ВЕРИФИКАЦИЯ ЦЕЛОСТНОСТИ МАКРОСТРУКТУРЫ WEB-САЙТА СРЕДСТВАМИ РЕЛЯЦИОННОЙ АЛГЕБРЫ

М.Р. Когаловский, Е.Н. Ефимова, Т.А. Рыбина, В.Б. Брахин
Институт проблем рынка РАН
Москва, Россия
e-mail: kogalov@cemi.rssi.ru

Абстракт

Рассматривается проблема верификации макроструктуры Web-сайта, информационные ресурсы которого представлены средствами языка HTML. Под *макроструктурой* сайта понимается структура взаимосвязей его ресурсов, основанная на гиперссылках. Предлагается подход, предусматривающий "инвентаризацию" информационных ресурсов сайта, синтаксический анализ представленных на нем HTML-файлов, выявление имеющихся в них гиперссылок и, тем самым, реконструкцию его макроструктуры (графа гиперсвязей ресурсов). Ее описание помещается в реляционную базу данных и анализируется далее формальными методами, основанными на реляционной алгебре и теории графов. В результате анализа обнаруживаются неактуальные ("висячие") гиперссылки, ресурсы, на которые не указывает ни одна внутренняя гиперссылка, а также гипертекстовые страницы, недостижимые из "домашней" страницы сайта. Описывается прототип программной системы, реализующей рассматриваемый подход. Работа частично поддержана грантом РГНФ 96-02-12016.

1 Введение

Одной из важных задач, связанных с разработкой и поддержкой Web-сайта, является обеспечение целостности структуры его ресурсов. Будем различать далее *микроструктуру* сайта - структуру содержания отдельных его страниц (документов) - и *макроструктуру*, т.е. структуру взаимосвязей информационных ресурсов сайта, основанную на гиперссылках. В этой работе предлагаются методы исследования целостности макроструктуры.

Верификации целостности макроструктуры сайта сводится, грубо говоря, к исследованию существования гиперссылок, указывающих на заданные ресурсы сайта, а также существования целевых ресурсов гиперссылок и навигационных путей, образуемых гиперссылками. Более точная формулировка задачи приведена в разд. 3.

Рассматриваемая здесь проблема аналогична проблеме *целостности по ссылкам* (Referential Integrity) в области баз данных. Однако, в отличие от гипермедийных информационных систем, основанных на Web-технологиях, в традиционных системах баз данных (реляционного типа) имеются специальные механизмы автоматической поддержки целостности по ссылкам. Для их активизации достаточно декларировать соответствующие ограничения в схеме базы данных средствами описания данных, и поддержка таких ограничений будет автоматически осуществляться системными механизмами. При этом проверка рассматриваемых ограничений осуществляется динамически, непосредственно в процессе выполнения операций манипулирования данными в базе данных.

Ограничения целостности по ссылкам для Web-сайтов не могут проверяться в процессе порождения новых информационных ресурсов, поскольку язык HTML не обладает средствами декларации таких ограничений. Web-технологии, основанные на языке HTML, оперируют слабоструктурированными данными [2], для которых отсутствуют возможности типизации и интенсионального определения данных. Поэтому спецификация каких-либо абстрактных ограничений целостности макроструктуры сайтов и использование автоматических механизмов их поддержки оказываются невозможными. По существу, ограничения целостности по ссылкам представляются здесь неявно. При корректном проектировании сайта предполагается, что если имеется ссылка, то должен существовать и целевой ее ресурс, если существует ресурс, то на него должна быть хотя бы одна ссылка. Однако на практике эти ограничения часто нарушаются как для ссылок на ресурсы самого сайта, так и для ссылок на ресурсы других сайтов. Это обстоятельство обнаруживается лишь непосредственно в процессе навигации и доступа пользователей к ресурсам сайта.

Поэтому верификация целостности макроструктуры Web-сайтов является важной задачей Web-мастера. Регулярное ее осуществление при изменениях ресурсов сайта позволяет в значительной мере уменьшить количество случаев возникновения раздражающих пользователей отказов в получении запрашиваемых ими информационных ресурсов, дает возможность обнаружить избыточные ресурсы сайта.

