

УДК 004.738.52

С.С. Кондрашов (6 курс, каф. АиВТ), В.М. Ицыксон, к.т.н., доц.

ОРГАНИЗАЦИЯ ИНТЕРФЕЙСОВ МЕЖДУ ПРИЛОЖЕНИЯМИ И INTERNET-РЕСУРСАМИ

В настоящее время язык HTML является базовым для описания данных и их представлений на Web-клиентах. Файл описания какого-либо документа с помощью данного языка содержит два типа информации: контент (сами данные) и информацию, необходимую для его отображения. При этом не происходит четкого разделения типов информации, и этого вполне достаточно для задач графического представления. Широкое распространение Internet привело к широкой поддержке языка HTML как основы для описания web-страниц. Существуют множество различных Web-клиентов на различных платформах. Однако теперь появляются задачи не только по отображению клиентами web-страниц, но и по программному анализу данных, которые хранятся на web-серверах. К сожалению, выделить необходимый контент из HTML документа достаточно трудно. Существует ряд технологий, позволяющих решать данную задачу. Одной из ведущих технологий является XML (eXtensible Markup Language).

Технология XML позволяет описывать иерархические структуры данных. Принцип построения Internet-ресурсов на базе данной технологии представляет собой четкое разделение между контентом и графическим представлением. Для описания документа создается не один файл, как в HTML, а два, первый из которых содержит XML-описание структуры данных документа, а второй - XSLT-описание по отображению данных на клиенте. На базе XML стали появляться различные технологии, такие как WSDL (Web Services Description language) - для описания расположения web-сервиса и операций, предоставляемых им и SOAP (Simple Object Access Protocol) - для задания ряда правил преобразования программных типов в XML. Использование данных технологий позволяет производить продуктивный анализ контента web-ресурсов в различных клиентских программах. К сожалению, большинство организаций не торопятся представлять Internet-ресурсы в виде XML. Это можно объяснить следующими причинами: web-браузеры поддерживающие XML еще не получили достаточно широкого распространения на большинстве платформ и необходимы дополнительные затраты на переработку ресурса, а они могут оказаться достаточно большими.

В качестве возможности работы с Internet-ресурсами на базе HTML предлагается организовать единый программный интерфейс для приложений. При этом задачу можно разделить на несколько этапов. Первый этап - создание языка описания процессов анализа HTML документов. Одной из главных задач данного этапа является рассмотрение различных методов построения языка и выбора оптимального. Не менее важной, является задача формирования правил выделения контента рассматриваемого документа. Второй этап – разработка программного интерфейса формирования выходного XML файла как результата анализа документа. В рамках данного этапа решаются вопросы выбора способа реализации программного интерфейса и методов представления полученной информации, разрабатывается программная библиотека, поддерживающая различные языковые шаблоны. Третий этап - разработка программного интерфейса для работы с полученным XML файлом. Необходимость данного этапа может быть обусловлена тем, что данный интерфейс будет содержать функции, которые будут более удобно позволять работу с конкретным типом файла, генерируемого после анализа.

Реализация рассмотренного программного интерфейса позволит унифицировать работу различных программ с Internet-ресурсами на базе широко распространенного языка HTML, и будет обеспечивать возможность анализа данных до тех пор, пока большинство web-серверов не перейдут на использование XML технологий.