На правах рукописи

Чеснавский Александр Александрович

ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА ИНТЕГРАЦИИ КОНТЕНТА УНАСЛЕДОВАННЫХ ВЕБ-ПРИЛОЖЕНИЙ В ЕДИНОЕ ИНФОРМАЦИОННОЕ ПРОСТРАНСТВО ПРЕДПРИЯТИЯ

05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата технических наук

Автор: 

Москва – 2009

Работа выполнена в Московском инженерно-физическом институте (государственном университете)

Научный руководитель: кандидат технических наук, доцент

Скворцов Владимир Иванович

Официальные оппоненты: доктор технических наук, профессор

Шелупанов Александр Александрович

кандидат технических наук

Сиротюк Олег Владимирович

Ведущая организация: Всероссийский институт научной и технической информации РАН (ВИНИТИ РАН)

Защита диссертации состоится 29 апреля 2009 г. в 15 часов 00 минут на заседании диссертационного совета Д 212.130.03 в Московском инженерно-физическом институте (государственном университете) по адресу: 115409, г. Москва, Каширское шоссе, 31.

С диссертацией можно ознакомиться в библиотеке института.

Автореферат разослан “_____” марта 2009 г.

Ученый секретарь

диссертационного совета д.т.н., профессор  Шумилов Ю.Ю.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследований

В последние десятилетия одной из основных характеристик бизнеса, вне зависимости от отрасли, географической или культурной принадлежности, стала глобализация. Уже не является удивительным, что для создания определенного продукта (например, автомобиля) комплектующие поставляются сотнями компаний из десятков стран. Интернет и веб-технологии, активно развивающиеся в последнее время, являются одним из катализаторов глобализации. Более того, интернет всего десять лет назад открыл новые возможности для ведения бизнеса и способствовал существенному росту экономики за счет организации коммуникаций между предприятиями, государственными учреждениями, населением. Более того, интернет активно используется конечными пользователями. Так, по данным Internet World Stats, доля проникновения интернета от общей численности населения в 2008 г. в США составила 73,6%, в Европе 48,1%, в России 23,2%. Количество пользователей интернета в России выросло в 10 раз за последние 8 лет и составляет порядка 33 млн.

Объем данных в сети интернет растет высокими темпами, и все чаще необходимая информация доступна в виде веб-страниц.. Это могут быть биржевые котировки, информация о публичных тендерах, курсы валют, новинки и изменения цен на продукцию конкурентов и т.п. Соответственно возникает задача получения данных с внешних веб-сайтов и использования полученных данных в бизнес-процессах. Однако HTML – язык разметки гипертекста – изначально не предназначался для автоматизированной обработки, это лишь средство для представления данных в браузере конечному пользователю. Таким образом, задача интеграции данных унаследованных веб-приложений в единое информационное пространство предприятия является нетривиальной.

В настоящее время задача веб-интеграции, создания унифицированного информационного пространства предприятия на основе веб-технологий, решается с помощью так называемых портальных платформ. Однако, даже самые развитые и функциональные портальные платформы предлагают ограниченный набор инструментов для интеграции унаследованных веб-приложений в единое информационное пространство. Ключевое ограничение связано с тем, что существующие портальные платформы ориентированы, в основном, на статичное отображение отдельных частей веб-страниц в виде портлетов, оставляя задачу обработки и интерпретации данных конечному пользователю. Такой подход, конечно, укладывается в классическую трактовку портала как интеграционного решения, в котором публикуются данные из различных источников, и большая часть их обработки возлагается на самого пользователя, но налагает существенные ограничения на построение единого интеграционного решения. Между тем, на практике

необходимо не только отображать данные из внешних веб-ресурсов, но и использовать их в различных бизнес-процессах. А для решения этой задачи уже недостаточно традиционного отображения HTML-данных унаследованного веб-приложения. Необходим анализ структуры исходной веб-страницы, отделение данных от элементов форматирования, составление иерархии данных на основе структуры тегов и предоставление полученной иерархии в унифицированном виде, удобном для дальнейшей автоматизированной обработки.

Целью диссертационной работы является построение методов, моделей и программных средств интеграции данных произвольных веб-страниц в единое информационное пространство. Использование результатов диссертационного исследования должно сократить временные и ресурсные затраты на реализацию задачи интеграции данных, предоставив разработчикам портальных решений адаптивное инструментальное программное средство для получения и представления в унифицированном формате данных внешних веб-страниц с целью их дальнейшей обработки. Для достижения этой цели в работе необходимо решить следующие задачи:

- исследовать современные модели и методы получения значимых данных с произвольных веб-сайтов, современные подходы к интеграции приложений для выявления основных проблем в области интеграции унаследованных веб-приложений;
- разработать модель унифицированного представления значимых данных веб-страниц;
- разработать алгоритм преобразования произвольной веб-страницы в унифицированное представление;
- разработать алгоритм анализа изменений иерархии значимых данных веб-страниц;
- разработать адаптивное инструментальное программное средство интеграции контента унаследованных веб-приложений;
- экспериментально проверить работоспособность разработанных методов и программных средств.

Методы исследования. При разработке математического аппарата в диссертационной работе используются методы теории графов, теории алгоритмов, методы обработки текстовой информации. При разработке программного обеспечения используются методы объектно-ориентированного, Web-ориентированного и клиент-серверного программирования, в т.ч. с использованием XML, XSLT, XPath-технологий.

