Корректное тестирование качества антиспам-продуктов

Алексей Тутубалин, lexa@lexa.ru
ЗАО «Ашманов и Партнеры»

Ашманов и партнеры

Постановка задачи

- Зачем тестировать:
  - оценка возможного эффекта от внедрения антиспама;
  - сравнение продуктов разных вендоров и выбор между ними;
  - подбор оптимальных настроек антиспам-системы.

- Требования к методике:
  - числовые метрики;
  - воспроизводимость;
  - возможность самостоятельного получения результатов;

Определения

- **СПАМ -** анонимные незапрошенные массовые рассылки электронной почты (как правило, имеющие рекламный характер).
- **Нежелательная почта** – более широкий класс сообщений, кроме спама это:
  - подписные рассылки от которых забыли отписаться;
  - квитанции от почтовых систем и другие технические сообщения;
  - ошибочные письма;
  - вирусы (для них есть антивирусы);
  - и так далее....

Нежелательную почту часто тоже называют спамом.

Потери от спама и антиспама

- Потери от спама:
  - лишний сетевой трафик;
  - потери времени и концентрации сотрудников при разборе спама;
  - потери нормальной почты «в груде спама».

- Потери от антиспам-фильтров:
  - ошибочная классификация сообщений как спама (такие письма не будут прочитаны);
  - затраты на закупку и внедрение;
  - затраты на поддержку (включая трафик).

- Цель фильтрации спама – минимизация общих потерь организации.

Критерии оценки качества

- Нужно **одновременно** использовать два критерия:
  - **Доля ложных тревог** (false positive, FP) – нормальных сообщений отклассифицированых как спам.
  - **Доля пропусков спама** (false negative).
- Идеального решения (0% ложных тревог и 100% определения спама) на сегодня не существует, даже отсутствие фильтрации приводит к потерям почты из-за ошибок пользователей.
- Для деловой переписки, уровень ложных тревог является определяющим критерием.
- Разумный уровень ложных тревог: ниже чем количество ошибок у человека (т.е. ниже 0.1-0.01%).

Ашманов и партнеры

Ложные тревоги

- Отношение числа ложно отклассифицированных как спам сообщений к общему числу не-спам сообщений (ошибка: считать долю от всей почты, включая спам).

- Уровень серьезности ложных тревог:

  - **Критические**: ложная классификация важной личной и деловой почты.

  - **Некритические**: ложная классификация массовых новостных рассылок, открыток и подобной «неважной» почты.

Требования к методике тестирования

1. Тестировать нужно на реальном потоке почты, а не тестирование на архивах почты (интересует скорость реакции антиспам-систем на реальные спам-рассылки)

2. Достаточная протяженность тестирования по времени, 2-3 недели и более (для сглаживания колебаний спам-трафика).

3. Достаточный объем трафика: десятки тысяч сообщений и более (для оценки уровня ложных срабатываний).

4. Достаточное количество почтовых ящиков, принимающих участие в тестах: десятки и более (спам у всех разный).

Ашманов и партнеры

Требования к методике тестирования (продолжение)

5. Нельзя пересылать (forward) почту (потеря части данных исходного сообщения).

6. Нельзя отвергать (reject) почту (невозможно проверить ложные срабатывания).

7. Нужен ручной анализ результатов (т.к. нет определения «спама», пригодного для автоматического анализа).

8. Равные условия для сравниваемых продуктов.

9. Единое определение спама и критичности ложных срабатываний для сравниваемых продуктов.

