УДК 004: 519.876

ОПТИМИЗАЦИЯ СТРУКТУР РАСПРЕДЕЛЕННЫХ БАЗ ДАННЫХ НА РАННИХ ЭТАПАХ ПРОЕКТИРОВАНИЯ

Д.т.н. В.В. Бескоровайный[1], к.т.н. О.Н. Замирец[2], А.О. Сбитнев[1]

1. Харьковский национальный университет радиоэлектроники.

2. Государственное предприятие Научно-исследовательский технологический институт приборостроения, г.Харьков

Рассмотрены и проанализированы методы оптимизации структур распределённых баз данных на ранних этапах проектирования по критериям минимума затрат, минимального времени доступа и минимального объема передаваемой информации. Произведена сравнительная оценка эффективности методов, даны рекомендации по их эффективному использованию.

Розглянуто та проаналізовано методи оптимізації структур розподілених баз даних на ранніх етапах проектування за критеріями мінімуму витрат, мінімального часу доступу і мінімального обсягу інформації, що передається. Проведена порівняльна оцінка ефективності методів, надано рекомендації щодо їх ефективного використання.

Considered and analyzed methods of optimization of structures distributed database in the early stages of the design criteria for the minimum cost, minimum access time and the minimum amount of information transmitted. A comparative evaluation of the effectiveness of methods, recommendations for their effective use.

Ключевые слова: *распределенная база данных, проектирование, физическая структура, синтез, многокритериальная оптимизация.*

Введение

В системах управления государственного, регионального, муниципального и корпоративного уровня используется информация из множества территориально рассредоточенных источников. Это приводит к необходимости создания соответствующих рассредоточенных систем информационных инфраструктур, основными элементами которых являются распределенные базы данных (РБД).

При декомпозиции процесса проектирования РБД выделяют множества взаимосвязанных задач, условно объединяемых в этапы концептуального, логического и физического проектирования, посвященных соответственно синтезу их концептуальной, логической и физической структур [1– 3].

Большинство задач проектирования РБД являются слабо формализованными, имеют многокритериальный, комбинаторный характер, высокую размерность, что делает проблематичным поиск эффективных и устойчивых проектных решений. При этом стоимостные и функциональные характеристики РБД во многом определяются их физическими структурами, что определяет важное место задач синтеза физических структур в процессах их автоматизированного проектирования.

Для повышения эффективности технологии проектирования РБД производится разработка математических моделей и методов, учитывающих специфику большинства практически интересных задач.

Анализ предметной области и постановка задачи

Физическая структура РБД определяет физическую реализацию ее информационной структуры на существующей или создаваемой компьютерной сети [4].

Среди задач синтеза физической структуры РБД, описанных в работах [5 – 9] выделяются две большие группы: задачи, решаемые в условиях ограничений на функциональные характеристики (время доступа, надежность, живучесть) по стоимостным критериям, и задачи, решаемые в условиях ограничений на затраты по функциональным критериям.

Основной целью физического проектирования РБД является выбор таких методов физической структуризации данных, при которых обеспечивается экстремум заданного критерия эффективности ее функционирования.

Проведенный анализ проблемы синтеза физических структур РБД показал, что в подавляющем большинстве работ рассматриваются однокритериальные задачи, а существующие математические модели и методы, как правило, ориентированы на конкретный вид компьютерной сети, что не позволяет использовать их для решения задач синтеза физической структуры РБД в сетях с другими структурами [5 – 9].

Большинство задач синтеза физических структур РБД сводится к задачам целочисленного математического программирования с булевыми переменными. Среди методов решения задач синтеза физической структуры РБД наибольшее распространение получили комбинаторные (метод неявного перебора, метод ветвей и границ, метод Балаша), позволяющие получать оптимальные решения однокритериальных задач. Однако, они имеют экспоненциальную временную сложность, требуют значительных временных и вычислительных затрат, что делает их малопригодными для решения практических задач с размерностью в несколько десятков переменных.

Среди приближенных методов наибольшее применение для решения задач синтеза физических структур получили методы, построенные на эволюционных алгоритмах [9 – 10]. Они имеют гораздо меньшую временную сложность, но не гарантируют получение оптимальных решений.

Для оценки показателей оперативности РБД используются аналитические и имитационные модели теории массового обслуживания [9 – 13], которые в большинстве своем также ориентированы на конкретный вид компьютерных сетей.

