Записки Ленинградского ордена Ленина,
ордена Октябрьской Революции и ордена Трудового Красного Знамени
горного института им.Г.В.Плеханова, т.(У1, вып.3, 1975 г.

ПРИНЦИПЫ РАЦИОНАЛЬНОЙ ОРГАНИЗАЦИИ ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМ

А.Д.ШКОЛЬНИКОВ
А.Н.ЛИВШИЦ

Разработка развитого информационного обеспечения является одной из важнейших проблем, возникающих при создании автоматизированных систем управления (АСУ). Необходимость организации чрезвычайно больших массивов разнообразной информации и оперативного обращения к ним (запись, чтение, информационный поиск, обработка) требует включения в состав АСУ сложной и многофункциональной информационно-поисковой системы (ИПС).

Неоднородность циркулирующей в АСУ информации носит не только семантический, но и языковый характер и обусловлена, в первую очередь, необходимостью использования многих языков программирования, а также естественных языков.

Современное состояние развития математического обеспечения таково, что объем памяти ЭВМ, отводимый под библиотеки стандартных и рабочих программ, трансляторы и операционную систему может достигать несколько десятков миллионов байт (Мбайт).

Для ряда подсистем АСУ и автономных ИПС необходимо формировать и запоминать информационные тексты на основе обширных словарей естественного языка. Так, например, объем одного только словаря для подсистемы материально-технического снабжения Норильского горно-металлургического комбината составляет около 200000 слов. Предварительный анализ множества информационных текстов, построенных на базе этого словаря дает оценку необходимого объема памяти 80-100 Мбайт.

Объемы цифровой информации, обрабатываемой ежегодно на горном предприятии даже средней производительности (например, на Сорском горнообогатительном комбинате) составляет 50-60 млн. десятичных знаков.

Как видно из приведенных примеров, необходимость формирования

столь больших текстовых и числовых информационных массивов предъявляет особые требования не только к техническим средствам АСУ (например, наличие весьма значительных резервов машинной памяти), но также к системному и информационному обеспечению. Последнее вызвано тем, что решение большинства задач АСУ связано с постоянным обращением к информационным массивам, хранящимся на внешних запоминающих устройствах. Это влечет за собой значительное снижение фактической производительности ЭВМ. При этом абсолютное время обращения к внешним запоминающим устройствам составляет весьма значительную долю всего машинного времени и может превышать чистое время вычислений в 3–4 раза. Достаточно эффективным способом борьбы с такими "непроизводительными" затратами машинного времени служит создание развитой операционной системы и комплекса технических средств, обеспечивающих в совокупности реализацию системы разделения времени, многопрограммную работу процессора, мультиплексорный и селекторный каналы и т.д. Такая структурная и функциональная организация ЭВМ, в частности, позволяет использовать более 100 различных внешних устройств и обеспечивает их параллельную работу.

Однако при организации региональных вычислительных центров (ВЦ), например, АСУ отрасли, ИВЦ геологического региона, банк данных (множество информационных текстов) столь велик, что и этот уровень развития, достигнутый современными ЭВМ третьего поколения, оказывается недостаточным. Поэтому становится особенно актуальной разработка специальных методов организации ИПС, обеспечивающих существенное сжатие объема информационных массивов.

Так, сжатие информационных массивов в 2 раза может привести в региональном ВЦ, имеющем 2–3 и более ЭВМ, к экономии одной вычислительной машины, или, по крайней мере, к сокращению необходимого комплекса внешних запоминающих устройств (ВЗУ).

Таким образом, может быть получена прямая экономия капитальных затрат от нескольких десятков до сотен тысяч рублей. В отдельных случаях (с учетом стоимости вычислительных машин класса "ЕС ЭВМ") эта экономия может составлять более 1 млн.руб.

Для подавляющего большинства задач АСУ характерна относительная малость времени вычислений по сравнению с временем выполнения информационных процедур (поиск, сортировка, формирование массивов и т.п.). Этот фактор создает предпосылки для повышения фактической производительности ЭВМ по крайней мере на 10–12%, что высвобождает 400–500 ч машинного времени ежегодно для решения дополнительных задач.

Наконец, такое сжатие информационных массивов, в силу одновременного повышения быстродействия информационного поиска почти на 50%, обеспечивает возможность эффективного решения класса оперативных задач АСУ в реальном масштабе времени и в режиме диалога.

Информационная структура АСУ и архитектоника ЭВМ третьего поколения позволяют выделить 4 основных сегмента машинной памяти: 1) операционная система; 2) рабочее поле; 3) поле, отводимое под математическое обеспечение; 4) поле, отводимое под запись информации по всему множеству задач АСУ.