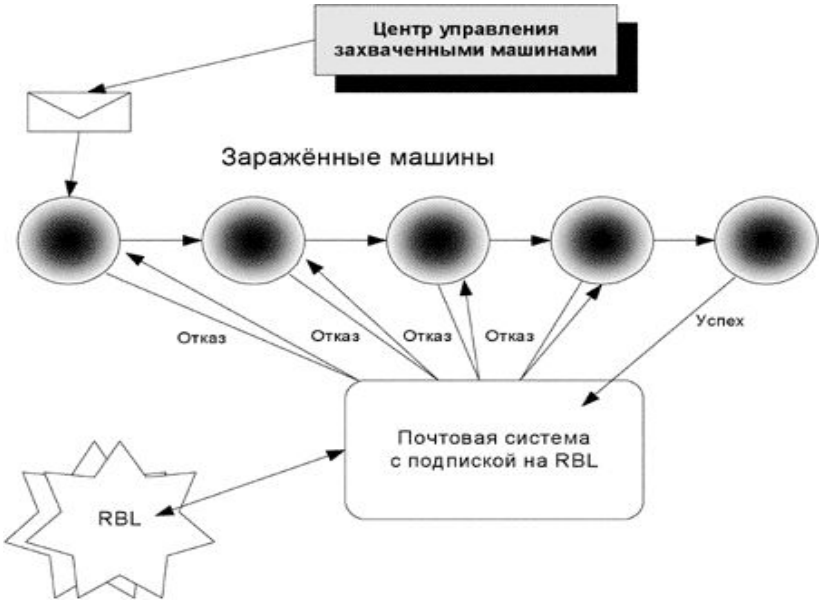
Ашманов и партнеры

«Черный ящик» или понимание принципов работы фильтра ?

- Принцип «черного ящика» - не требует знаний о методах работы фильтра.
  Можно притвориться «наивным пользователем»
- Недостатки подхода:
  - Все антиспам-системы нуждаются в настройке.
  - У каждого метода (фильтра, решения) свои особенности, которые нужно учитывать.
  - При тестировании нескольких систем одновременно, нужно дать всем им необходимые для работы данные;
  - возможны наведенные ошибки, влияющие на результаты.

Особенности: RBL-системы

- Одно спам-письмо отправляется многократно и многократно отвергается, потом все-таки доставлено.
- Результат: показатели качества (отношение числа reject-ов к количеству доставленных писем) завышены, хотя речь про единственное письмо.
- При отвержении (reject) писем нельзя оценить уровень ложных срабатываний.
- RBL – системы реального времени, нельзя тестировать на архивах.
- RBL-системы нельзя тестировать с пересылкой почты.



Ашманов и партнеры

Особенности: фильтры с обновлениями

- Нельзя тестировать на архивах:
  - образец (сигнатура) письма может быть уже удален из БД;
  - образец мог отсутствовать на момент прихода письма, но имеется в БД на момент тестирования;
  - следовательно, тестируя не в реальном времени мы получим искаженные представления о качестве.
- Обновления должны быть включены и доходить до антиспама

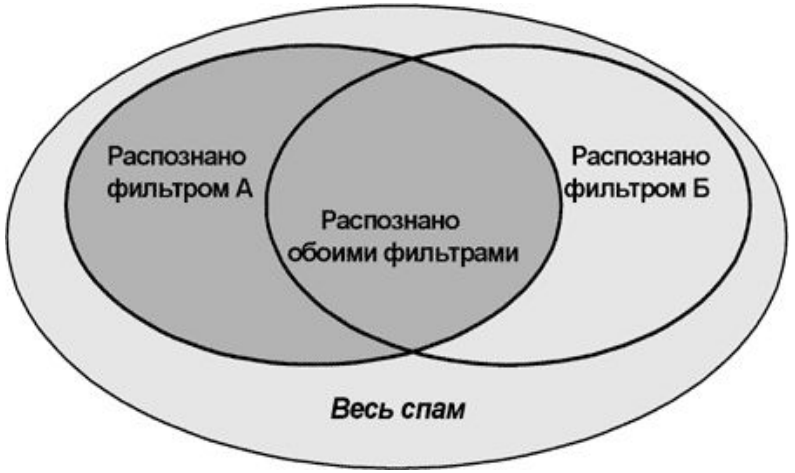
Особенности: статистические фильтры

- Статистическим фильтрам свойственно «переобучение», следовательно тестирование должно быть длительным (недели-месяцы)
- Нужно переобучение в режиме реальной эксплуатации
- Наведенные эффекты: статистические фильтры могут обучаться по меткам других антиспам-систем, нужно учитывать при одновременном тестировании
- Тестирование на архивах может завысить качество за счет ошибок при составлении обучающей выборки.

