "Наука", 1977.

И.С. Ризаев

СЖАТИЕ БАЗЫ ДАННЫХ

Запоминающие устройства с прямой записью (ЗУПД) позволяют создавать информационные системы, базирующиеся на единой интегрированной базе данных. Для эффективной работы с базой данных необходимо, чтобы прямой доступ был обеспечен в любое время ко всей базе. Но ввиду ограниченности дисковой памяти и больших объемов информации такой доступ, как правило, не обеспечивается. Решением этой проблемы может быть сокращение избыточности хранимой информации в базе данных путем уплотнения или сжатия этих данных.

Пусть $M=\{m_i\}$ - множество массивов информации, составляющих базу данных до сжатия; $M'=\{m_i'\}$ - множество полученных массивов информации после сжатия данных; $A=\{a_i\}$ - алгоритмы преобразования, $i=\overline{1,n}$. Видимо, сжатие данных можно проводить только в том случае, если будет выполняться соотношение

$$M' + \bar{A} < \bar{M},$$

где $\bar{M}$, $\bar{M}'$, $\bar{A}$ - объемы соответствующих массивов и алгоритмов.

Таким образом, задача сводится к созданию или подбору соответствующих алгоритмов сжатия. Существуют различные методы сжатия: машинные, немашинные, методы оптимального размещения, комбинированные. Хотя методы сжатия разрабатываются давно, применительно к базам данных они стали рассматриваться сравнительно недавно [I]. Большинство известных методов сводятся к рассмотрению сжатия информации, расположенной непосредственно в памяти ЭВМ. Чаще используют лексическое кодирование [2], суть которого заключается в том, что более длинному сообщению ставится в соответствие более короткое, а именно, порядковый номер данного сообщения. В справочниках или в памяти ЭВМ хранятся два словаря:

словарь слов или сообщений и словарь их номеров, представленных в виде двоичных кодов. При уплотнении записи с символьной информацией заменяются их двоичными номерами.

В работе рассматривалось сжатие информации применительно к базам данных, используемым в системах интегрированной обработки типа СИОД-1, СИОД-2 [3]. База данных, создаваемая ППП СИОД-1 и СИОД-2, состоит из массивов двух типов: главных и связующих. Каждый массив состоит из набора записей фиксированной длины. Формат записей Z представляет собой набор полей, часть из которых образуют системные поля $\{c\}$, а часть представляют поля пользователей $\{p\}$:

$$Z = <c_1, c_2, \ldots, c_m, p_1, p_2, \ldots, p_n>.$$

Поля записей могут быть обязательными и необязательными. Обязательные поля в любом входном массиве могут быть расположены произвольно, но их длина и местоположение задаются. Поля пользователей могут быть произвольной длины и произвольного содержания. Поскольку поля пользователей имеют значительно большую длину, чем системные поля, то уплотнение необходимо проводить именно по отношению к этим полям. Выбор тех или иных алгоритмов сжатия зависит от вида полей записей. В СИОД-1, СИОД-2 встречаются поля, включающие только символьную информацию, например наименование предмета. Встречаются поля, включающие только цифровую информацию, хранящуюся в упакованном формате, поля с двоичной информацией и смешанные поля, включающие как символьную так и цифровую информацию. Естественно, что для различных видов полей должны применяться различные виды сжатия. Например, для малосимвольных полей, длиной в один, два, три байта, могут подойти обычные таблицы перекодировок. Для полей длиной в четыре и более символов, таблицы перекодировок могут не подойти в связи с большими затратами времени на поиск информации. Для таких полей, как "наименование", выгоднее в памяти хранить набор дескрипторов (ключевых слов). В некоторых случаях для ускорения процесса преобразования исходного слова в сжатый код можно использовать методы однозначного преобразования без промежуточного обращения к таблицам перекодировок. Цифровую десятичную информацию выгоднее хранить в виде двоичных чисел путем непосредственного преобразования десятичных чисел в двоичные.

