Обнаружение почтового спама на основе сигнатур электронных писем

Е.В. Шарапова[1], Р.В. Шарапов[1]

[1]Владимирский государственный университет, Муромский институт, Орловская 23, Муром, Россия, 602264

Аннотация. В настоящее время в Интернет активно используются нежелательные почтовые сообщения. Миллионы копий электронных писем одновременно отправляются различным пользователям. Часто электронные письма подвергаются незначительным изменениям, чтобы усложнить обнаружение спама. В статье предлагаются варианты определения сигнатур электронных писем, позволяющих идентифицировать письма с одинаковым содержанием и структурой. Сигнатура содержания письма включает в себя основные фразы в тексте электронного письма, за исключением имен, числовых кодов, подозрительных слов, которые не включены в словарь. Сигнатура структуры включают в себя однотипные структурные элементы электронного письма, такие как абзацы, таблицы, изображения. В статье приведены результаты использования сигнатур для обнаружения почтового спама.

1. Введение

В настоящее время в Интернет активно используются нежелательные почтовые сообщения (спам). Эти сообщения содержат рекламу различных товаров и услуг, политическую рекламу, используются для фишинга и распространения вирусов. В начале 2018 года доля спама в почтовом трафике в России составила 56,72%. Другими словами – более половины почтовых писем является спамом.

Спам – анонимная незапрошенная массовая рассылка электронной почты. Миллионы копий электронных писем одновременно отправляются различным пользователям. Часто копии отличаются друг от друга приветствием (например, автоматическим указанием имени отправителя из словаря – Леонтий Людвигович, Ядвига Святославовна) или цепочкой символов (например, 1c3790b4b8ad11e8aa21e41d2d101530). Уникальность сообщений обеспечивается автоматическим путем, то есть случайными последовательностями символов, приветствиями и так далее [1]. Таким образом, подобные сообщения можно считать нечеткими дубликатами [2], обнаружение которых является не тривиальной задачей.

2. Анализ состояния проблемы

Существует множество путей обнаружения (фильтрации) спама. Крупные почтовые сервисы и компании, занимающиеся информационной безопасностью, используют распределенные методы борьбы со спамом [3]. Компании собирают информацию о проходящем через них почтовом трафике и обмениваются этими данными между собой. Таким образом они получают полную картину о действиях лиц, рассылающих спам, и могут разрабатывать и выбирать эффективные средства защиты от спама [4].

Другая группа методов борьбы со спамом – локальная. Она не использует данные с внешних сервисов, а работает только с полученными сообщениями. Локальные методы применяются как почтовыми серверами, так и конечными получателями. Часто их используют для фильтрации почты организаций [5, 6, 7].
Для поиска спама осуществляется проверка подлинности отправителя и анализ заголовков почтовых сообщений. Для этого проверяется информация о хосте-отправителе письма, его IP-адресе, анализируются коды ответов сервера и т.д. [8].
Часто для проверки используются адреса-ловушки – почтовые ящики, предназначенные исключительно для получения почтового спама и не используемые в обычной жизни.
Успешно используются в борьбе со спамом методы машинного обучения. Так распространение получила байесовская фильтрация [9], деревья решений [10], метод опорных векторов [11], методы, основанные на правилах [12] и т.д.
Много работ посвящено извлечению и последующему анализу отличительных свойств и характеристик почтовых отправлений [13, 14, 15]. Рассматриваются различные характеристики сообщений - визуальные, структурные, системные. В работе [16] предлагается использовать динамическое пространство свойств почтовых сообщений.
Ряд работ связана с анализом текстового содержания письма [17, 18]. В работе [19] анализируется текстовая информация, помещенная в изображения. В работе [20] для борьбы со спамом предлагается использовать социальные сети.
Одно из направлений борьбы со спамом основано на использовании различных. Направление основано на подсчете различными способами контрольных сумм почтовых сообщений, позволяющих обнаружить повторяющиеся письма. Известность получили сигнатуры Rabin [21], Winnowing [22], Nilsimsa [23], позволяющие подсчитывать нечеткие контрольные суммы писем. Тем не менее, совершенствующиеся методы рассылки спама делают существующие сигнатуры мало эффективными. Таким образом, возникает необходимость модификации структуры сигнатур, для более эффективного обнаружения повторяющихся писем.

