УДК 004.042

С.В. Корелов, Л.Ю. Ротков, А.А. Рябов

Вероятностный метод идентификации спама

Анализируется метод определения вероятностных характеристик текстов и их использование для идентификации спама. Исследовались вероятностные характеристики для двух типов текста: спам и легальные рассылки. По значениям рассчитанных характеристик принималось решение о применении процедуры фильтрации.
Ключевые слова: информационная безопасность, спам, идентификация.

Борьба со спамом является актуальной задачей в сфере компьютерной безопасности. Спам превратился из легкого раздражающего фактора в одну из самых серьезных угроз информационной безопасности. Непрошеные почтовые сообщения переполняют индивидуальные почтовые ящики и парализуют работу корпоративных серверов. Время, которое сотрудники вынуждены тратить на разбор и чтение спама, постоянно растет, а с ним и финансовые потери компаний, составляющие уже, по разным оценкам, от 50 до 200 долл. в год в расчете на одного сотрудника [1].

Спам является одним из важных каналов распространения вредоносных программ. Наблюдается слияние деятельности спамеров и хакеров, которые захватывают компьютеры пользователей при помощи вирусов-троянцев и продают спамерам доступ к захваченным компьютерам для рассылки спама. Подавляющее большинство спамерских рассылок осуществляется именно с таких компьютеров-«зомби», причём только в США еженедельно обнаруживается от 80 000 до 100 000 таких машин [2].

Одним из основных классов задач, связанных с защитой компьютерных систем от различных угроз, является проблема различения «своих» и «чужих» элементов, присутствующих в этих системах. В данной работе рассматривается проблема борьбы со спамом, точнее, с нелегальными рассылками по электронной почте. Основной целью является исследование методов цифровой обработки текстовых сообщений и выяснение наиболее эффективных методов различения «хороших» (т.е. «своих») и «спамерских» («чужих») электронных писем.

Сегодня на рынке существует ряд программ (спам-фильтров): Anti-Spam Filter (ASF), SpamKiller, Spam Eater Pro, Spamassassin, Spam Buster, фильтры, интегрированные в антивирусные программы. Основные принципы работы антиспамерских программ таковы: «черные» и «белые» списки, фильтрация по заголовкам письма, эвристический анализ, вычисления IP-адреса отправителя и ответ «нет такого адреса».

Каждый из этих принципов имеет очень серьёзные недостатки.

Ведение чёрных и белых списков, а также вычисление IP-адреса отправителя полезно лишь в тех случаях, когда спамеры используют одни и те же серверы для отправки писем. С учётом нынешнего распространения технологии троянцев, делающих из обычных компьютеров «зомби», эффективность принципа ведения чёрных и белых списков как минимум сомнительна. Фильтрация по заголовкам письма и эвристический анализ любой программы становятся бесполезными, как только алгоритм продукта оказался в руках «интересующихся» людей. А ответ «нет такого адреса» может быть просто проигнорирован.

В данной работе рассматривается метод вероятностного анализа текстовых сообщений для различения спама и легальных рассылок.

В первом подходе анализируется частота использования заглавных и строчных букв кириллицы, цифр, а также знаков препинания. Каждой букве и знаку препинания ставится в соответствие число и рассчитывается относительная частота ρ_n использования n-го символа в текстовом документе:

$$\rho_n = \frac{M_n}{M_\Sigma}, \qquad (1)$$

где M_n – число вхождений n-го символа; M – общее число символов в документе.

Для каждого типа документов (спамерское письмо или легальная рассылка) на основе обучающих образцов формируется свое распределение вероятности ρ. При этом нумерованный алфавит состоит из строчных и прописных букв кириллицы, знаков препинания и символа, соответствующего всем остальным.

На рис. 1, *а* и *б* приведены диаграммы относительных частот использования символов в спамерских письмах и легальных рассылках.

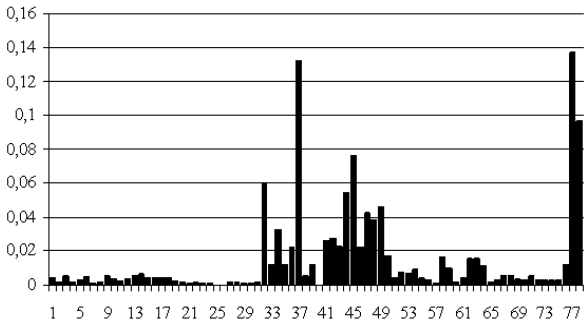

Рис. 1а. Диаграмма частоты букв в спамерских письмах

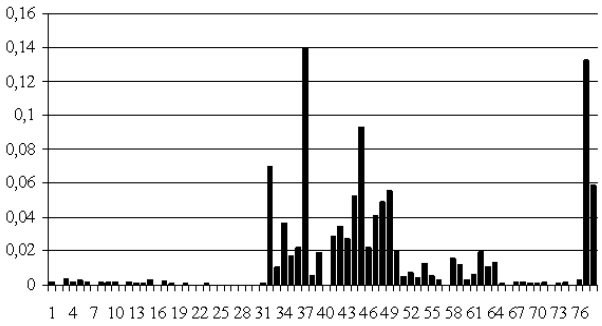

Рис. 1б. Диаграмма частоы букв в легальных рассылках

Получив данные распределения, можно рассчитать абсолютную и относительную разность частот использования букв в спамерских и легальных письмах:

$$\Delta_abs_n = \rho_{nS} - \rho_{nL}\,, \tag{2}$$

$$\Delta_otn_n = \frac{\rho_{nS} - \rho_{nL}}{w_S\rho_{ns} + w_L\rho_{nL}}\,, \tag{3}$$

где ρ_{nS} – вероятность символа в спаме; ρ_{nL} – вероятность символа в легальной рассылке; w_S и w_L – весовые коэффициенты, учитывающие разную длину сообщений соответственно спама и легальной рассылки.

При данном подходе для каждого тестового текста вычислялось распределение ρ, и с помощью метода наименьших квадратов принималось решение о принадлежности текста к спаму или легальной рассылке.

Кроме распределения вероятностей, рассматривались другие вероятностные характеристики текста.

Текстовый документ может быть представлен в виде отсчётов x_n, где каждый из отсчётов представляет собой код символа соответствующего номера.

Для обучающих и тестовых текстов спама и рассылок определялось значение среднего и дисперсии символов алфавита, определенного ранее:

$$M \equiv < x > = \frac{\sum_{n=1}^{N} x_n}{N}\,, \tag{4}$$

где N – количество символов в тексте, и

$$D = < x^2 > - < x >^2\,. \tag{5}$$

По величине отклонения среднего значения и дисперсии символов тестового письма от определенных для спама и рассылок значений принималось решение о принадлежности текста к спаму или легальной рассылке. Предельное значение этого отклонения может выбираться из величины вероятности ошибочного решения.

Разработана тестовая программа, осуществляющая проверку текстового документа произвольной длины и вырабатывающая решение об отнесении письма к тому или иному классу документов. В ходе проведения тестирования программа сортировала документы по двум классам: Legal (легальная рассылка) и Spam (спамерское письмо).

Для спама и легальной рассылки введены вероятности ошибки:

$$P_S = \frac{N_{Legals}}{N_{Spam}}\,, \tag{6}$$

$$P_L = \frac{N_{Spaml}}{N_{Legal}}\,, \tag{7}$$

где N_{Legals} – число сообщений, отнесенных к легальным, N_{Spam} – общее число спам–сообщений; N_{Spaml} – число сообщения, отнесенных к спаму; N_{Legal} – общее число легальных сообщений.

Для определения вероятности ошибки идентификации спама использовалась выборка из 204 спам-писем и 232 писем легальных рассылок. В результате были получены значения вероятности ошибки идентификации. Для подхода, использующего распределение вероятности, $P_S = 0{,}222$, $P_L = 0{,}0395$; для подхода с использованием среднего $P_S = 0{,}32$, $P_L = 0{,}25$, и, если в качестве идентифицирующего параметра использовалась дисперсия, $P_S = 0{,}198$, $P_L = 0{,}105$.

Применение данного метода не накладывает ограничений на возможность модификации и создания новых объектов, соответствующих спаму и легальным рассылкам. Нет ограничений также и на накопление, отслеживание и постоянное обновление (для новых или не похожих на существующие) образцов рассылок и спама.

Литература

1. Спам [Электронный ресурс]. – Режим доступа: http://www.kaspersky.ru/spam, свободный (дата обращения: 11.05.2010).

Корелов Сергей Викторович
Ст. преп. центра «Безопасность информационных систем и средств коммуникаций» радиофизического факультета Нижегородского госуниверситета, г. Нижний Новгород
Тел.: (+7 831) 462-32-84
Эл. адрес: korsvik@rf.unn.ru

Ротков Леонид Юрьевич
Кандидат техн. наук, доцент, проректор по безопасности Нижегородского госуниверситета, г. Нижний Новгород
Тел.: (+7 831) 462-30-06
Эл. адрес: rtv@rf.unn.ru

Рябов Аркадий Анатольевич
Ассистент центра «безопасность информационных систем и средств коммуникаций» радиофизического фагультета Нижегородского госуниверситета, г. Нижний Новгород
Тел.: (+7 831) 462-32-84
Эл. адрес: ark@rf.unn.ru

S.V. Korelov, L.Yu. Rotkov, A.A. Ryabov

Probabilistic method of spam detection

The method of definition of probabilistic characteristics of texts and their use for spam identification is analyzed. Probabilistic characteristics for two text types were investigated: a spam and legal dispatches. On values of the calculated characteristics the decision on application of procedure of a filtration was made.

Keywords: information security, spam, identification.