В рамках агрегативно-декомпозиционного и блочно-иерархического подходов в работе [11] выполнена декомпозиция проблемы на комплексы задач

макро- и микроуровня. Это позволило определить место задач синтеза физических структур среди всего комплекса задач синтеза РБД и разработать информационные модели технологии проектирования РБД на макро- и микроуровнях. Информационные модели отражают схемы взаимосвязей задач по входным и выходным данным и служат эскизами технологии и процедур проектирования РБД.

Для практической реализации технологии и процедур проектирования РБД необходима разработка соответствующего математического и программного обеспечения.

Целью работы является исследование эффективности методов оптимизации структур централизованных распределенных баз данных на ранних этапах их проектирования.

Математическая модель задачи проектирования структур РБД

Основные задачи проектирования структур РБД решаются на микроуровне [11]:

$Task^2 = \{ Task_i^2 \},\ i = \overline{1,5}$,

где $Task_1^2$ – определение количества и состава информационных ресурсов; $Task_2^2$ – выбор типа и архитектуры базы данных; $Task_3^2$ – размещение информационных ресурсов (ИР) в базе данных; $Task_4^2$ – определение характеристик каналов связи; $Task_5^2$ – определение объемов запоминающих устройств для локальных баз данных (ЛБД).

При этом входными данными для комплекса задач микроуровня являются выходные данные задачи макроуровня.

В рамках предложенной декомпозиции проблемы общая задача многокритериального синтеза физических структур РБД рассматривается в следующей постановке [14].

Заданы: множество потенциальных пользователей базы данных $I = \{ i \}$, $i = \overline{1,n}$, связанных однородной компьютерной сетью $G = (I, R)$ (где $R = [r_{ik}]$, $i,k = \overline{1,n}$ – матрица смежности, определяющая множество каналов связи между узлами сети); множество ИР $J = \{ j \}$, $j = \overline{1,m}$; $X = \{ x \}$ – множество допустимых реализаций физических структур.

Необходимо определить наилучший вариант физической структуры РБД $x^o \in X$ (количество ЛБД, распределение ИР по ЛБД $x = [x_{ij}]$, размещение ЛБД по узлам сети, объемы запоминающих устройств для хранения ЛБД $b = [b_i]$, $i = \overline{1,n}$ и пропускные способности каналов связи между узлами сети $h = [h_{ik}]$, $i,k = \overline{1,n}$) по показателям затрат, времени доступа к данным и объему передаваемых данных.

Затраты на реализацию физической структуры РБД определяются соотношением :

$$c(x) = \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}(x) x_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} c_t (\alpha_{ij} + \beta_{ij})\, x_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} c_t\, l'_j\, z_{ij}, \quad (1)$$

где $c_{ij}(x)$ – затраты на хранение j-го ИР в i-м узле сети; c_t – затраты на передачу единицы информации; α_{ij} – суммарный объем запросов к j-му ИР из i-го узла; β_{ij} – суммарный объем ответов на запросы к j-му ИР из i-го узла сети; $x = [x_{ij}]$ – матрица размещения ИР по узлам сети (x_{ij} – булева переменная: $x_{ij} = 1$, если j-й ИР хранится в i-м в узле сети; $x_{ij} = 0$ – в противном случае); l'_j – суммарный объем информации, передаваемой при обновлении j-го ИР; $z = [z_{ij}]$ – матрица обновлений ИР (z_{ij} – булева переменная: $z_{ij} = 1$, если j-й ИР обновляется из i-го узла сети; $z_{ij} = 0$, в противном случае).

Время доступа к информационным ресурсам РБД определяется соотношением:

$$t(x) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} [\, t_{ij}^{tr}(x) + t_{ij}^{pr}(x) + t_{ij}^{qp}(x) + t_{ij}^{rp}(x) \,]\, x_{ij}}{n \cdot m}, \quad (2)$$

где $t_{ij}^{tr}(x)$ – время передачи запроса из i-го узла сети к j-му ИР; $t_{ij}^{pr}(x)$ – время ожидания запроса из i-го узла сети по j-му ИР; $t_{ij}^{qp}(x)$ – время обработки запроса из i-го узла сети по j-му ИР; $t_{ij}^{rp}(x)$ – время передачи ответа на запрос из i-го узла сети к j-му ИР; $x = [x_{ij}]$ – матрица размещения ИР по узлам сети (x_{ij} – булева переменная ($x_{ij} = 1$, если j-й ИР хранится в i-м в узле сети; $x_{ij} = 0$, в противном случае); n – количество узлов компьютерной сети; m – количество ИР.