Научная новизна работы заключается в следующем:

- разработана модель унифицированного представления иерархии значимых данных веб-сайтов;

- разработан алгоритм получения иерархии значимых данных произвольной веб-страницы и метод идентификации узлов полученной иерархии значимых данных;
- разработан алгоритм анализа изменений иерархии значимых данных на основе дистанции редактирования между двумя иерархиями значимых данных веб-сайтов;
- разработано адаптивное инструментальное программное средство для интеграции контента унаследованных веб-приложений.

Практическая значимость. Разработанные модели и методы извлечения значимых данных и анализа изменений в иерархии значимых данных веб-страниц могут быть использованы в следующих областях:

- интеграция унаследованных веб-приложений;
- создание композитных приложений;
- создание в среде Веб 2.0 новых сервисов на основе существующих веб-ресурсов;
- мониторинг изменений данных на веб-сайтах (например, мониторинг котировок акций, курсов валют, информации о продукции конкурентов, аукционах и т.п.);
- эффективное кэширование веб-страниц.

Реализация результатов. Предложенные в диссертации модели и методы получения и представления иерархии значимых данных веб-сайтов реализованы в виде адаптивного инструментального программного средства для интеграции контента унаследованных веб-приложений в среде портальной платформы. Разработанное инструментальное программное средство было использовано в проектах «Автоматизация процесса поставок» в компании ООО «Хайтиан» (российское представительство HAITIAN INTERNATIONAL Hlds., Ltd) и «Организация процесса продаж» в компании ООО «Умный софт», что подтверждается актами о внедрении.

На защиту выносятся:

- модель представления иерархии значимых данных веб-страницы;
- метод индексации элементов иерархии значимых данных;
- алгоритм получения иерархии значимых данных с произвольной веб-страницы;
- алгоритм анализа изменений иерархии значимых данных веб-сайтов;
- адаптивное инструментальное программное средство интеграции контента унаследованных веб-приложений в среде портальной платформы.

Апробация работы. Теоретические положения и практические

результаты были доложены на следующих конференциях и семинарах:

- Научные сессии МИФИ 2003, 2004, 2006 – 2008 (г. Москва, 2003, 2004, 2006 – 2008 гг.);
- XVII Международный научно-технический семинар «Современные технологии в задачах управления, автоматизации и обработки информации» (г. Алушта, 2008 г.);
- Семинар «Современные ИТ-решения для повышения эффективности работы предприятия» (г. Санкт-Петербург, 2005 г.).

Публикации. Результаты диссертации опубликованы в 14 печатных трудах, в том числе в шести статьях в журналах, которые включены ВАК РФ в перечень ведущих рецензируемых научных журналов и изданий, в статье в журнале и тезисах докладов в сборниках трудов конференций.

Структура работы. Диссертация содержит четыре главы, раздел терминологии, введение и заключение, 65 рисунков, 6 таблиц, 2 приложения. Общий объем без приложений: 138 с. (с приложениями 144 с.). Список использованных источников литературы содержит 53 наименования.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность работы, определяются цели и задачи работы.

В первой главе проводится анализ современных методов и средств интеграции унаследованных приложений и, в частности, унаследованных веб-приложений, существующих алгоритмов анализа изменений в текстовых документах.

В настоящее время все сложнее найти предприятие, не использующее информационные технологии. Автоматизация проникает все глубже в бизнес-процессы компаний во всех отраслях экономики. Однако зачастую процесс автоматизации является неупорядоченным, что приводит к так называемой «лоскутной автоматизации», когда на предприятии нет единых автоматизированных бизнес-процессов, а есть обособленные информационные системы, в ряде случаев дублирующие друг друга. В результате возникает задача интеграции унаследованных приложений. Особо следует отметить, что в последнее время в связи с активным ростом Интернета и веб-технологий все более актуальной становится интеграция унаследованных веб-приложений.

В ходе анализа существующих подходов и технологий интеграции унаследованных приложений была предложена классификация систем, предназначенных для интеграции унаследованных приложений. На основе

данной классификации показано, что в контексте интеграции унаследованных веб-приложений наиболее подходящим способом интеграции является интеграция на уровне пользовательских интерфейсов. Однако использование традиционных технологий и методов интеграции веб-приложений в единое информационное пространство предприятия на базе портальной платформы не является результативным, т.к. обладает существенными ограничениями.

Во-первых, активно продвигаемая OASIS и Java Community Process концепция интероперабельных портлетов на базе таких стандартов, как Web Services for Remote Portlets (WSRP) и Java Specification Request 168 (JSR-168), подходит исключительно для интеграции веб-контента между порталами, поддерживающими эти стандарты. Учитывая крайне низкую распространенность таких решений (особенно вне корпоративных сетей), не представляется возможным использование данного подхода для интеграции абсолютного большинства унаследованных веб-приложений.

Во-вторых, использование механизма Web-clipping, который представлен в таких программных продуктах, как IBM WebSphere Portal, Oracle Portal, Microsoft SharePoint Server, также является затруднительным на практике, т.к. не учитывается такая специфика языка HTML, как совмещение в одном документе непосредственно данных и элементов форматирования, что, в конечном итоге, позволяет корректно работать лишь со статичными HTML-страницами. Это ограничение является существенным ввиду того, что большинство современных веб-ресурсов (представленных в среде интернет или интранет) являются динамическими. Это, соответственно, накладывает требования по интеграции значимых данных веб-сайтов, которые не могут быть решены с помощью традиционных механизмов портальных платформ.

