УДК 002:004:056

М. А. Семёнова, В. А. Семёнов

МЕТОД АВТОМАТИЧЕСКОЙ ФИЛЬТРАЦИИ ПРИ БОРЬБЕ СО „СПАМОМ“

Рассматривается метод борьбы с нежелательной корреспонденцией, распространяемой через сеть Интернет. Предлагаемый метод основан на автоматической фильтрации электронных сообщений с использованием байесовской теории.

Ключевые слова: *спам, нежелательная электронная корреспонденция, обеспечение информационной безопасности, автоматическая фильтрация, формулы Байеса.*

Расширение возможностей использования информационных ресурсов, доступных через сеть Интернет, привело, в частности, к широкому распространению нежелательной и бесполезной корреспонденции — так называемого „спама“. Спам является одной из наиболее острых проблем Интернета. Распространение такой корреспонденции сопряжено не только с потерями сетевых ресурсов, но и с временными затратами, необходимыми пользователю сети для обработки подобной информации. Потери времени на просмотр таких сообщений и, что более важно, затраты средств и ресурсов, необходимых для приобретения и обслуживания программ, фильтрующих почту, наносят ущерб более значительный, чем сетевые вирусы.

Распространение спама, в частности таких его видов, как реклама, антиреклама, так называемые нигерийские письма и фишинг [1], письма религиозного содержания и пр. [2], опасно еще и тем, что зачастую рассылаемые сообщения содержат компьютерные вирусы. Особую опасность представляют вредоносные программы определенного типа (почтовые черви), распространяющиеся с помощью электронной почты. При этом способы распространения спама также весьма разнообразны: это, например, кроме электронной почты, мгновенные и сетевые сообщения или SMS-сообщения.

Самый большой поток спама распространяется через электронную почту. В настоящее время доля вирусов и спама в общем трафике электронной почты составляет по разным оценкам от 70 до 95 %.

Распространители спама копируют электронные адреса с помощью специального робота или вручную, используя Web-страницы, конференции Usenet, списки рассылки, электронные доски объявлений, гостевые книги, чаты или другие способы. При этом рассылка спама обходится его распространителям практически бесплатно, тогда как получателю спама приходится оплачивать своему провайдеру время (или трафик), затраченное на получение непрошеной корреспонденции. Кроме того, массовый характер почтовых рассылок затрудняет работу информационных систем и ресурсов, создавая повышенную нагрузку на каналы.

В такой ситуации, с учетом перечисленных факторов, особую важность приобретает способ создания фильтров, препятствующих распространению нежелательной электронной корреспонденции. Одним из таких способов является автоматическая фильтрация — программное обеспечение (так называемые спам-фильтры), которое не требует вмешательства человека и может быть использовано как на стороне клиента (получателя письма), так и на стороне сервера. Известна также и неавтоматическая фильтрация — используемая пользователем фильтрация по ключевым словам, маскам или регулярным выражениям. Однако данный способ мало применяется, так как требует от пользователя определенных навыков, и не является особенно актуальным вследствие быстрого видоизменения спама, что пользователь зачастую не в состоянии учесть.

Рассмотрим метод автоматической фильтрации спама более подробно. Это программное обеспечение предусматривает два основных подхода [3].

Первый заключается в том, что на основе анализа содержания письма определяется, спам это или нет. Письмо, классифицированное как спам, отделяется от прочей корреспонденции: оно может быть помечено, перемещено в другую папку, удалено. Второй подход заключается в том, чтобы опознать отправителя как распространителя спама, не читая текст письма. Этот подход может быть использован только на сервере, который непосредственно принимает письма.

Проблемой при автоматической фильтрации является возможность ошибочно отметить как спам полезные сообщения. Поэтому многие почтовые сервисы и программы по желанию пользователя могут не стирать те сообщения, которые фильтр счел спамом, а помещать их в отдельную папку.

Существует множество алгоритмов поиска нежелательной корреспонденции во входящем потоке сообщений. Некоторые алгоритмы реализуются в программных средствах, позволяющих фильтровать сообщения удаленно или после копирования на компьютер пользователя. При этом анализируются заголовки сообщений, их содержание и присоединенные файлы. Наиболее эффективным из существующих является алгоритм на основе теоремы Байеса.

В основе метода автоматической фильтрации лежит механизм разбиения входящих писем на условные слова (так называемые „токены“). На основе этих токенов составляется частотный словарь, и к полученным наборам слов применяется теорема Байеса. Далее, архив прежних, вручную отсортированных, сообщений передается программе обучения. Она вычисляет частотные словари для каждого типа сообщений (папки: спам — не-спам): сколько раз определенное слово встречалось в письмах этой папки. Когда словари заполнены, вычисление вероятности принадлежности конкретного нового письма к тому или иному типу (папке) производится по формуле Байеса для каждого слова нового письма. Суммированием и нормализацией вероятностей определения спама в сообщениях получают общую оценку письма. Как правило, вероятность принадлежности сообщения к одному из типов (к папке) намного (на порядки) выше, чем его принадлежность к другому типу. Вот в эту папку (1-ю) сообщение и отправляется.

