ОБЪЕДИНЕННЫЙ ИНСТИТУТ ЯДЕРНЫХ ИССЛЕДОВАНИЙ

11 - 10387

ЯНЕВ
Никола Иванов

МЕТОДЫ ЦЕЛОЧИСЛЕННОГО ПРОГРАММИРОВАНИЯ ДЛЯ ОПТИМИЗАЦИИ СТРУКТУРЫ ДАННЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

Специальность 01.01.10 - Математическое обеспечение вычислительных комплексов и АСУ

Автореферат диссертации на соискание ученой степени кандидата физико-математических наук

Дубна 1977

Работа выполнена в Лаборатории вычислительной техники и автоматизации Объединенного института ядерных исследований, г.Дубна.

Научный руководитель: доктор технических наук, старший научный сотрудник АРНАУДОВ Д.Д.

Официальные оппоненты: доктор физико-математических наук, доцент КАРМАНОВ В.Г.

кандидат физико-математических наук, старший научный сотрудник КОНДУРОВ И.А.

Ведущее предприятие: Вычислительный центр АН СССР, Москва.

Защита диссертации состоится "___"_____1977 г. в _____часов на заседании специализированного совета по защите диссертации при ЛВТА ОИЯИ - Д-56/4.

Автореферат разослан "___"_____1977 г.

С диссертацией можно ознакомиться в библиотеке ОИЯИ.

Ученый секретарь специализированного совета
кандидат физико-математических наук Пузынина Т.П.

Пузынина

Практика информационного обслуживания все чаще выдвигает задачи разработки больших автоматизированных информационно-поисковых систем (ИПС). Резкое возрастание информационных фондов этих систем требует развития таких структур данных, которые способствуют достижению определенного быстродействия системы при разумном расходовании памяти ЭВМ. Отыскание лучших в определенном отношении структур данных может быть сведено к решению задачи оптимизации, что и является предметом настоящей диссертации. Дискретный характер исследуемых задач определил и выбор средств для их решения - целочисленное программирование (ЦП). Несмотря на относительно большой объем публикаций по ЦП, в литературе почти отсутствуют данные относительно его применения для решения подобных задач. Это объясняется тем, что исключительно трудный характер целочисленных задач требует расходования многих сил и средств не только для их решения, но даже и для проверки самых скромных идей. Поэтому автор считает для себя особой удачей наличие таких благоприятствующих обстоятельств для выполнения работы, как:
1) реальность проблематики - создание современной ИПС ОИЯИ, в процессе которого не только возникают оптимизационные задачи, но и ощущается эффект от их решений; 2) современная электронно-вычислительная техника и программное обеспечение; 3) широчайшие

возможности, которые предоставляются для их использования. Все проблемы, рассмотренные в диссертации, связаны между собой, как в области применения - автоматизированный информационный поиск, так и в использовании математических методов - целочисленное программирование. Вне этой связи каждая проблема является более общей - в области применения большинство задач относится к машинной обработке данных, а методы и алгоритмы, которые даются для решения отдельных задач, могут быть использованы для класса задач, описываемых заданными математическими моделями. Поэтому для изложения материала выбран следующий принцип: содержательная сторона всех задач, решаемых в диссертации, описывается в отдельной главе (первая), на основе рассмотрения определенной области информационно-логических процессов - дескрипторных ИПС. Математическая трактовка возникших задач дается в отдельных главах диссертации, как-то:

- общая задача частично-целочисленного линейного программирования (ЧЦЛП);

- специальная задача ЧЦЛП с большой размерностью;

- нелинейная задача ЦП.

