OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 004.7

ЖИВУЧЕСТЬ ИНФОРМАЦИОННЫХ ОБЪЕКТОВ В СЕТИ ИНТЕРНЕТ

Додонов А.Г.*, Ландэ Д.В.*

** Институт проблем регистрации информации НАН Украины*
г. Киев, Украина
dodonov@ipri.kiev.ua
dwlande@gmail.com

В работе рассматривается понятие живучести информационных объектов в сети Интернет, приводятся механизмы обеспечения живучести информационных объектов, дается формальное определение и формулы для расчета живучести информационных объектов в случае степенного распределения потерь информации на серверах.
Ключевые слова: информационный объект; Интернет; живучесть; степенное распределение

Введение

В связи с развитием сети Интернет в последнее время особое место среди задач, получивших актуальность, занимают задачи, связанные с обеспечением живучести информационных объектов и систем, которые связываются с моделированием их жизненного цикла: формирования и развития, реакции на деструктивные воздействия, восстановления, разрушения.

Под живучестью понимают способность системы (или ее части, объекта) адаптироваться к новым непредусмотренным условиям функционирования, противостояния нежелательным влияниям при одновременной реализации основной функции.

Существует несколько механизмов, обеспечивающих живучесть информационных объектов в Интернете.

В статье рассматриваются некоторые наиболее распространенные механизмы обеспечения живучести, которые в реальности применяются не в чистом виде, а как правило, комбинируются.

Для изучения проблем, связанных с живучестью необходимо четко определить как само это понятие, так и привести формальную модель, на основании которой можно рассчитывать уровень живучести для таких трудно формализуемых сущностей, как информационные объекты.

Кроме того, с живучестью информационных объектов сегодня связывают такие социально важные проблемы, как обеспечение информационной безопасности, приватности в сети. Этим вопросам посвящена заключительная часть данной статьи.

1. Механизмы обеспечения живучести информационных объектов

Понятие живучести информационной составляющей сети Интернет подразумевает способность информационных объектов (новостных сообщений, статей, документов, видеороликов и т.д.) своевременно выполнять свои функции (информирования) в условиях действия дестабилизирующих факторов. Такими факторами могут быть устранение отдельных объектов из информационного пространства, потеря ими свойств актуальности, доступности [Додонов, 2011], [Knight, 2003] рассмотрим некоторые из них.

1. Копирование данных при размещении их на целевой ресурс. То есть автор размещает информацию, которая копируется хостинг-провайдером на некоторое количество зеркальных серверов. Пример – скандально известная служба WikiLeaks (несколько сотен серверов, на которых хранятся фрагменты копий).

2. Перепечатка информации (републикации, «копипаст») на другие сайты с целью их информационного наполнения. В качестве примера приводится соотношение оригинальной информации и общего объема информации, сканируемой системой InfoStream [Григорьев, 2007] за первые четыре месяца 2012 г. по дням (рисунок 1). При этом следует отметить, что наиболее важная и интересная информация перепечатывается сотни раз, в то время как неактуальная, неинтересная информация практически не дублируется.

3. Размещенная однажды информация навсегда попадает в архивные службы Интернета типа Архив Интернета (http://archive.org), который накапливает сетевую информацию. Библиотека Конгресса США (www.loc.gov) купила права на хранение всех публичных сообщений социальной сети Twitter с 2006 года и всех твитов, которые будут опубликованы впредь. Библиотека Конгресса также реализует и национальный проект сохранения и распространения цифрового контента Digital Preservation (www.digitalpreservation.gov – 1400 коллекций данных).

