Мирошник С.Н.

Оптимизация времени доступа модулей к информации в базе данных реального времени

Аннотация: Рассматривается задача минимизации избыточности базы данных (БД), связанной с неиспользуемыми информационными полями модулями в БД. Предлагается алгоритм, решающий эту задачу. Используются понятия внутрифайловой, внутримодульной избыточностей, а также понятия вторичной внутрифайловой избыточности.

Ключевые слова: избыточность БД, неиспользуемые поля, рейтинги полей, критерий использования алгоритма

Одной из проблем БД является избыточность информации. Избыточность БД подробно исследована, что отражено в многочисленных публикациях. Это понятие используется в теории информации как превышение количества информации над ее информационной энтропией и как мера неопределенности информации в БД.

Во многих исследованных случаях под избыточностью понимается неоднозначность значений информационных полей, особенно когда они упоминаются в БД многократно. Чтобы избавиться от такой избыточности используется нормализация БД определенного типа. Избыточность приводит к дополнительным расходам памяти, и иных ресурсов компьютера. Это влияет на время доступа запросов к информации в БД в реальном времени. Последние замечания требуют определенной структуризации информации в виде специфической конструкции БД с минимальной избыточностью информации в БД.

В настоящей работе под избыточностью информации понимается число неиспользуемые модулями информационных полей. Предлагается алгоритм, минимизирующий эту избыточность.

Постановка задачи

Задан набор программных модулей $M_1, \dots, M_s$. Эти модули используют информацию из полей $\phi_1, \dots, \phi_r$. Для минимизации избыточности БД в работе [4] предлагается разделить все модули на файлы специальным образом с тем, чтобы минимизировать в этом случае межфайловую избыточность. Файлы могут использовать одинаковые поля. Каждый модуль принадлежит только одному файлу. В настоящей постановке задачи модули вместе с используемыми этими модулями полями объединим в двумерную таблицу, в котором поле ϕ помечается символом «1», если это поле используется каким-либо модулем, и «0», если поле не используется. Поля, помеченные «1», заполняются фактической информацией. Все поля пронумерованы натуральным рядом чисел. Каждый модуль *M* может использовать часть или все поля. Модуль в таблице задан номером своего первого поля, а также длиной своей записи *l*. Предполагается, что запись модуля занимает связный сегмент полей, помеченных символом «1».

В работе [5] введены понятия межфайловой, внутрифайловой и внутримодульной избыточностей. Пусть построены файлы $F_1, \dots, F_k$, образованные модулями $M_1, \dots, M_s$, использующие поля числом L_i, (L_i – запись файла F_i, $i = 1, \dots, k$). Файлы могут использовать одинаковые поля. Каждый модуль используется только в одном файле.

Межфайловая избыточность есть число общих полей для всех файлов $F_1, \dots, F_k$. Процедура построения файлов с минимальной межфайловой избыточности описана в работе [4] и обозначена I_2. Также в работе приведена верхняя оценка допустимого количества файлов для заданного числа модулей и числа полей.

Далее, пусть построен файл *F* с числом полей *L* (запись файла), образованный модулями $M_1, \dots, M_n$. Процедура построения файла использует понятие близости пары модулей или близость модуля к набору близких модулей. Близость модулей определяется сравнением числа совпадающих полей этих модулей и несовпадающих.

Внутрифайловая избыточность I_1, описанная в работе [5], образуется от разности длины *L* модуля, входящего в файл *F* с длиной *L* этого файла. Эта разность вычисляется по формуле

$I_1(F) = L \cdot n - \sum_{i=1}^{n} l_i$, где l_i – длина записи модуля M_i, n – число модулей в файле *F*. Внутримодульная избыточность I_3 есть число неиспользуемых модулем M полей в записи l. В данной работе эта избыточность в модулях появляется в результате работы алгоритма оптимизации. Здесь и далее индекс *i* используется только как индекс модуля M_i, $i = 1, \dots, n$.

Определение.

Рейтинг *P* поля ϕ есть число модулей, использующих это поле, и вычисляется по формуле $P_j = \sum \phi_j, j = 1, \dots, L$.

Пусть $N(\phi_j), j = 1, \dots, L$ есть номер поля ϕ_j в пронумерованной последовательности полей $\phi_1, \dots, \phi_L$. Эти же номера используются для нумерации рейтингов: $N(P_1), \dots, N(P_L)$, причем $N(P_j) = N(\phi_j)$, $j = 1, \dots, L$.

Модуль M_i использует поля $\phi_1^i, \dots, \phi_{l_i}^i$, где l_i – длина записи модуля M_i. Определим исходную избыточность $I_1(M_i)$ модуля M_i. Номер $N(\phi_1^i)$ есть число полей, предшествующих первому полю модуля, отсчитанного от $N(\phi_1)$.