В качестве обобщающего информационного массива будем рассматривать: тексты на любых естественных, алгоритмических и машинных языках (доступных современным ЭВМ), числовые массивы, смешанные тексты.

На рис. 1 показана структура машинной памяти и в незаштрихованной области - обобщенный информационный массив. Сложность структуры и чрезвычайно большой объем обобщенного информационного массива АСУ предопределяет актуальность разработки эффективных методов организации ИПС. Основными показателями эффективности являются: кодовая длина представления обобщенного информационного массива в памяти ЭВМ и быстродействие автоматического кодирования, поиска и декодирования информации.

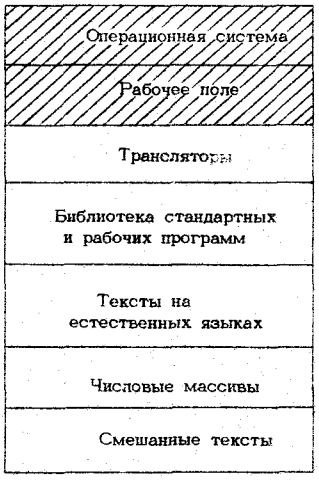


Рис. 1

В соответствии с особенностями задач АСУ выделяются следующие основные требования к функциональной организации ИПС: а) абсолютная однозначность декодирования; б) функционально-управляющая часть ИПС должна быть выполнена на уровне операционной системы; в) доступность и простота общения с ЭВМ потребителей (источников) информации в режиме диалога.

Выше уже отмечалось, что сжатие обобщенного информационного массива уменьшает количество ВЗУ. В то же время, благодаря компактной записи информационных текстов в память, повышается быстродействие поиска, находящееся в прямой зависимости от числа обращений к ВЗУ. Следовательно, значительный прогресс в этом направлении может быть получен в результате решения проблемы оптимального кодирования обобщенных информационных текстов.

Эффективное решение этой проблемы в значительной степени предопределяет выбор и развитие метода информационного поиска, структуру информационных массивов и, следовательно, функциональную организацию ИПС в целом.

Исследованиями установлено [1], что поставленная проблема сводится к решению двух основных задач.

1. Найти оптимальное расширение естественных алфавитов для заданного обобщенного информационного текста.

2. Разработать метод оптимального кодирования обобщенного информационного текста на построенном расширенном алфавите.

Возможность эффективного решения сформулированных задач обусловлена значительной избыточностью как естественных, так и алгоритмических языков, оцениваемой в 50-80%. В связи с этим, перспективным является использование статистических особенностей информационных текстов. Хорошо известно, что для любого текста существует некоторое множество упорядоченных наборов символов (букв) естественного (исходного) алфавита, встречающихся достаточно часто. Очевидно, в зависимости от приро-

ды текста в качестве таких устойчивых буквосочетаний могут оказаться не только пары, тройки и другие сочетания букв, но также целые слова и предложения.

Статистический анализ текста позволяет расширять естественный алфавит за счет наиболее часто встречающихся буквосочетаний. Однако известные алгоритмы такого расширения являются эвристическими.

Тексты на заданном алфавите могут кодироваться с помощью равномерных кодов (постоянной длины), а также кодов переменной длины (например, код Шеннона-Фано, Хаффмена). Использование кодов Хаффмена обеспечивает минимум средней кодовой длины на символ принятого алфавита, т.е. является оптимальным для заданного информационного текста [2]. К несколько худшим результатам приводит использование неравномерного кода Шеннона-Фано.

Основной трудностью применения неравномерных кодов является необходимость знания распределения вероятностей символов расширенного алфавита.

В общем случае использование кодов переменной длины обеспечивает более высокий коэффициент сжатия текста по сравнению с равномерными кодами. Однако использование последних допустимо при наличии свободных кодовых комбинаций в стандартной машинной кодовой системе. Например, широко распространенные коды *ASCII*, *EBCDIC* содержат 8 бит и следовательно 256 различных кодовых комбинаций. Фактически же используется, включая латинский алфавит, знаки операций и служебные знаки, около 110. Оставшиеся кодовые комбинации (около 150) оказываются свободными. Это обстоятельство использовано в работе [4], где на основании статистического анализа французских текстов было получено расширение исходного алфавита примерно на 150 символов. В результате тексты были сжаты в среднем на 10%.

В работе [5] для информационных текстов на алгоритмических языках (Кобол, Алгол, Фортран) естественный алфавит, состоящий из алфавита английского языка и служебных знаков, расширяется основными символами языков программирования (*begin* , *do* , *procedure* и т.д.). Для кодирования этих текстов используются коды Хаффмена, в результате чего достигается сжатие в среднем 20%.

Работы [4,5] являются концентрированным выражением сложившегося подхода к решению сформулированной выше проблемы.