Для каждой задачи в соответствующих главах приводится алгоритм ее решения, исследуется эффективность алгоритма, обсуждаются вопросы машинной реализации и прилагаются результаты машинного эксперимента. Программная реализация алгоритмов осуществлена на языках ФОРТРАН и КОБОЛ с максимальным стремлением к машинной независимости. Все рассмотренные оптимизационные задачи задаются как задачи о нахождении минимума целевой функции $f(x)$ векторного аргумента x, для $x \in M$, где M является подмножеством n-мерного евклидова пространства. Термин "специальная"

задача ЧЦЛП употребляется для обозначения того, что в алгоритме ее решения существенно учитываются конкретные особенности множества M и логическая связь между компонентами вектора x. Необходимость рассмотрения специальных задач ЦП возникает из-за того, что на основе учета специфики решаемой задачи зачастую оказывается возможным создать алгоритмы, дающие возможность получить решение задачи намного быстрее, чем с использованием общих методов. Типичным примером в этом отношении является задача об уплотнении узловых списков многоуровневой ИПС, которая рассматривается в третьей главе. Поскольку это задача ЧЦЛП, то она может быть решена при помощи алгоритма решения общей задачи ЧЦЛП, рассмотренной во второй главе. Однако специфика задачи (бинарная матрица ограничения, сильно выраженная логическая связь между переменными, большая размерность) потребовала построения специального алгоритма, который оказался (как и следовало ожидать) многократно превосходящим по быстродействию алгоритм для решения общей задачи.

Полученные в диссертации результаты практически апробированы и внедрены в ИПС ОИЯИ, в разработке которой автор принимал непосредственное участие. Так как практическое применение связано с разработкой алгоритмов и программ, учитывающих конкретные особенности данной системы, то эти вопросы в диссертации не рассматриваются, за исключением обсуждения эффекта от внедрения.

Ниже дается краткое изложение содержания отдельных глав диссертации.

Диссертация состоит из введения, четырех глав и заключения. Во введении дается краткое представление о содержании отдельных глав.

В первой главе описывается содержательная сторона решаемых задач оптимизации структуры данных. Для обеспечения четкости изложения строится теоретико-множественная модель дескрипторной ИПС, на основе которой значительно упрощается анализ структуры данных и обосновывается идея иерархической структуры организации информации.

Сформулированы как задачи ЦП проблемы: уплотнения мультисписковой структуры многоуровневой ИПС, машинной реализации дескрипторных словарей ИПС с учетом использования дисковой памяти, оптимального заполнения полей информационного массива, освободившихся вследствие стирания информации. Построенные модели относятся соответственно к следующим классам задач ЦП: общая задача ЧЦЛП, специальная задача ЧЦЛП с большой размерностью и нелинейная задача ЦП. Решение именно этих задач рассматривается в следующих главах диссертации.

Во второй главе рассматривается алгоритм для решения общей задачи частично-целочисленного линейного программирования, модель которой задается как

$$\min Z(x,y) = c_1 x + c_2 y \ , \ (x,y) \in M \ , \qquad (1)$$

где

$$M = \{(x,y) \mid A_1 x + A_2 y = b \ , \ x \geq 0 \ \text{целое} \ , \ y \geq 0\} \qquad (2)$$

для заданных $A_1(m \times n_1)$, $A_2(m \times n_2)$, $c_1(1 \times n_1)$, $c_2(1 \times n_2)$ и $b(m \times 1)$.

Рассмотренный алгоритм принадлежит к классу алгоритмов направленного перебора и его конечность гарантируется при условии, что $x \leq \beta$, где β – целочисленный вектор. Выбор базисной переменной, на основе которой делается ветвление, осуществляется с вычислением нижних границ наращивания целевой функции, при по-

следующей двойственной симплексной итерации. Эффективность алгоритма эмпирически показана на основе решения большого числа тестовых задач. Машинные эксперименты с программой, реализующей данный алгоритм, проводились на ЭВМ ЕС-1040, IBM 360/30, CDC-6400 и опубликованы в [1,2]. Программная реализация алгоритма осуществлена в виде пакета прикладных программ для ЭВМ ЕС-1040 и написана на языке ФОРТРАН IУ. Автоматическое вычисление размеров рабочих массивов, в зависимости от входных данных (m , n_1, n_2) и наличного объема памяти, выполняется специальным модулем, написанным на языке АССЕМБЛЕР для ЕС. Здесь показаны результаты машинного эксперимента, позволяющие сравнить эффективности предложенной программы и подобных программ фирм IBM и CDC.

