УДК 681.3.016

АВТОМАТИЗАЦИЯ ПЕРЕНОСА WEB-ДАННЫХ С РЕГУЛЯРНОЙ СТРУКТУРОЙ В РЕЛЯЦИОННУЮ БАЗУ ДАННЫХ

Аласвад Салех

Научный руководитель - доц. каф. «Системное программное обеспечение», канд. техн. наук Блажко А.А.

Сегодня сеть Интернет является быстро развивающимся источником информации, который находится в сложноструктурированных или неструктурированных HTML-страницах. Процесс извелечения данных из сложноструктурированных источников данных в основном применяется для выполнения информационных запросов к этим источникам и при информационной интеграции данных [1-3]. Большинство существующих работ предполагает значительный ввод данных со стороны операторов, которые должны настраивать описания шаблонов страниц и описаний соответствий между элементами страниц и структурой БД.

Анализ сложноструктурированных источников данных, хранящихся в HTML, XML документах позволяет упростить следующие процессы, которые не достаточно освещены в литературе:

1) интеграция WEB-сайта с информационной системой управления предприятием;

2) перевод WEB-сайта со статической архитектурой в систему управления содержимым сайта (CMS - Control Managment System).

Необходимость в выполнении указанных процессов возникает при развитии сайта предприятия, когда объемы данных, которые необходимо располагать на сайте, и оперативность обновления этих данных возрастают и не позволяют продолжать ручное редактирование страниц.

При интеграции WEB-сайта с информационной системой управления предприятием необходимо определить данные, хранящиеся на страницах сайта, которые можно в дальнейшем заполнять через программное обеспечение, установленное в информационной системе предприятия, использующего данный сайт.

В результате интеграции сайт преобразуется в стандартную систему с динамическим созданием страниц на основании содержимого базы данных ИС. В процессе интеграции необходимо согласовать структуру БД ИС предприятия и структуру данных страниц сайта.

С увеличением числа страниц сайта и возникновением сложной иерархии их зависимости, возникает необходимость перевода старого WEB-сайта со статической

архитектурой в CMS. В процессе перевода сайта под управление CMS-системой необходимо заполнить БД со структурой CMS-системы в соответствии с сформированной структурой данных страниц сайта, что требует процесса согласования.

Описанные выше требования по автоматизации процессов ведения WEB-сайтов явились основой для создания методики и ее программного обеспечения по переносу содержимого HTML-документов в реляционную базу данных.

Предлагаемая методика включает следующие этапы:

1) преобразование HTML-документа в шаблон;

2) определение повторяющихся структур в шаблоне;

3) выбор повторяющейся структуры;

4) определение атрибутов повторяющейся структуры;

5) формирование и заполнение таблицы в реляционной БД.

На первом этапе выполняется сканирование символом документа с целью поиска ограничителей разметки типа < и >. Шаблон содержит условные обозначения элементов разметки (Т-элементы) и элементов текста (С-элементы). Условное обозначение элементов разметки не учитывает тип разметки (<table>, </table>,
, <U> , </U> „y „t„„.), что не требует дополнительного процесса обучения со стороны оператора, как в существующих методиках.

На втором этапе выполняется поиск повторяющихся подстрок, входящих в шаблон. Для каждой подстроки определяется число ее последовательных повторений и число С-элементов.

На третьем этапе выбирается подстрока шаблона, характеризующаяся максимальным количеством повторений и максимальным числом С-элементов.

На основании предложенной методики разработано программное обеспечение на языке ANSI-C и языке хранимых процедур PL/pgSQL СУБД PostgreSQL.

В настоящий момент разрабатывается программное обеспечение по переносу содержимого WEB-сайта со статичными страницами в CMS-систему, которое использует разработанное программное обеспечение по переносу содержимого HTML-документов в реляционную базу данных.

СПИСОК ЛИТЕРАТУРЫ

1. C. Change and S. Lui. IEPAD: Information extracting based on pattern discovery. In proc. Of 2001 Intl. World Wide Web Conf. , 2001. - pp. 681-688.

2. N. Kushmerick, D. Weld, and R. Doorenbos. Wrapper induction for information extraction. In Proc. Of the 1997 Intl. Joint Conf. On Artificial Intelligence, 1997. - pp. 729-737.

3. L. Lui, C. Pu, and W. Han. XWRAP : An XML -enabled wrapper construction system for web information sources. In Proc. Of the 2000 Intl. Conf. On Data ENGINEERING , 2000. - pp. 611-621.