В-третьих, в работе показано, что, несмотря на популярность использования в последнее время микроформатов как инструмента выделения семантики на веб-страницах, их применение как части языка разметки неизбежно приводит к внесению изменений в интегрируемый HTML-документ, что зачастую недопустимо.

Таким образом, для интеграции произвольных унаследованных веб-приложений необходимо получить с веб-страницы иерархию значимых данных, свободную от несущественных элементов форматирования и устойчивую к изменению на самой веб-странице, иметь возможности манипулирования полученными данными и анализа изменений в иерархии. В работе показано, что одной из ключевых сложностей является то, что существует довольно ограниченное число алгоритмов, подходящих для анализа изменений в иерархических документах. Если же рассматривать класс алгоритмов для анализа изменений в HTML-документах, то все известные автору алгоритмы ориентированы на синтаксический анализ изменений, что имеет невысокую применимость в более общей задаче

интеграции унаследованных веб-приложений в силу того, что необходимо, прежде всего, анализировать значимые изменения на веб-страницах. Под синтаксическим анализом изменений понимается анализ изменений в HTML-документах, не делающий различия между значимыми данными и элементами форматирования этих данных, а также не учитывающий структуру значимых данных.

В работе показано, что наибольшую сложность в теоретических исследованиях и практических реализациях интеграции унаследованных веб-приложений представляют вопросы, связанные с задачами получения иерархии значимых данных с произвольной веб-страницы и анализа изменений в полученной иерархической структуре. Последняя задача не может быть решена с помощью существующих алгоритмов анализа изменений в текстовых документах на основе расстояния Левенштейна, Хэмминга и т.п. в виду необходимости учитывать иерархию данных. Однако показано, что эти алгоритмы могут применяться как часть общей задачи анализа изменения в иерархии значимых данных веб-сайтов.

В данной части работы сформулированы ключевые проблемы интеграции контента унаследованных веб-приложений и поставлены детальные задачи диссертационного исследования.

Во второй главе даются формальные описания алгоритма построения иерархии значимых данных, модели представления иерархии значимых данных веб-сайтов, метода идентификации узлов иерархии значимых данных, алгоритма анализа изменений иерархии значимых данных веб-сайта.

Структурно HTML-документ состоит из одной или более секций, которые:

- находятся друг относительно друга на одном уровне иерархии, например, Section 1, Section 2, и т.д.;
- одна секция структурно включает другую, например, Section 1 и Section 1.2;
- две секции находятся на разных уровнях и одна из них не включает другую, например Section 1.3 и Section 4.

Основная задача – определить иерархию секций в HTML-документе, используя HTML-теги. Поскольку язык HTML был создан не столько для структурирования данных, сколько для их отображения конечному пользователю посредством веб-браузера, данные и элементы форматирования на веб-странице смешаны, отсутствуют требования к обязательному наличию закрывающих тегов и т.п. Таким образом, более детально задача заключается в идентификации того, какие HTML-теги могут быть использованы для конструирования иерархической структуры данных

HTML документа (Тип 1), а какие служат для представления данных конечному пользователю (Тип 2). Список тегов с разделением по типам представлен на рис. 1.