Что касается проблем обеспечения целостности микроструктуры сайта, то здесь открываются радикально новые возможности благодаря принятому W3C новому стандарту языка разметки - *Extensible Markup Language (XML)* [1] и разработке его инфраструктуры. В частности, комплекс деклараций *Document Type Definition (DTD)* в языке XML позволяет описывать структурные

Первая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
19 - 21 октября 1999 г., Санкт-Петербург

свойства XML-документов. Используя эти метаданные, различные приложения XML, например, программы-браузеры, могут контролировать целостность микроструктуры отдельных XML-документов. Дальнейшее развитие средств описания структуры и других свойств XML-документов обеспечивают разрабатываемые W3C (World Wide Web Consortium) спецификации языка определения схемы для XML-документов [5, 6].

В среде языка XML для декларации гиперссылок между документами и/или их фрагментами должны использоваться разрабатываемые в настоящее время языки XLink [3] и XPointer [4]. Проблемы анализа и верификации макроструктуры сайта в этом случае принципиально не отличаются от случая Web-сайтов, основанных на языке HTML, хотя задача в этом случае несколько усложняется, поскольку указанные языковые средства обладают существенно более развитой функциональностью в структурообразовании, чем язык HTML. Возникающие дополнительные (по отношению к рассматриваемому в данной работе подходу) задачи относятся к стадии реконструкции макроструктуры сайта (см. разд. 4.1). Описываемые же в данной работе формальные методы ее анализа полностью применимы и в этом случае.

В данной работе предлагается подход к решению задачи верификации макроструктуры Web-сайтов, основанных на HTML-технологии, который не требует для этой цели непосредственного обхода элементов структуры сайта по гиперссылкам. Используя программные средства, реализующие рассматриваемый подход, Web-мастер может исследовать макроструктуру сайта после каждой модификации его содержания и тем самым постоянно поддерживать ее целостность.

Заметим, что особую проблему составляет поддержка актуальности ссылок на внешние ресурсы. Ресурсы этого рода не контролируются Web-мастером данного сайта, и при отсутствии регулярного мониторинга их состояния такие ссылки часто могут становиться неактуальными.

В следующих разделах работы предлагаемый подход рассматривается более подробно.

2 Основные понятия

Прежде всего, необходимо ввести некоторые дополнительные понятия, которые мы будем использовать далее наряду с общеупотребительными в области World Wide Web (см., например, [10, 12, 14]).

Будем рассматривать *Web-сайт* как совокупность взаимосвязанных гипертекстовых (гипермедийных) ресурсов Web, обладающих единством содержания.

Информационные ресурсы Web-сайта - это содержащиеся в нем файлы допустимых для HTML-среды форматов (HTML, TXT, PDF, ZIP, GIF, JPEG, CLASS и др.), а также идентифицируемые якорными точками (тег $< A\ NAME = ... >$) фрагменты HTML-файлов.

Взаимосвязи ресурсов Web обеспечиваются *гиперссылками*, содержащимися в его HTML-файлах. Каждая гиперссылка задает бинарную ориентированную связь между исходным и целевым ресурсами. Исходным ресурсом при этом всегда является HTML-файл, содержащий данную ссылку. Целевым ресурсом ссылки может быть не только ресурс данного сайта (*внутренний ресурс*), но и ресурс какого-либо другого сайта или других информационных служб Internet - FTP, Telnet, Gopher и т.д. (*внешний ресурс*). В соответствии с этим, будем различать *внутренние* и *внешние* гиперссылки.

Только такие ресурсы Web-сайтов, которые могут потенциально являться исходными или целевыми для каких-либо гиперссылок в HTML-файлах, принимаются во внимание в данной работе. К их числу не относятся, например, системы баз данных или другие приложения, доступные пользователям через CGI-интерфейс Web-сервера, на котором поддерживается данный сайт, или иным образом.

Будем различать две составные части гиперссылки - имя гиперссылки и ее значение. *Именем гиперссылки* будем называть ту ее часть (обычно отражающую смысл целевого ресурса), которую пользователь видит на экране, когда браузер воспроизводит на экране дисплея содержимое запрашиваемой HTML-страницы. Точнее говоря, имя гиперссылки - это строка, содержащаяся между открывающим и закрывающим тегом гиперссылки. В случае, если роль имени ссылки играет какой-либо графический образ, то будем далее использовать в качестве имени такой ссылки имя соответствующего графического файла. *Значением гиперссылки* будем называть заданный в ней URL целевого ресурса. Именно благодаря тому, что значение гиперссылки может оставаться скрытым от пользователя, осуществляющего навигацию в гиперсреде WWW, распределенность этой среды является для него прозрачной.