Ашманов и партнеры

Ошибки: последовательное соединение фильтров

- Фильтр Б проверяет поток спама после фильтра А. Если пропускает, то «распознал не все»
- Необходимость «зеркальной» схемы (А после Б)
- Таблица результатов:
  - распознано/пропущено А, Б, обоими фильтрами;
  - ложные тревоги А, Б, у обоих.



Ашманов и партнеры

Ошибки: пересылка спама

- При пересылке теряются данные почтовой сессии:
  - IP-адрес посылающей стороны
  - Параметры SMTP-сессии (HELO, MAIL FROM)
  - Заголовки письма (добавляется лишний Received, метки антиспам-систем)

- Если тестируемая система учитывает эти данные, то результаты тестирования будут ошибочными

Ошибки: тестирование на коллекциях

- Трудность в сборе коллекций:
  - Вручную много не набрать, у разных пользователей разные критерии
  - Отбор программой аналогичен последовательному соединению фильтров
  - Поток спама на ловушках может отличаться от среднего
  - Трудно набрать хорошую базу легитимной почты
- Данные антиспам-программ быстро меняются (RBL, базы обновлений), архивы могут быть устаревшими. На коллекциях невозможно оценить скорость реакции.
- Невозможно воспроизвести окружение при приеме письма

Ошибки: статистические системы

- Проблемы с обучающей выборкой:
  - Обучение и тестирование на одной коллекции (даст замечательные результаты).
  - Деление архива пополам, учим на одной половине, тестируем на другой: дубли сообщений попадут в обе выборки, качество будет завышенным.
  - Наличие в сообщениях меток от других антиспам-программ.

Пример тестирования

- Тестовая площадка: @lexa.ru:
  - персональный домен с 9-летней историей;
  - «засвеченный» адрес lexa@lexa.ru, кроме него ~30 почтовых ящиков, включая ловушки;
  - 2500-2700 спам-сообщений в сутки (рост в 3-4 раза за год);
  - разнообразная почта: разные языки, разные темы, много разовых корреспондентов, 150-600 сообщений в сутки (включая рассылки);
  - много спама, пересылаемого в качестве внутрифирменной переписки;
  - установлено 3 антиспам-системы (одна из них – в трех вариантах настроек одновременно),
  - тестирование ведется постоянно уже более двух лет.

Анализ пропусков и ложных срабатываний

- Пропуски и ложные срабатывания:
  - Вся почта, распознанная как спам (т.е. не попавшая в Inbox), пересылается в лингвистическую лабораторию «Лаборатории Касперского», где просматривается вручную.
  - Весь попавший в Inbox спам просматривается вручную и архивируется.
  - Таким образом, все результаты работы программных фильтров просматриваются человеком.

- Статистика может быть получена за все время хранения архива.

Особенности настройки

- Система с использованием RBL и анализом данных SMTP-сесии: установлена первой в цепочке фильтров.
- Система с использованием статистики: явно запрещено использовать метки от других фильтров.
- Собственные черные и белые списки: не используются, но при анализе ложных срабатываний учитывается, что письмо от постоянного корреспондента могло бы быть пропущено белым списком.

Пример результатов

S1,S2,S3 – три антиспам-системы (S2 – в трех конфигурациях), настроены так, чтобы не зависеть друг от друга.

	S1	S2a	S2b	S2c	S3
обнаруже-ние спама	91%	59%	63%	65%	98.6%
Ложные тревоги	<0.005%*	~0.01%	~0.01%	~0.1%**	4.3%***

* без учета срабатывания на пересылаемых образцах спама;

** - основные ложные срабатывания на сообщениях от «роботов» (системы регистрации, системы заказа товара и т.п.), небольшая часть – на рассылках (втч. рабочих);

*** - из всех FP: – половина на письмах от роботов, 10% - подписные рассылки, остальное – важная почта.

На примере S2 видно, что уровень обнаружения и ложные срабатывания – противоречивые требования

Аштанов и партнеры

Спасибо за внимание

Пожалуйста, задавайте вопросы.

Текст статьи доступен на WWW:
Spamtest.RU -> Публикации -> Аналитика