УДК 681.3.06

ПРОГРАММНЫЙ КОМПЛЕКС ТЕКОНВ ДЛЯ АВТОМАТИЗАЦИИ ПРЕОБРАЗОВАНИЙ БОЛЬШИХ КОЛЛЕКЦИЙ ТЕКСТОВЫХ ФАЙЛОВ

О. Б. Арушанян[1], **Н. А. Богомолов**[1], **Н. И. Волченскова**[1], **И. В. Зюряева**[1], **А. Д. Ковалев**[1]

В НИВЦ МГУ в рамках работ по созданию и сопровождению Библиотеки программ решения типовых задач численного анализа был разработан программный комплекс ТЕКОНВ, предназначенный для автоматизации преобразований коллекций файлов. Рассматривается постановка задачи, архитектура и основные возможности комплекса. Обсуждается использование комплекса для сжатия исходных текстов JavaScript-сценариев и для автоматизированного перевода документации Библиотеки в версии на языке Фортран в документацию для версии на языке Си.

1. Архитектура и основные возможности комплекса ТЕКОНВ. Программный комплекс ТЕКОНВ предназначен для преобразования больших коллекций текстовых и двоичных файлов в диалоговом и пакетном режимах на основе задания сложных сценариев массовой обработки такого рода файлов.

Комплекс ТЕКОНВ имеет модульную структуру и включает в себя:

— модуль подготовки наборов файлов для групповых преобразований,

— модули групповых преобразований файлов и файловых структур,

— редактор файловых структур,

— редактор файлов,

— модуль сравнения содержимого двух файлов.

Комплекс ТЕКОНВ создан в системе программирования Delphi. Общий объем исходного кода комплекса составляет более 60 тысяч строк. Комплекс снабжен развитой встроенной системой помощи. Подробное описание комплекса и рекомендации по его применению приведены в [1]. Сайт комплекса зарегистрирован в НТЦ “ИНФОРМРЕГИСТР” Федерального агентства по информационным технологиям как электронное научное издание (номер государственной регистрации 0320601828, номер свидетельства 9124 от 22.12.2006).

На базе комплекса ТЕКОНВ, в частности, созданы специальные конверторы, обеспечивающие сжатие JavaScript-сценариев и автоматизированное порождение документации для программ Библиотеки решения типовых задач численного анализа [2], реализованных на языке Си, на основе имеющейся документации для программ на языке Фортран.

1.1. Групповые преобразования файлов и файловых структур. Преобразования файлов и файловых структур, выполняемые комплексом ТЕКОНВ, организованы в виде групп однотипных операций:

— преобразование файловых структур (копирование, перемещение, удаление), а также объединение нескольких файлов в один и разбиение одного файла на несколько файлов;

— преобразование регистра символов как для содержимого тестовых файлов, так и для имен выбранных файлов;

— преобразование кодировки текстовых файлов (Win1251, KOI8-R, DOS), включая возможность использования таблицы перекодировок, заданной пользователем, а также выполнение прямой и обратной транслитераций;

— преобразование с помощью развитых средств контекстного поиска и замены, основанных на использовании заданных выражений.

Комплекс ТЕКОНВ позволяет осуществлять управление выполнением групповых преобразований в следующих пошаговых режимах:

[1] Научно-исследовательский вычислительный центр, Московский государственный университет им. М.В. Ломоносова, Ленинские горы, 119991, Москва; e-mail: arush@srcc.msu.ru, nbogom@srcc.msu.ru, iraz@srcc.msu.ru, kovalev@srcc.msu.ru

— с остановкой по каждому найденному контексту (или строке в случае изменения регистра или кодировки);

— с остановкой по каждому файлу;

— с остановкой по каждой группе файлов набора.

Средства контроля групповых преобразований файлов позволяют вести детальный протокол выполненных действий, включая сборку всех найденных и измененных фрагментов в отдельный файл.

Комплекс ТЕКОНВ содержит развитые средства подготовки сценариев групповых преобразований файлов. Сами сценарии, а так же наборы параметров отдельных операций (списки файлов, наборы контекстов поиска и замены) можно сохранять во внешних файлах. Сохраненные ранее сценарии, списки файлов и наборы контекстов можно использовать как в интерактивном, так и в пакетном режимах. В пакетном режиме имя файла со сценарием задается в командной строке вызова. Средства подготовки сценариев включают в себя возможность автоматической записи в форме сценария действий пользователя по групповому преобразованию файлов, выполняемых в интерактивном режиме.