Объем передаваемой информации определяется соотношением:

$$v(x) = \sum_{i=1}^{n}\sum_{j=1}^{m} (\alpha_{ij} + \beta_{ij})\, d_{ij}(x)\, x_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} l_j\, d_{ij}(x)\, z_{ij}, \quad (3)$$

где α_{ij} – суммарный объем запросов к j-му ИР из i-го узла; β_{ij} – суммарный объем ответов на запросы к j-му ИР из i-го узла сети; $d_{ij}(x)$ – расстояние от места хранения j-го ИР до i-го узла сети; $x = [x_{ij}]$ – матрица размещения ИР по узлам сети (x_{ij} – булева переменная: $x_{ij} = 1$, если j-й ИР хранится в i-м в узле сети; $x_{ij} = 0$, в противном случае); l – объем ИР ($l = [l_j]$, $j = \overline{1,m}$); $z = [z_{ij}]$ – матрица обновлений ИР (z_{ij} – булева переменная: $z_{ij} = 1$, если j-й ИР обновляется из i-го узла сети; $z_{ij} = 0$, в противном случае).

Компромиссное решение многокритериальной задачи $x^o \in X$ формально определяется системой трех частных критериев

$$c(x) \to \min_{x \in X},\ t(x) \to \min_{x \in X},\ v(x) \to \min_{x \in X}. \quad (4)$$

Используемые показатели качества физических структур (1) – (3) являются разнородными, имеют различную размерность и интервал измерения. Для формирования обобщенного критерия предлагается использовать функции полезности частых критериев вида

$$\xi_i(x) = \left(\frac{k_i(x) - k_i^-}{k_i^+ - k_i^-} \right)^{\beta_i} \quad (5)$$

где $k_i(x)$, k_i^+, k_i^- – соответственно текущее, наихудшее и наилучшее значения i-го частного критерия; β_i – параметр, определяющий вид зависимости (5).

Выбор наилучшего компромиссного решения $x^o \in X$ предлагается производить в рамках кардиналистического подхода с использованием аддитивно-мультипликативной функции общей полезности:

$$P(x) = \delta \cdot \sum_{i=1}^{3} \eta_i \, \xi_i(x) + (1-\delta) \prod_{i=1}^{3} [\xi_i(x)]^{\eta_i}, \quad (6)$$

где $\xi_i(x)$ – функция полезности частного критерия $k_i(x)$; δ – параметр модели, определяющий конкретный вид схемы компромиссов, $0 \le \delta \le 1$: при $\delta = 1$, функция (6) принимает форму аддитивной, а при $\delta = 0$ – форму мультипликативной функции общей полезности.

Для выбора наилучшего компромиссного варианта физической структуры РБД требуется решить задачу оптимизации вида:

$$x^o = \arg\max_{x \in X} P(x). \quad (7)$$

Решение задачи осуществляется в системах автоматизации проектирования с использованием достоинств ординалистического и кардиналистического подходов [13].

Методы решения задачи

Проектируемые РБД существенно различаются по количествам потенциальных пользователей, узлов компьютерных сетей, в которых они функционируют, ИР, локальных баз. Это требует использования в системах автоматизации проектирования множества методов синтеза физических структур, различающихся по точности и временной сложности.

В процессе оптимизации методом перебора сочетаний при заданных *I, G, J*, необходимо определить размещение ИР в РБД x^o, а также значения пропускных способностей каналов связи $h = [h_{ij}]$, $i, j = \overline{1,n}$ и объем запоминающих устройств узлов $b = [b_i]$, $i = \overline{1,n}$.

С учетом этого, при решении задачи формируется множество возможных размещений ИР по ЛБД $X = \{x\}$, которая представляет собой множество сочетаний C_n^m (где n – количество узлов в компьютерной сети, m – количество ИР в РБД).

Метод основан на упорядочении области допустимых решений $X = \{x\}$ таким образом, что оптимальное решение задачи находится путем направленного перебора по количеству мест размещения локальных баз и наборам ИР в них [12 – 13]. Поиск наилучшего варианта размещения ИР в РБД идет путем последовательного увеличения количества мест размещения ИР (локальных баз) до достижения экстремума функции цели.

