РАЗРАБОТКА МЕТОДА ДЛИТЕЛЬНОГО ХРАНЕНИЯ ЦИФРОВЫХ МАТЕРИАЛОВ НА ИНТЕРНЕТ-РЕСУРСАХ

Яковлев Б.С., Проскуряков Н.Е., Архангельская Н.Н.
Тульский государственный университет

Abstract
The method of long-term storage and check of the integrity of digital content on the internet resources, including the periodic downloading of files from the server to the computer of designer and comparing them the checksum with the copy of files offered. The program was created to perform these actions in manual and automatic modes. Research and analysis of the speed of check of data were conducted and the results allow us to recommend this method for widespread use.
***Keywords**: digitization, electronic edition, Internet, protection, storage, site*

В настоящее время наблюдается рост активности в оцифровке и размещении в сети Интернет материалов образовательных учреждений, архивов, библиотек, музеев и других организаций.

Сегодняшнее состояние цифровых и телекоммуникационных технологий дает возможность дополнить классический процесс некоторыми специфическими этапами из-за потребности организаций выкладывать свои материалы в общий доступ, для рекламы и привлечения новых посетителей. Все основные и дополнительные этапы обычно выполняются параллельно, поэтому они не сильно замедляют процесс подготовки Интернет-порталов к окончательному запуску, но в них содержатся главные проблемы защиты и будущее коллекций музеев, архивов др. организаций.

Первый важной проблемой является выбор итогового формата для длительного хранения. В России архивные учреждения взяли за основу JPEG, хотя работа по их оцифровке осуществляется по программе с 2010-2020 г. и ориентировалась при создании на уже существующие методики в других странах, в том числе на Национальное управление архивов и документации (NARA - National Archives and Records Administration, USA) [1, 2].

Вторая по значимости проблема заключается в проектировании платформы сайта, которой удобно пользоваться посетителям и которая подходит для воспроизведения и защиты любого типа контента.

Проблема безопасности сайтов и долговременных хранилищ библиотек, архивов, издательств, университетов в настоящее время усугубилась

также из-за появления «вирусов-шифровальщиков», которые поражают локальные и сетевые диски, шифруя все популярные файлы, при этом каждый раз меняя пароль и алгоритм шифрования. Из-за этого на сегодняшний день данные после заражения восстановлению не подлежат.

Чтобы определить, как бороться с описанными выше угрозами необходимо понимать, как устроена работа WEB-серверов и что они из себя представляют.

Общая схема работы оборудования представлена на рис. 1. Уязвимостям заражения подвержено то оборудование, которое участвует в обработке сигнала, т.е. «управляющий компьютер», получающий запросы от пользователей, обрабатывающий их, отправляющий уже свои запросы по локальной сети в основное хранилище к «стойкам».