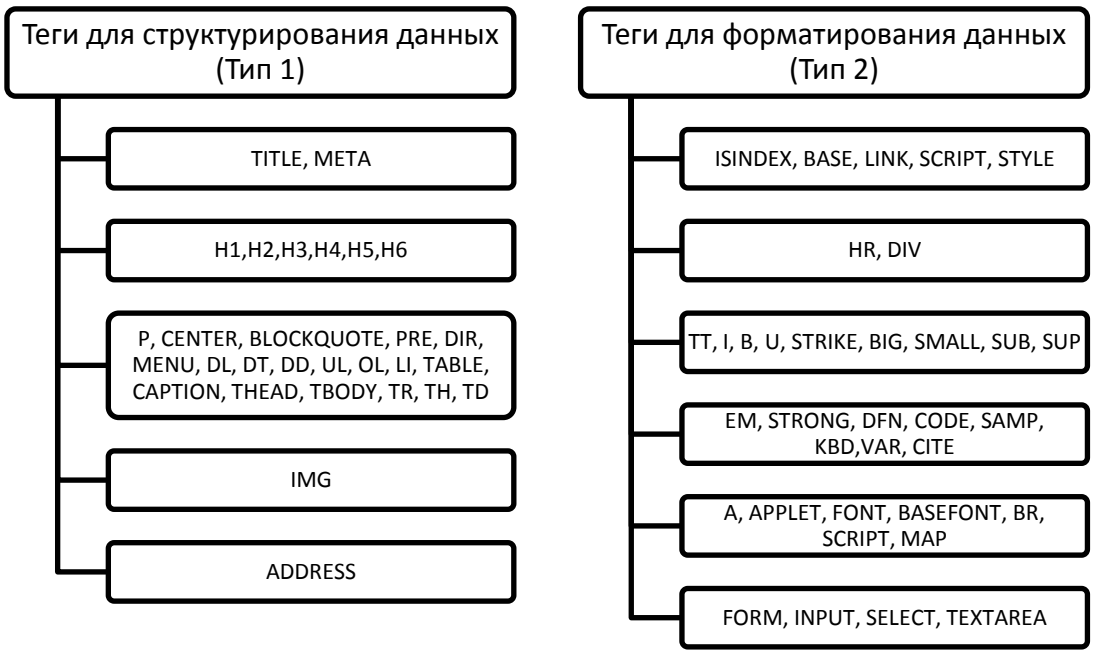

Рис. 1 Группы HTML тегов

С точки зрения построения иерархии значимых данных веб-сайтов можно выделить два основных типа данных на веб-страницах: табличные, т.е. данные, которые заключены в тег TABLE, и нетабличные. Конструирование иерархии для нетабличных данных состоит из двух шагов. На первом шаге все теги типа 2 удаляются из исходного HTML документа. На втором шаге иерархия значимых данных конструируется на основе отношения предшествования нетабличных HTML тегов так, как это изображено на рис. 2. Предшествование между двумя HTML элементами A и B, обозначаемое $A \gg B$, показывает, что данные, содержащиеся в A, находятся выше в соответствующей иерархии, чем данные, содержащиеся в B.

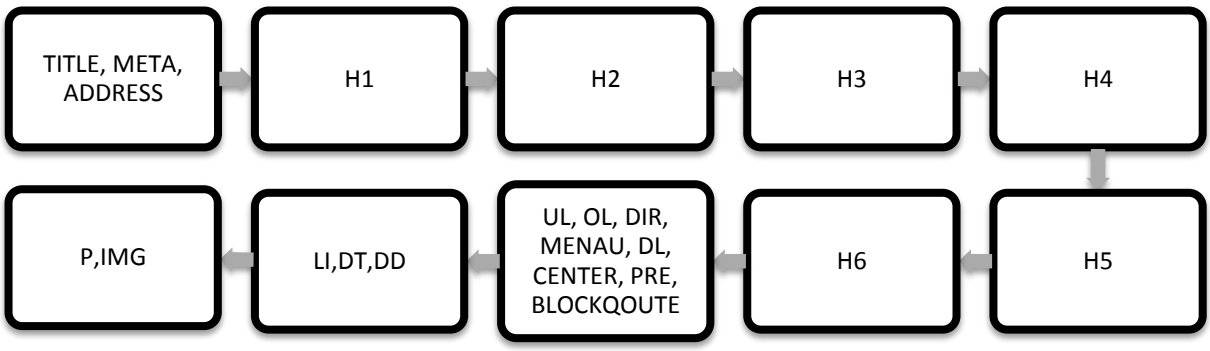

Рис. 2 Порядок предшествования нетабличных элементов (тип 1)

После определения порядка предшествования среди тегов типа 1 (за исключением тегов, предназначенных для создания таблиц) в HTML документе H применяются соответствующие правила к H для конструирования иерархии S.

Обработка табличных данных принципиально отличается от процедуры построения иерархии значимых нетабличных данных. Типовая HTML-таблица имеет как минимум один столбец-заголовок в верхней части таблицы и как минимум одну строку-заголовок в левой части. Такой тип таблиц в работе называется строчно-столбцовым. Другой тип таблицы содержит как минимум один столбец-заголовок (одну строку-заголовок) и называется в этом случае столбцовым (строчным соответственно) типом таблицы. Заголовки в строчных и столбцовых таблицах задают схему таблицы. Для любых таблиц, которые не имеют элементов TH, в ходе анализа было выявлено, что первая строка или столбец обычно используется как заголовок. Кроме того, такие атрибуты табличных элементов, как ROWSPAN и COLSPAN, играют существенную роль при построении иерархии значимых данных, т.к. объединяют соответствующие строки и столбцы конечной HTML-таблицы.

Для представления иерархии значимых табличных данных в работе вводится понятие псевдотаблицы, которая может рассматриваться как особый тип HTML-таблицы и может быть использована для выражения строчно-столбцовых, строчных и столбцовых таблиц. Общая схема построения иерархии значимых табличных данных – это, в первую очередь, отображение таблицы T на псевдотаблицу и, затем, получение из нее иерархии значимых данных. HTML-грамматика определяет иерархию HTML-документа отношением контейнер-содержимое между тегами и данными, что отлично от иерархии в псевдотаблице, поскольку в псевдотаблице нет тегов. Рис.3 иллюстрирует псевдотаблицу и соответствующую иерархию значимых данных. Основная задача в конструировании псевдотаблицы – это определить каждую строку, т.е. $a_{i1} \dots a_{in} (1 \le i \le m)$, и столбец, т.е. $a_{1j} \dots a_{mj} (1 \le j \le n)$, из соответствующей HTML-таблицы.

