УДК 519.254

Исследование функции распределения почтового трафика для подтверждения гипотезы о его сезонности

О.С. Терновой, Е.В. Данько

Алтайский государственный университет (Барнаул, Россия)

Study of Mail Traffic Distribution Functions for Seasonality Hypothesis Confirmation

O.S. Ternovoy, E.V. Danko

Altai State University (Barnaul, Russia)

С помощью различных статистических методов изучается функция распределения почтового трафика. Отдельно исследуются потоки нежелательных почтовых сообщений (спам) и легитимного почтового трафика. Рассматриваются текущие состояние вопроса и аналогичные научные работы, выполненные на данную тему. Делается вывод о значительном увеличении почтового трафика в последние годы и изменении его структуры. В качестве гипотезы высказывается предположение о существовании «сезонности» (повторяющихся сходных периодов активности и спада) поступления почтовых писем, которое частично подтверждается. Для подтверждения данной гипотезы используются R/S-анализ и метод Херста, на основании которого рассчитываются показатели Херста для различных периодов. Анализ показывает наличие устойчивых трендонаправленных сезонных периодов для легитимного почтового трафика и стремление к блуждающему ряду для нежелательного почтового трафика. Описываются практические задачи, для которых могут быть применены полученные результаты.

Ключевые слова: R/S-анализ, метод Херста, почтовый трафик, сезонность почтового трафика.

In this paper, mail traffic distribution functions are studied using various statistical methods. Streams of unwanted e-mail messages (spam) and legitimate mail traffic are studied separately. The current state of the problem and similar studies are considered. It is concluded that there has been a significant increase in mail traffic in recent years and a change in its structure. As a hypothesis, an assumption is made about the "seasonality"(repetition of rise and fall time periods) of incoming e-mails. This hypothesis has been partially confirmed. The R/S analysis and the Hurst method are used to confirm the hypothesis. The analysis shows the presence of stable trends of legitimate mail traffic and wandering series tendency for unwanted e-mail traffic. In conclusion, there is a discussion of practical problems for which the obtained results can be applied.

Key words: R/S-analysis, Hurst method, mail traffic, seasonality of mail traffic.

DOI 10.14258/izvasu(2017)4-28

Введение. Одним из эффективных способов коммуникации продолжает оставаться электронная почта. По данным компаний, занимающихся мониторингом глобальной сети, количество существующих на сегодняшний день почтовых аккаунтов приближается к пяти миллиардам. При этом ежедневное количество почтовых сообщений, отправляемых с этих аккаунтов, составляет более 200 миллиардов [1, с. 140; 2]. Примечательно, что около 90% всего почтового трафика являются нежелательными почтовыми сообщениями, как правило, содержащими рекламную информацию, вредоносное программное обеспечение, различные мошеннические уловки и т.д. [3]. Существующие на сегодняшний момент решения не обеспечивают полноценной защиты и, как правило, имеют различные минусы [4, с. 31]. Так, например, внедрение черных списков дает большой процент ложных срабатываний, применение серых списков влияет на время доставки почты, контентная фильтрация и проверка заголовков почтовых сообщений теряют свою эффективность в случае, если злоумышленники пытаются имитировать легитимное сообщение электронной почты. Увеличение эффективности существующих решений и спам-фильтров,

направленных на блокировку нежелательного почтового трафика, может быть осуществлено в соответствии с подходом, применяемым для анализа трафика, поступающего в результате DDoS-атак, и учитывающим сезонные колебания в нагрузке сетевого ресурса (периоды активности и спада) [5, с. 57], но предварительно необходимо провести анализ функции распределения почтового трафика.

Далее, под почтовым трафиком в данной статье будет пониматься количество писем электронной почты, поступающих к почтовому серверу за определенный период времени. Принято считать, что функция распределения почтового трафик относится к распределению Пуасона, при этом функция плотности распределения вероятностей выражается формулой $f(t) = \lambda e^{-\lambda t}$, где $\lambda > 0$ — параметр интенсивности потока [6, с. 2410; 7, с. 108].

Также существует мнение о стационарности потока почтового трафика, при котором вероятностные характеристики потока не будут зависеть от времени и $\lambda = \text{const}$ [8, с. 93–94]. При этом указанные выводы были сделаны в работах, опубликованных в то время, когда интенсивность почтового трафика была значительно ниже и другими были показатели и структура спам-трафика, что может не соотноситься с сегодняшними условиями [3].

Таким образом, задачи, поставленные в работе, сводятся к исследованию функции распределения, актуализации её классификации и подтверждения гипотезы о существовании периодов активности и спада (сезонности) в почтовом трафике.