Гиперссылку будем называть *актуальной* в данный момент, если ее целевой ресурс действительно в этот момент существует, доступен, и ему соответствует заданный значением ссылки URL. В противном случае ссылка называется *неактуальной*. Поддержка актуальности внутренних ссылок - процесс, полностью контролируемый администратором данного Web-сайта. Что касается внешних ссылок, то, как уже отмечалось, Web-мастер может лишь осуществлять мониторинг их актуальности, поскольку внешние информационные ресурсы сайта им не контролируются. Нарушение актуальности гиперссылки требует от Web-мастера исключения ее вхождений в HTML-страницах данного сайта или изменения должным образом ее значения.

Информационный ресурс данного сайта будем называть *несвязанным* (или автономным), если он не является целевым ни для каких гиперссылок в этом сайте. Если следовать этому определению, к числу несвязанных ресурсов иногда может быть отнесена домашняя страница сайта. Однако это - вырожденный случай, не представляющий практического интереса.

Строго говоря, несвязанность ресурса не может трактоваться как его избыточность, поскольку всегда возможен доступ к таким ресурсам непосредственно по их URL, как и к домашней странице сайта, например, через посредство обращения к поисковой машине WWW. Тем не менее, разработчики сайтов обычно стремятся не допускать наличия такого рода ресурсов.

В соответствии с существующими традициями проектирования, предполагается, что каждый Web-сайт имеет единственную точку входа - *домашнюю страницу*. Ее URL рассматривается как адрес сайта. Каждый HTML-файл (за исключением, возможно, файла-носителя домашней страницы) должен быть *достижимым* из домашней страницы с помощью навигационных операций по гиперсвязям, т.е. должен существовать хотя бы один путь к нему по гиперссылкам из домашней страницы.

Теперь мы можем ввести понятие целостной макроструктуры сайта. Макроструктура данного сайта является *целостной*, если все содержащиеся в нем гиперссылки актуальны, он не содержит несвязанных ресурсов, и все

его HTML-файлы доступны из домашней страницы.

Для дальнейшего рассмотрения важно сделать следующее замечание. Мы предполагаем, что совокупность внутренних ресурсов сайта представляет собой множество всех файлов, содержащихся в некотором поддереве дискового каталога Web-сервера. Поэтому для задания ресурсов сайта достаточно задать соответствующий подкаталог корневого каталога сервера.

3 Постановка задачи и общий подход

Пусть для некоторого Web-сайта известны его URL, например, *SiteAddr*, подкаталог каталога Web-сервера *SitePath*, содержащий информационные ресурсы сайта, а также имя HTML-файла его домашней страницы *SiteHome*.

Задача заключается в том, чтобы проверить целостность макроструктуры данного сайта и диагностировать обнаруженные случаи ее нарушения, не прибегая при этом к непосредственному обходу структуры сайта по навигационным путям, образуемым гиперссылками, которые содержатся в HTML-файлах.

Предлагаемый общий подход к решению указанной задачи сводится к следующему.

Прежде всего проводится инвентаризация информационных ресурсов сайта. После этого осуществляется синтаксический анализ содержимого HTML-файлов и выявляются имеющиеся в них гиперссылки. Тем самым реконструируется в явном виде макроструктура сайта (граф гиперсвязей его ресурсов). Сведения о ней помещаются в реляционную базу данных. Далее с помощью формальных методов, основанных на реляционной алгебре [11, 13] и теории графов [7, 8], осуществляется собственно верификация целостности макроструктуры сайта. Выявляются имеющиеся неактуальные ссылки, несвязанные ресурсы, а также недостижимые из домашней страницы HTML-файлы, если они имеются.

Методы решения этой задачи рассматриваются в следующем разделе.

4 Формализация предлагаемого подхода

Процесс верификации макроструктуры сайта состоит из нескольких описанных ниже этапов.