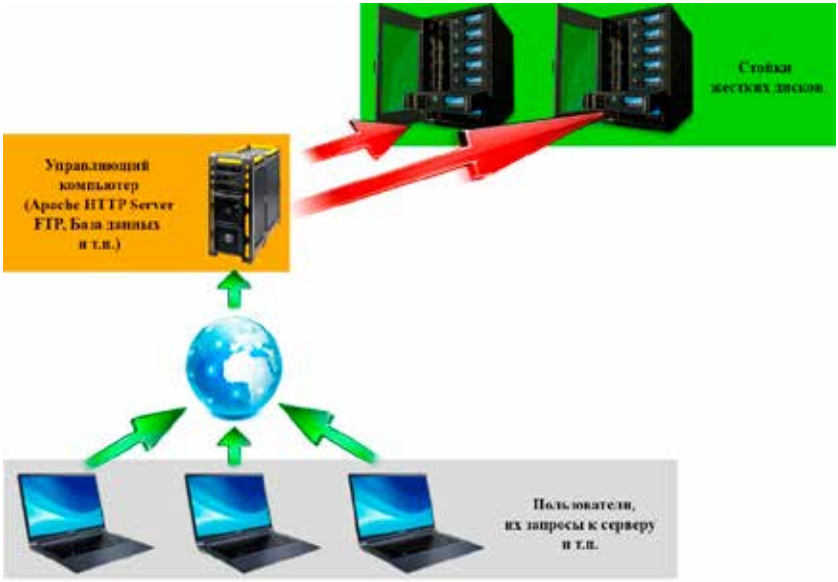

Рис. 1. Общая схема оборудования для WEB-сервера

В классической схеме на «управляющем компьютере» установлено программное обеспечение по резервному копированию, антивирусная система и программы, регламентирующие права пользователей, доступа к файлам и т.п. Он получает, проверяет, обрабатывает запросы от пользователей и после этого принимает решение давать доступ к файлам, отдавать ли ответ на запросы или нет.

«Стойки» представляют из себя корпус с отсеками под жесткие диски. В каждом отсеке есть micro-материнская плата с предустановленной операционной системой Linux, LAN разъемом. Кроме того, данное оборудование поддерживает работу RAID-массивов. Обмен данными между «стойками» и «управляющим компьютером» осуществляется через запросы, текстовыми командами.

В этой компоновке заразиться может «управляющий компьютер», т.к. гипотетически он может быть создан и не на системах UNIX. На жестких дисках в «стойках» хранится относительно статичная информация и автоматически создается резервное копирование при каждой записи, перезаписи файлов, т.к. используется RAID-массив. Поэтому мы имеем отличную систему хранения файлов на сервере, позволяющую почти на 100% защитить файлы от потери, но они могут быть заражены программным образом. Связано это с тем, что в «стойках» не может работать и обновляться антивирус, т.к. система относительно замкнута и большинство фирм, работающие на рынке Интернет-услуг имеют огромные объемы данных и проверять файлы пользователей обычным сканером в реальном времени они просто не в состоянии.

Поэтому в современных реалиях появились «антивирусы для сайтов». Что касается угроз, от которых защищают WEB-антивирусы, то это классические вирусы перенаправления, рекламные баннеры и прочие визуальные эффекты, мешающие пользователям использовать Ваш сайт.

К сожалению, есть и другая проблема в области защиты данных на WEB-серверах: сам сервер может быть заражен, если он создан не на основе системы UNIX. Тогда как бы Вы не лечили сайт, какие бы уязвимости не нашли и исправили, все равно будете постоянно заражены, хоть и будете лечить свой Интернет-ресурс. Исправить это нельзя из-за того, что WEB-антивирусы не могут проверять файлы, находящиеся на уровень выше, чем находится сам сайт.

Для того чтобы понять это необходимо вспомнить как устроено взаимодействие программ на WEB-сервере и какое ПО на нем установлено. Все программное обеспечение установлено на «управляющем компьютере» (рис. 1). В комплект программ входит: ядро (Apache); серверный язык (PHP, ASP); FTP-сервер (обычно FileZilla FTP-Server); база данных (MySQL или Oracle) и модуль отправки писем (SMTP-server). При этом ядро связывает все программы воедино, путем отправки запросов в соответствующие службы. Система конфигурации Apache основана на текстовых конфигурационных файлах. Имеет три условных уровня конфигурации: конфигурация сервера (httpd.conf); конфигурация вир-

туального хоста (httpd.conf, начиная с версии 2.2 - extra/httpd-vhosts.conf); конфигурация уровня директории (.htaccess).

Доступ к файлу конфигурации уровня директории (.htaccess) может получить владелец сайта, но конфигурация сервера (httpd.conf) и конфигурация виртуального хоста (httpd.conf) в иерархии выше, и до них владелец сайта доступа не имеет.

Чтобы осуществить такое разделение прав разработчики Apache создали 2 области - системную и пользовательскую. Формально в пользовательской области находятся все файлы вашего сайта и всех клиентов данной фирмы, а в служебной - только файлы Apache и его настройки. При попытке зайти на сервер при помощи FTP-протокола пользователи просто не увидят каталоги служебной области Apache.

