УДК 004.773.3

О.С. Волковский, Е.В. Выборов

ЭВРИСТИКО-ЛОГИЧЕСКИЙ МЕТОД ФИЛЬТРАЦИИ СПАМ-СООБЩЕНИЙ

Постановка проблемы

Электронная почта представляет собой современное и высокотехнологическое средство коммуникации. В настоящее время этот коммуникационный канал активно используется не только для обмена информацией, но и для продвижения товаров и услуг, в том числе и для рассылки спама.

Сейчас уже никого не надо убеждать в том, что спам – это негативное явление. Выпущены многочисленные спам-фильтры, позволяющие отфильтровывать до 85-95% нежелательной корреспонденции. Но данная проблема остается актуальной, так как даже при таком высоком проценте фильтрации большое количество спам-писем “пробивается” через фильтры и доходит до конечного пользователя, увеличивая время его работы с электронной почтой и принося тем самым большие финансовые убытки.

В данной статье описывается метод фильтрации почтовых спам-сообщений, основанный на статистическом анализе прямых и не прямых признаков спама, локализованный для русско(украино)-язычного спама и способный адаптироваться к тематике сообщений конкретного пользователя.

Обзор существующих методик.

В результате изучения существующей методологической базы фильтрации спама были выделены следующие основные методы:

1. Лингвистические методы, основанные на анализе содержания и автоматической классификации темы документа;

2. Лингвистические сигнатуры (анализ текстового содержания письма);

3. Формальные методы, основанные на анализе оформления технических заголовков писем и на наиболее надежных черных, белых и серых списках;

4. Обновления базы данных сигнатур и эвристик с базового сервера;

© О.С. ВОЛКОВСКИЙ, Е.В. ВЫБОРОВ, 2007

 ISSN 1562-9945

5. Обработка графических вложений в письма;

6. Вероятностные методы (метод Байеса, метод Фишера)

7. Статистические методы, фильтрующие спам на основе общих признаков.

Существуют также другие методы, которые не будут обсуждаться в данной статье.

Последнее время стали очень распространены вероятностные фильтры, основанные на теореме Байеса. Подобные фильтры отлично работают, но имеют ряд недостатков. В первую очередь это проблема переобучаемости. После определенного порога обучения их эффективность резко уменьшается (в среднем с 90% до 70-75%) и появляется большое количество ложных срабатываний (до 5-10%). Путей решения проблемы переобучения до сих пор не найдено. Кроме того, как показывает практика, для качественной работы вероятностных фильтров необходимо выделять дополнительные группы классификации для сообщений, таких как рассылки, а еще лучше рассылки определенной тематики. Иначе велика вероятность того, что такие сообщения будут отправляться в спам. Также такие фильтры не могут работать с графическим спамом, который можно отсеять только либо с использованием процессоров графического анализа, либо по непрямым признакам; и с различными трюками спамеров, такими как замаскированные слова, номера телефонов и т.д. Очень часто Байесовские фильтры называют «наивными» Байесовскими классификаторами. Слово «наивный» присутствует по той причине, что в используемой Теореме Байеса все признаки изначально принимаются статистически независимыми, что само по себе в спам-сообщениях не так.

Большинство недостатков фильтров, работающих на основе теоремы Байеса, отсутствуют в фильтрах, определяющих спам на основе общих его признаков. Но последние имеют свои минусы. Например, и в том, и в другом случае результатом оценки является, так называемый, "вес" письма. Однако при применении метода с использованием признаков спама "вес" письма вычисляется только на основе "плохих" признаков, что приводит к "обвинительному уклону" фильтра, и, как следствие, появляются ложные срабатывания.

Исходя из вышеописанного, существует необходимость создания фильтрующей системы, комбинирующей в себе преимущества вероятностных методов и методов, определяющих спам на основе его общих признаков.

Обзор системы

Целью исследований являлась разработка системы фильтрации спам сообщений, основанной на статистическом анализе прямых и не прямых признаков спама, локализованной для русско(украино)-язычного спама и способной адаптироваться к тематике сообщений конкретного пользователя.