3. Сигнатуры

Для идентификации писем с одинаковым содержанием и структурой могут использоваться различные сигнатуры электронных писем.
Сигнатура содержания письма SigData включает в себя основные фразы в тексте электронного письма, за исключением имен, числовых кодов, подозрительных слов, которые не включены в словарь. Сложность заключается в степени фильтрации содержания. При слабой фильтрации в тексте могут остаться элементы, используемые для уникализации теста письма. При сильной фильтрации (например, учитывать только существительные или наиболее частотные слова), различные письма могут ошибочно признаваться идентичными.
По результатам экспериментов было принято решение проводить нормализацию текста и включение в сигнатуру словоформ, полученных после обработки модулем LEMMATIZER пакета АОТ. При этом из электронного письма программно формировался пакет слов-кандидатов для включения в сигнатуру и для каждого слова проводилась лемматизация с использованием API-функций пакета АОТ. При отсутствии слова-кандидата в словаре оно в сигнатуру не включалось. В качестве словаря использовался русский морфологический словарь А.А.Зализняка, включающий 161 тысяч лемм. Таким образом, удается выявлять сообщения, прошедшие уникализацию (то есть нечеткие дубликаты писем). Сигнатура содержания письма SigData (см. рис. 1) представляет собой хэш-код, подсчитанный для обработанного выше указанным способом текста электронного сообщения.
Массово рассылаемые письма могут иметь незначительные отличия в содержании, но при этом не отличаются оформлением и расположением текстовых элементов. Другими словами, структура таких писем одинакова.

Результативное продвижение сайтов в Яндексе и Гугле. Мы поможем вам: Увеличить посещаемость сайта Увеличить количество заявок с сайта Улучшить CTR в рекламной кампании Улучшить технические и юзабилити показатели сайта Комплексное SEO продвижение (улучшение показателей проекта) продвижение по всей России продолжительный эффект и увеличение позиций Быстрая настройка контекстной рекламы Яндекс.Директ мгновенный запуск рекламы снижаем цену за клик Сделаем аудит вашего

↓

SigData: c7296428a36b7c01fbdd3fae7e41b649

Рисунок 1. Сигнатура содержания.

Сигнатура структуры SigStr включает в себя однотипные структурные элементы электронного письма, такие как абзацы, таблицы, изображения. При этом содержательная часть письма не учитывается. Для полученной таким образом структуре подсчитывается хэш-код (см. рис. 2). Письма с одинаковой внутренней структурой будут иметь одинаковые хэш-коды.

<html><head><title></title></head><body><div style=3D"text-=align:center; font-size:100%; font-family:Arial;background-color:=#ffffff !important;" class='topmessage'>

 <table border="0" cellspacing="1" cellpadding="0" width="820" height="1200"> <tbody><tr><td colspan="2" style="width: 820px; "></td></tr><tr><td style="border-right-width: 1px; border-right-color: rgb(79, 129, 189); border-right-style: solid; width: 180px; text-align: center; vertical-align: top; " rowspan="2">

<font face="Arial"><font size=

↓

SigStr: d1b37003288e83c5fdf5e34f0af0a252

Рисунок 2. Сигнатура структуры.

Надо заметить, что сигнатуры для почтовых сообщений высчитываются один раз. Дальнейшая проверка осуществляется по подсчитанным сигнатурам.

4. Использование сигнатур для обнаружения почтового спама

Предложенные сигнатуры были использованы для обнаружения почтового спама, приходящего на адреса почтового сервера Муромского института Владимирского государственного университета mivlgu.ru и адреса интернет ресурсов, расположенных на коммерческом хостинге Majordomo.ru (при отключенном фильтре спама). Почтовые сообщения, приходящие на адреса популярных почтовых сервисов (gmail.com, yandex.ru, mail.ru и т.д.), успешно проходят фильтрацию спама и не могут использоваться как источник данных для исследований.