Степень распознавания спама измеряется по шкале от 0 до 1, причем 1 означает полную уверенность в том, что сообщение является спамом, а 0,5 — отсутствие какой-либо определенной оценки.

Формулы Байеса иногда используются при проверке статистических гипотез. Рассмотрим возможность получения формул Байеса с применением формулы полной вероятности. Требуется найти вероятность P наступления события A_i, если известно, что событие B произошло. Согласно теореме умножения

$$P(A_iB)=P(B)P(A_i\,|\,B)=P(A_i)P(B\,|\,A_i),$$

следовательно,

$$P(A_i\,|\,B)=\frac{P(A_i)P(B\,|\,A_i)}{P(B)}. \qquad (1)$$

Используя формулу полной вероятности для знаменателя, находим:

$$P(A_iB)=\frac{P(A_i)P(B\,|\,A_i)}{\sum_{i=1}^{k}P(A_i)P(B\,|\,A_i)}. \qquad (2)$$

Формулы (1) и (2) называются формулами Байеса. Общая схема их использования такова. Пусть событие B может происходить в различных условиях, относительно которых может быть сделано k гипотез A_1, A_2,…, A_k. Априорные вероятности этих гипотез есть $P(A_1)$, $P(A_2)$,…, $P(A_k)$. Известно также, что при справедливости гипотезы A_i вероятность

осуществления события B равна $P(B|A_i)$. Естественно, после наступления события B следует уточнить оценки вероятностей гипотез.

В прикладной статистике существует направление „байесовская статистика“, которая, в частности, на основе априорного распределения параметров после проведения измерений, наблюдений и т.п. позволяет вычислять их уточненные оценки.

К преимуществам метода автоматической фильтрации с использованием байесовской теории относятся следующие факторы:

— просмотр полного нежелательного сообщения, а не только ключевых слов или известных подписей;

— изучение исходящих сообщений электронной почты (приемлемых для получателя), что позволяет достичь заметного снижения ошибочных результатов;

— непрерывное определение новых нежелательных и новых приемлемых сообщений;

— наличие набора терминов, характерных для каждой конкретной организации, что делает невозможным обход фильтра;

— возможность распознавания нежелательной корреспонденции вне зависимости от языковой принадлежности сообщения.

При всех явных достоинствах данной теории существуют и некоторые недостатки, такие как необходимость переобучения программы, ложные срабатывания и др.

Ложные срабатывания — недостаток любых спам-фильтров. Различают два вида ложных срабатываний: false positive — неверное зачисление письма в спам, т.е. собственно ложное срабатывание, и false negative — неверное причисление письма к не-спаму, т.е. ложное не-срабатывание. При использовании байесовского фильтра ложное не-срабатывание не является проблемой: достаточно один раз указать фильтру, что сообщение является спамом, и впоследствии подобных писем не будет. Наличие же ложного срабатывания практически сводит на нет эффект борьбы со спамом: приходится просматривать папку „спам“ в поисках возможно ошибочно занесенных туда важных писем. Далее, можно дообучить спам-фильтр посредством ввода команды „это не спам“, и в будущем ложных срабатываний станет меньше.

Использование байесовской теории при создании фильтров, препятствующих распространению спама, позволяет с достаточно большой вероятностью определять принадлежность письма к спаму на основе анализа его заголовка и текста с учетом ранее полученных конкретным пользователем сообщений. Каждый владелец почтового ящика в данном случае „обучает“ программу распознавать заведомо ненужные сообщения и отсеивать их в отдельную папку.

СПИСОК ЛИТЕРАТУРЫ

1. RAZOR: [Электронный ресурс]: <http://razor.sourceforge.net>.

2. Спам — Википедия: [Электронный ресурс]: <http://ru.wikipedia.org/wiki/Спам>.

3. [Электронный ресурс]: <http://www.nsu.ru/mmf/tvims/chernova/tv/lec/node14.html>.

Сведения об авторах

Мария Александровна Семёнова — аспирант; Санкт-Петербургский государственный университет информационных технологий, механики и оптики, кафедра безопасных информационных технологий; E-mail: semeonova-maria@rambler.ru

Вениамин Александрович Семёнов — канд. техн. наук; Главное управление Банка России по Санкт-Петербургу, отдел управления безопасностью и защитой информации, эксперт; E-mail: semenov-veny@yandex.ru

Рекомендована кафедрой безопасных информационных технологий СПбГУ ИТМО

Поступила в редакцию 24.03.08 г.