1.2. Редакторы файлов и файловых структур. Помимо модулей групповых преобразований файлов и файловых структур, комплекс ТЕКОНВ содержит модули редакторов для "ручного" редактирования отдельных файлов и каталогов.

Редактор файловых структур (редактор каталогов файлов) позволяет создавать, удалять, переименовывать, копировать и перемещать выбранные каталоги и файлы, в том числе и в режиме "перетаскивания" (Drag-and-Drop).

Редактор файлов ориентирован на "нестандартную" работу и исследование содержимого текстовых и двоичных файлов. Редактор позволяет одновременно работать с несколькими файлами. Открываемые файлы не загружаются полностью в оперативную память, что позволяет обеспечить "быструю" работу с файлами большого размера, в том числе с файлами, превосходящими по размеру оперативную память компьютера.

Редактор позволяет в динамике изменять визуальное представление содержимого файла:

— текст,

— текст с автоматической разбивкой на строки,

— текст в различной кодировке (Win1251, KOI8-R, DOS),

— шестнадцатеричное представление.

Поиск по содержимому текстовых и двоичных файлов выполняется с использованием средств контекстного поиска, применяемых в режиме групповой обработки файлов. Имеется возможность использовать для поиска ранее подготовленные наборы контекстов.

Комплекс ТЕКОНВ содержит также модуль сравнения двоичных и текстовых файлов в построчном и пофрагментном режимах. Подробный протокол сравнения отображается в окне редактора файлов.

2. Использование комплекса ТЕКОНВ для сжатия текста JavaScript-сценариев. Язык JavaScript является широко используемой WEB-технологией, позволяющей "оживлять" статичные HTML-страницы, а также снимать "лишнюю" нагрузку с сервера за счет выполнения на компьютере клиента таких задач, как проверка вводимых пользователем данных, динамическое изменение содержимого HTML-страниц без обращения к серверу и т.д. Иногда JavaScript-сценарии превращаются в достаточно объемные и изощренные программы, разработку которых невозможно осуществлять без тщательного документирования программного кода как за счет использования развернутых интуитивно понятных имен переменных и функций, так и за счет подробных комментариев. Все это может значительно увеличить размер файлов, содержащих программы на языке JavaScript.

Специальная обработка HTML-страниц, содержащих JavaScript-сценарии, для сжатия исходных текстов программ представляется весьма полезной из-за сокращения объема передаваемой по сети информации. Кроме того, сжатие исходных текстов существенно усложняет "читаемость" программ для желающих воспользоваться чужими результатами с целью создания собственных сценариев.

Сжатие JavaScript-сценариев с помощью комплекса ТЕКОНВ основано на удалении из текста сценария фрагментов, не влияющих на алгоритм его выполнения, а также на автоматическом изменении исходных имен функций и переменных на имена, имеющие минимальный размер.

При удалении "лишних" фрагментов текста сценария выполняются следующие действия:

— удаление комментариев,

— удаление пустых строк,

— удаление "лишних" пробелов в операторах,

— удаление концов строк (удаление разбивки сценария на отдельные строки).

Автоматическая генерация коротких имен программных элементов основана на анализе иерархи-

ческой структуры пространства имен JavaScript-сценария, позволяющем определить минимально допустимое количество Nmin различных имен, необходимых для идентификации программных элементов, и выявить области сценария, допускающие повторное использование имен. Уникальные имена программных элементов формируются из текстовых представлений целых чисел от 0 до Nmin–1, дополненных односимвольным префиксом.

На первом уровне иерархии имен находятся идентификаторы глобальных функций и переменных, объявленных вне тела функций. Обращение к этим объектам возможно из любого места сценария. Доступ к формальным параметрам любой функции, ее локальным переменным и локальным (вложенным) функциям возможен только в теле самой функции, поэтому имена этих программных элементов, уникальные внутри функции, можно повторно использовать для обозначения программных элементов других функций того же уровня. Аналогичным образом осуществляется обработка вложенных функций следующего уровня.

Подобную оптимизацию имен нельзя применять к идентификаторам полей структур данных JavaScript, поскольку в языке JavaScript отсутствует возможность статического (до начала выполнения) описания структур и состава полей конкретных экземпляров данных.

Описанный алгоритм формирования уникальных имен программных элементов применим для сценариев, размещенных в одном HTML-файле. Однако в ряде случаев "связанные" сценарии могут размещаться в нескольких независимо загружаемых файлах. Язык JavaScript не предусматривает формального статического описания информационных связей между глобальными программными элементами разных файлов. В этих условиях приходится либо оказаться от оптимизации имен глобальных переменных и функций, ограничившись оптимизацией имен формальных параметров, локальных переменных и вложенных функций, либо согласованно оптимизировать все файлы проекта.

