Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.021

Алексеев Юрий Геннадьевич

ИНТЕЛЛЕКТУАЛЬНЫЙ ФИЛЬТР ЭЛЕКТРОННЫХ СООБЩЕНИЙ

АВТОРЕФЕРАТ
диссертации на соискание степени магистра технических наук

по специальности 1-40 80 02
«Системный анализ, управление и обработка информации»

_____
(подпись магистранта)

Научный руководитель

Герман Олег Витольдович
(фамилия, имя, отчество)
Кандидат техн. наук, доцент
(ученая степень, ученое звание)

_____
(подпись научного руководителя)

Минск 2017

ВВЕДЕНИЕ

При работе систем в той или иной степени связанных с обработкой естественного языка в компьютерно-опосредованной среде, возникает ряд вопросов по контролю потоков текстовой информации, в том числе передаваемых средствами электронной почты. Одним из важнейших является вопрос определения категории (семантики) поступающих сообщений на предмет их классификации, в частности, для выявления спама. Использование интеллектуальных фильтров является одним из возможных подходов к решению подобного рода задач, поскольку они позволяют не только более надежно отбраковывать нежелательную корреспонденцию, но и выявлять важные сообщения с целью их последующей сортировки. Фильтр считается интеллектуальным, если его работа связана в той или иной мере со смыслом обрабатываемых сообщений. В крайнем случае, спам рассматривается как бессодержательная информация для пользователя. В этом отношении все, что не является спамом, имеет для конечного пользователя смысл.

Электронные сообщения рекламного характера, отправляемые автоматически большому количеству адресатов, занимают около 75% от общего почтового трафика. Программные средства фильтрации спама, устанавливаемые на почтовых серверах или на компьютерах получателей, решают задачу классификации (категоризации) текстов. При фильтрации спама наиболее широкое распространение получили байесовские классификаторы. Эти классификаторы используют байесовскую формулу для оценки апостериорной вероятности принадлежности сообщения определенному классу документов на основании имеющейся статистики почтовых отправлений и частот слов, которые в них использовались. В магистерской диссертации реализован подход к решению проблемы фильтрации почтовой корреспонденции, использующий теорему Байеса.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Тема диссертации является актуальной в связи с постоянной потребностью в использовании средств электронной рассылки сообщений и, как следствие, поиском механизмов их обработки и сортировки.

Цель работы

Найти и реализовать эффективные способы классификации входящей электронной корреспонденции по категориям смысловой нагрузки.

Задача исследования

Научиться распознавать структурно-функциональные характеристики текстов по принципу нахождения ключевых маркеров смысловой направленности.

Объект исследования

Текстовые единицы различных уровней смысловой иерархии и сами тексты электронной корреспонденции.

Предмет исследования

Структурные и функциональные характеристики стиля электронного письма. Сходства и различия электронных писем разной смысловой направленности.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Для осуществления классификации нужно создать модель, которая построена данных статистики. Классификация письма будет строиться следующим образом: выбираем класс, значение которого максимально, исходя из выражения, посчитанного для всех классов по следующей формуле:

$$\log\frac{D_c}{D}+\sum_{i\in Q}\log\frac{W_{ic}+1}{|V|+L_c},$$ где

Dc – число текстов из обучающей выборки относящихся к классу *c*;

D – число всех текстов в выборке;

|*V*| – число слов, являющихся уникальными со всех текстов обучающей выборки;

Lc – общая сумма слов в текстах выборки относящихся к классу *c*;

Wic – показатель, с которым *i*-ое слово попадается в текстах класса *c*;

Q – множество слов (с учётом повторений) исследуемого текста.

Пример:

Используем несколько текстов, классы которых заведомо известны (NSP означает не спам, SP – спам):

[SP]: бесплатная юридическая консультация;

[SP]: торопитесь купить лотерею;

[NSP]: нужно купить молоко.

Модель классификатора выглядит так:

	NSP	SP
частоты классов	1	2
общая сумма слов	3	6

Таблица 1. Перечень классификаций

	NSP	SP
бесплатная	0	1
юридическая	0	1
консультация	0	1
торопитесь	0	1
купить	1	1
лотерею	0	1
нужно	1	0
молоко	1	0

Таблица 2. Классификация

Запустим классификацию предложения «надо купить книгу». И произведём расчеты для класса SP:

$$\log\frac{2}{3}+\log\frac{1}{8+6}+\log\frac{2}{8+6}+\log\frac{1}{8+6}\approx -7,629$$

Для класса NSP:

$$\log\frac{1}{3}+\log\frac{2}{8+3}+\log\frac{2}{8+3}+\log\frac{1}{8+3}\approx -6,906$$

В этом примере класс NSP оказался превалирующим, что говорит о том, что сообщение не содержит спам.

В простейшем случае выбираем класс, который получил максимальную оценку. Но если, например, надо помечать сообщение как спам только если соответствующая вероятность больше 80%, то сравнение логарифмических оценок ничего не даст. Оценки, которые выдает алгоритм, не удовлетворяют двум формальным свойствам, которым должны удовлетворять все вероятностные оценки:

– они все должны быть в диапазоне от нуля до единицы;

– их сумма должна быть равна единице.

Для того чтобы решить эту задачу, необходимо из логарифмических оценок сформировать вероятностное пространство. А именно: избавиться от логарифмов и нормировать сумму по единице.

$$P(c \mid d) = \frac{e^{q_c}}{\sum_{c' \in C} e^{q_{c'}}}$$

Здесь q_c – это логарифмическая оценка алгоритма для класса c, а возведение её (основание натурального логарифма) в степень оценки используется для того чтобы избавиться от логарифма ($a^{\log_a x} = x$). Таким образом, если в расчётах использован не натуральный логарифм, а десятичный, необходимо использовать не число e, а 10.

