9 МЕТОДЫ И СИСТЕМЫ ЗАЩИТА ИНФОРМАЦИИ

УДК 002:004:056

МЕТОД ГРАДУИРОВАННОЙ ФИЛЬТРАЦИИ НЕЖЕЛАТЕЛЬНОЙ КОРРЕСПОНДЕНЦИИ («СПАМА»)

М.А. Семёнова, В.А. Семёнов

Получение «спама» сопряжено с потерями сетевых ресурсов и времени получателей. В такой ситуации приобретает особую важность правильный способ создания фильтров от такой нежелательной корреспонденции. В статье представлен метод градуированной фильтрации «спама», позволяющий наиболее точным образом определять «спам» в поступающем потоке сообщений.

Ключевые слова: «Спам», противоспамные фильтры, коэффициенты «спамерности», фильтрующее ПО, алгоритм фильтрации, градуированная фильтрация.

Введение

Распространение писем в сети приняло угрожающие размеры и стало серьезно мешать работе Интернета. «Спам» составляет до 90% полного объема почтовых сообщений. «Спам» делает бизнес менее эффективным [1]: он вызывает раздражение сотрудников, потери рабочего времени, ресурсов на приобретение и обслуживание фильтрующих программ, часть «спама» часто заражена вирусами, «червями» или «троянскими программами» [2]. Противоспамные фильтры могут не пропустить и уничтожить важное сообщение, сочтя его за «спам». Впрочем, и человек, вынужденный просматривать десятки рекламных сообщений в день, тоже легко может пропустить среди них нужную корреспонденцию.

В настоящее время существует несколько алгоритмов фильтрации нежелательной корреспонденции. В частности, фильтр, основанный на алгоритме Байеса, имеет следующие достоинства:

- уникальный для каждой организации набор данных, что делает невозможным обход фильтра;
- просмотр полного нежелательного сообщения, а не только ключевых слов или известных подписей;
- многоязычность.

В других алгоритмах для расчета «спамерности» применяются следующие правила:

- для анализа сообщений используются ограниченный набор (15, 25, 27 в зависимости от алгоритма) наиболее «интересных» слов, для которых коэффициент «спамерности» слов наиболее сильно отклоняется от нейтрального значения (0,5).
- если ранее слово встречалось менее чем в пяти письмах, оно игнорируется.

Главными недостатками всех существующих методов являются ложные тревоги, пропуск «спама», фиксированное количество слов, участвующих в оценке письма. Представленный в данной статье метод градуированной фильтрации «спама» позволяет более точным образом определять «спам» во входящем потоке сообщений за счет иного способа расчета коэффициентов «спамерности», а также гибкой настройки определения количества слов, участвующих в оценке письма.

Борьба с нежелательной электронной корреспонденцией

В последнее время в почтовых программах, кроме стандартных папок «Входящие», «Исходящие», «Отправленные», «Корзина», у каждого ящика появилась папка

«Спам» («Junk mail»), в которую должна отсортировываться вся нежелательная или сомнительная корреспонденция. Эта сортировка возможна как вручную, так и при помощи самодельной системы фильтров. Действительно, все рекламные письма имеют в своем тексте что-то общее: в одних предлагают что-нибудь купить, в других – что-то посетить; в одних оставляют свой адрес, в других – телефон [3]. Совсем не сложно выбрать около десятка признаков, по которым можно отнести письма к категории «спам» [4]. Используемый при этом метод заключается в следующем: анализируется содержание письма, производятся расчеты коэффициентов «спамерности», и на основании рассчитанных коэффициентов производится общая оценка письма и делается вывод, «спам» это или нет. Письмо, классифицированное как «спам», отделяется от прочей корреспонденции: оно может быть помечено, перемещено в другую папку, удалено. Пользователь не видит отфильтрованного «спама», но продолжает нести издержки, связанные с его приемом, так как фильтрующее программное обеспечение (ПО) получает каждое письмо и только потом решает, показывать его или нет.

Проблемой при автоматической фильтрации является то, что она может по ошибке отмечать 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.

Метод фильтрации сообщений на основе градации сообщений

Метод позволяет более детально оценивать письмо, перед тем как отнести его к определенной категории («спам» или «не спам»). В основе метода градуированной фильтрации лежит механизм разбиения входящих писем на слова («токены»), на основе которых составляются частотные словари. Блок-схема составления частотных словарей представлена на рис. 1.

