С.С Ахмедзянов
Научный руководитель – В.Г. Жуков
Сибирский государственный аэрокосмический
университет имени академика М. Ф. Решетнева,
Красноярск

Алгоритм фильтрации нежелательной электронной корреспонденции с использованием теоремы Байеса

Рассматривается применение системы обнаружения «спама», основанной на использовании вероятностной теоремы Байеса. Сильные и слабые стороны подобного фильтра. Перспективы развития.

«Спам» - массовая рассылка коммерческой, политической и иной рекламы или иного вида сообщений лицам, не выражавшим желания их получать. Опасность такой корреспонденции в том, что помимо потраченного времени на ее просмотр и удаление, в ней могут содержаться вредоносные программы различного характера. Так же массовая рассылка сообщений может использоваться для вывода из строя почтовой системы (DoS-атака).

Актуальность проблемы обнаружения спама сегодня ни у кого не вызывает сомнений. Достаточно лишь привести цифру, доля спама в почтовом трафике в феврале 2010 года в среднем составила 86,1% [1]. Исходя из этого разрабатываются способы обнаружения нежелательной электронной корреспонденции. Перечислим некоторые из них:

1. Черные списки - включают перечни IP-адресов отправителей спама;
2. Формальные правила – проверяют служебную информацию о письме (способ посылки электронного письма, протокол, время отправки, обратный IP-адрес отправителя). К типичным признакам нежелательного письма относятся отсутствие адреса отправителя, отсутствие или наличие слишком большого числа получателей, отсутствие IP-адреса;
3. Сигнатуры – для каждого нежелательного письма может быть автоматически создана сигнатура (образец оформления письма и его содержание) позволяющая распознать это письмо, иногда даже с небольшими модификациями;
4. Байесовские фильтры – позволяют с помощью статистических методов охарактеризовать письмо как «спам» или «не спам»;
5. Обучаемые системы – предназначены для обнаружения «спама» с использованием искусственного интеллекта и нейронных сетей.

Сигнатурный подход к обнаружению нежелательной электронной корреспонденции не обеспечивает необходимую эффективность, т.к. путем перестановки слов, словосочетаний и предложений в письме, его сигнатура изменится, и оно не будет помечено как «спам». Чтобы избегать подобных ситуаций, необходимо использовать системы статистического анализа содержимого писем. Примером такой системы является обнаружение нежелательной электронной корреспонденции с использованием теоремы Байеса. Теорема Байеса - является одной из основных теорем используемых в теории вероятностей, которая определяет вероятность наступления того или иного события, когда после проведенных наблюдений известна лишь некоторая частичная информация о событиях [2]. Формула Байеса: $P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$, где:

$P(A)$— априорная вероятность наступления события A;

$P(A|B)$— вероятность наступления события A при наступлении события B (апостериорная вероятность);

$P(B|A)$— вероятность наступления события B при наступлении события A;
$P(B)$— вероятность наступления события B [3].

При обучении системы обнаружения для каждого встреченного в письмах слова высчитывается и сохраняется его «вес» - вероятность того, что письмо с этим словом – «спам» (в простейшем случае - по классическому определению вероятности: «появлений в спаме / появлений всего»). При проверке вновь пришедшего письма вычисляется вероятность того, что оно – «спам», по указанной выше формуле для множества событий. В данном случае «события» - это слова, и для каждого слова «достоверность события» $P(A_i) = \frac{N_{слов_i}}{N_{слов_{всего}}}$ - процент этого слова в письме, а «зависимость одного события от другого» $P(B|A_i)$ - вычисленный ранее «вес» слова [4]. То есть «вес» письма в данном случае — не что иное, как усредненный «вес» всех его слов. Отнесение письма к нежелательной электронной корреспонденции производится по тому, превышает ли его «вес» некую границу, заданную пользователем. После принятия решения по письму в базе данных обновляются «веса» для вошедших в него слов.

Данный способ обнаружения «спама» прост в реализации и достаточно эффективен (после обучения на достаточно большой выборке исключает до 95-97 процентов «спама»).

Впрочем, у метода есть и принципиальный недостаток: он базируется на предположении, что одни слова чаще встречаются в нежелательной почте, а другие - в обычных письмах, и неэффективен, если данное предположение неверно [2]. Еще один, не принципиальный, недостаток, связанный с реализацией - метод работает только с текстом. Зная об этом ограничении, распространители спама используют графические изображения для оформления письма, текст же в письме либо отсутствует, либо не несет смысла. Против этого приходится пользоваться либо интеллектуальными средствами анализа и распознавания изображений, либо старыми методами фильтрации - «черные списки» и регулярные выражения (так как такие письма часто имеют стереотипную форму).

Перспективой развития данного способа обнаружения «спама» является его совместное использование с системами искусственного интеллекта анализа регулярных выражений. Такой симбиоз позволит анализировать словосочетания и предложения в электронных сообщениях, исходя из их контекста. Что позволит избежать ошибочного отнесения к «спаму» письма, не являющегося таковым.

Библиографические ссылки

1. [Электронный ресурс] Лаборатория Касперского www.securelist.com/ru
2. [Электронный ресурс] веб-сайта www.science.wikia.com
3. Чистяков В.П. Курс теории вероятностей / Чистяков В.П. М.: Наука, 1982. 112с.
4. [Электронный ресурс] веб-сайта www.computerra.ru