Созданная система, как альтернативу применению правила Байеса, использует собственную метрику признаков. Каждый признак имеет 2 показателя вероятностей:

- вероятность обнаружить этот признак в спам сообщении;
- вероятность обнаружить признак в корректном сообщении.

Эти вероятности задаются по умолчанию, но могут изменяться администратором системы вручную, либо путем обучения фильтра на готовом наборе писем пользователя. Весовой коэффициент i-го признака вычисляется по следующей формуле:

$$P_i = \frac{P_{спам} - P_{не\_спам}}{P_{спам}} \quad (1)$$

По умолчанию в системе заданы 3 результирующие категории сообщений: спам, нормальное письмо, сомнительное письмо. Эти категории разделяются в зависимости оттого, в какой интервал попала результирующая вероятность проанализированного письма. Интервалы заданы по умолчанию, но могут и должны изменяться администратором системы для получения более корректных результатов фильтрации.

Результирующая вероятность определяется по следующей формуле:

$$P = \frac{\sum_{i=1}^{N} P_i}{N} \quad (2)$$

где N – количество проанализированных признаков.

Письмо проверяется, начиная с наиболее важных признаков. Если один из важных признаков найден и его вероятность превышает

 ISSN 1562-9945

порог спама, то проверка останавливается и сообщение относится к категории спама.

Система использует следующие основные признаки, позволяющие классифицировать сообщение:

1. Набор слов, наиболее часто встречающихся в спаме. Список этих слов задан по умолчанию, но может переформировываться из существующей базы электронных сообщений пользователя, что позволяет адаптировать фильтр к тематике корректных/спам сообщений пользователя. Слова автоматически разделяются на группы однокоренных. Каждое из слов имеет как характеристику 3 вероятности: вероятность появления слова с спам сообщении; вероятность появления слова в корректном сообщении; вероятность того, что это спам сообщение, если слово встретилось в замаскированной (преднамеренно искаженной) форме. Некоторым словам, таким как «рассылка», по умолчанию завышаются коэффициенты, так как их появление в сообщении с большой долей уверенности гарантирует, что это спам.

2. По умолчанию заданный набор словосочетаний, характеризующий данное письмо как спам. Этот набор так же может изменяться и добавляться администратором.

3. Наличие обратной связи. Это могут быть номера телефонов, адреса электронной почты, ссылки на веб-сайты, реже, почтовые адреса. Часто способы обратной связи маскируются, для того, что бы обойти вероятностные фильтры. Поэтому созданный фильтр ищет и распознает несколько уровней маскировки обратной связи (в частности телефонов), каждому из уровней присваивается своя вероятность того, что рассматриваемое сообщение - спам.

4. Визуальное выделение текста и фрагментов текста. Для обычных текстовых сообщений это выделение прописными буквами. Для HTML писем это выделение цветом, размером, стилем шрифта, цветом фона и т.д. Важным признаком является написание темы письма (поле Subject) прописными буквами.

5. Наличие замаскированного текста в HTML сообщениях, например, цвет которого задан цветом фона. Подобные трюки необходимы спамерам для обхода вероятностных фильтров, так как могут содержать в себе текст, характерный для корректных писем, либо делать не читаемыми для фильтров некоторые слова или даже

все письмо (например, невидимые символы ставятся вместо пробелов).

6. Отсутствие в списке получателей письма реального адреса получателя. Этот признак практически гарантирует, что сообщение является спамом.

7. Список получателей письма содержит адреса, начинающиеся с одинаковой комбинации букв. Этот признак обусловлен тем, что рассылка спам-сообщений часто производится по выборке из базы данных электронных адресов будущих получателей, которая упорядочена по алфавиту.

8. Другие, менее значительные, лингвистические и формальные признаки.