При помощи этого пакета очень быстро решались практические задачи об оптимальном заполнении освободившихся мест в основном информационном массиве ИПС ОИЯИ вследствие стирания устаревшей информации. Математическая модель этой задачи получается из (1),(2) для $n_2 = 0$, A_I – булевая матрица, X – вектор двоичных переменных. Эффективность алгоритма в этом случае объясняется спецификой задачи, для которой известно оптимальное значение целевой функции, при достижении которого прекращается дальнейший поиск оптимума. Таким образом, выигрывается самая существенная часть времени, которая уходит на доказательство оптимальности.

В третьей главе рассматривается задача об уплотнении узловых структур как задача ЧЦЛП [3] с большой размерностью. Формально эту задачу можно определить как:

найти $\min\limits_{M} |\varphi(M)|$

для $M \subset T$, $|M| = m$,

где T – конечное множество объектов t_i (для ИПС это множество документов), D – конечное множество признаков (для ИПС – мно-

жество дескрипторов), при помощи которых описываются объекты из T, т.е. известно однозначное соответствие $\varphi: T \to 2^D$ (2^D - множество всех подмножеств множества D), $\varphi(M) = \bigcup_{t_i \in M} \varphi(t_i)$, m - заданное натуральное число и $|A|$ означает мощность множества A.

Математическая модель задачи после ее приведения к задаче ЧЦЛП задается как

$$\min \; V = \sum_{i=1}^{K} v_i \qquad (3)$$

при ограничениях

$$\sum_{j=1}^{n} a_{ij} x_j = \ell_i (v_i - z_i) \;, \quad i = \overline{1,K}$$

$$\sum_{j=1}^{n} x_j = m$$

$$0 \le z_i \le 1 - \frac{1}{\ell_i} \;, \quad i = \overline{1,K}$$

$$x_j, v_i \in \{0,1\} \;, \quad i = \overline{1,K}; \; j = \overline{1,n} \;,$$

где двоичные числа a_{ij} суть элементы матрицы, которой задается отображение φ, а ℓ_i - числа, удовлетворяющие $\ell_i \ge \sum_j a_{ij}$. Выборка M из T определяется при помощи булевых переменных x_j .

Доказывается, что оптимальное решение непрерывной задачи, которая получается из (3) опусканием требований для целочисленности переменных x_j , v_i , целочисленно для x_j и может быть получено посредством решения задачи о выборе m -ого минимального элемента из множества чисел $\{c_1, c_2, \ldots, c_n\}$, где $c_j = \sum_{i=1}^{K} \frac{a_{ij}}{\ell_i}$.

Приведен алгоритм (на языке АЛГОЛ-60), который решает эту задачу в среднем за $n + \min(m, n-m)$ операций сравнения.

На основе этих результатов предложен комбинаторный алгоритм для решения задачи (3) с большой размерностью, использующий левостороннее ветвление только по переменным v_i , вместе с вычислением эффективной нижней оценки для целевой функции. При этом не решается задача линейного программирования. Для проведения анализа

эффективности алгоритма был поставлен ряд машинных экспериментов на ЭВМ CDC-6400, в которых кроме искусственно генерированных задач успешно решена и реальная задача, где число целочисленных переменных превышало 4000. Конечным результатом от решения задачи (3) является создание алгоритма, применяемого для уплотнения сегментов списков в зонах информационного массива ИПС ОИЯИ, в результате чего значительно уменьшается время поиска и объем требуемой памяти для многоуровневой ИПС.