В общем случае задача сжатия должна решаться как оптимизационная. Критериями оптимизации являются заданный объем оперативной памяти и время перекодировок.

Пусть $\{a_j\}$ – множество алгоритмов сжатия информации p_i поля, $j = \overline{1,n}$, $i = \overline{1,m}$;

b_j – число ячеек памяти, требуемое для алгоритма a_j;

t_j – время сжатия, затрачиваемое алгоритмом a_j;

K_j – степень сжатия информации алгоритмом a_j;

m – число полей (реквизитов) в z записи.

Таким образом, запись z включает набор полей, предназначенных для сжатия p_1, p_2, ..., p_i, ... p_m.

Соответственно запишем матрицу алгоритмов сжатия:

$$\begin{matrix} a_{11} & a_{21} \dots & a_{i1} \dots & a_{m1} \\ a_{1j} & a_{2j} \dots & a_{ij} \dots & a_{mj} \\ a_{1n} & a_{2n} \dots & a_{in} \dots & a_{mn} \end{matrix} \Bigg\| A = \{a_{ij}\}$$

– множества алгоритмов сжатия.

Задача выбора вариантов сжатия будет состоять в том, что необходимо выбрать такой набор алгоритмов a_{ij}, чтобы их объем не превосходил S ячеек оперативной памяти, а время счета T было бы минимальным.

Введем переменные

$$x_{ij} = \begin{cases} 1, \text{ если } a_{ij} \text{ входит в } A \\ 0, \text{ в противном случае} \end{cases}$$

Тогда

$$\sum_{i=1}^{m} \sum_{j=1}^{n} b_{ij} x_{ij} \leqslant S;$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} K_{ij} x_{ij} \to max;$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} t_{ij} x_{ij} \to min.$$

Элементы x_{ij} должны удовлетворять следующему условию:

$$\sum_{i=1}^{m} x_{ij} = 1, \quad j = \overline{1,n}.$$

Если ввести ограничение на время T, то данную задачу можно свести к задаче линейного программирования.

При сжатии важно провести предварительную оценку временных соотношений. Такую оценку можно провести путем сравнения временных затрат на обращение к записям в ЗУПД и временных затрат на перекодировки.

Рассмотрим время доступа к записям, расположенным на магнитных дисках. Время, требуемое на доступ к данным в ЗУПД, и время на пересылку данных состоит из четырех частей: времени перемещения механизма доступа, время на выбор головки, задержки от вращения носителя информации и время на пересылку данных [4].

Время перемещения механизма доступа к цилиндру t', содержащему нужную запись, является функцией от числа цилиндров и может быть определено по следующей приближенной формуле:

$$t' = (40 + 0{,}5\,d)\,10^{-3}\ c\ .$$

Здесь d -количество пересекаемых дорожек ($1 \ll d \ll 200$). При этом t_{min} = 25 мс, t_{max} = 135 мс. Среднее время доступа составит $t'_{ср}$ = 90 мс.

Время на выбор головки очень мало и им пренебрегают. Задержка от вращения носителя составит t''_{max} = 25 мс, $t''_{ср}$ = 12,5мс. Время на передачу данных $t'''_{ср} \approx 6$ мкс $\approx 0{,}006$ мс. Таким образом, общее среднее время доступа к записи составит:

$$T_{ср} = t'_{ср} + t''_{ср} + t'''_{ср}\ .$$

К N записям $T = N T_{ср}$.

Сжатие информации будет в основном влиять на время $t'_{ср}$, так как при этом будет изменяться число пересекаемых дорожек. При сжатии в K раз получим:

$$t'^{сж}_{ср} = (40 + 0{,}5\frac{d}{K})\,10^{-3};$$

$$T^{сж}_{ср} = t'^{сж}_{ср} + t''_{ср} + t'''_{ср};$$

$$T^{сж}_{пол} = N T^{сж}_{ср}\ .$$

При оценке временных затрат на перекодирование необходимо исходить из того, что наибольшее время будет затрачиваться на поиски информации в таблицах уплотнения. Пусть система сжатия включает n таблиц с длинами L_1, L_2, ..., L_n. Средняя длина таблицы составит $L_{ср} = \frac{1}{n}\sum_{i=1}^{n} L_i$.