4.1 Реконструкция структуры сайта

Для реконструкции макроструктуры анализируемого сайта прежде всего необходимо осуществить инвентаризацию его информационных ресурсов. (В соответствии с нашим предположением, к их числу относятся все файлы, содержащиеся в соответствующем данному сайту подкаталоге дискового каталога Web-сервера, на котором поддерживается рассматриваемый сайт.) С этой целью в реляционной базе данных системы верификации строится отношение $Resources(P, FN, FE, A)$, в котором каждому файлу сайта или фрагменту HTML-файла, идентифицируемому якорной точкой (тегом $< A\ NAME = \ldots >$), соответствует отдельный кортеж, а атрибуты этого отношения имеют следующий смысл:

P - путь доступа к данному файлу

FN - имя файла

FE - тип файла (расширение имени)

A - имя якорной точки, идентифицирующей фрагмент файла. Кортежи со значениями этого атрибута, отличными от неопределенного (Null-Value), соответствуют якорным точкам и записываются в базу данных позднее, в процессе анализа содержимого HTML-файлов при обнаружении в них этих якорных точек. В остальных кортежах этот атрибут имеет неопределенное значение.

Теперь можно приступить к реконструкции макроструктуры сайта. Для этого необходимо обнаружить все гиперссылки, которые содержатся в HTML-файлах сайта. С этой целью проводится синтаксический анализ содержимого всех HTML-файлов, зарегистрированных в отношении *Resources*. Для каждой гиперссылки, встретившейся в процессе анализа, в отношение $Links(SP, SFN, LN, Pr, SA, TP, TFN, TFE, TA)$ базы данных записывается соответствующий кортеж. Атрибуты отношения *Links* имеют следующий смысл:

SP - путь доступа к HTML-файлу сайта, содержащему данную гиперссылку

SFN - имя исходного HTML-файла ссылки

LN - имя данной ссылки

Pr - протокол доступа к целевому ресурсу ссылки

SA - адрес сервера целевого ресурса ссылки

TP - путь доступа к целевому ресурсу на сервере

TFN - имя файла целевого ресурса ссылки

TFE - тип этого файла (расширение имени)

TA - имя якорной точки, идентифицирующей фрагмент целевого HTML-файла гиперссылки.

Кроме того, как указывалось выше, выявленные в процессе синтаксического анализа HTML-файлов якорные точки регистрируются в отношении *Resources*.

Построенные таким образом отношения *Resources* и *Links* полностью представляют макроструктуру рассматриваемого сайта, и можно приступить к ее анализу.

4.2 Выявление неактуальных внутренних ссылок

Эта проблема решается теперь достаточно просто. Сначала с помощью реляционной операции селекции (σ) декомпозируем отношение *Links* на два, одно из которых *ILinks* будет содержать только внутренние ссылки, а другое *ELinks* - только внешние:

$ILinks = \sigma_{SA=SiteAddr}(Links)$;

$ELinks = \sigma_{SA\neq SiteAddr}(Links)$;

Теперь можно с помощью реляционных операций внутреннего соединения ($\bowtie$), проекции (π) и разности ($-$) отношений построить новое отношение *NoLinks* с той же схемой, что и у *Links*, содержащее сведения о неактуальных внутренних ссылках в сайте:

$$NoLinks = Links - \pi_{LAttr}(ILinks \underset{CritL}{\bowtie} Resources)$$

где:

$LAttr$=$\{SP, SFN, LN, Pr, SA, TP, TFN, TFE, TA\}$ - множество атрибутов отношения *ILinks*,

$CritL$=$(TP{=}P)\&(TFN{=}FN)\&(TFE{=}FE)\&(TA{=}A)$ - критерий соединения отношений в формуле вычисления отношения *NoLinks*.

Актуальность обнаруженных внешних ссылок, представленных в отношении *ELinks*, может быть проверена для каждой из них только с помощью попытки непосредственного доступа к ее целевому ресурсу.

4.3 Нахождение несвязанных внутренних ресурсов сайта

Для нахождения несвязанных ресурсов сайта строится отношение $UnRef$ с той же схемой, что и у отношения *Resources*, в котором каждому несвязанному внутреннему ресурсу сайта соответствует один кортеж:

$$UnRef = Resources - \pi_{RAttr}(Resources \underset{CritR}{\bowtie} ILinks)$$

где:

$RAttr$=$\{P, FN, FE, A\}$ - множество атрибутов отношения $Resources$,

$CritR$=$(P$=$TP)\&(FN$=$TFN)\&(FE$=$TFE)\&(A$=$TA)$ - критерий соединения отношений в формуле для вычисления отношения $UnRef$.