Всего было вручную подобрано 30000 электронных сообщений, являющихся почтовым спамом. Надо заметить, что более половины сообщений (18638) были представлены несколькими копиями. Задача была в обнаружении таких писем – писем, являющихся нечеткими копиями других документов. Кроме этого, в базу сообщений было добавлено 30000 писем от реальных отправителей (то есть, не являющихся спамом).

В начале была предпринята попытка сравнить письма по телу письма – содержанию за исключением системного заголовка, содержащего отправителя, получателя, адрес почтового сервера и прочую системную информацию. Для каждого почтового сообщения были подсчитаны хэш-коды. Сообщения с одинаковыми хэш-кодами признавались за дубликаты. Число одинаковых сообщений оказалось невелико – всего 130 писем. Остальные письма имеют отличия в структуре и содержании.

При использовании сигнатуры содержания SigData было обнаружено 12237 похожих сообщений. Кроме того, из-за особенностей фильтрации содержания при подсчете сигнатуры

(удаления неинформативных элементов) 42 сообщения ошибочно были посчитаны копиями других сообщений.
При использовании сигнатуры структуры SigStr было обнаружено 14226 похожих сообщений. Из-за использования схожих шаблонов при формировании почтовых сообщений, а также сообщений в виде неформатированного текста 844 сообщения ошибочно были посчитаны копиями других сообщений.
При использовании связки сигнатур содержание-структура SigData+SigStr было обнаружено 15244 похожих сообщения и 886 сообщений ошибочно были посчитаны копиями других сообщений.
Для оценки качества работы использовались следующие метрики:

- полнота:

$$Recall = \frac{\text{Число спам} - \text{писем, отмеченных как спам}}{\text{Общее число спам} - \text{ писем}}$$

- точность:

$$Precision = \frac{\text{Число спам} - \text{писем, отмеченных как спам}}{\text{Число писем, отмеченных как спам}}$$

- число ошибок:

$$Error = \frac{\text{Число спам} - \text{писем, не отмеченных как спам}}{\text{Общее число спам} - \text{ писем}} + \frac{\text{Число не спам} - \text{писем, отмеченных как спам}}{\text{Общее число не спам} - \text{ писем}}$$

- F-мера:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Результаты приведены в таблице 1.

Таблица 1. Результаты использования сигнатур для обнаружения спама.

Сигнатура	Полнота	Точность	Число ошибок	F-мера
Содержание	0,007	1	0,993	0,014
SigData	0,656	0,996	0,332	0,791
SigStr	0,763	0,944	0,311	0,844
SigData+ SigStr	0,818	0,945	0,260	0,877

Как можно заметить, наибольшая полнота 0,818 и наименьшее число ошибок 0,260 достигается при использовании связки сигнатур содержание-структура. Наибольшие показатели точности достигаются при полном сравнении содержания писем, но при этом не определяются нечеткие дубликаты.

5. Заключение

Предложенные сигнатуры содержания и структуры могут использоваться обнаружения массовых рассылок спама, даже в случае проведения уникализации почтовых сообщений. Сигнатуры могут использоваться как по отдельности, как и в паре друг с другом. В последнем случае достигается наилучший результат с точки зрения полноты и наименьшего числа ошибок.
Для повышения качества фильтрации спама сигнатуры могут использоваться совместно с другими методами определения нежелательных сообщений. Так же предложенные сигнатуры могут служить отдельными свойствами сообщений, используемыми в качестве компонентов при применении методов машинного обучения.
В качестве практической реализации было предложено использовать сигнатуры SigData и SigStr в спам-фильтре почтового сервера управляемого авторами сервиса, размещенного на коммерческом хостинге. Для этих целей были реализованы скрипты подсчета сигнатур и добавлены новые правила в спам-фильтр. Анализ показал, что спам-письма одного содержания приходят различным получателям сервера с периодичностью от нескольких долей секунды по нескольких дней. Кроме того, многие рассылки повторяются с периодичностью от нескольких

недель до нескольких месяцев. По этой причине было принято решение хранить сигнатуры каждого письма в течении трех месяцев и использовать их для принятия решения о принадлежности к спаму вновь получаемых писем. Надо заметить, что письма помечаются фильтром как спам при совпадении хотя бы одной из сигнатур SigData и SigStr.

