УДК 004.021

Алексеев Юрий Геннадьевич
студент
Белорусский Государственный университет информатики и радиоэлектроники
Alekseyev Y. G.
student
Belarusian State University of Informatics and Radioelectronics

ИНТЕЛЛЕКТУАЛЬНЫЙ ФИЛЬТР ЭЛЕКТРОННЫХ СООБЩЕНИЙ

INTELLIGENT FILTER FOR THE ELECTRONIC MESSAGES

Аннотация. *В данной работе предложен алгоритм фильтрации сообщений электронной корреспонденции. Решение основано на использовании теоремы Байеса. Анализ содержимого задействует базу данных лингвистической лаборатории, которая регулярно обновляется. По результатам анализа сообщение удаляется, либо помечается специальным образом (если оно содержит спам) либо же остаётся на сервере и проходит дальше.*

Ключевые слова: *спам-фильтр, условная вероятность, теорема Байеса.*

Abstract. *This work presents an algorithm for filtering e-mail. In solving the problem applies linguistic signatures and text analysis. Content analysis involves base of linguistic laboratory, which updates regularly. After the message is analyzed, there is a choice of what to do with it: if it contains spam then it gets a special mark or just removes. If it doesn't then the message passes on.*

Key words: *spam filtering, the conditional probability, Bayes' theorem.*

Введение

Задача, которую предстоит решить, заключается в анализе текста сообщения на предмет обнаружения в нём информации относящейся к нежелательной рекламе, т.е. спаму. Обработка текстов сообщений будет строиться на сортировке и подсчёте слов.

- В случае появления несуществующего слова будем учитывать аппроксимацию вероятности P(слово | класс).
- В анализе будет задействована таблица, которая содержит все слова использующиеся для фильтрации письма. К каждому слову будет привязано три числа обозначающие количество вхождений слова в письма не являющиеся спамом, второе — в письма являющиеся спамом, третье — идентификатор слова, данный самой программой.

Работа основана на применении теоремы Байеса, используя которую и строится алгоритм фильтрации.

Чтобы реализовать систему байесовской фильтрации необходимо получить выборку, в которой будут представлены соответствия текстовых фрагментов классам. Далее, из полученной выборки надо извлечь:

– частоту повторения текстов того или иного класса,
– общее количество слов, найденное в каждом наборе рассматриваемых текстов определённого класса,
– частоту повторения слов внутри отдельного класса,
– общий объём слов из выборки и количество уникальных слов.

Описание

Для осуществления классификации нужно создать модель, которая построена данных статистики. Классифицировать письмо будем 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 D_c — 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;

L_c — общая сумма слов в текстах выборки относящихся к классу c;

W_{ic} — 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:

Таблица 1

Перечень классификаций

	NSP	SP
частоты классов	1	2
общая сумма слов	3	6

Таблица 2

Классификация

	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

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.

Для решения задачи, когда оценки выдаваемые алгоритмом, не будут удовлетворять следующим свойствам: а именно находиться в диапазоне от 0 до 1 (при этом их сумму должна быть раной 1), формируется вероятностное пространство, т.е. мы отбрасываем логарифмы и нормируем сумму по единице.

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

где q_c — логарифмическое значение оценки для класса c.

Для избавления от логарифма воспользуемся $a^{\log_a x}=x$, т.е. возведением основания натурального логарифма в степень значения оценки. Если в расчетах задействован десятичный логарифм, то будем использовать 10, а не значение степени. Итого, получаем, что вероятность наличия спама в письме в рамках данного условия, составит:

$$\frac{e^{-7{,}629}}{e^{-7{,}629}+e^{-6{,}906}}=0{,}327\ =32.7\%.$$

Далее, выражение можно оптимизировать, сократив экспоненту по знаменателю и числителю. Тогда получим:

$$P(c\,|\,d)=\frac{1}{1+\sum_{c'\in C\setminus\{c\}}e^{q_{c'}-q_c}}\,.$$

Сумма в знаменателе считается только по классам отличным от того, для которого считалась вероятность. Но в каждом из слагаемых есть логарифмическая оценка рассматриваемого класса.

Литература

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.