При решении задачи методом Балаша необходимо определить нижнюю границу целевой функции *min P(x)* (недопустимое значение критерия (6)), которая соответствует наихудшему варианту. Значение критерия (1) вычисляется для случая размещения всех ИР в каждом узле компьютерной сети, что соответствует *max c(x)*. Наихудшие значения критериев (2) – (3) определяются путем задания максимально допустимых уровней времени доступа, что соответствует *max t(x)* и объема передаваемых данных, что соответствует *max v(x)*.

С учетом распределения ИР в РБД, первый уровень дерева решений определяется путем размещения одного ИР в одном узле компьютерной сети. Для решаемой задачи количество ветвей дерева равно количеству уровней ветвлений, т.е. $n = m$. На каждом последующем уровне ветвления происходит доразмещение ИР до тех пор, пока не будут распределены все ИР.

Выбор направления ветвления в дереве решений осуществляется на основании максимального значения выражения (6). В результате работы алгоритма получаются n максимальных значений обобщенного критерия *P(x)*, среди которых необходимо выбрать максимальное, соответствующее оптимальной физической структуре РБД.

Эволюционный метод решения задачи, реализуемый с помощью генетического алгоритма, представляет комбинацию переборного и локально-градиентного методов: механизмы кроссовера и мутации реализуют переборную часть метода, а отбор лучших хромосом – градиентный подъем [13].

Формирование начального набора хромосом осуществляется путем генерации случайной последовательности нулей и единиц. Количество генов в хромосоме соответствует количеству узлов компьютерной сети *n*, а аллели генов принимают значения *0* или *1*, в зависимости от наличия или отсутствия ИР в данном узле сети. Локус определяет номер узла в компьютерной сети. Проверка веса каждой хромосомы *u* осуществляется путем анализа сгенерированного битового массива, с учетом задаваемого количества ИР ($u = m$). При невыполнении условия $u = m$, сгенерированная с использованием генератора случайных чисел хромосома отбрасывается и генерация продолжается до выполнения этого условия.

Технологию решения задачи синтеза физической структуры РБД с использованием этого метода можно представить последовательностью из трех этапов: генерация размещений ИР по ЛБД; формирование вариантов размещения ИР по ЛБД; оценка эффективности сформированных вариантов и их ранжирование. Результатом решения является физическая структура РБД с наибольшим значением обобщенного критерия (6).

Суть предлагаемой модификации метода направленного перебора вариантов состоит в использовании схемы покоординатной оптимизации для выбора наилучшего размещения заданного количества ИР *m* на заданном количестве узлов компьютерной сети *n*.

Покоординатная оптимизация состоит в улучшении некоторого начального размещения ИР в узлах компьютерной сети путем последовательного перемещения одного из ресурсов при фиксированном размещении *m–1* остальных. При наличии достаточных

вычислительных ресурсов точность этого метода может быть повышена путем использования процедуры мультистарта (многопроходности).

Практическая реализация и эксперименты

Для комбинаторных методов определены их временная сложность (время решения задачи как функция от размерности задачи: количества узлов компьютерной сети n и количества размещаемых информационных ресурсов m). В виду быстрого увеличения количества вариантов размещения m информационных ресурсов на n узлах компьютерной сети (сочетаний C_n^m) с увеличением m и n время решения задачи также резко возрастает. Этот метод имеет неполиномиальную временную сложность, его целесообразно применять для $n \le 20$ и $m \le 20$.

Метод Балаша позволяет за приемлемое время получать решения задач большей размерности (до $n = 30$ и $m = 30$). По сравнению с методом перебора сочетаний для $m = 30$ и $n = 30$ уменьшение времени счета для метода Балаша составляет порядка 70 %.

При значительно меньшем времени решения задачи методом покоординатной оптимизации средняя относительная погрешность относительно глобального оптимума составила $\bar{\sigma} = 6\%$, а максимальная относительная погрешность $\sigma_{max} = 9{,}8\%$.

Для эволюционного метода были установлены относительные погрешности решения, которые составляют 22,1% и 40,5% соответственно.

При анализе временной сложности получаемых решений аддитивным алгоритмом при $n = m$ время счета алгоритма меньше времени счета методом перебора сочетаний примерно в 3 раза, что обеспечивает эффективность его применения для фиксированных $n = m$. При размерности задач $n \times m \ge 150$ и однокритериальной их постановке время решения методом Балаша сравнимо с временем решения методом перебора сочетаний.

