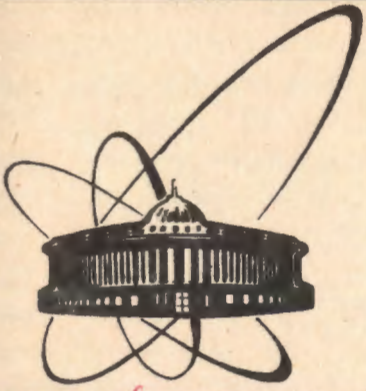

сообщения
объединенного
института
ядерных
исследований
дубна

10-90-562

А.М.Ершов, О.Г.Мельникова, Е.А.Пащенко,
П.П.Сычев

ПРОЦЕССОР СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДЛЯ СУБД С ИНВЕРТИРОВАННЫМИ ФАЙЛАМИ

1990

Современный опыт разработки и эксплуатации банков данных показывает, что наиболее важными моментами являются обеспечение мощного и простого интерфейса для общения с пользователем, создание гибкого программного окружения систем управления базами данных (СУБД)[1]. В последнее время широкое распространение получил ряд СУБД, основанных на инвертировании файлов. Для хранимых в базе данных сведений фактографического типа эти системы позволяют организовать эффективный поиск и выборку информации. Отдельные компоненты СУБД обеспечивают формирование и выдачу отчетных форм по задаваемым пользователем запросам. Для более детального анализа количественных значений хранимой информации, их группировки, обобщения и наглядного представления требуется дополнительное программирование и разработка специальных средств.

Программный процессор статистической обработки PROSTO был разработан как универсальное инструментальное средство для построения сложных статистических отчетов по фактографической информации, содержащейся в базах данных. Пакет ориентирован на работу в среде СУБД типа ADABAS, ДИСОД, КВАНТ и аналогичных им по физической структуре хранения баз данных. Процессор PROSTO не является в чистом виде пакетом прикладных программ для статистического анализа данных[2] и соответственно не служит для выполнения таких специфических функций, как вычисление статистических характеристик, восстановление регрессии и т.п. Результатом работы процессора являются двухмерные многоуровневые статистические таблицы, в наглядной форме содержащие сведения о количестве тех или иных фактографических значений в базе данных. Исходной информацией для получения таблицы служит составляемый пользователем запрос.

В простейшем случае программным процессором PROSTO может быть выдана одномерная таблица в форме строки или столбца с заголовком. При формировании более развернутых статистических отчетов допускается задание в общей сложности до 6 уровней размещения граф и подграф. При этом предоставляется воз-

можность произвольного распределения подуровней по левому и верхнему заголовкам (шапкам) статистической таблицы. При выдаче на печать будет соответствующим образом динамически сгруппирован собранный статистический материал.

В клетках получаемой таблицы размещаются числовые значения, отражающие количества фактографических сведений в базе данных для соответствующих пересечений граф и подграф. Каждая графа и подграфа имеет двухстрочный заголовок, содержащий идентификатор фактографического данного, например, имя поля базы данных, и его конкретное значение для этого столбца или строки. Пользователю предоставляются некоторые средства для формирования заголовков таблицы. Если среди хранимых в базе данных фактографических сведений есть такие, которые представляют собой значения некоторых справочников, то имеется возможность использования для них отдельного файла кодов. В этом случае при выдаче статистической таблицы процессором будет автоматически произведена раскодировка указанных значений.

Язык запросов процессора PROSTO

Каждый пользовательский запрос на получение статистической таблицы заключает в себе всю информацию, необходимую как для осуществления информационного поиска, так и для последующей обработки. Для программного процессора PROSTO разработан входной язык запросов форматного типа[3]. На этапе разбора запроса для него формируется внутреннее представление, и в дальнейшем все функциональные модули работают только с этим представлением.