4.4 Нахождение ресурсов, недостижимых из домашней страницы

Если представить макроструктуру анализируемого сайта в виде ориентированного графа (орграфа), вершины которого соответствуют его гипертекстовым файлам, а дуги - связывающим их гиперссылкам, то рассматриваемая проблема сводится к хорошо известной - к построению матрицы достижимости для орграфа [7, 8].

Напомним, что две вершины графа находятся в бинарном отношении достижимости, если в графе существует путь из первой вершины во вторую. Матрица достижимости орграфа описывает указанное отношение на множестве его вершин. Она представляет собой матрицу смежности транзитивного замыкания данного орграфа. Единичные элементы в некоторой строке этой матрицы соответствуют тем вершинам графа, которые достижимы из вершины, соответствующей данной строке. Один из возможных алгоритмов построения матрицы достижимости предложен в [8].

Поскольку рассматриваемая здесь система построена на основе реляционной базы данных, мы полагаем, что более естественно воспользоваться для указанных целей иным алгоритмом, который эффективно реализуется в терминах реляционной алгебры. Рассмотрим предлагаемый нами алгоритм.

Пусть D - множество имен всех файлов, содержащих гипертекстовые ресурсы данного сайта. Как нетрудно видеть, отношение, которое содержит кортежи, соответствующие этим файлам, можно легко построить с помощью реляционной операции селекции из отношения $Resources$.

Построим отношение $SRC(F_s, F_t)$ такое, что $dom(F_s) = D$, $dom(F_t) = D$ и кортеж $< f_s, f_t >$, где $f_s, f_t \in D$, тогда и только тогда принадлежит SRC, когда в f_s существует хотя бы одна гиперссылка, указывающая на f_t. Ясно, что в качестве f_s может выступать только HTML-файл. Отношение SRC легко строится из $Links$ с помощью операции селекции и исключения дубликатов кортежей.

При этих условиях проблема анализа достижимости страниц сайта решается следующим итеративным алгоритмом (записанным на смеси языка Паскаль с нотацией реляционной алгебры):

$CUR := ExcludeDupl(\sigma_{TFE=''htm''}(ILinks))$;

$x := 0$;

$CUR := \sigma_{F_s=SiteHome}(SRC)$;

while $x < card(CUR)$ *do*

begin

$x := card(CUR)$;

$CUR_1 := \pi_{CUR.F_s, SRC.F_t}(CUR \underset{CUR.F_t=SRC.F_s}{\bowtie} SRC)$;

$CUR := ExcludeDupl(CUR \cup CUR_1)$

end;

где:

$card(R)$ - функция, возвращающая текущее количество кортежей в отношении R,

$ExcludeDupl()$ - функция, которая удаляет из отношения-операнда дубликаты кортежей,

σ, π, $\bowtie$, $\cup$ - символы реляционных операций селекции, проекции, внутреннего соединения и объединения отношений.

В результате исполнения этого алгоритма будет фактически построена единственная нужная нам строка матрицы достижимости для исходного графа, а именно строка, соответствующая его вершине, которая представляет домашнюю страницу исследуемого Web-сайта. Нетрудно видеть, что число возможных итераций не превышает числа гипертекстовых страниц сайта, т.е. заведомо не превышает $card(D)$.

Пусть теперь $D^* = adom(F_t, CUR)$ - актуальный домен атрибута F_t относительно результирующего отношения CUR [8]. Тогда, если $D - D^* = \emptyset$, то все HTML-файлы сайта достижимы из домашней страницы. В противном случае $D - D^*$ содержит имена HTML-файлов, не достижимых из домашней страницы.

5 Реализация прототипа

Для реализации описанного подхода авторами разработан прототип программной системы, использующий реляционную СУБД Clipper и алгоритмическим путем осуществляющий анализ и верификацию макроструктуры Web-сайтов.