Подводя промежуточный итог можно выделить 2 проблемы:

- антивирусы для сайтов малоэффективны и недейственны;
- хостинги не проводят постоянный контроль за вирусами на своих серверах из-за больших объемов данных пользователей;

Эти проблемы можно решить несколькими способами:

1. Постоянно визуально проверять состояние и поведение Вашего сайта в Интернет частыми захода на него;

2. Как угодно часто загружать файлы с Вашего сайта к себе на PC во временную директорию и проводить проверку антивирусной программой. При обнаружении угроз перезаписывать данные, содержащиеся на сервере, их незараженными копиями;

3. Сверять файлы на сервере с копиями на Вашем жестком диске.

Стоит пояснить, что первым способом результат заражения вирусами визуально выявляются быстрее, чем обход всех файлов программным образом, но он не может быть автоматизирован, а значит и не может относительно часто применяться, что делает его практически бесполезным.

Второй способ более действенен, но требует больших временных затрат и, что более важно, очень сильно связан с реакцией на зараженный файл со стороны установленного антивируса. Известно, что корпорации ESET и Kaspersky Lab при обнаружении зараженного файла вирусами перенаправления удаляют или перемещают эти файлы, не производя их лечения. Dr.WEB более лоялен и производит автоматическое лечение этих файлов. Но данный способ все равно содержит один явный проблемный пункт - антивирусные системы не распознают источник заражения, т.е. само тело вируса, т.к. для них это обычный код страниц Интернет. Данную проблему вручную решает сам человек.

Третий способ подразумевает использование проверки контрольных сумм файлов (CRC). Он более универсален, т.к. может быть применен в большинстве случаев, не зависит от описаний вирусных сигнатур в базах антивирусных программ и, главное, может обезвредить тело вируса, потому что в случае заражения файлов на сервере код так или иначе изменит размер файла.

На наш взгляд наиболее действенным способом проверки файлов WEB-серверов на вирусы является третий метод - проверка контрольных сумм файлов. Однако программным образом получить CRC по запросам к файлам по FTP-соединению невозможно. Поэтому можно предложить 2 альтернативных метода проверки:

1. Опрос файлов сервера средствами серверных языков (PHP, ASP).
2. Последовательная загрузка файлов с сервера на PC и проверка контрольных сумм файлов.

Первый метод будет работать более быстро, но в нем заложены некоторые недостатки: при осуществлении такого опроса сам файл может стать переносчиком заражения, и при этом есть вероятность ошибки определения контрольной суммы из-за разницы файловой системы Unix с пользовательской (чаще всего FAT32, NTFS).

Поэтому в работе предлагается использовать последний вариант. Чтобы избежать ошибки определения контрольных сумм, планируется загрузка файлов с сервера на PC пользователя и последующее сравнение контрольных сумм с копией файлов сайта.

С этой целью была разработана программа, выполняющая данные действия в ручном (одноразовая проверка по требованию пользователя) и автоматическом режимах (постоянный режим проверки через установленное пользователем время), интерфейс которой представлен на рис. 2.

Суть работы ПО заключается в следующем:

1. Указывается каталог где хранится оригинальная (эталонная) копия проверяемого сайта на PC.
2. В указанном каталоге, проводится опрос существующих файлов.
3. Производится подсчет контрольных сумм файлов и перевод этих данных в MD5.
4. Производится проверка файлов на сервере.
5. В случае обнаружения несовпадения файл из эталонного каталога загружается на сервер с полной заменой. Также данный файл не удаляется из временного каталога программы проверки сумм, для дальнейшего ручного анализа человеком. В случае отсутствия несовпадений временный файл удаляется с компьютера пользователя и происходит повторение пункта 4 для следующего файла.

6. В случае активации автоматического режима работы ПО будет повторять пункты 4 и 5 через заданное пользователем время. Если же был активирован ручной режим, то по окончании проверка остановится.

7. По окончании проверки можно просмотреть сохраненный отчет работы.