Условно запрос может быть разбит на три компоненты. Первая из них содержит поисковые критерии для выделения подмножества записей, интересующих пользователя. При этом записи должны принадлежать одному файлу базы данных, они могут включать в себя периодические группы и многозначные поля. Отобранное подмножество записей будет участвовать в дальнейшей статистической обработке.

Для выделения подмножества интересующих записей используется фраза SELECT, в которой в нормальной дизъюнктивной

форме задаются условия отбора информации. При этом обеспечивается формирование достаточно гибких условий отбора записей из базы данных. Важным моментом является тот факт, что фигурирующие в запросе переменные могут быть любыми полями записи файла, в том числе и не дескрипторами.

Вторая компонента запроса определяет набор переменных, по значениям которых будет собираться статистика. В качестве этих переменных выступают отдельные поля обрабатываемого файла базы данных. При указании имен переменных (полей) производится их дискретизация в виде задания конечного числа принимаемых ими значений, интересующих пользователя. Эта компонента запроса представляется фразой CORRELATION. Она определяет переменные, по заданным значениям которых будет производиться набор статистического материала. В общем случае в качестве значения переменной может выступать:

- конкретное значение;
- сумма значений;
- интервал значений;
- сумма интервалов значений;
- линейная шкала значений.

Следует подчеркнуть, что фраза CORRELATION еще не определяет структуру выдаваемой на печать таблицы, и имеется определенная свобода в оформлении собранного статистического материала. Конкретное выбранное размещение переменных по верхней и левой шапкам таблицы задается в фразе PRINT, реализующей третью компоненту запроса. В ней описывается структура результирующей двухмерной статистической таблицы, задаются заголовки всей таблицы и отдельных граф, указываются переменные, помещаемые соответственно в строках (левая шапка) и столбцах (верхняя шапка) таблицы, а также переменные, значения которых в заголовках граф требуется раскодировать из файла кодов. Для ссылки на переменные используются их порядковые номера в фразе CORRELATION, здесь же задается размещение переменных по подуровням таблицы. Реализовано наглядное представление накопленного статистического материала в виде графиков и гистограмм, выдаваемых на печатающее устройство.

Функциональная реализация пакета PROSTO

Головной модуль PSOMAIN выполняет инициализацию процессора статистической обработки и осуществляет управление работой остальных модулей пакета. Первым рабочим модулем, которому передается управление, является модуль синтаксического и лексического анализа - PSOQUERY. Им производится разбор получаемых во входном потоке запросов и генерируется ряд управляющих таблиц, на основе которых в дальнейшем осуществляется взаимодействие всех модулей пакета.

Результатом работы модуля PSOFIND является выделение из обрабатываемого файла базы данных подмножества записей, которые будут рассматриваться процессором при формировании статистических таблиц. Другими словами, отбирается (в виде набора внутрисистемных номеров) совокупность записей, релевантных поисковым условиям фразы SELECT входного запроса. При этом в каждом поисковом критерии отыскиваются поля-дескрипторы (если они есть) и на основе заданной их комбинации выполняется поиск в базе данных. Затем совокупность релевантных записей просматривается на предмет удовлетворения недескрипторным элементам поисковых критериев. Полученные таким образом группы внутрисистемных номеров записей объединяются с отбрасыванием повторяющихся номеров.

Статистическая обработка отобранного подмножества записей производится модулем PSOCORR. В качестве исходной информации используются управляющие таблицы, сгенерированные при разборе фразы CORRELATION входного запроса. Для каждой из отобранных записей на основе операций быстрого сравнения строятся внутренние логические таблицы, идентифицирующие удовлетворение корреляционным критериям первого уровня (по первой переменной в фразе CORRELATION), второго уровня и т.д. При несоответствии записи одному из критериев любого уровня она отбрасывается и происходит переход к обработке следующей записи. Чтобы обеспечить универсальность и независимость внутреннего представления накапливаемых статистических данных от дальнейшего их расположения в печатаемых таблицах, реали-

зована специальная списковая структура хранения. Каждый элемент этой списковой структуры содержит два указателя: на следующий элемент данного уровня корреляционных критериев и на следующий элемент более нижнего уровня. На последнем нижнем уровне вместо указателя хранится накапливаемая статистическая сумма.

