УДК 81' 332.2

О ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ СПАМ-ФИЛЬТРА ДЛЯ АНАЛИЗА ТЕКСТА

А.А. Солдатова
Военная Академия Воздушно-космической Обороны, г. Тверь

Обсуждается междисциплинарный метод исследования (на стыке лингвистики и математики), определяющий использование числовых данных, получаемых в результате применения спам-фильтра. Внимание фокусируется на возможных способах применения методов статистического исследования.

Ключевые слова: *спам-фильтр, статистический анализ, научный поиск, метод классификации, распознавание текста*.

Важнейшим показателем качества исследования, его научной ценности является точность, которая зависит от применяемых исследователем методов. Так, математические и статистические методы повышают надежность выводов и дают основания для теоретических обобщений. Подчас такие методы имеют весьма специфический характер, однако позволяют дать точные, количественно определённые сведения об исследуемом феномене.

Представляется весьма интересной возможность применения в исследовании статистического спам-фильтра DSPAM, написанного Джонатаном А. Здзиарски [2], автором книги Ending Spam. Результаты, обработанные с помощью данного метода, позволяют показать количественную зависимость в виде диаграмм, графиков или таблиц.

Суть данного метода проста. Любой пользователь электронной почты периодически сталкивается с необходимостью отсеивания spam-сообщений (от англ. *spam* – «мусорное сообщение») путём их удаления или перемещения в специальную предназначенную для таких сообщений папку. Избегать получения нежелательных сообщений помогает масштабируемый спам-фильтр, точность результата которого после его обучения составляет 99.5% – 99.95%. Данное программное обеспечение, в целом, предназначено для крупных многопользовательских систем, однако возможно и более узкое его применение, когда необходимо проанализировать, например, частоту использования тех или иных лексических единиц и их суперпозиций (положения лексических единиц относительно друг друга). При этом действует предположение, что одни слова чаще встречаются в «спам-сообщениях», а другие – в «обычных письмах». Следует заметить, что несомненным достоинством DSPAM является то, что это регулируемый и обучаемый фильтр, который можно адаптировать под нужды пользователя / исследователя.

В основе работы системы лежат 4 алгоритма статистического анализа, которые, в свою очередь, опираются на теорему английского математика Байеса (Thomas Bayes). Формула Байеса даёт возможность оценить вероятность некоторого события эмпирическим путём, отталкиваясь от того, какова была вероятность данного события в прошлом. Применительно к исследуемому тексту принцип работы классификатора Байеса можно описать формулой:

$$P= S / (S + G),$$

где P – вероятность того, что сообщение является спамом,
S – суммарный коэффициент «спамности» сообщения,G – суммарный коэффициент «неспамности» сообщения.
S и G рассчитываются по следующим формулам:

$$S = p(w1) \cdot p(w2) \cdot \ldots \cdot p(wn)$$
$$G = (1 – p(w1)) \cdot (1 – p(w2)) \cdot \ldots \cdot (1 – p(wn))$$

Весьма наглядный пример работы спам-фильтра приведён С. Супруновым в статье «Настраиваем DSPAM – ваш личный спам-фильтр» [1], где анализируется краткое сообщение: «*Привет! Купи меня!*»

Автор допускает, что ранее указанные слова встречались в следующих письмах:

Слово	Спам	Не спам
Привет	35	64
Купи	187	19
меня	9	11

p("Привет") = 35 / (35 + 64) = 0,35
p("Купи") = 187 / (187 + 19) = 0,91
p("меня") = 9 / (9 + 11) = 0,45
S= 0,35 · 0,91 · 0,45 = 0,14
G= (1 – 0,35) · (1 – 0,91) · (1 – 0,45) = 0,03
P= 0,14 / (0,14 + 0,03) = 0,82

Таким образом, фраза «*Привет! Купи меня!*» за счёт высокой «спамности» слова «купи» будет с вероятностью 82% нежелательным сообщением.

Мы не будем описывать остальные алгоритмы, на основании которых работает данная программа, а лучше опишем возможные способы применения методов статистического исследования на основе теоремы Байеса и использования доступных программ.