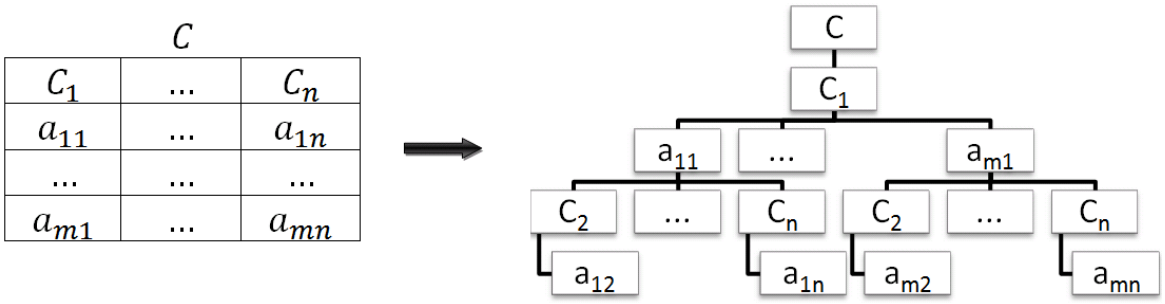

Рис. 3 Псевдотаблица T и соответствующая иерархия значимых данных

Разработана модель унифицированного представления иерархии значимых данных в XML- и RDF-формате, которая повышает интероперабельность результатов работы адаптивного инструментального программного средства интеграции контента унаследованных веб-приложений. Одной из основных задач, связанных с обработкой полученной иерархии, является корректная идентификация узлов. В работе показано, что традиционные техники обхода дерева (например, префиксный, суффиксный) возможны как способ нумерации узлов дерева, но обладают существенными недостатками в контексте изменения структуры иерархии значимых данных. В качестве решения предлагается использовать XPath-нотацию для идентификации узлов. Данный подход позволяет обеспечить навигацию и манипулирование отдельными элементами иерархии значимых данных, а также повышает устойчивость индексации элементов к изменениям в структуре иерархии значимых данных.

Еще одной немаловажной задачей, связанной с обработкой полученной иерархии, является анализ изменений в иерархии значимых данных. Другими словами, возникает задача определения степени соответствия между двумя HTML-страницами или отдельными их частями. В работе показано, что в данном случае можно применить аппарат анализа дистанции редактирования между двумя помеченными упорядоченными ориентированными деревьями. На основе данного математического аппарата разработан алгоритм анализа изменений иерархии значимых данных веб-сайтов, который может применяться для мониторинга изменений данных на веб-сайтах, эффективного кэширования веб-страниц за счет анализа степени отличия исходной и текущей иерархии веб-сайта.

Операции редактирования образуют т.н. отображение, которое является графическим представлением операций редактирования, применяемым к обоим деревьям. Рассмотрим преобразование, представленное на рис. 4.

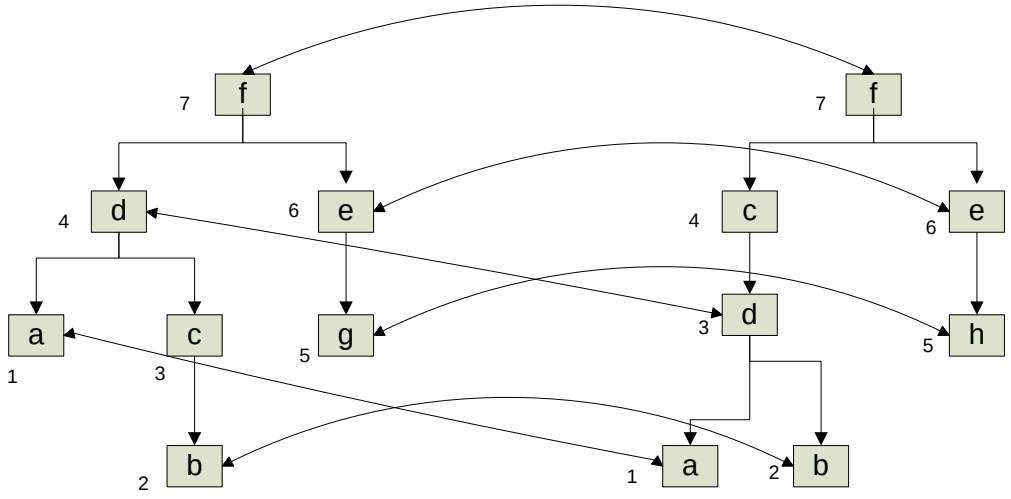

Рис. 4 Графическое представление отображения исходного дерева T_1 в результирующее дерево T_2

Это преобразование соответствует последовательности операций редактирования (delete(узел c), change($g \to h$), insert (узел c)). Каждый узел дерева представляет собой строку символов из алфавита Σ. Пусть $\lambda \notin \Sigma$ – уникальный null-символ. Операция редактирования представляется в виде $a \to b, a, b \in \Sigma \cup \lambda$. Определяется три вида операций $a \to b$: редактирование ($a \neq \lambda$ и $b \neq \lambda$), удаление ($a \neq \lambda$ и $b = \lambda$) и вставка ($a = \lambda$ и $b \neq \lambda$).

Пусть S – это последовательность $s_1, \dots, s_k$ операций редактирования. Тогда S-выводом дерева B из дерева A называется последовательность деревьев $A_1, \dots, A_k$, такая что $A = A_0, B = A_k, A_{i-1} \to A_i$ с помощью операции s_i для $1 \leq i \leq k$.

В работе показано, что мера стоимости дистанции редактирования $\gamma(M)$ отображения из T_1 в T_2может быть выражена следующей формулой:

$$\gamma(M) = \sum_{(i,j)\in M} \gamma'(t_1[i] \to t_2[j]) + \sum_{\{i \mid \nexists j,(i,j)\in M\}} \gamma(t_1[i] \to \lambda) + \sum_{\{j \mid \nexists i,(i,j)\in M\}} \gamma(\lambda \to t_2[i])$$

Мера γ' определяется либо как стандартная операция редактирования узла дерева, либо на основе дистанции редактирования Левенштейна между строками значений узлов $t_1[i]$ и $t_2[i]$, что позволяет в случае динамически изменяющихся значений узлов иерархии значимых данных определять изменения в самой структуре и определять меру схожести исходного и целевого дерева, а также в ряде задач производить поиск части дерева по шаблону.