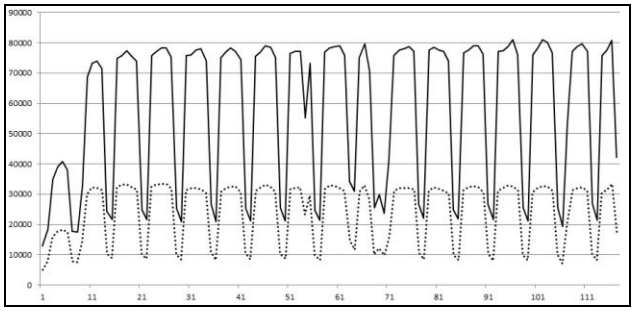

Рисунок 1 – Соотношение оригинальной информации (пунктирная линия) и общего объема информации (сплошная линия), сканируемой системой InfoStream

4. Информация часто остается в кешах поисковых систем, даже если она удалена с веб-страницы или страницы социальной сети. Информация индексируется глобальными информационно-поисковыми системами и остается в их кеш-памяти, откуда она доступна пользователям. Лишь относительно недавно у администраторов веб-ресурсов появилась возможность самостоятельного удаления своего контента из кешей Google и Яндекс. Часто многое, например, о человеке можно узнать в его блоге, онлайн-репутация – сегодня модный бренд. Что касается социальной сети Twitter, то twitFlink (http://www.twitflink.com), к примеру, который быстро соберет и выдаст твиты пациента. Сервис Google Replay позволяет находить и просматривать тематические сообщения в микроблогах за указанный период времени.

5. И наконец, информация с веб-сайта может сохраняться на локальных компьютерах конечных пользователей, которые получили к ней доступ либо непосредственно, либо через интеграторов информации.

2. Формальные модели живучести информационных объектов

Организационный комитет считает своим долгом указать на ряд часто встречающихся ошибок при оформлении статей:

Известно, что живучесть информационного объекта оценивается как вероятность того, что объект будет неповрежденным в течение определенного периода времени *t* при определенных условиях [Li, 2012].

Если информационный объект сохраняется на n серверах (носителях информации), то вероятность разрушения этого объекта оценивается как:

$$F_{lost}(t) = \prod_{i=1}^{n} F_i(t).$$

В этом произведении $F_i(t)$ – вероятность потери информационного объекта на i –м сервере за время t.

Соответственно живучесть оценивается как:

$$S_n(t) = 1 - F_{lost}(t) = 1 - \prod_{i=1}^{n} F_i(t).$$

Допуская, что вероятность уничтожения информационных объектов пропорциональна времени их существования, и то, что время их разрушения имеет степенное распределение (в соответствии с законом Парето), можно считать целесообразным и обоснованным исследование модели со степенным распределением потерь информационных объектов, что принципиально отличается от подходов, в которых используется пуассоновский поток ошибок (теория систем массового обслуживания) и распределение ошибок по Вейбулу. В этом случае, живучесть можно оценивать как:

$$S_n(t) = 1 - \prod_{i=1}^{n} F_i(t) = 1 - \prod_{i=1}^{n} Ct^{-\beta} = 1 - C^n t^{-n\beta},$$

где C, β – некоторые константы.

Закономерности статистического распределения времени жизни информационных объектов позволяют делать выводы, связанные с их живучестью, а именно учитывать явления самого подобия, нерегулярности потери информации, наличие «тяжелого хвоста» в распределении, которое характеризует чрезвычайно большое количество фактически устаревших информационных объектов и т.п.

При анализе жизненного цикла информационного объекта можно использовать еще два больших класса моделей: булевы и марковские.

В булевой модели можно предположить, что копии информационного объекта содержатся на *n* серверах, при этом *i*-му серверу соответствует булева переменная x_i, которая может принимать значения {0,1}, то есть $x_i = 1$, если информационный объект на сервере *i* активен, и 0 – в противном случае. Состояние информационного объекта определяется структурной функцией его доступности (булевой функцией) $S(x_1, x_2, \ldots, x_n)$, которая принимает значения 1, если информационный объект доступен, и 0 в противном случае.

Если доступность информационного объекта рассматривать как функцию времени, то состояние информационного объекта на *i*-м сервере можно рассматривать как случайный процесс $x_i(t)$, принимающий в произвольные моменты времени $t \geq 0$ значения 0 и 1. Для системы определяется вероятность ее безотказной работы по приведенным выше.

Среди недостатков булевых моделей можно назвать предположение только о двух состояниях информационного объекта – активности (доступности) и неактивности. Кроме того, в общем случае характер отказов отдельных копий информационного объекта зависит от состояния других копий.