Как видно из приведенных результатов, достигнутое сжатие текстов значительно меньше их избыточности. Это вызвано существенными недостатками имеющихся методов кодирования и расширения естественного алфавита.

Представление слов над расширенным алфавитом не является однозначным. Иначе говоря, каждое слово может быть записано несколькими различными способами (генерациями). От выбора генерации для каждого слова текста зависит распределение вероятностей символов расширенного алфарита, и следовательно их оптимальных кодов Хаффмена. Более того, при кодировании на расширенном алфавите необходимо учитывать кодовую длину не только текста, но и расширенного алфавита.

Таким образом, задача оптимального кодирования информационных текстов T на фиксированном расширенном алфавите B_r может быть сфор-

мулирована следующим образом:

$$L(T^G)_B + L(B)_A \xrightarrow{G} min,$$

где $L(T^G)_B$ – кодовая длина G -ой генерации текста T над B ; $L(B)_A$ – кодовая длина расширенного алфавита B над естественным алфавитом A .

Эффективное решение этой задачи, заключающееся в определении оптимальной генерации текста, рассмотрено ранее [1] . Разработанный итеративный процесс позволил получить на экспериментальных текстах сжатие кодового представления на 41%. При этом расширенный алфавит выбирался на основании эвристического осмысливания статистических испытаний над заданным текстом.

С целью дальнейшего повышения эффективности кодирования необходимо решить задачу оптимального расширения естественного алфавита A :

$$L(T^G)_B + L(B)_A \xrightarrow{B,G} min.$$

Полное решение этой задачи, включающей построение оптимального расширения алфавита и определение оптимальной генерации текста, рассматривалось А.Н.Лившицем [3] .

На основе предложенных методов [1,3] разработана функциональная организация сегмента операционной системы АСУ (рис. 2) обеспечивающая постоянное накапливание информации, изменение и расширение математического обеспечения и библиотеки стандартных и рабочих программ, динамическое формирование оптимального расширенного алфавита и оптимальное кодирование изменяющегося во времени обобщенного информационного массива.

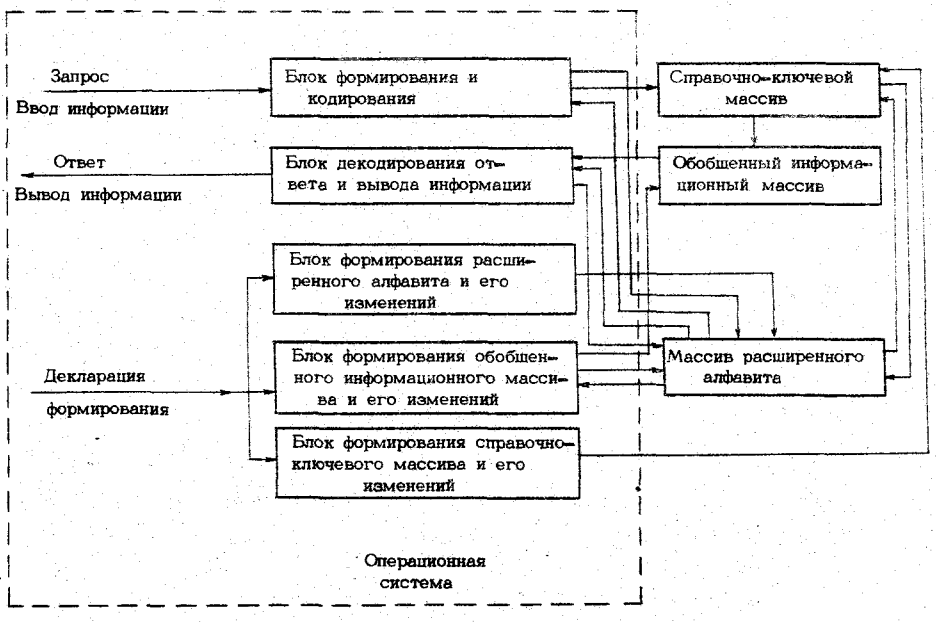


Рис. 2

Такая организация ИПС позволит значительно экономить машинную память, увеличить быстродействие информационного поиска и, как следствие, повысить эффективность АСУ в целом.

Литература

1. Школьников А.Д., Лившиц А.Н. Эффективный метод записи информации в автоматизированных системах управления горных предприятий. "Управление горным производством", вып. 1, 1973.

2. Новик Д.А. Эффективное кодирование. "Энергия", 1965.

3. Лившиц А.Н. Метод эффективного кодирования информационных текстов, основанный на оптимальном расширении алфавита. Зап.ЛГИ, т. 62 , вып. 1 , 1974.

4. Torino-Crozatier R., Chasseux I.I., Frecon L. Dussauchoy A. TA informs. 1971.

5. Wells M. Comput.J., 1972, 15.