Печать итоговых статистических таблиц выполняется модулем PSOPRINT. На основе управляющей информации, получаемой при разборе фраз CORRELATION и PRINT из входного запроса, производится оформление верхней и левой шапок статистического отчета, разметка всей таблицы и ее перенос в случае нехватки места на печатной странице. Правая и нижняя строки отчета содержат итоговые суммы по каждой графе и в целом по всей статистической таблице. Выдача графического представления статистического материала на печатающем устройстве осуществляется модулем PSOGRAF.

Заключительные замечания

Наряду с использованием в пакетном режиме обработки программный процессор PROSTO может служить удобным средством для оперативного получения различного рода статистических справок по содержащейся в базе данных фактографической информации. С этой целью реализован интерактивный режим работы. Приемлемое время ответа обеспечивается при обработке в запросе в среднем до 1 тыс. записей. Это подмножество базы данных может быть специфицировано критериями отбора информации в фразе SELECT входного запроса.

Инструментальное приложение интерактивного режима зависит от используемого на вычислительной установке монитора телеобработки или диалоговой системы. Для того чтобы обеспечить максимальную независимость от конкретной диалоговой среды, все операции взаимодействия с терминалом выделены в отдельный программный модуль управления сообщениями. Основными функциями модуля являются чтение и выдача информации на экране терминала.

Интерактивный режим работы с процессором PROSTO реализован в среде диалоговой системы ТЕРМ[4]. Модуль управления сообщениями обеспечивает построчный вариант обмена с терминалом, что дает возможность обращения к процессору с терминалов, подключаемых к диалоговой системе при помощи средств локальной сети ОИЯИ JINET[5]. Так как процессор PROSTO является обычной прикладной программой по отношению к СУБД, он может взаимодействовать с программой управления базой данных при организации многопользовательского режима обработки.

В заключение приведем пример простой 3- уровневой статистической таблицы из автоматизированной кадровой подсистемы.

	ПОЛ Мужской		ПОЛ Женский		ИТОГО
	ОКЛАД 200	ОКЛАД 250	ОКЛАД 200	ОКЛАД 250	
ОБРАЗОВАНИЕ Высшее	130	84	68	46	328
ОБРАЗОВАНИЕ Ср.-специальное	64	42	38	18	162
ИТОГО	194	126	106	64	490

Быстрое и с минимальными затратами получение сложных многоуровневых статистических отчетов с возможностью представления информации в графической форме делает программный процессор PROSTO удобным инструментом, который может быть использован в различных автоматизированных информационных системах, построенных на основе СУБД с инвертированными файлами.

ЛИТЕРАТУРА

1. Бойко В.В., Савинков В.М. Проектирование баз данных информационных систем. - М., Финансы и статистика, 1989, 351 с.
2. Кощеев В.А. Автоматизация статистического анализа данных: пакеты прикладных программ. - М., Наука, 1988, 232 с.
3. Криницкий Н.А., Миронов Г.А., Фролов Г.Д. Автоматизированные информационные системы. - М., Наука, 1982, 384 с.
4. Гончаков В.С., Кореньков В.В., Шириков В.П. Диалоговая система ТЕРМ для ЕС ЭВМ, совместимая по входному языку с диалоговыми подсистемами ЭВМ фирмы CDC и БЭСМ-6. - В кн.: Тезисы докладов Всесоюзной конференции "Диалог "Человек - ЭВМ". - Л., ЛИАП, 1982.
5. Говорун Н.Н., Дорохин А.Т., Заикин Н.С. и др. Локальная вычислительная сеть ОИЯИ: аппаратное и программное обеспечение. - Дубна, ОИЯИ, Д11-86-702, 1986, 6 с.

Рукопись поступила в издательский отдел
17 декабря 1990 года.