Сжатие JavaScript-сценариев с помощью комплекса ТЕКОНВ было практически использовано для обработки библиотек "связанных" JavaScript-сценариев, размещенных в нескольких файлах. В результате удалось уменьшить размер файлов со сценариями на 10–50 процентов.

3. Использование комплекса ТЕКОНВ для преобразований документации Библиотеки программ решения типовых задач численного анализа НИВЦ МГУ. Научно-образовательный Интернет-ресурс НИВЦ МГУ по численному анализу [3] содержит методические материалы по использованию Библиотеки численного анализа. Первоначально Библиотека была реализована на языке Фортран. В связи с переводом программ Библиотеки на язык Си возникла необходимость подготовки соответствующих методических материалов для ее Си-версии. Библиотека численного анализа содержит свыше двух тысяч программ, охватывающих следующие основные разделы вычислительной математики:

— простейшие вычислительные операции;
— элементарные статистики, обработка данных;
— статистические критерии;
— алгебра полиномов;
— линейная алгебра;
— специальные функции;
— численное интегрирование;
— обыкновенные дифференциальные уравнения;
— интерполяция, аппроксимация, сглаживание, численное дифференцирование;
— анализ и синтез рядов, быстрые преобразования;
— решение уравнений и систем общего вида;
— математическое программирование;
— интегральные уравнения;
— генерация случайных чисел.

В этих условиях весьма актуальной является задача автоматизации подготовки описаний программ для Си-версии Библиотеки на основе соответствующих методических материалов для версии Библиотеки на языке Фортран. Данная задача была решена средствами инструментального комплекса ТЕКОНВ.

Документация Библиотеки для каждого алгоритмического языка представляет собой иерархическую коллекцию файлов, в которой для каждой программы имеется файл с ее описанием, а также несколько ZIP-архивов текстовых файлов на алгоритмическом языке, содержащих текст описываемой программы, текст программы выдачи диагностических сообщений и текст примера использования программы.

Описание каждой программы Библиотеки представляет собой документ, содержащий следующие разделы:

— имя программы,

— назначение программы,
— математическое описание алгоритма,
— прототип программы на языке программирования,
— описание параметров,
— перечисление версий программ для переменных разной точности,
— перечисление используемых программ,
— замечания по использованию,
— пример использования программы на алгоритмическом языке.

Конвертор, созданный на основе комплекса ТЕКОНВ, позволяет находить в тексте описания программ текстовые фрагменты, подлежащие замене, формировать их новые значения и заменять найденные фрагменты, обеспечивая следующие возможности:

— преобразование имен библиотечных программ на Фортране в нотацию, принятую для Си-версии (перевод в нижний регистр и добавление постфикса "_c"; например AM01R преобразуется в am01r_c);

— преобразование найденных в тексте описаний идентификаторов переменных и параметров программ (кроме математического описания алгоритма) в нижний регистр;

— замена прототипа программы на языке Фортран на прототип Си-версии, найденный в соответствующем файле с текстом программы на языке Си;

— замена текста примера использования программы на языке Фортран на текст, найденный в соответствующем файле с программой примера использования на языке Си.

Описания программ, в которых осуществляется поиск, оформлены в виде HTML-файлов; в этих файлах отсутствует специальная разметка требуемых разделов документа кроме обычных тегов форматирования HTML. При выполнении преобразований учитывалась возможность появления в идентификаторах программных элементов символов русского алфавита, имеющих начертание, сходное с соответствующими символами английского алфавита. Для поиска и извлечения нужных фрагментов текста из файлов программ на языке Си необходимо было предварительно извлекать их из соответствующих ZIP-архивов.

Генерация документации для Си-версии Библиотеки проводилась поэтапно по мере появления Си-версий программ, поэтому в конвертор, созданный на основе комплекса ТЕКОНВ, были включены средства задания фрагментов общей коллекции файлов документации, подлежащих преобразованию.

Более подробное описание средств автоматизированной генерации документации для Си-версий программ и результаты их применения приведены в [1, 3].

СПИСОК ЛИТЕРАТУРЫ

1. http://teconv.srcc.msu.ru
2. *Арушанян О.Б., Волченскова Н.И.* Библиотека программ НИВЦ МГУ для решения типовых задач численного анализа // Вычислительные методы и программирование. 2002. **3**, № 2. 158–163.
3. http://num-anal.srcc.msu.ru/

Поступила в редакцию
09.01.2007