Использование сигнатур для обнаружения массовой рассылки нечетких копий электронных сообщений

Е.В. Шарапова, Р.В. Шарапов, С.Н. Середа

Федеральное государственное бюджетного образовательного учреждения высшего образования "Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых"

Для обнаружения массовой рассылки нечетких копий электронных сообщений предложено использовать сигнатуры структуры и содержания. Приводится описание структуры сигнатур. Даются результаты использования сигнатур для обнаружения массовых рассылок почтового спама.

To detect the mass distribution of fuzzy copies of electronic messages, it was suggested to use structure and content signatures. A description of the signature structure is given. The results of the use of signatures to detect mass mailings of spam are given.

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 или цепочкой. 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.

Сигнатуры

Для идентификации писем с одинаковым содержанием и структурой могут использоваться различные сигнатуры электронных писем [3, 4].

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 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.

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. Письма с одинаковой внутренней структурой будут иметь одинаковые хэш-коды.

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

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.

Таблица 1. Результаты использования сигнатур для обнаружения почтового 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

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.

Выводы

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.

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

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. Мироненко, А.Н. Модель фильтрации спам-сообщений в потоке электронной почты / А.Н. Мироненко, С.В. Белим // Вестник компьютерных и информационных технологий. – 2011. – №11. – С. 34-36.
4. 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.