Информационный объект можно описать также марковской моделью. Пусть система (множество копий информационного объекта) имеет *m* возможных состояний. Обозначим множество состояний через $M = \{z_1, z_2, \ldots, z_m\}$. Для любого фиксированного момента времени $t \geq 0$ состояние системы $z(t)$ интерпретируется как случайная величина. Заданы множество всех состояний *M*, вектор распределения начальных вероятностей $p(0)$ и функция переходных вероятностей. Определяется вероятность активности, «жизни» системы в заданный момент времени *t* (готовность системы). Применимость марковских моделей также имеет свои границы. Интенсивности переходов между отдельными состояниями системы могут быть нестационарными, принимаемые при расчете допущения относительно распределения интенсивности отказов могут значительно снизить точность полученных результатов; число состояний системы может быть так велико, что расчет становится практически невозможным.

Оценка живучести информационных объектов может проводиться на всех этапах их жизненного цикла. Существует несколько подходов, к проведению оценки живучести, имеющих наиболее общий характер. Живучесть можно оценить относительно некоторого стандартного внешнего воздействия либо относительно множества внешних воздействий. В этом случае решается задача нахождения множества характеристических векторов состояний информационного объекта (в простейшем случае – распределение по серверам), в которых реализуется конфигурация, обеспечивающая выполнение цели функционирования. Мощность этого множества может служить мерой живучести всего информационного объекта.

При анализе живучести информационных объектов рассматривается проблема информирования по их различным аспектам, независимо от наличия или отсутствия неблагоприятных факторов. В связи с этим, в качестве количественного критерия оценки живучести целесообразно использовать отношение количества функций, выполняемых объектом при наличии определенных неблагоприятных воздействий либо множества таких воздействий, к общему количеству функций информационного объекта, с учетом критичности выполняемых и не выполняемых функций. Критичность каждой конкретной функции определяется индивидуально для каждого конкретного информационного объекта исходя из его специфики. Количественный показатель живучести конкретного информационного объекта в заданных условиях можно вычислять по формуле:

$$S = \sum_{i \in \Delta} \alpha_i \Big/ \sum_{j \in \Theta} \alpha_j ,$$

где Θ – множество всех функций информирования, Δ – множество функций информационного объекта, выполняемых в заданных условиях ($\Delta \subseteq \Theta$), α_n – критичность *n*-й функции. Таким образом, количественная оценка живучести информационного объекта будет изменяется в интервале [0, 1], живучесть тем выше, чем больше ее количественная оценка.

3. Цифровые следы и тени

Удаление информационного объекта с веб-ресурса не может гарантировать его исчезновения из Интернета. Остаются не только «цифровые следы» и «цифровые тени».

Выражение «цифровые следы» («digital footprint») относятся к той информации, которая оставляется самим пользователем при работе в Сети и по которой можно не только его идентифицировать, но и «привязать» к определенным действиям, событиям, восстановить какие-то фрагменты биографии.

Часто пользователи по доброй воле указывает свои ФИО, «привязывая» дальнейшую информацию к собственной личности, дату рождения, семейное положение, образование, профессию, места предыдущей работы и много чего еще, включая и контактные телефоны, и адреса электронной почты. Кроме «цифровых следов», которые мы оставляем сами, информация о пользователях постоянно тиражируется и без всякого его участия.

Информация о пользователе, создаваемая без его участия, получила название «цифровой тени» («digital shadow»), которые возникают и накапливаются всякий раз, когда кто-то ищет пользователя через поисковые системы, когда происходит электронная почтовая рассылка по спискам, в которых он фигурирует и во многих других случаях. Индексация роботами поисковых машин страниц с информацией пользователя и их последующее кеширование – это тоже создание «цифровой тени», доступной каждому. Кроме «цифровых теней открытого доступа», создаются и копятся «цифровые тени ограниченного доступа» – записи камер наблюдения, банковские транзакции,

биллинги Интернет-магазинов, сервисов продажи билетов, телефонных звонков и др.