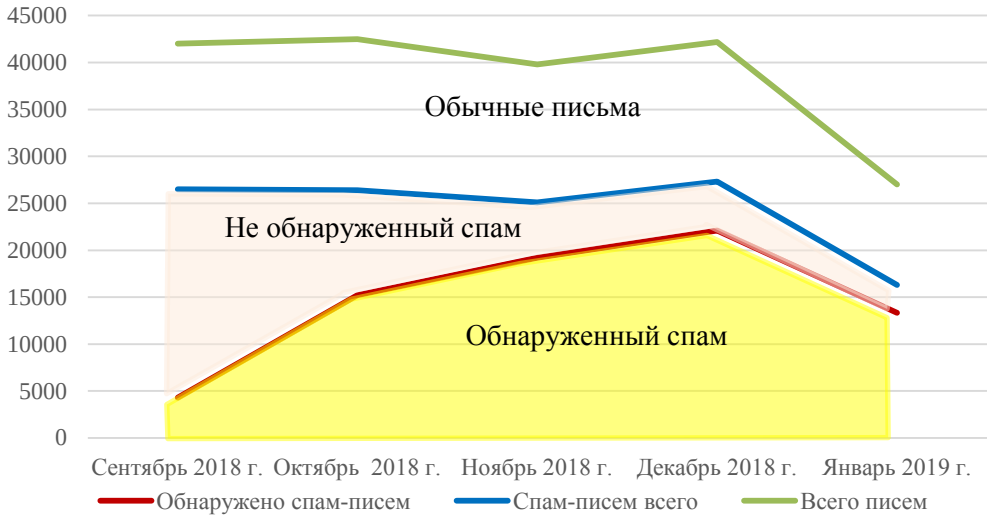

Рисунок 3. Результаты фильтрации спама на почтовом сервере.

Результаты практического использования (см. рис. 3) показали жизнеспособность предложенной методики борьбы с почтовым спамом. Сигнатуры SigData и SigStr начали использоваться с октября (в сентябре использовался другой спам-фильтр). По мере накопления информации и адаптации системы удалось существенно снизить число не обнаруженных спам-сообщений (с 42% в октябре до 18% в январе).

6. Литература

[1] Ляпичева, Н.Г. Проблемы защиты от почтового спама: влияние облачных технологий // Вестник ЦЭМИ РАН. – 2018. – Выпуск 1. DOI: 10.33276/S0000044-1-1.

[2] Sharapov, R. The problem of fuzzy duplicate detection of large texts / R. Sharapov, E. Sharapova // CEUR Workshop Proceedings. – 2018. – Vol. 2212. – P. 270-277.

[3] Ковалев, С.С. Современные методы защиты от нежелательных почтовых рассылок / С.С. Ковалев, М.Г. Шишаев // Труды Кольского научного центра РАН. – 2011. – Т. 7. – С.100-111.

[4] Терентьев, А.М. Корпоративный вариант реализации антивирусных пакетов Doctor Web в научных учреждениях: реализация // Национальные интересы. Приоритеты и безопасность. – 2013. – Т. 19, № 208. – С.40-45.

[5] Баранчикова, Е.А. Способ фильтрации электронных почтовых сообщений // Вестник РГРТУ. – 2009. – Т. 2, № 28. – С. 56-60.

[6] Мироненко, А.Н. Многоуровневая система фильтрации спама / А.Н. Мироненко, С.В. Белим // Информационные системы и технологии. – 2011. – Т. 3. – С. 125-128.

[7] Мироненко, А.Н. Модель фильтрации спам-сообщений в потоке электронной почты / А.Н. Мироненко, С.В. Белим // Вестник компьютерных и информационных технологий. – 2011. – Т. 11. – С. 34-36.