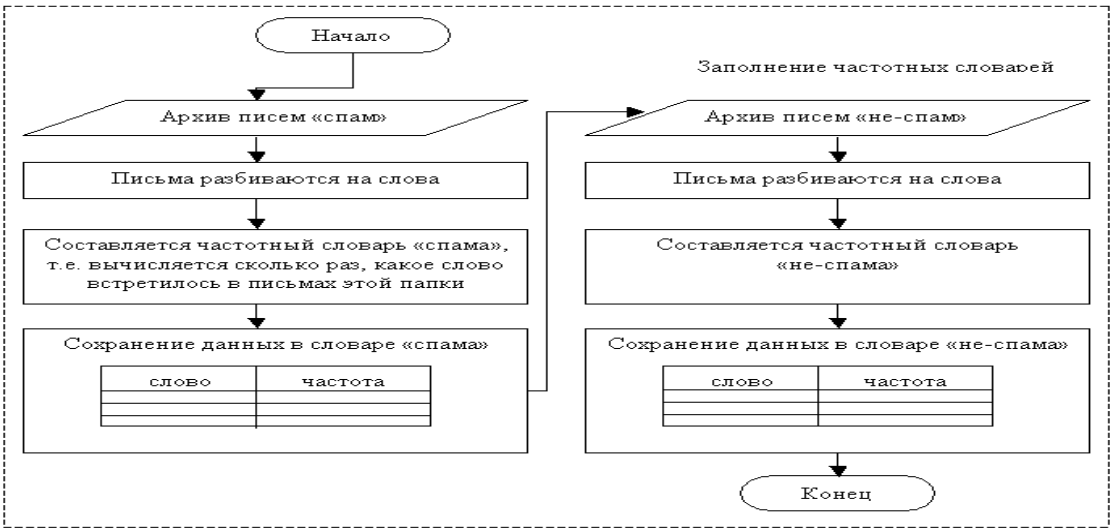

Рис. 1. Составление частотных словарей

Для обучения используется архив старых сообщений, отсортированных вручную (почти все пользователи ведут свои почтовые архивы, в которых «спам» хранится либо в папке «удаленные», либо в отдельной папке «спам»). Программа обучения для каждого типа (папки) сообщений вычисляет частоту встречаемости слов в письмах этой папки. Когда словари заполнены, производится вычисление вероятности принадлежности конкретного нового письма к той или иной категории («спам» или «не спам») (рис. 2). Письмо, поступающее через фильтр, разбивается на слова, и каждому слову сопоставляется коэффициент встречаемости из частотного словаря, причем при отсутствии сло-

ва в словарях устанавливается коэффициент, равный 0,5. После этого рассчитывается коэффициент «спамерности» (коэффициент, по которому письмо относится к категории «спам» или «не спам») по формуле Байеса, но с подстановкой новых вероятностей нахождения «спама» в письме. Далее следует выборка показателей, наиболее «интересных» с точки зрения оценок. Уровень «интересности» определяется тем, насколько оценка «токена» отличается от нейтральной. Эвристическим параметром для данной модели статистической фильтрации писем будет количество «токенов», по которым оценивается то или иное письмо.

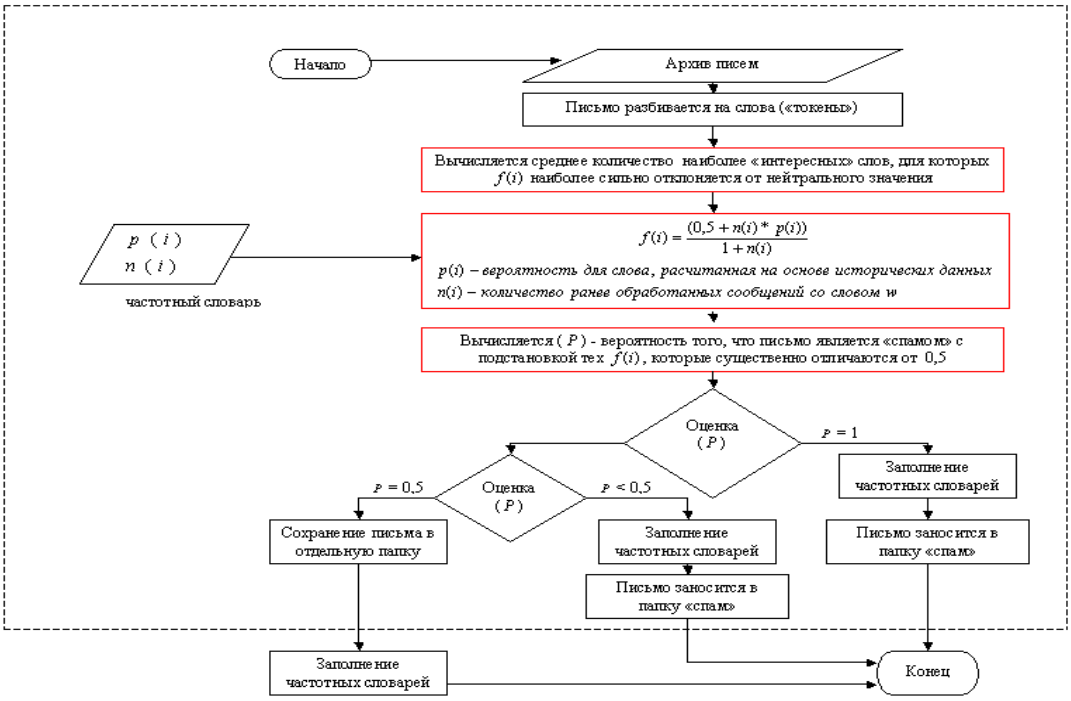

Рис. 2. Процесс фильтрации писем на основе метода градуированной фильтрации «спама»