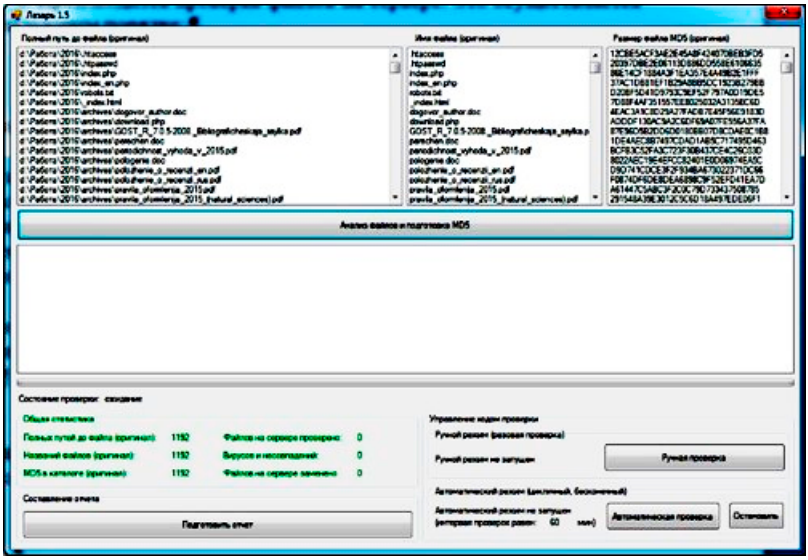

Рис. 2. Внешний вид программы проверки контрольных сумм

Данное ПО было разработано и апробировано для контроля целостности данных сайта журналов серии «Известия Тульского государственного университета» («Известия ТулГУ»), входящих в перечень ВАК РФ.

В ходе работы было установлено, что вирусы или ошибки WEB-сервера могут приводить к удалению файлов с сайта. Встречались случаи обнаружения сбойных файлов, в случае, когда происходил разрыв соединения с сетью. Предлагаемый метод и разработанное ПО справляются с данными проблемами.

Используя результаты предыдущих исследований [3], изучалась зависимость времени проверки от объемов данных и степени их заражения (табл. 1). В качестве испытуемого файла использовался архив, разбитый на части по 100 Мб. Определялось время проверки данных объемом 100 Мб, 500 Мб и 1 Гб. Также определялось время, затрачиваемое на лечение сайта в случае его полного заражения.

Для этого видеофайл был разбит на части по 100 Мб, выбрана одна из его частей. Проверяемый объем создавался за счет добавления этой части на сервер в каталоги с порядковыми номерами от 1 до 10. Это позволило строго контролировать объем данных. В случае проверки сайта на заражение, в каталоги помещался архив объемом 75 Мб с названием файла оригинала. Этим создавалось искусственное несоответствие объемов оригинального файла с проверяемым на сервере.

Стоит обратить внимание на то, что, если бы использовались малые по объему файлы, время проверки было бы меньшим, т.к. процесс опроса файлов проходил бы более динамично. Результаты исследования представлены на рис. 3 и 4.

Таблица 1. Время проверки CRC в зависимости от объемов

Объем файла	Время проверки контрольных сумм, с					Среднее значение
	опыт 1	опыт 2	опыт 3	опыт 4	опыт 5	
заражение отсутствует						
100 Мб	10	11	10	9	10	10
500 Мб	78	74	73	75	75	75
1 Гб	164	163	162	163	163	163
файлы отсутствуют						
100 Мб	47	48	47	47	47	47,2
500 Мб	230	232	232	231	232	231,4
1 Гб	430	428	428	428	427	428,2
файлы заражены						
100 Мб	45	42	44	44	45	44
500 Мб	260	266	260	261	262	261,8
1 Гб	559	580	578	578	576	574,2

Анализ скорости проверки данных (рис. 3) показал, что при отсутствии заражения файлов на сервере скорость сравнения контрольных сумм в 3…5 раз больше, чем в ситуации, когда файлы отсутствуют или были заражены. Это обусловлено тем, что по предлагаемому алгоритму файлы должны быть сохранены с сайта, а после, при несовпадении контрольных сумм, загружены обратно на сервер.

В случае отсутствия файлов на сервере (кривая 2 на рис. 4), скорость проверки данных равна скорости их прямой загрузки на сервер, т.е. равна реальной скорости работы Интернет-подключения.