Методы исследования. Для решения поставленной задачи в работе используются статистические методы, включающие в себя:

- анализ основных статистических характеристик;
- методы проверки на нормальность, основанные на критериях согласия;
- метод наименьших квадратов;
- методы R/S-анализа;
- метод Хёрста.

Также предварительно был выполнен анализ почтового трафика, основанный на визуальных способах оценки его нормальности и наличия повторяющихся периодов роста и спада количества поступающих писем.

В качестве исходных данных для анализа были использованы данные, полученные с одного из почтовых серверов Алтайского государственного университета — mail.asu.ru. Это данные, характеризующие поток почтового трафика на основании количества поступающих почтовых сообщений (писем электронной почты) к серверу. В целях оптимизации расчетов и учитывая тот факт, что для решения основной задачи — выявления сезонности почтового трафика — не требуется большой точности, было решено количество писем, поступающих к почтовому серверу, суммировать в рамках каждого часа. При этом было учтено, что наличие в этих данных информации о почтовом спаме может быть стихийным и генерировать различные статистические выбросы, влияющие на результаты расчетов. Во избежание такой ситуации было решено рассмотреть отдельно поток легитимного почтового трафика и поток спама. Временной период, выбранный для анализа, соответствовал одному календарному месяцу, включающему в себя данные о выходных и праздничных днях. Таким образом, итоговые данные представляют собой два массива, каждый из 744 значений (24 значения для каждых суток). В одном массиве хранится число писем за каждый час, полученных сервером и доставленных в итоге пользователю, в другом — число писем за каждый час, которые были помечены сервером как спам и не доставлены пользователю. Сумма значений двух этих массивов характеризует общее количество писем, поступивших к серверу.

После проведенного визуального анализа к этим данным был применен метод нормированного размаха (1), позволяющий определить размах, сравнимый с колебаниями количества сообщений электронной почты:

$$X(t,N) = \sum_{i=1}^{t} (r_i - \bar{r}_N),\ t \le N, \qquad (1)$$

где $X(t,N)$ — накопленное отклонение за N периодов, r_i — прирост количества сообщений в i-м периоде, $\bar{r}_N$ — среднее r_i за N периодов. Тогда размах есть разность между максимальным и минимальным значениями, достигнутыми в (1).

$$R(N) = \max_{1 \le t \le N} X(t,N) - \min_{1 \le t \le j} X(t,N). \qquad (2)$$

Далее, на основании закономерности, выведенной Гарольдом Херстом и записанной здесь в общем виде (3), можно рассчитать так называемый показатель Хёрста [9, с. 85].

$$\frac{R}{S} = (a \cdot N)^H, \qquad (3)$$

где R/S — нормированный размах; N — число наблюдений; a — константа; H — показатель Хёрста. Принято считать, что значение H = 0.5, свидетельствует о том, что ряд является случайным блужданием; при H>0.5 ряд является трендонаправленным. Таким образом, на основании расчета и последующего анализа показателя Хёрста для различных периодов временного ряда можно будет сделать выводы о существовании «сезонности».

Полученные результаты. Для проведения визуального анализа был рассмотрен график, отража-