Принцип работы классификатора «спама», построенного на основе метода градуированной фильтрации, основан на вычислении для каждого «токена» оценки

$$P = \frac{S}{S+G} > K \ , \qquad (1)$$

где P – вероятность того, что сообщение является «спамом», S – суммарный коэффициент «спамерности» сообщения, G – суммарный коэффициент «неспамерности» сообщения, K – заданный пользователем порог. Для вычисления вероятностных оценок p_i используется описанный выше процесс обучения, во время которого анализируются заранее классифицированные документы.

Для корректного соотношения «спама» и «не спама» и дальнейшей оценки категории письма (при градуировании) будем рассчитывать «спамерность» по формуле вычисления оценок вероятностей, учитывающей специальные случаи редких характеристик, что при нулевой частоте дает нейтральный результат, а при увеличении частоты асимптотичности приближается к реальной оценке. Тогда суммарный коэффициент «спамерности» можно представить в следующем виде:

$$S = \frac{0{,}5+n_1 \cdot p_1}{1+n_1} \cdot \frac{0{,}5+n_2 \cdot p_2}{1+n_2} \cdot \frac{0{,}5+n_3 \cdot p_3}{1+n_3} \cdot \ldots \cdot \frac{0{,}5+n_{i-1} \cdot p_{i-1}}{1+n_{i-1}} \cdot \frac{0{,}5+n_i \cdot p_i}{1+n_i},$$

а суммарный коэффициент «неспамерности» – в виде

$$G=(1-\frac{0{,}5+n_1\cdot p_1}{1+n_1})\cdot(1-\frac{0{,}5+n_2\cdot p_2}{1+n_2})\cdot(1-\frac{0{,}5+n_3\cdot p_3}{1+n_3})\cdot\ldots\cdot(1-\frac{0{,}5+n_{i-1}\cdot p_{i-1}}{1+n_{i-1}})\cdot(1-\frac{0{,}5+n_i\cdot p_i}{1+n_i}).$$

Здесь n – общее число «токенов» в письме с оценками $p_1 \ldots p_n$, n_i – количество ранее обработанных сообщений с «токеном» i, p_i – вероятность для «токена», рассчитанная на основе исторических данных.

Использование метода позволяет решить две проблемы: 1) оценки «токенов», впервые встретившихся в проверяемом письме и не существовавших до этого в базе; 2) повышения качества оценки данных (градуирования). Если анализируемый «токен» ранее не встречался, то, как уже упоминалось, оно автоматически получает коэффициент 0,5, а по мере накопления статистики это значение будет выходить на свой естественный уровень.

Метод был реализован в разработанном программном обеспечении «антиспам». В процессе испытания через фильтр были пропущены 480 писем, половина из которых являлась «спамом». В результате система не смогла распознать лишь 0,5% сообщений типа «спам», а количество ложных срабатываний фильтра оказалось равным 1,1%.

Заключение

«Антиспамерские» фильтры, основанные на статистической оценке, позволяют с достаточно большой вероятностью определять принадлежность письма к «спаму» на основе анализа его заголовка и текста с учетом сообщений, ранее полученных конкретным пользователем. При этом каждый владелец почтового ящика имеет возможность индивидуальной настройки характеристик распознавания в процессе обучения фильтра.

Литература

1. Distributed Checksum Clearinghouse [Электронный ресурс]. – Режим доступа: http://www.rhyolite.com/anti-spam/dcc, свободный.
2. Спам – Википедия [Электронный ресурс]. – Режим доступа: http://ru.wikipedia.org/wiki/Спам, свободный.
3. Визначення терміну СПАМ у законодавстві України. Журнал «Інформаційні технології. Аналітичні матеріали». [Электронный ресурс] – Режим доступа: http://it.ridne.net/uaspamdef, свободный.
4. RAZOR. [Электронный ресурс] – Режим доступа: http://razor.sourceforge.net, свободный.

Семенова Мария Александровна – Санкт-Петербургский государственный университет информационных технологий, механики и оптики, аспирант, semeonova-maria@rambler.ru

Семенов Вениамин Александрович – Главное управление Банка России по Санкт-Петербургу, кандидат технических наук, эксперт I категории, veny-semenov@yandex.ru