Нахождение всех лингвистических признаков в системе основано на регулярных выражениях (RegEx). Например, для нахождения замаскированных слов в сообщении используется генератор регулярных выражений, который строит их на основе так называемых алиасов (похожие по начертанию символы) знаков кириллицы. Алиасы могут добавлятся/изменятся администратором системы. Например, для слова «рассылка» будет сгенерированно регулярное выражение:

```
[pPpPrR](\w?|\W?|\s?)[aAaA@](\w?|\W?|\s?)[cCcC\\(](\w?|\W?|\s?)[cCcC\\(](\w?|\W?|\s?)[ыЫ](\w?|\W?|\s?)[лЛlL](\w?|\W?|\s?)[кКkK](\w?|\W?|\s?)[aAaA@](\w?|\W?|\s?)
```

В результате в наборе сообщений будут найдены следующие слова: р@ссылк@, рассылka, R@ссыka и т.д. При этом, если в тех же местах текста регулярное выражение для поиска НЕ замаскированных слов ничего не нашло, то к общей вероятности прибавляется весовой коэффициент замаскированного слова, что практически гарантирует отнесение его к спам-сообщениям.

Также регулярные выражения используются для поиска замаскированных телефонных номеров. Нахождение телефонного номера, подобного этому «+38 (044) 4-5.5-9.9-9.9» практически гарантирует отнесение сообщения к спаму.

Результаты экспериментов

Изначально, проводилось тестирование на 600 писем (500 – спам, 100 – корректные сообщения) без обучения и настройки. Все

 ISSN 1562-9945

значения использовались по умолчанию. Были получены следующие результаты:

Таблица 1

Результаты эксперимента без обучения

	Корректные сообщения	Сообщения зоны неопределенности	Спам-сообщения
Спам (500)	1,2%(6)	5,6%(28)	93,2%(464)
Не спам (100)	92%(92)	6%(6)	2%(2)

Затем система была обучена на оттестированных сообщениях (500 – спам, 100 – корректные сообщения) и через неё был пропущен новый набор сообщений (также 500 – спам, 100 – корректные сообщения). В результате эксперимента были получены следующие результаты:

Таблица 2

Результаты эксперимента после обучения

	Корректные сообщения	Сообщения зоны неопределенности	Спам-сообщения
Спам (500)	0,6%(3)	4,8%(24)	94,6%(473)
Не спам (100)	95%(95)	4%(4)	1%(1)

В первом и втором экспериментах был использован интервал классификации сообщения, заданный по умолчанию. Для устранения ложных срабатываний фильтра, эти значения были адаптированы под имеющийся набор сообщений: 0 – 0,4 – корректное сообщение; 0,41 – 0,75 – сомнительное сообщение; 0,76 – 1 – спам.

Кроме того, были повышены весовые коэффициенты для таких признаков, как слабо замаскированный номер телефона и заголовок письма прописными буквами. Изменения были обусловлены тем, что в корректных сообщениях из рассматриваемого набора такие признаки практически не встречаются, в отличие от спама.

В итоге, через систему был пропущен полный набор сообщений: 1000 спам сообщений и 200 корректных сообщений и в результате эксперимента были получены следующие результаты:

Таблица 3

Результаты эксперимента после обучения и настройки

	Корректные сообщения	Сообщения зоны неопределенности	Спам-сообщения
Спам (1000)	0,5%(5)	5%(50)	94,5%(473)
Не спам (200)	96,5%(193)	3%(6)	0,5%(1)

Выводы

Предложенный метод показал свою пригодность в задачах фильтрации спам-сообщений. Несомненным преимуществом метода является простота и гибкость реализации, дальнейшего администрирования и обслуживания, а также отсутствие необходимости привлечения сторонних лицензионных программных продуктов.

ЛИТЕРАТУРА

1. Paul Graham, Better Bayesian filtering, 2003.
2. Gary Robinson, A statistical approach to the spam problem, 2003.
3. А. Власова, К. Зоркий, Проблема намеренных искажений письменного текста в электронных рекламных рассылках (спаме). 2004.
4. И.С. Ашманов,А.Е. Власова, К.П. Зоркий, А.П. Иванов, А.Л. Калинин Технология фильтрации содержания для Интернет // Труды Международного семинара Диалог’2002 по компьютерной лингвистике и ее приложениям. Том 2. Москва 2002.

Получено 22.03.07

 ISSN 1562-9945