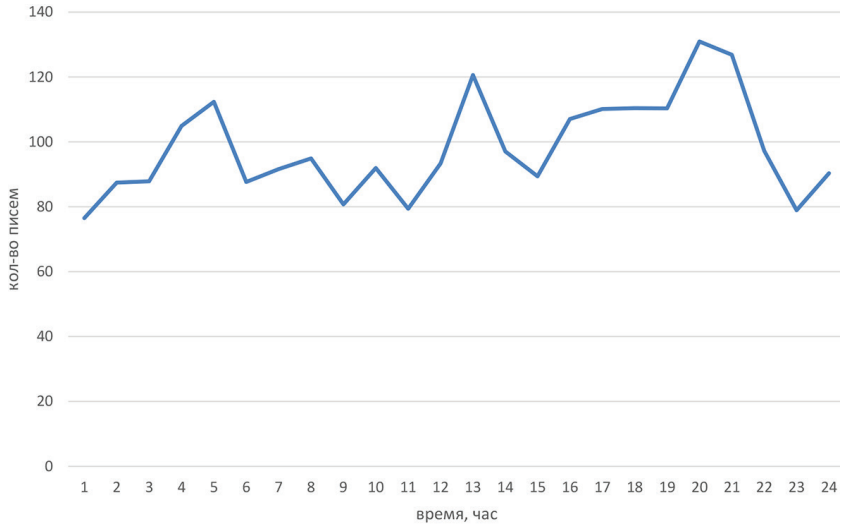

Рис. 1. Среднесуточное количество сообщений, распознанных фильтром как спам

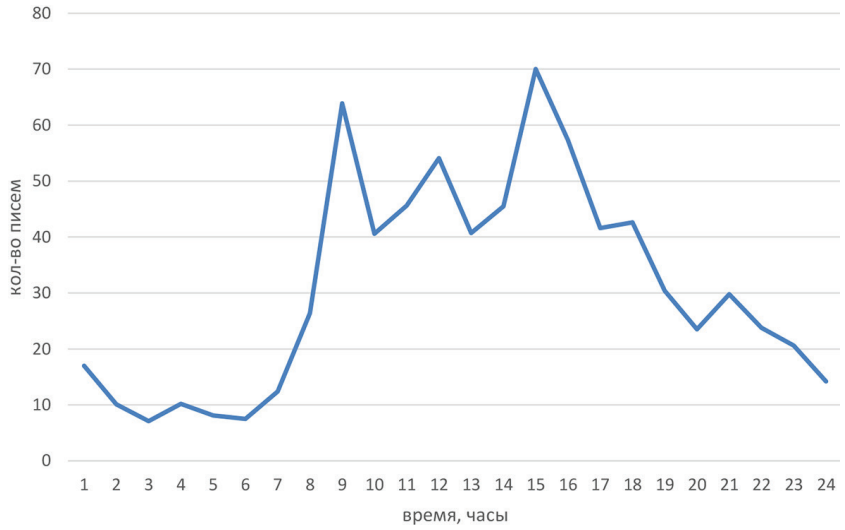

Рис. 2. Среднесуточное количество легитимных сообщений, доставленных пользователям

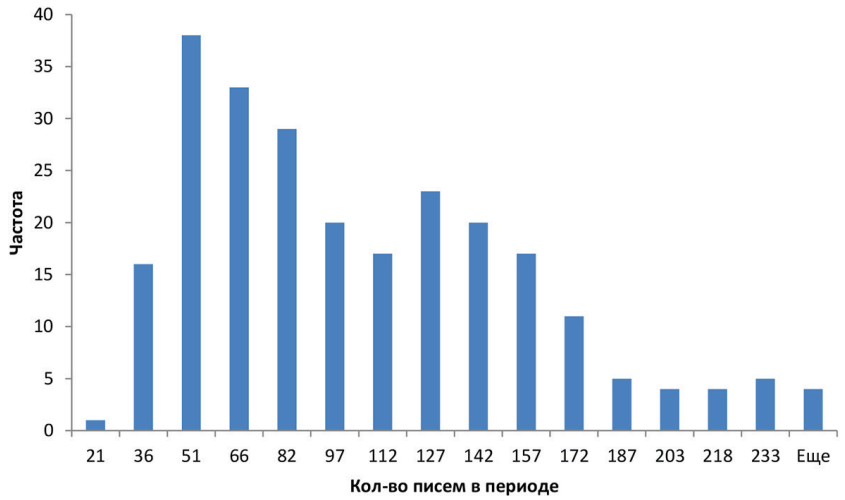

Рис. 3. Частотная диаграмма для почтового трафика, помеченного как спам

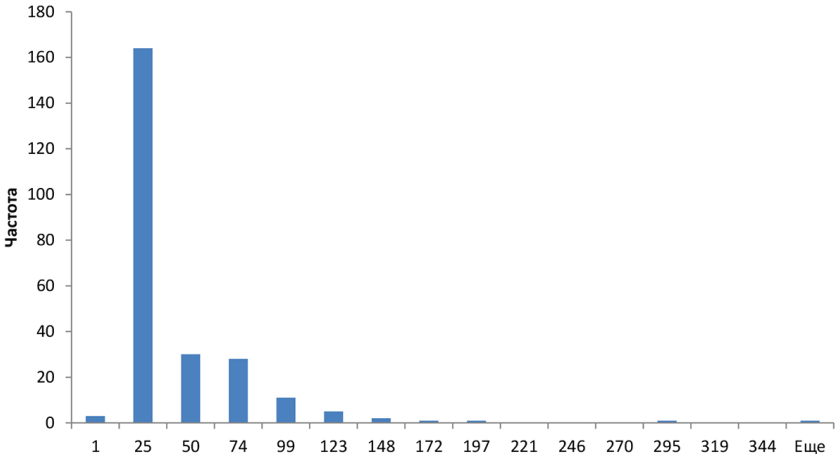

Рис. 4. Частотная диаграмма для легитимного почтового трафика

ющий среднесуточную нагрузку почтового сервера в виде полученных сообщений электронной почты. Этот график оказался малоинформативным. Однако отдельное рассмотрение графиков, характеризующих спам и легитимные почтовые сообщения, позволяет сделать несколько выводов. Так, например, спам-трафик не имеет выраженных сезонных периодов и поступает к серверу почти равномерно, независимо от времени суток, а также рабочих и выходных дней (рис. 1). Трафик легитимных почтовых сообщений, по всей видимости, имеет внутрисуточную сезонность, при которой наибольшее количество почтовых сообщений поступает в течение рабочего времени с двумя пиками, приходящимися на начало и конец рабочего дня (рис. 2).