В третьей главе рассмотрены подходы к проектированию и разработке адаптивного инструментального средства интеграции контента унаследованных веб-приложений. Приведены основные результаты и особенности разработки системы, функционирующей в рамках Microsoft SharePoint Server 2007 в среде .NET.

Объектная модель данных разработана с использованием шаблонов (паттернов) проектирования. Основная особенность разработанной модели – модульность, возможность замены отдельных компонентов системы в зависимости от требований к конкретному интеграционному решению.

Разработана архитектура системы интеграции контента унаследованных веб-приложений, включающая в себя компонент построения иерархии значимых данных, идентификации и анализа изменений узлов, визуализации и пользовательской настройки результирующих данных. Система реализована как Web-приложение, разработанное в среде ASP.NET, в качестве репрезентационного компонента используется портальный сервер Microsoft SharePoint Server 2007. Выбор репрезентационного компонента

обоснован сравнительным анализом таких программных продуктов, как IBM WebSphere Portal , Oracle Application Server 10g, Microsoft SharePoint Server 2007, SunOne Portal, SAP Enterprise Portal на основе методики Refined Hierarchical Analysis, разработанной компанией Gartner.

Архитектура разработанного адаптивного инструментального программного средства для интеграции унаследованных веб-приложений представлена на рис. 5.

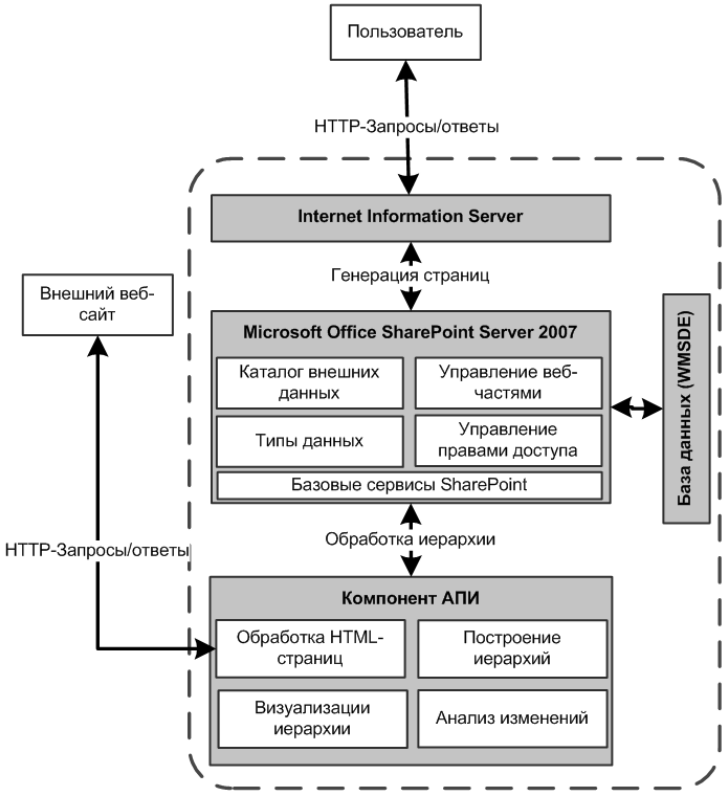

Рис. 5 Архитектура адаптивного инструментального программного средства

Общую схему интеграции унаследованных веб-приложений с помощью разработанного инструментального средства можно разделить на три основных блока:

- внешние информационные ресурсы (веб-сайты), на которых находится необходимый для интеграции контент;
- компонент построения иерархии значимых данных, который на вход получает веб-страницу, а в качестве выхода формирует унифицированное XML- или RDF-представление иерархии значимых данных веб-страницы с возможностью адресации и отслеживания изменений отдельных элементов;
- компонент обработки иерархии значимых данных. В работе показано, что в качестве данного компонента может быть не только портальный сервер, как это указано на рис. 5, но и любая информационная система, например, сервер приложений,

позволяющий публиковать веб-страницы в виде XML-иерархий, компонент трансформации иерархии в другие форматы ее представления (например, RDF и т.п.), сервер интеграции бизнес-процессов (например, на основе WSBPEL - Web Services Business Process Execution Language) для интеграции унаследованных приложений (в т.ч. и веб-приложений), создания композитных приложений или создания в среде Веб 2.0 новых сервисов на основе существующих веб-ресурсов.

В четвертой главе приводятся результаты экспериментальной проверки работоспособности инструментального программного средства, выделены ключевые особенности функционирования решения, а также предложены перспективы дальнейшего развития.

Для экспериментальной проверки работоспособности инструментального программного средства были сформированы две выборки тестовых данных: выборка существующих сайтов различной тематики и структуры представленных на них данных и выборка специально сгенерированных HTML-документов, содержащих табличные данные. В результате были получены фактические данные и проведен анализ зависимости времени выполнения основных этапов работы инструментального программного средства (получение HTML-страницы, первичная обработка полученной веб-страницы, построение иерархии значимых данных, индексация иерархии значимых данных) от таких факторов, как размер исходной страницы, наличие табличных данных, количество узлов иерархии значимых данных (рис. 6).