Число операций сравнения составит величину $Q_{ср} = L_{ср} q$, где q – цикл, выраженный количеством операций, затрачиваемых на выборку кода слова из словаря в стандартные ячейки и на сравнение с кодом исходного слова, $q \approx 20$ операций [2]. Время, затрачиваемое на один поиск, составит

$$t = \frac{1}{2} Q_{ср} t,$$

где t – время одной операции сравнения. При обращении к N записям $E_{пол} = NE$.

Время доступа к записям, расположенных на магнитных дисках, после сжатия составит величину

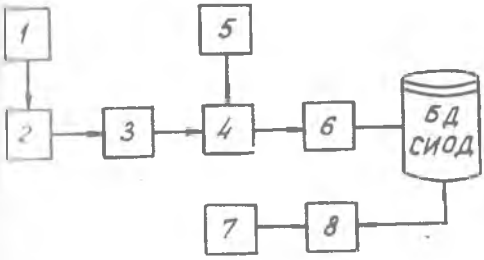
$$T^{сж} = T^{сж}_{пол} + E_{пол}.$$

Целесообразность применения алгоритмов сжатия определится из сравнения величин $T^{сж}$ и T.

Рассмотрим схему сжатия базы данных. Сжатие базы данных можно проводить тремя методами: ручным, полуавтоматическим и автоматическим. При ручном методе заранее составляются таблицы и методы уплотнения, которые хранятся в специальных альбомах. При вводе исходной информации производится поиск соответствующих сжатых кодов в альбоме и их ввод в ЭВМ. Декодирование производится аналогичным образом.

При полуавтоматическом методе таблицы и схемы уплотнения, составленные предварительно вручную, размещаются в памяти ЭВМ. При вводе исходной информации производится автоматическое обращение к таблицам и замена исходных кодов слов сжатыми кодами.

При использовании автоматической системы из исходной необработанной информации автоматически создаются таблицы и схемы уплотнения и производится автоматическое обращение к ним (рис. I).



Р и с. I.

Из исходных неуплотненных данных (блок I), находящихся на перфокартах или на МЛ, создается сводная ведомость (блок 2), в которой по каждому реквизиту записей будут представлены все значащие характеристики. Из сводной ведо-

мости создаются таблицы перекодировок (блок 3) или различные схе мы уплотнения. Блок оптимизации (5) предназначен для настройки программ уплотнения (4) на выбор тех или иных алгоритмов сжатия соответствии с заданными критериями оптимизации. Уплотненные дан ные (блок 6) размещаются на любом из внешних носителей, после че го, используя стандартную программу генерации, можно сформироват базу данных СИОД с уплотненной информацией.

Функциональные пакеты (блок 7) смогут работать с уплотненно базой данных СИОД через программы перекодировок (блок 8).

Проведенная подобным образом работа с системой интегрирован ной обработки СИОД позволила получить сжатие базы данных боле чем в два раза без существенных временных затрат.

Литература

1. *Alsberg Peter A. Space and time savings through large data base compression and dynamic restructuring.„ Proc. IEEE", 1975, 63, N8.*

2. Курбаков К.Н. Кодирование и поиск информации в автоматическом словаре. М., "Советское радио", 1968

3. Келехсаев А.А., Беляев А.П. Системы интеграции и обработки данных СИОД-1, СИОД-2. М., "Статистика", 1977.

4. Система ИБМ/360. Введение в запоминающие устройства прямого доступа и методы организации данных. М., "Статистика", 1974.