Для вышеприведенного примера вероятность того, что сообщение суть спам, равно:

$$\frac{e^{-7,629}}{e^{-7,629} + e^{-6,906}} = 0,327$$

то есть сообщение является спамом с вероятностью 32.7%.

Теперь покажем, как перейти от вероятностной записи формулы Байеса к нечеткому варианту. Перепишем формулу Байеса следующим образом:

Проводя аналогию между вероятностями и нечеткими мерами, получим:

$$P(c) \Leftrightarrow \mu(c), \quad P(a \square b) \Leftrightarrow \min(\mu(a), \mu(b)),$$
$$\mu(A \mid B) = \begin{cases} \dfrac{\mu(A)}{\mu(B)} & \text{ï ðè } \mu(A) \le \mu(B), \\ 1 \text{ ï ðè } & \mu(A) > \mu(B). \end{cases}$$

(1.1)

С учетом (1.1) формула Байеса в нечетком варианте перепишется так

$$\mu(A \mid B) = \frac{\mu(AB)}{\mu(B)}.$$

(1.2)

Формула Байеса в определении (1.1, 1.2) выгодно отличается тем, что нечеткие меры соответствуют экспертным оценкам, так что, в принципе, можно не хранить базу данных ранее полученных почтовых сообщений, а использовать экспертные оценки и проводить «дообучение» модуля фильтрации. Прологарифмировав (1.2), получаем:

$$Ln\mu(A \mid B) + Ln\mu(B) = Ln\mu(A \cap B),$$
$$Ln\mu(A \mid B) = Ln\mu(A \cap B) - Ln\mu(B). \qquad (1.3)$$

Теперь можно поставить вопрос об обучении системы фильтрации спама. Будем искать линейную распознающую функцию в форме

$$Y = \alpha_1 x_1 + \alpha_2 x_2,$$
$$\text{ãäå}$$
$$Y \Leftrightarrow Ln\mu(A \mid B); \quad x_1 = Ln\mu(A \cap B); \quad x_2 = Ln\mu(B). \qquad (1.4)$$

В (1.4) символ $\Leftrightarrow$ означает соответствие (но не буквальное равенство)

Задача обучения – правильно настроить коэффициенты α_1, α_2. До обучения будем считать, что $\alpha_1 = \alpha_2 = 1$. Как известно, обучение выполняется по обучающей выборке. Разумеется, для этой цели нужна база данных с ранее полученными сообщениями. Однако когда обучение завершится, то необходимость в базе данных отпадет и ее можно не хранить на сервере. Алгоритм обучения кратко сводится к следующему: для каждого письма имеем значение Y, определяемое учителем из простой классификации: спам- не спам. Если $Y > 0$, то имеем спам, если Y меньше 0, то получили не спам. Вычисляем значения x_1 и x_2 и далее Y из (1.4). Для вычисления x_1 и x_2 используем частотный аналог нечеткого значения, т.е

$$P(c) \Leftrightarrow \mu(c).$$

Если на рассматриваемом почтовом сообщении с характеристиками

$$x_1 = Ln\mu(A \cap B); \quad x_2 = Ln\mu(B).$$

значение *Y* посчитано правильно, то никакой коррекции не выполняем. В противном случае корректируем коэффициенты согласно следующему:

– Если письмо распознано как спам, но спамом не является, то полагаем

$$\alpha_1 = \alpha_1 - x_1; \quad \alpha_2 = \alpha_2 - x_2.$$

– Если письмо не распознано как спам, но является спамом, то полагаем

$$\alpha_1 = \alpha_1 + x_1; \quad \alpha_2 = \alpha_2 + x_2.$$

Описанная техника является стандартной техникой обучения нейроэлемента. В качестве входных сигналов такого нейроэлемента взяты логарифмы от соответствующих нечетких мер.

Таким образом, техника Байеса нами улучшена в плане дообучения системы фильтрации с возможностью «выбросить» и более не пополнять базу сообщений, а также существенно ускорить алгоритм распознавания спама.

ЗАКЛЮЧЕНИЕ

В процессе работы над диссертацией были изучены основные особенности проектирования и разработки программного средства фильтрации сообщений, проведён анализ предметной области, изучены основные подходы к проектированию и разработке, выполнено проектирование системы, выбраны оптимальные средства разработки и разработана программа, реализующая фильтр электронных сообщений. Таким образом, цель работы достигнута.

Функционал разработанной системы позволяет фильтровать сообщения от спама и постоянно пополнять базу данных.

Пользовательский интерфейс разработан так, что доступ к функциональным возможностям приложения осуществляется через графическое окно и интуитивно понятен любому пользователю.

В ходе работы над диссертацией все поставленные задачи по проектированию и разработке выполнены в полном объёме.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Эндрю Джелман, Джон Б. Карлин, Халь С. Штерн, Дональд Б. Рубин, «Байесовский анализ данных», Второе издание. – 2012. – С. 50-58.

2. Питер Ли, «Байесова статистика: введение», Вайли. – 2012. – С. 281-297.

3. Смирнов И. В., Шелманов А. О. Семантико-синтаксический анализ естественных языков. Часть I. Обзор методов синтаксического и семантического анализа текстов // Искусственный интеллект и принятие решений. – 2012. – С. 41-74.

СПИСОК СОБСТВЕННЫХ ПУБЛИКАЦИЙ

1. Алексеев, Ю.Г., «Интеллектуальный фильтр электронных сообщений» / Международный научный журнал «Интернаука», Киев, Выпуск №1 Январь 2017 г.