Избыточность модуля M_i есть $I_1(M_i) = N(\phi_1^i)$.

Общее число полей записи модуля M_i, определяющее время запроса модуля M_i к БД, есть $I_1^i = I_1(M_i) + l_i$.

Требуется построить алгоритм, который минимизирует I_1^i, то есть вычисляет избыточность $\tilde{I}_1^i$ такую, что $I_1 \geq \tilde{I}_1^i$.

Алгоритм решения задачи основан на изменении порядка следования полей. Это изменение определяется упорядочиванием по убыванию связанных с этими полями рейтингами, помеченные в данном случае символом «~»: $\tilde{P}_1 \geq \tilde{P}_2 \geq, \dots, \geq \tilde{P}_L$.

Пусть $P = \max P_j$, $j = 1, \dots, L$. Рейтингу *P* соответствует поле *T* и $N(P) = N(T)$. Теперь исходный упорядоченный ряд полей $\phi_1, \dots, \phi_L$ становится неупорядоченным с номерами $N(\tilde{\phi}_1), \dots, N(\tilde{\phi}_L)$.

Установим связь между полями $\phi_1, \dots, \phi_L$ и $\tilde{\phi}_1, \dots, \tilde{\phi}_L$. Разделим модули M_i на две группы: V_i – модули, содержащие в своей записи поле *T*, W_i – модули, не содержащие поле *T*. Рассмотрим группу модулей V_i. Для этих модулей $P = n$, и тогда $T = \tilde{\phi}_1$, $N(T) = 1$. Для группы модулей W_i: $N(\tilde{\phi}_1^i) \neq N(\tilde{\phi}_1)$. Это означает, что $N(\tilde{\phi}_1^i)$ есть вторичная внутрифайловая избыточность.

Изменение порядка следования полей в записях модулей приводит к появлению внутримодульной избыточности $I_3(M_i)$. Запись модуля теперь не образует единого сегмента полей, то есть состоит из «1» и «0». Отсюда $\tilde{l}_i \geq l_i$, таким образом, вторичная внутрифайловая избыточность вместе с внутримодульной избыточностью $I_3(M_i)$ есть избыточность модуляM_i: $\tilde{I}_1(M_i)$.

Вычислим $\tilde{I}_1^i$.

Величина $\tilde{I}_1^i$ определяется номерами конечных полей $\tilde{\phi}_{\tilde{l}_i}^i$ модулей M_i с начальным полем $\tilde{\phi}_1 = 1$. Пронумерованные рейтинги $\tilde{P}_1, \dots, \tilde{P}_L$ есть $N(\tilde{P}_1), \dots, N(\tilde{P}_L)$. Неупорядоченный ряд номеров полей $N(\tilde{\phi}_1), \dots, N(\tilde{\phi}_L)$ пронумеруем. Получим ряд номеров: $N(N(\tilde{\phi}_1))$, ..., $N(N(\tilde{\phi}_L))$. Здесь $N\left(N(\tilde{\phi}_1)\right) = 1$ и т.д.

По-прежнему, $N(\phi_1^i), \dots, N(\phi_{l_i}^i)$ номера исходных полей модуля M_i, и $N(P_1^i), \dots, N(P_{l_i}^i)$ порядковые номера рейтингов этих полей. Здесь $N(\phi_j^i) = N(P_j^i), j = 1, \dots, l_i$.

Пусть $N(N(\tilde{\phi}_1^i))$, $N(N\left(\tilde{\phi}_{\tilde{l}_i}^i\right))$ есть начальные и конечные порядковые номера полей в записях модулей M_i длиной записи $\tilde{l}_i$. Здесь $N(N(\tilde{\phi}_1^i))$ определяет число вторичной избыточности модуля M_i. Получаем: $N\left(N\left(\tilde{\phi}_{\tilde{l}_i}^i\right)\right) = N\left(N(\tilde{\phi}_1^i)\right) + \tilde{l}_i$, где $\tilde{l}_i = l_i + I_3(M_i)$. Заметим, что поля $\tilde{\phi}_1^i$, $\tilde{\phi}_{\tilde{l}_i}^i$ а также их номера неизвестны. Номер $N(N\left(\tilde{\phi}_{\tilde{l}_i}^i\right))$ определяет общее число используемых и неиспользуемых полей модулей для реализации запросов к БД. Получаем: $\tilde{I}_1^i = N(N\left(\tilde{\phi}_{\tilde{l}_i}^i\right))$.