Применение метода эволюционного поиска оправдано при необходимости получения приближенного решения задачи синтеза физических структур РБД с $n \times m > 100$ за короткое время, а также использовании совокупности разнородных и противоречивых критериев.

Сравнительная оценка эволюционного метода и метода покоординатной оптимизации проводилась для заданных значений времени поиска решения. При этом определялись средняя и максимальная относительные погрешности получаемых решений.

На основании этих оценок можно сделать вывод, что для размерности задачи $n \le 20$, $m \le 20$, метод покоординатной оптимизации является более эффективным по комплексному показателю «точность-сложность». Применение метода генетических алгоритмов эффективно при размерности задачи синтеза при $n > 20$, $m > 20$.

Выводы

На основе анализа постановки проблемы проектирования физических структур РБД выбрана математическая модель задачи оптимизации физических структур по показателям затрат, оперативности и объема передаваемой информации. Для ее решения выбраны методы Балаша, направленного перебора и эволюционного синтеза на основе генетического алгоритма. Приведены сравнительные оценки точности и временной сложности методов.

Использование в системе проектирования множества методов позволит выбирать наиболее рациональный из них исходя из размерности решаемой задачи, требуемой точности решения и имеющихся временных и вычислительных ресурсов.

СПИСОК ЛИТЕРАТУРЫ:

1. Арсеньев В.П. Интегрированные распределенные базы данных / В.П. Арсеньев. – СПб.: Изд.-полигр. центр СПбГЭТУ (ЛЭТИ), 2004. – 498 с.

2. Коннолли Т. Базы данных: проектирование, реализация, сопровождение. Теория и практика / Т. Коннолли, К. Бегг; пер. с англ. В.А. Иванов. – М.: Вильямс, 2003. – 720 с.

3. Теоретические основы проектирования оптимальных структур распределенных баз данных / В.В. Кульба, С.С. Ковалевский, С.А. Косяченко, В.О. Сиротюк. – М.: Синтег, 1999. – 660 с.

4. Зарецкий, К.А. Модель оптимального размещения информационных ресурсов в неоднородной компьютерной сети [Текст] / К.А. Зарецкий, В.И. Мейкшан // Вестник СибГУТИ. – 2006. – № 1. – С. 18-21.

5. Танянский С.С. Технология построения крупномасштабной базы данных / С.С. Танянский // Бионика интеллекта. – 2006. – № 2(65). – С. 53 – 56.

6. Теоретические основы проектирования оптимальных структур распределенных баз данных / [Кульба В.В., Ковалевский С.С., Косяченко С.А., Сиротюк В.О.]. – М.: Синтег, 1999. – 660 с.

7. Begg C. Distributed databases / C. Begg. – Los Angeles: California Institute of Technology Edition, 2003. – 489 p.

8. Силин А.В. Методы и модели проектирования структур территориально-распределенных баз данных / А.В. Силин, В.И. Воробьев, Г.И. Ревунков // Деп. рук. ВИНИТИ № 3282-00В. – 2005. – С. 21 – 25.

9. Лаздынь С.В. Оптимизация распределенных корпоративных информационных сетей с использованием генетических алгоритмов и объектного моделирования / С.В. Лаздынь, С.Ю. Землянская // Наукові праці ДонНТУ. – 2009. – № 147. – С. 83 – 95.

10. Лаздынь С.В. Оптимизация распределенных баз данных использованием генетических алгоритмов / С.В. Лаздынь, А.О. Телятников // Вестник Херсонского государственного технического университета. – 2004. – № 1(19). – С. 236 – 239.

11. Бескоровайный В.В. Системологический анализ проблемы автоматизированного проектирования распределенных баз данных / В.В. Бескоровайный, В.В. Евсеев, О.С. Ульянова // Вестник Херсонского национального технического университета. – 2010. – № 38. – С. 120 – 125.

12. Бескоровайный, В.В Методы синтеза физических структур распределенных баз данных / В.В. Бескоровайный, О.С. Ульянова // Открытые информационные и компьютерные интегрированные технологии. – 2010. – № 47. – С. 136 – 146.

13. Петров Э.Г. Территориально распределенные системы обслуживания / Э.Г. Петров, В.П. Писклакова, В.В. Бескоровайный. – К.: Техника, 1992. – 208 с.

14. Бескоровайный В.В. Математические модели многокритериального синтеза физических структур распределенных баз данных / В.В. Бескоровайный, О.С. Ульянова // Восточно-европейский журнал передовых технологий. – 2010. – № . – С.44 – 48.