Для предварительного заключения о нормальности функции распределения также рассмотрены частотные диаграммы для спам-трафика (рис. 3) и легитимного почтового трафика (рис. 4). Принадлежность трафика к генеральной совокупности нормального распределения не подтверждается.

Расчет показателя Хёрста для различных периодов, для легитимного и спам почтового трафика (табл.) подтверждает наличие внутрисуточной сезонности для легитимного почтового трафика, а также её отсутствие для спам-трафика.

Значение показателя Хёрста для почтового трафика

	Легитимный почтовый трафик		Не желательный почтовый трафик	
	1 день	30 дней	1 день	30 дней
Показатель Херста	0.76	0.62	0.52	0.54

Заключение Подтверждение гипотезы о существовании повторяющихся периодов интенсивного поступления почтового трафика и периодов спада открывает дополнительные возможности для его анализа и блокировки нежелательных сообщений. Так, например, одним из популярных методов контекстной фильтрации является способ, при котором каждому письму начисляются баллы за обнаруженные подозрительные объекты. В случае, если сумма баллов превышает определенное значение, письмо считается спамом. В качестве гипотезы можно высказать предположение, что в период активности потока нежелательной сообщений и, наоборот, спада потока легитимного почтового трафика, добавление большего количества штрафных баллов может уменьшить количество нераспознанного спама. Это заключение является вполне очевидным, если посмотреть на него с бытовой точки зрения. Например, ночная активность на закрытом предприятии будет больше привлекать внимание охраны, чем в часы его дневной работы.

Для формального обоснования этого заключения была модифицирована и применена функция оценки субъективной полезности принимаемого решения в условиях неопределенности [10, с. 24], в основу которой положена мера А.А. Харкевича:

$$V = \log_2 \frac{p_1}{p_0},$$

где P_0 — вероятность достижения цели до получения информации о наличии сезонных периодов; P_1 — вероятность достижения цели после получения информации о наличии сезонных периодов.

С помощью данного подхода можно улучшить эффективность и других методов борьбы со спамом. Например, в периоды активности спама увеличить время повторного запроса при использовании фильтров, основанных на «серых» списках.

В качестве следующего шага научного исследования планируется разработка методики увеличения защищённости почтового сервера от спам-трафика по аналогии с принципами, выработанными для DDoS-атак [11, с. 123].

Библиографический список

1. Будников К.И., Курочкин А.В., Лубков А.А., Яковлев А.В. Оценка датчиков мониторинга электронной почты с помощью синтетического теста transmail // Актуальные проблемы вычислительной и прикладной математики : труды Международной конференции, посвященной 90-летию со дня рождения академика Г. И. Марчука. — Новосибирск, 2015.

2. Email Statistics Report, 2010 // The Radicati Group, inc. [Electronic resourse]. — URL: www.radicati.com (дата обращения: 10.05.2017).

3. State of Spam & Phishing. A Monthly Report // Symantec corp. Report #53, [Electronic resourse]. — URL: http://www.symantec.com/content/en/us/enterprise/other_resources/b-state_of_spam_and_phishing_report_05-2011.en-us.pdf (дата обращения: 10.05.2017).

4. Назаров Д. Режем спам. Дополнительные методы // Системный администратор. — 2005. — № 2 (27).

5. Терновой О.С., Жариков А.В., Шатохин А.С Применение метода Хёрста для определения сезонности сетевого трафика с целью раннего обнаружения DDOS-атак // Динамика систем, механизмов и машин. — 2016. — Т. 4, № 1.

6. Song Luo, Gerald A. Marin. Realistic internet traffic simulation through mixture mode ling and a case study // Winter Simulation Conference. 2005.

7. Jena K., Popescu A., Nilsson A. Modeling and Evaluation of Internet Applicati ons // International Teletraffic Congress ITC18. 2003.

8. Калашников С.Г. Анализ характеристик почтового трафика на примере мэи (ТУ) // Вестник МЭИ. — 2010. — № 2.

9. Эдгар Э. Петерс. Хаос и порядок. — М., 2000

10. Данько Е.В. Функция субъективной полезности инвестиционных решений в условиях информационной неопределенности и метод оценки ее параметров // Вестник Новосибирского гос. ун-та. Серия: Информационные технологии. – 2015. – Т. 13, вып. 3.

11. Терновой О.С. Методика и средства раннего выявления и противодействия угрозам нарушения информационной безопасности в результате DDOS-атак // Известия Алтайского гос. ун-та. — 2013. — № 1/2 (77). DOI:10.14258/izvasu(2013)1.2-24.