[8] Subramaniam, T. Overview of textual anti-spam filtering techniques / T. Subramaniam, H.A. Jalab, A.Y. Taqa // International Journal of. Physical Sciences. – 2010. – Vol. 5. – P. 1869-1882.

[9] Metsis, V. Spam Filtering with Naive Bayes - Which Naive Bayes? / V. Metsis, I. Androutsopoulos, G. Paliouras // Proceedings of the 3rd Conference on Email and Anti-Spam CEAS, 2006. – 9 p.

[10] Carreras, X. Boosting trees for anti-spam email filtering / X. Carreras, L. Márquez // Proceedings of 4th international conference on recent advances in natural language processing, 2001. – P. 1-8.
[11] Drucker, H. Support vector machines for spam categorization / H. Drucker, D. Wu, V. Vapnik // IEEE Transactions on Neural Networks. – 1999. – Vol. 10(5). – P. 1048-1054.
[12] Cohen, W. Learning rules that classify e-mail // Proceedings of the 1996 AAAI spring symposium on machine learning in information access, 1996. – P. 18-25.
[13] Lee, S.M. Spam detection using feature selection and parameters optimization / S.M. Lee, D.S. Kim, J.H. Kim, J.S. Park // International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), 2010. – P. 883-888.
[14] Wu, C.T. Using visual features for anti-spam filtering / C.T. Wu, K.T. Cheng // Proceedings of IEEE International Conference on Image Processing. – 2005. – P. 509-512.
[15] Beiranvand, A. Spam Filtering By Using a Compound Method of Feature Selection / A. Beiranvand, B. Shadgar // Journal of Academic and Applied Studies. – 2012. – Vol. 2. – P. 25-31.
[16] Zhou, Y. Adaptive spam filtering using dynamic feature space / Y. Zhou, M.S. Mulekar, P. Nerellapalli // Proceedings of 17th IEEE international conference on tools with artificial intelligence (ICTAI), 2005. – P. 302-309.
[17] Sasaki, M. Spam detection using text clustering / M. Sasaki, H. Shinnou // Proceedings of international conference on cyberworlds, 2005. – P. 316-319.
[18] Chirita, P.A. Mailrank:using ranking for spam detection / P.A. Chirita, J. Diederich, W. Nejdl // Proceedings of the 14th ACM international conference on information and knowledge management (CIKM), 2005. – P. 373-380.
[19] Fumera, G. Spam filtering based on the analysis of text information embedded into images / G. Fumera, I. Pillai, F. Roli // The Journal of Machine Learning Research. – 2006. – Vol. 7. – P. 2699-2720.
[20] Boykin, P. Leveraging social networks to fight spam / P. Boykin, V. Roychowdhury // Computer. – 2005. – Vol. 38(4). – P. 61-68.
[21] Rabin, M. Digitalized signature as intractable as factorization. Technical Report MIT/LCS/TR212 // MIT Laboratory for Computer Science, 1978.
[22] Schleimer, S. Winnowing: Local Algorithms for Document Fingerprinting / S. Schleimer, D.S. Wilkerson, A. Aiken // Proceedings of the ACM SIGMOD International Conference on Management of Data, 2003.
[23] Damiani, E. An open digest-based technique for spam detection / E. Damiani, S. De Capitani di Vimercati, S. Paraboschi, P. Samarati // Proc. of the International Workshop on Security in Parallel and Distributed Systems, San Francisco, CA USA, 2004.

Благодарности
Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00692.

Detection of spam using email signatures

E.V. Sharapova[1], R.V. Sharapov[1]

[1]Vladimir State University, Murom Institute, Orlovskaya street 23, Murom, Russia, 602264

Abstract. Currently on the Internet actively send unsolicited e-mail messages. Millions of copies of emails are sent simultaneously to various users. Often emails undergo minor modifications to complicate the detection of spam. The paper proposes options for determining the signature of e-mails that allow identify letters with the same content and structure. Content signature of the letter includes the basic phrases in the text of the email with the exception of names, numeric codes, suspicious words that are not included in the dictionary. Structure signatures incorporate the same type of emails, such as paragraphs, tables, images. The paper shows the results of using signatures to detect email spam.