Помимо решения задачи собственно верификации целостности структуры сайта с помощью методов, описанных в разд. 4, прототип диагностирует обнаруженные в HTML-файлах синтаксические ошибки, генерирует список встретившихся адресов электронной почты, формирует HTML-страницу, содержащую полный список имеющихся в сайте внешних ссылок. Актуальность таких ссылок автоматически в прототипе не проверяется, однако Web-мастер может, используя сформированную страницу, проверить далее существование нужных внешних ресурсов вручную с помощью обычного Web-браузера, последовательно "прозванивая" внешние ссылки.

Поскольку рассматриваемый прототип реализован на платформе IBM PC в среде MS-DOS, при его использовании должно выполняться ограничение на именование файлов сайта и каталогов на диске - он не работает с "длинными" именами. Кроме того, при реконструкции макроструктуры сайта не анализируются некоторые сравнительно редко используемые конструкции языка HTML, которые могут содержать в общем случае структурообразующие элементы.

В последнее время появились некоторые коммерческие и свободно доступные программные средства, предназначенные для поддержки работы Web-мастера (см. например [9]), которые позволяют синтезировать по запросу графическое представление структуры сайта и тем самым дают возможность визуального ее анализа. Нам неизвестны случаи реализации решения рассматриваемой в данной работе задачи формальными методами и автоматической диагностики нарушений целостности макроструктуры сайта.

6 Заключение

Предложенные в данной работе формальные методы анализа макроструктуры Web-сайтов, основанных на

языке HTML, являются достаточно эффективными. Об этом свидетельствует наш опыт практического использования разработанного прототипа.

Верификация макроструктуры сайта для случая XML-технологий должна стать предметом самостоятельного исследования. Однако ясно, что новые проблемы, связанные с более развитыми функциональными возможностями механизма гиперссылок в среде XML, возникают здесь лишь на стадии реконструирования макроструктуры сайта. Использованные же в данной работе формальные методы ее анализа остаются, как нам представляется, в полной мере применимыми.

Библиография

[1] *Extensible Markup Language (XML) 1.0.* W3C Recommendation 10-February-1998. [http://www.w3.org/TR/1998/REC-xml-19980210]"

[2] Florescu D., Levy A., Mendelzon A. *Database Techniques for the World-Wide Web: A Survey.* SIGMOD Record, Vol. 27, No. 3, September 1998. Есть русск. пер.: Флореску Д., Леви А., Мендельсон А. *Технологии баз данных для World-Wide Web: обзор.* СУБД, 4-5/1998.

[3] *XML Linking Language (XLink).* WWWC Working Draft, 3-March-1998. [http://www.w3.org/TR/1998/WD-xlink-19980303]"

[4] *XML Pointer Language (XPointer).* WWWC Working Draft, 3-March-1998. [http://www.w3.org/TR/1998/WD-xpointer-19980303]"

[5] *XML Schema Part 1: Structures.* W3C Working Draft 6-May-1999. [http://www.w3.org/TR/xmlschema-1]"

[6] *XML Schema Part 2: Datatypes.* W3C Working Draft 6-May-1999. [http://www.w3.org/TR/xmlschema-2]"

[7] Берж К. *Теория графов и ее применения* /Пер. с фр. под ред. И.А. Вайнштейна. - М.: ИЛ, 1962.

[8] Евстигнеев В.А. *Применение теории графов в программировании.* - М.: Наука, Гл. ред. физ-мат. литературы, 1985.

[9] Кеплер Ф. *Стратегия управления сервером Web.* LAN/Журнал сетевых решений, Октябрь 1998.

[10] Клименко С., Уразметов В. *Internet. Среда обитания информационного общества.* - Протвино: Российский центр физико-технической информатики, 1995.

[11] Мейер Д. *Теория реляционных баз данных* /Пер. с англ. под ред. М.Ш. Цаленко. - М.: Мир, 1987.

[12] Рассохин Д., Лебедев А. *World Wide Web - Всемирная Информационная Паутина в сети Internet.* - М.: Химический факультет МГУ, 1997.

[13] Ульман Дж. *Основы систем баз данных* /Пер. с англ. с предисл. и под ред. М.Р. Когаловского. - М.: Финансы и статистика, 1983.

[14] Храмцов П. *Лабиринт Internet. Практическое руководство.* - М.: Электронинформ, 1996.