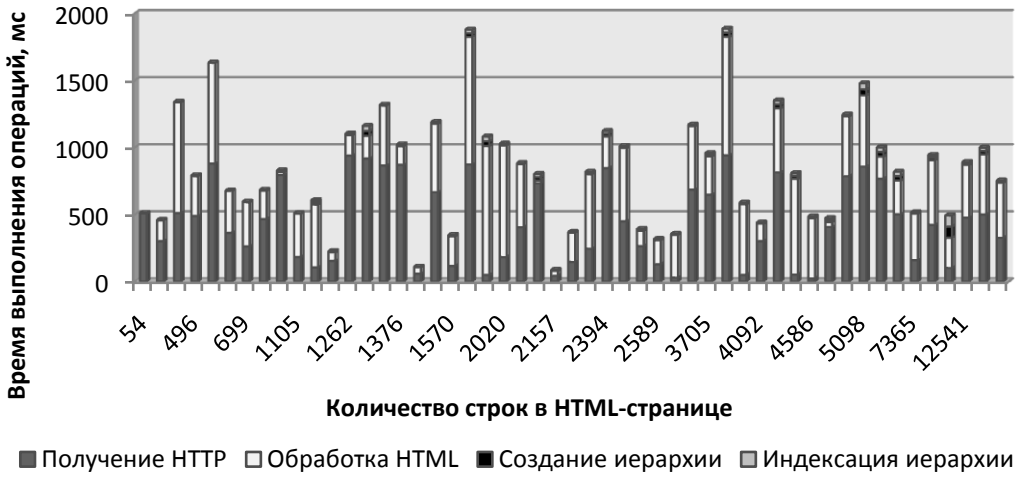

Рис. 6 Распределение времени выполнения основных операций

Для апробации адаптивного инструментального программного средства было проведено моделирование на основе разработанного

прототипа программного комплекса для интеграции контента унаследованных веб-приложений. В работе подробно рассматриваются все этапы получения и визуализации данных с внешних сайтов в среду портальной платформы Microsoft SharePoint Server 2007 с помощью разработанного инструментального программного средства.

С точки зрения дальнейших перспектив развития разработанного адаптивного инструментального программного средства показано, что можно не только проводить визуализацию данных унаследованных веб-приложений, но использовать такие механизмы портальной платформы, как списковые типы данных и Business Data Catalog для реализации логики обработки данных с нескольких веб-сайтов и публикации на портале агрегированной информации.

Следует отметить, что алгоритм построения иерархии значимых данных может быть использован не только в задаче создания единого унифицированного информационного пространства пользователя, но и в более общей задаче интеграции унаследованных веб-приложений с помощью систем класса Enterprise Application Entegration (EAI). В этом случае результаты работы инструментального программного средства могут рассматриваться как сервис (в терминологии систем EAI – адаптер) получения значимых данных произвольной веб-страницы в унифицированном XML- или RDF-представлении. Иерархия значимых данных (или отдельные узлы иерархии) может быть использована при автоматизации бизнес-процессов благодаря тому, что и сама иерархия, и язык интеграции бизнес-процессов WSBPEL (Web Services Business Process Execution Language, язык для интеграции бизнес-процессов) основаны на XML и в своей работе используют соответствующие специализированные технологии (например, XPath, XSLT и т.п.).

Результаты диссертационного исследования использованы в проектах «Автоматизация процесса поставок» в компании ООО «Хайтиан» (российское представительство HAITIAN INTERNATIONAL Hlds., Ltd) и «Организация процесса продаж» в компании ООО «Умный софт». В соответствии с актом о внедрении инструментального программного средства в ООО «Хайтиан» использование указанных результатов позволяет: сократить срок поставок продукции заказчикам за счет своевременного информирования сотрудников об изменении состояния заказа, уменьшить количество ошибок за счет интеграции разрозненных информационных систем, повысить удовлетворенность и лояльность клиентов благодаря предоставлению полной и актуальной информации о заказе. По результатам опытной эксплуатации разработанного интеграционного решения для ООО «Умный софт» были достигнуты следующие результаты: увеличилось число поступивших первичных заявок на продажу и внедрение программных решений компании ООО «Умный софт» (в среднем на 142%), увеличилось

число первичных встреч с клиентами (на 38%), повысилась эффективность работы менеджеров по продажам.

В заключении отражены основные результаты, полученные в данной работе.

В приложениях содержатся таблицы с информацией о выборке веб-сайтов для экспериментальной проверки разработанного адаптивного инструментального программного средства, копии актов о внедрении системы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

Основные результаты, полученные автором диссертационного исследования, состоят в следующем:

- Проанализированы существующие подходы к интеграции приложений для выявления основных современных проблем в области интеграции унаследованных приложений. В результате анализа предложена классификация систем интеграции приложений.
- Проведено исследование современных моделей и методов получения значимых данных с произвольных веб-сайтов, выявлены ключевые особенности и требования к разработанному алгоритму получения иерархии значимых данных унаследованных веб-приложений.
- Разработан алгоритм получения иерархии значимых данных произвольной веб-страницы с учетом структуры тегов и степени их влияния на иерархию данных.
- Разработана модель унифицированного представления иерархии значимых данных в XML- и RDF-формате, которая повышает интероперабельность результатов работы адаптивного инструментального программного средства интеграции контента унаследованных веб-приложений. Разработанный метод индексации элементов иерархии с помощь XPath-нотации предоставляет возможности получения и манипулирования отдельными элементами иерархии данных, а также повышает устойчивость индексации элементов к изменениям в структуре иерархии значимых данных.
- Разработан алгоритм анализа изменений и сопоставления иерархий значимых данных на основе дистанции редактирования в упорядоченных помеченных деревьях.
- На основе портальной платформы Microsoft SharePoint Server 2007 и технологии .NET разработано инструментальное программное средство для интеграции контента унаследованных веб-приложений, позволяющее получить унифицированное представление значимых данных с произвольной веб-страницы, выделить отдельные элементы и отобразить их в специализированном портлете на странице портала.