В четвертой главе исследуются поисковые свойства наращиваемых признаковых деревьев, использующих для организации ветвления только один адрес связи /5/. Наращиваемое дерево строится для одного поискового признака, который имеет количественное выражение и значение которого является идентификатором узлов дерева. Поиск объекта по признаку в информационном массиве, построенном в виде дерева, происходит в процессе сравнения со значениями признака, идентифицирующего узлы дерева, при спуске от корня к терминальному уровню. Поисковые свойства дерева исследуются в двух предложениях: каждый узел несет объектную информацию, размеры пучков могут изменяться от уровня к уровню. При этом наибольшее из значений идентификаторов терминальных узлов каждого пучка всегда меньше идентификатора корневого узла пучка.

При обусловленных предположениях структура исследуемого дерева однозначно определяется вектором $Q = (m, n_0, n_1, \ldots, n_{m-1})$, где m - количество уровней дерева, n_i - размер пучка узлов i -ого уровня. Задача оптимизации структуры наращиваемого дерева сводится к нахождению вектора Q^* такого, чтобы

$$\varphi(Q^*) = \min_{Q \in \Omega} \frac{\Phi(Q)}{N(Q)} = \frac{1 + \sum_{K=1}^{m} \left\{ \prod_{j=0}^{K-1} n_j \left(1 + \sum_{i=0}^{K-1} \frac{1+n_i}{2}\right)\right\}}{1 + \sum_{j=0}^{m-1} \prod_{i=0}^{j} n_i}$$

$$\Omega = \{ Q \mid E \le N(Q) \le \delta E \} \quad , \qquad (4)$$

где E – ожидаемое потребное число объектных узлов дерева, δ – допуск на превышение E, разрешенный наличным резервом памяти, а компоненты вектора Q должны удовлетворять требованиям целочисленности.

Доказывается, что целевая функция задачи (4) задает среднюю длину поискового пути в дереве Q. Для решения этой задачи нелинейного целочисленного программирования разработан алгоритм /4/, укладывающийся в общую схему метода ветвей и границ. Программа реализации алгоритма сделана на языке ФОРТРАН IУ. При помощи этой программы получены оптимальные параметры деревьев с числом узлов $E = 2^{K}$, для $K = \overline{8,17}$. Из полученных результатов /4/ видно, что множество значений размеров пучков в оптимальных структурах равно $\{4,3,2\}$ и что имеет место эффект редукции размеров пучков от корня к терминальному уровню.

Рассмотрены также вопросы, связанные с машинной реализацией узлов дерева. Если ограничения оперативной памяти таковы, что доступ к узлам уровня $K > x$ происходит через операцию чтения из внешнего устройства, то возникает задача: определить оптимальную структуру дерева при условии, что трудоемкость узлов K-ого уровня для $K > x$ увеличена на ρ единиц. Решение этой задачи в принципе может быть получено с использованием алгоритма для решения задачи (4).

Полученные результаты по синтезу оптимальных сбалансированных односвязных признаковых деревьев нашли практическое при-

ложение при машинной реализации дескрипторного словаря ИПС ОИЯИ. Это позволило значительно уменьшить среднее время кодирования поступающих дескрипторов и объем внешней памяти по сравнению с использованием стандартных методов для организации файлов с прямым доступом.

<u>В заключении</u> приведены основные результаты, полученные в диссертации:

I) Сформулированы как задачи целочисленного программирования проблемы: уплотнения мультисписковой структуры иерархической ИПС; организации дескрипторного словаря с учетом использования дисковой памяти; оптимального заполнения свободных полей информационного массива с активной динамикой.

Построенные математические модели, не будучи слишком сложными, обеспечивают адекватное описание рассмотренных задач.

2) Создан, исследован и программно реализован алгоритм типа "ветвей и границ" для решения общей задачи ЧЦЛП со следующими особенностями:

- стратегия ветвления – односторонняя;
- стратегия выбора переменной для фиксации – максимальное наращивание целевой функции;
- метод решения задач ЛП – компактный двойственный симплекс-метод, позволяющий существенно уменьшить размерности оптимальных таблиц.