Метод 1. Данный метод удобен, когда есть возможность создать на почтовом сервере такое количество новых антиспам-профайлов, которое соответствует числу исследуемых классов текста. Технически антиспам-профайл представляет собой отдельную запись в базе данных, которая служит для накопления частоты вероятностей принадлежности к конкретному классу токенов. Токен – это слово, или группа слов в определённой последовательности или без учёта последовательности их расположения (в зависимости от используемого метода, основанного на теореме Байеса). Создаётся необходимое количество профайлов и производится заполнение статистик (обучение классификатора). Необходимо создать несколько почтовых ящиков на любом известном почтовом сервере (например, почтовый сервер ТвГУ или gmail.com), осуществить отправку множества писем с текстом, соответствующим и несоответствующим классу, закреплённому за данным ящиком. Например, можно создать два ящика: «Доброта» и «Злость». В ящик «Доброта» необходимо отправить 100 «добрых» писем и 100 «злых». В ящик «Злость» следует отправить 100 «злых» писем и 100 «добрых». Затем следует зайти в ящик «Доброта» и пометить все «добрые» письма как «спам», т.е. это те письма, которые принадлежат к классу Доброта. После этого зайти к ящик «Злость» и обозначить соответствующие «Злости» письма как «спам». Через несколько минут программа на сервере увидит произведенные отметки и заполнит базу данных статистик. После этого, если необходимо определить вероятность принадлежности какого-то текста к классу Доброты или Злости, достаточно просто отправить письмо в соответствующий ящик и увидеть, помечено ли оно как «спам». Если письмо окажется помеченным как «спам», значит оно принадлежит соответствующему классу. В зависимости от настроек почтового сервера в исходном тексте письма будет вставлено численное значение вероятности принадлежности текста письма данному классу.

Метод 2. Программа DSPAM распространяется с открытыми исходными текстами. Любой специалист с хорошей квалификацией в области программирования способен взять исходный текст этой программы и скомпилировать его для своего компьютера. При этом возможно собрать программу так, что она будет представлять собой программную библиотеку, которую можно будет с использованием программирования связать с оконными интерфейсами ОС Windows или Linux. В этом случае это будет программа с графическим интерфейсом, которая запускается с рабочего стола и с которой можно работать в режиме off-line. В этой программе можно предусмотреть опции выбора метода классификации и пороговое значение начального набора статистик (количество токенов, которое необходимо набрать, чтобы добротно классифицировать текст). Также можно было бы предусмотреть возможность создания новых классов одним щелчком мыши.

Метод 3. Метод опорных векторов. В этом случае используется SVM (Support Vector Machine). Support Vector Machine – классификатор текстов, написанных на языке C в техническом университете Дормондта. Он распространяется бесплатно в виде исходных текстов. Использует метод классификации, основанный на представлении документа в виде вектора. Сначала при помощи обучающего модуля создаётся словарь-коллекция. После этого любой новый документ может быть классифицирован при помощи программы-классификатора. Специализированные алгоритмы настройки SVM успешно справляются с выборками из десятков тысяч объектов.

Метод 4. При помощи *Google Prediction API*. Это экспериментальная разработка *Google*, представляющая собой программу для вероятностной классификации текста. Программа не может быть использована без создания программного интерфейса, использующего предоставленный компанией *Google API* для доступа к функциям обучения и запроса о классификации. Программа не может быть использована в режиме *off-line*. Однако можно обратиться к специалистам для написания программы, которая использовала бы функции, предоставляемые *Google*. Для этого необходимо иметь профайл в *Google* и быть зарегистрированным на сайте проектов *Google API* для разработчиков. На сайте проектов необходимо активировать опцию использования *Google Prediction API* и *Google Cloud Storage API*. В таблице приведена краткая характеристика методов классификации текста.

Таблица. Краткая характеристика доступных методов классификации текста

Метод	Уровень сложности (* / ** / ***)	On-line/off-line
1.	* (требуется обучение антиспам-фильтра)	On-line
2.	*** (требуется обращение к программисту для создания графической оболочки библиотеки DSPAM)	Off-line
3.	* (требуется скачать программу SVM и обучить её)	Off-line
4.	** (требуется обращение к программисту для создания оболочки, использующей Google API)	On-line

Таким образом, в задачу классификации текста входит обучение классификатора, а также сама классификация текста. Во время обучения классификатора используется обучающая выборка текстов, а в задачу обучающего механизма входит выявление общих элементов выделенных

классов. Исходя из обобщения строятся описатели классов, используемые в ходе работы классификатора.

Описанные методы могут быть использованы не только для тематической группировки входящих почтовых сообщений, в частности для отделения информативных сообщений от сообщений рекламного характера, но и для анализа любых других текстовых массивов.

Список литературы

1. Супрунов С. Настраиваем DSPAM – ваш личный спам-фильтр [Электронный ресурс] // Системный администратор. 2005. № 8. URL: http://samag.ru/archive/article/526 (дата обращения: 02.10.2014).
2. Zdziarski A. Jonathan. Ending Spam // No Starch Press. 2005. 312 p.

ON POSSIBLE APPLICATION OF SPAM-FILTERING FUNCTION TO TEXT ANALYSIS

A.A. Soldatova
Military Academy of Airospace Defence, Tver

The article focuses on a cross-disciplinary perspective aiming to improve the accuracy of the research data with the help of a baseline statistical technique. Spam-filtering helps to organize any text according to specific criteria.
***Keywords**: spam-filtering, statistical analysis, scientific search, classification method, text recognition.*

Об авторе:

СОЛДАТОВА Анастасия Анатольевна – кандидат филологических наук, преподаватель кафедры английского языка Военной Академии Воздушно-космической Обороны, Тверь, e-mail: sotto.voice@gmail.com