Способ вычисления $N(N\left(\tilde{\phi}_{\tilde{l}_i}^i\right))$ основан на использовании последовательности рейтингов $P_1, \dots, P_{l_i}$. Заданы ϕ_1^i, $\phi_{l_i}^i$ – начальные и конечные поля модуля M_i и их вычисленные рейтинги P_1^i, $P_{l_i}^i$. Пусть $P_H^i = \max(P_1^i, P_{l_i}^i)$, $P_H^i = \min(P_1^i, P_{l_i}^i)$. В ряду упорядоченных рейтингов $\tilde{P}_1, \dots, \tilde{P}_L$ с номерами $N(\tilde{P}_1)$, $N(\tilde{P}_L)$ рейтинги P_K^i, P_H^i именованы как $\tilde{P}_K^i$, $\tilde{P}_H^i$. Отметим, что $N(\tilde{P}_H^i) \neq 1$ для модулей из W_i. Для модулей из V_i: $P_H^i = P$ и потому $N(\tilde{P}_H^i) = 1$. Заметим, что если $N(P_j^i)$, $j = 1, \dots, L$ есть порядковый номер в

исходных полях, то $N(\tilde{P}_j^i)$ есть порядковый номер в упорядоченных номерах рейтингов и

$N(P_j^i) \neq N(\tilde{P}_j^i), j = 1, \dots, l_i$.

Задача свелась к поиску рейтингов $\tilde{P}_H^i$ и $\tilde{P}_K^i$ в ряду упорядоченных рейтингов $\tilde{P}_1, \dots, \tilde{P}_L$, а также их порядковых номеров $N(\tilde{P}_H^i)$ и $N(\tilde{P}_K^i)$. Получаем $N(\tilde{P}_K^i) = N(N\left(\tilde{\phi}_{\tilde{l}_i}^i\right))$, $N(\tilde{P}_H^i) = N\left(N(\tilde{\phi}_1^i)\right)$.

Окончательно: $\tilde{I}^i = N(\tilde{P}_K^i)$.

P.S. Поиск рейтингов $\tilde{P}_H^i$ и $\tilde{P}_K^i$ значительно упрощается в упорядоченном ряду рейтингов.

Далее, рассмотрим частный случай: $P_1^i = P_{l_i}^i$. Но в пронумерованных номерах рейтингов $N(\tilde{P}_1^i) \neq N(\tilde{P}_{\tilde{l}_i}^i)$.

Выберем $N(\tilde{P}_K^i) = \max\left(N(\tilde{P}_1^i), N(\tilde{P}_{\tilde{l}_i}^i)\right)$. Отсюда порядок следования одинаковых рейтингов в упорядоченном списке рейтингов не имеет значения, кроме их номеров.

Как отмечено выше, $N(\tilde{P}_1^i) = N(\tilde{P}_H^i)$ и $N(\tilde{P}_1^i) \neq 1$ для модулей из W_i, но для модулей из V_i: $N(\tilde{P}_1^i) = 1$. Получаем $N(\tilde{P}_K^i) = N(\tilde{P}_{\tilde{l}_i}^i)$. Отсюда $\tilde{I}_1^i = N(\tilde{P}_K^i)$. Кроме того определяем номер $N(\tilde{P}_H^i)$.

Теперь вычисляются значения для $\tilde{l}_i$ и $I_3(M_i)$:

$\tilde{l}_i = N(\tilde{P}_K^i) - N(\tilde{P}_H^i)$, $I_3(M_i) = \tilde{l}_i - l_i$.

Окончательно: $N(\tilde{P}_K^i) = N(N\left(\tilde{\phi}_{\tilde{l}_i}^i\right))$.

Неравенство $I_1 \geq \tilde{I}_1$ определяет возможность использования предложенного алгоритма для минимизации избыточности БД, образованной неиспользуемыми модулями полей.

Литература:

1. *Шеннон К.Э.* Работы по теории информации и кибернетики. – М.: ИЛ., 1963. – 829 с.

2. *Мартин Н., Ингленд Дж.* Математическая теория энтропии. – М.: Мир, 1988. – 251 с.

3. *Колмогоров А.Н.* Три подхода к определению понятия количества информации // Проблемы передачи информации. – 1965. –Т. 1. – № 1. – С. 3-11.

4. *Мирошник С.Н.* Построение верхней оценки межфайловой избыточности в БД реального времени. // Труды XXVI межд. конф. «Проблемы управления безопасностью сложных систем», 19 декабря 2018 г.– М.: ИПУ РАН, 2018. – С. 137-140.
5. *Мирошник С.Н., Гончар Д.Р.* Вычисление верхней оценки избыточности данных и её использование при определении времени доступа модулей БД реального времени. // Управление большими системами. – 2018. – Выпуск 76. Ноябрь – С. 254-265.