Программа для решения общей задачи ЧЦЛП оформлена по стандартам для пакетов прикладных программ для ЭВМ единой системы и принята Государственной комиссией НРБ по матобеспечению ЭВМ ЕС.

3) Создан и теоретически обоснован метод получения эффективных нижних оценок оптимального решения задачи об уплотнении мультисписковой структуры, что многократно сокращает количество анализируемых вариантов и ускоряет решение при использовании алгоритмов комбинаторного типа.

На основе полученных результатов создан и программно реализован алгоритм типа "ветвей и границ", обеспечивающий эффективное решение специальной задачи ЧЦЛП с большой размерностью. Практическое применение алгоритма для уплотнения мультисписковой структуры позволило значительно уменьшить время поиска в многоуровневой ИПС.

4) Получено формализованное описание структуры сбалансированного признакового дерева, на основе которого сформулирована и решена задача его оптимального синтеза как задача нелинейного ЦП.

Результаты синтеза оптимального дерева практически использованы при машинной реализации дескрипторного словаря ИПС ОИЯИ и могут быть использованы при организации поисковых массивов, для которых минимальная трудоемкость поиска должна сочетаться с небольшими дополнительными затратами памяти.

Практический эффект от внедрения полученных результатов в ИПС ОИЯИ показан в таблице I.

Результаты работы докладывались: на семинаре отдела математической обработки экспериментальных данных Лаборатории вычислительной техники и автоматизации Объединенного института ядерных исследований в г.Дубне, на семинаре "Дискретные модели в АСУ" ВЦ и ЦЭМИ АН СССР, на семинаре АСУ вычислительных комплексов факультета вычислительной математики и кибернетики МГУ, на семинаре кафедры "Исследования операций" Математического факуль-

Таблица I.

Задача оптимизации	Уплотнение узловых списков ИПС с параметрами: число уровней - 3, число документов - 70000.	Реализация дескрипторного словаря в виде признакового дерева с параметрами: размеры пучков: 4,4,3,3, 2,2,2,I5; число терминов в словаре - I4I98; длина одного термина - 3,5 6О-битных слов.
Редукция дисковой памяти	25% (2I0000 слов)	59% (9I400 слов)[I]
Редукция времени поиска	20%	50%
Относительное увеличение времени формирования ИПС за счет работы оптимизирующей программы	I%	Нет

[I] Замечание: Редукция выражается относительно использования стандартной СДС индексно-последовательной организации файла.

тета Софийского университета, на 3-ей весенней конференции Болгарского математического общества в г. Бургасе (НРБ) в 1974 г.

Основные результаты диссертации опубликованы в следующих работах:

1. Н.И.Янев, М.Д.Иванчев, Й.Г.Митев. "Един алгоритъм за решаване на задачата на смесено-целочислено оптимиране, удобен за машинна реализация", сб."Математика и математическо образование". Доклади на III пролетна конференция БМД Бургас 74 г.Изд.БАН 1975 г.
2. М.Д.Иванчев, Й.Г.Митев, Н.И.Янев. "Реализация метода "ветвей и границ" для решения общей задачи целочисленного программирования", ЖВМ и МФ, № 3, 1976 г.
3. Д.Д.Арнаудов, Н.И.Янев. "Об одном способе применения частично-целочисленного линейного программирования для оптимизации поиска в многоуровневой ИПС". Препринт ОИЯИ Р11-9770, 1976 (направлено в ж."Программирование").
4. М.Д.Иванчев, В.М.Сумароков, Н.И.Янев. "Метод"ветвей и границ" для направленного синтеза экономной структуры файла",препринт ОИЯИ Р11-9756 , Дубна, 1976.
5. М.Д.Иванчев, В.М.Сумароков, Н.И.Янев. "Комбинаторный анализ сбалансированных признаковых деревьев с переменным размером пучка". Препринт ОИЯИ Р11-9777, Дубна, 1976.

Рукопись поступила в издательский отдел
17 января 1977года.