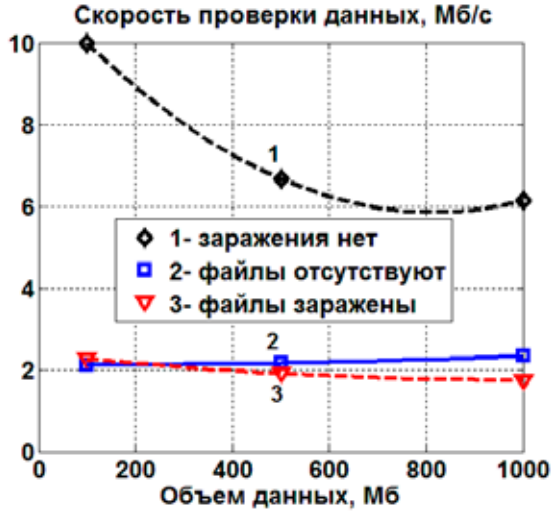

Рис. 3. График скорости проверки данных:
1 – заражения нет; 2 – файлы отсутствуют;
3 – файлы заражены

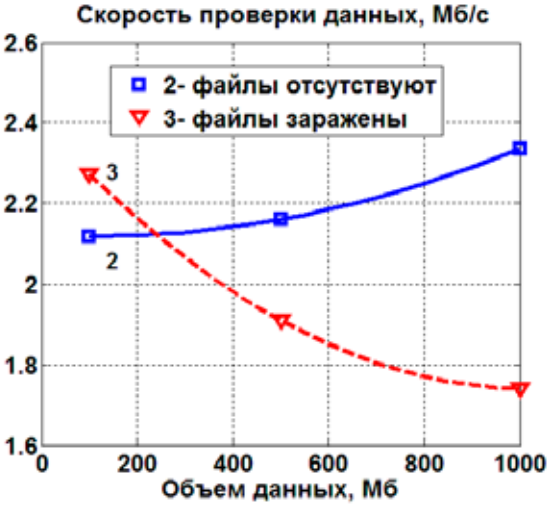

Рис. 4. График скорости проверки данных:
2 – файлы отсутствуют; 3 – файлы заражены

Минусом предложенной системы проверки сайтов является снижение скорости проверки при заражении файлов (кривая 3 на рис. 4), т.к. фактически это приводит к двукратному увеличению времени проверки. Однако, такой подход необходим, во-первых, из-за отсутствия явного метода опроса свойств файлов через FTP-соединение средствами языков программирования, во-вторых, из-за разницы файловых систем сервера и PC.

Основные выводы по работе:

1. В случае организации длительного хранения оцифрованных данных на WEB-сервере или локальной сети необходимо использовать только бинарный способ записи файлов.

2. В большинстве случаев WEB-антивирусы не в состоянии справиться с реальной угрозой для сайтов.

3. Метод борьбы с вирусными угрозами для WEB-ресурсов, основанный на проверке контрольных сумм файлов показывает хорошие результаты и может применяться для этих задач.

4. В современных условиях необходимо более активно использовать решения по резервному копированию данных, использовать облачные технологии и другие способы.

5. При разработке сайтов архивных, музейных, образовательных учреждений необходимо стараться группировать файлы сценариев по важности, т.к. это повлияет на быстроту проверки ресурса.

Список литературы

1. Малых В.В. Российская vs американская концепции развития госархивной отрасли. [Электронный ресурс]. – Режим доступа: URL: http://www.pcweek.ru/ecm/blog/ecm/6723.php#29642 (дата обращения: 10.03.2016).
2. Юмашева Ю.Ю. Методические рекомендации по электронному копированию архивных документов и управлению полученным информационным массивом. [Электронный ресурс]. – Режим доступа: URL: http://archives.ru/documents/rekomend_el-copy-archival-documents.shtml (дата обращения: 28.02.2016).
3. Яковлев Б.С. Исследование стойкости несетевых электронных изданий и основных видов контента // Б.С. Яковлев, Н.Н. Архангельская, Н.Е. Проскуряков / В сборнике: Информационное общество: образование, наука, культура и технологии будущего. Труды XVIII объединенной конференции «Интернет и современное общество» (IMS-2015). Университет ИТМО; Библиотека Российской академии наук. Санкт-Петербург, 2015. С. 153–166.