По оценке International Data Corporation (IDC), аналитической компании, которая специализируется на исследованиях IT-рынка, объем «цифровой тени», т.е., информации о пользователе Интернет, которая создается без его участия, уже в 2007 г. превысил объем информации, которую создает сам пользователь.

С проблемой репутации в Интернете ежедневно сталкивается все больше пользователей. Об этом свидетельствует и появление особых сайтов (например, www.suicidemachine.org), позволяющих разом удалить регистрацию и все сделанные записи на различных форумах и в социальных сетях. Такая операция называется «покончить с собой в Интернете». Однако эта система пока несовершенна. С недавнего времени эту заботу берутся специальные компании, так называемые «Интернет-чистильщики».

Заключение

Понятие живучести информационного объекта подразумевает его способность своевременно выполнять свои функции (в данном случае – информирования) в условиях действия дестабилизирующих факторов. Такими факторами могут быть устранение отдельных информационных объектов из информационного пространства, потеря их актуальности, доступности. Необходимо отметить, что привлечение внимания аудитории к другой теме, порождение другого информационного объекта также может снизить актуальность текущего информационного объекта.

При этом следует учитывать, что самая важная информация, попав в Интернет, остается там практически навсегда, и как показывает практика, рассчитывать на ее легкое удаление или изменение не приходится. Лучшим методом оказывается вытеснение нежелательной информации новыми сюжетами, проведение специальных мероприятий по содержательному исправлению ошибок [Додонов, 2010].

Живучесть информационных объектов и систем трудно заметить в нормальных условиях функционирования. Это свойство рельефно проявляется только в случаях потери информации, возникновения нарушений в структуре информационной системы, отказе ее составляющих, отдельных функций, целенаправленных деструктивных влияний. В зависимости от класса систем, их сложности, степени организованности, а также от выбранного уровня анализа свойство живучести может оцениваться как устойчивость, надежность, адаптивность, отказоустойчивость.

Наблюдаемый в настоящее время процесс в области интеллектуализации автоматизированных систем, перехода от простой обработки данных к процессам поддержки принятия решений требует новых подходов. Изначальная парадигма информационных систем, сформированная несколько десятилетий тому назад, уже не соответствует реальной ситуации – объемам и динамике информационных потоков, сетевой топологии. Необходим поиск новых принципов, в рамках которых оказалось бы возможным проектирование качественно новых систем обработки больших и динамичных массивов информации – информационных систем. Именно поэтому особое место среди задач, занимают задачи, связанные с обеспечением живучести информационных систем и их компонент.

Библиографический список

[Додонов, 2011] Додонов А.Г., Ландэ Д.В. Живучесть информационных систем. – К.: Наук. думка, 2011. – 256 с.

[Knight, 2003] Knight J.C., Strunk E.A., Sullivan K.J.Towards a Rigorous Definition of Information System Survivability // Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX'03), 2003.

[Григорьев, 2007] Григорьев А.Н., Ландэ Д.В., Бороденков С.А., Мазуркевич Р.В., Пацьора В.Н. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – Киев: ООО "Старт-98", 2007. – 40 с.

[Li, 2012] Li Y., Miller E.L., Long D.D.E. Understanding Data Survivability in Archival Storage Systems // Proceedings of the 5th Annual International Systems and Storage Conference (SYSTOR 2012), June 4–6, 2012, Haifa, Israel.

[Додонов, 2010] Додонов А.Г., Ландэ Д.В. Живучесть информационных сюжетов // Материалы XI Международной научно-практической конференции "Информационная безопасность". Ч. 2. – Таганрог: Изд-во ТТИ ЮФИ, 2010. – С. 179-183.

SURVIVABILITY OF INFORMATION OBJECTS IN THE INTERNET

Dodonov A.G. *, Lande D.V. *

******The Institute for Information Recording NAS Ukraine, Kiev, Ukraine*

dodonov@ipri.kiev.ua

dwlande@gmail.com

In work the concept of survivability of information objects in the Internet is considered. Mechanisms of ensuring survivability of information objects are given. Formal definition and formulas for calculation of survivability of information objects in case of power law distribution of losses of information on servers is given.

Keywords: information object; Internet; survivability; power law distribution