- Проведена экспериментальная проверка работы предложенных алгоритмов, моделей и методов на базе прототипа интеграционного решения, созданного с помощью адаптивного инструментального программного средства.
- Результаты диссертационного исследования использованы в проектах «Автоматизация процесса поставок» в компании ООО «Хайтиан» (российское представительство HAITIAN INTERNATIONAL Hlds., Ltd) и «Организация процесса продаж» в компании ООО «Умный софт», что подтверждается соответствующими актами о внедрении.

Результаты работы показывают, что поставленные цели построения, анализа и программной реализации интеграции контента унаследованных веб-приложений в единое информационное пространство предприятия можно считать достигнутыми. Практическое внедрение разработанного адаптивного инструментального программного средства подтвердило теоретические разработки, предложенные в данной работе и показало возможность практического использования в задаче интеграции унаследованных веб-приложений.

Основные положения диссертационной работы опубликованы в печатных работах [1-14].

Основные публикации по теме диссертации

1. Чеснавский, А.А. Семантическое отслеживание изменений на веб-сайтах [Текст] / А.А. Чеснавский // Вестн. НГУ. Сер. Информационные технологии. Т.5. – Новосибирск, 2008. – Вып. 5. – С. 87-94.
2. Чеснавский, А.А. Интеграция унаследованных веб-приложений [Текст] / А.А.Чеснавский // Вестн. компьютерных и информационных технологий. – М., 2009. – №3. – С. 31-36.
3. Чеснавский, А.А. Анализ изменений данных в html-документах [Текст] / А.А.Чеснавский // Вестн. компьютерных и информационных технологий. – М., 2008. – №4. – С. 37-44.
4. Чеснавский, А.А. Практическое применение алгоритма семантического анализа изменений в HTML-документах [Текст] / А.А.Чеснавский // Информационные Технологии. – М., 2009. – №1 – С.51-58.
5. Чеснавский, А.А. Семантическое отслеживание изменений на веб-сайтах [Текст] / А.А.Чеснавский // Информационные Технологии. – М., 2008. – №5 – С.16-22.
6. Чеснавский, А.А. Практическое применение алгоритма семантического анализа изменений в html-документах [Текст] / А.А. Чеснавский // Вестн. НГУ. Сер. Информационные технологии. Т.6. – Новосибирск, 2008. – Вып. 1. – С. 89-99.

7. Информационная система CACHE DOWNLOAD PAGE (CDP) / Визгалов Е.И., Кравцова А.Ю., Макаров П.А., Микушкин Д.И., Свеженцев Д.К., Чеснавский А.А. // Научная сессия МИФИ-2003. Сборник научных трудов. Т.13 Технологии разработки программных систем. Информационные технологии. – М.: МИФИ, 2003. – С. 25 -26.
8. Чеснавский, А.А.. Интеграция унаследованных веб-приложений [Текст] / А.А. Чеснавский // Современные технологии в задачах управления, автоматики и обработки информации: Труды XVII Международного научно-технического семинара. – СПб: ГУАП, 2008. – С. 237.
9. Чеснавский, А.А. Семантическое отслеживание изменений на веб-сайтах [Текст] / А.А.Чеснавский // Управление большими системами. – Вып. 19. – М., 2008. – С. 134-153.
10. Чеснавский, А.А. Унифицированный подход к интеграции унаследованных приложений [Текст] / А.А.Чеснавский // Научная сессия МИФИ-2006. Сборник научных трудов. Т.2 Технологии разработки программных систем. Информационные технологии. – М.: МИФИ, 2006. – С. 104-105.
11. Соловьев, Н.Г., Чеснавский, А.А. Механизм обмена данными в гетерогенных системах [Текст] / Н.Г. Соловьев, А.А.Чеснавский // Научная сессия МИФИ-2004. Сборник научных трудов. Т.2 Технологии разработки программных систем. Информационные технологии. – М.: МИФИ, 2006. – С. 97-98.
12. Чеснавский, А.А.. Анализ семантических изменений на веб-сайтах [Текст] / А.А. Чеснавский // Современные технологии в задачах управления, автоматики и обработки информации: Труды XVII Международного научно-технического семинара. – СПб: ГУАП, 2008. – С. 238.
13. Чеснавский, А.А. Интеграция унаследованных веб-приложений [Текст] / А.А.Чеснавский // Научная сессия МИФИ-2007. Сборник научных трудов. Т.2 Технологии разработки программных систем. Информационные технологии. – М.: МИФИ, 2007. – С. 90-91.
14. Чеснавский, А.А. Семантическое отслеживание изменений на веб-сайтах [Текст] / А.А.Чеснавский // Научная сессия МИФИ-2008. Сборник научных трудов. Т.11 Технологии разработки программных систем. Информационные технологии. – М.: МИФИ, 2008. – С. 89-91.