КЛАССИФИКАЦИЯ И ОБРАБОТКА ЭЛЕКТРОННЫХ СООБЩЕНИЙ

Богдан А. А., Лапицкая Н. В.
Кафедра программного обеспечения информационных технологий,
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: alexeybogdan@icloud.com, lapan@bsuir.by

В данной статье обсуждаются методы и подходы для решения проблемы классификации сообщений с дальнейшим распределением по папкам и также их дальнейшей обработкой без участия пользователя.

Введение

Постоянно увеличивающийся поток входной информации, который является отличительной чертой настоящего времени, требует решения задач ее классификации даже на бытовом уровне. Скорости обработки входных данных привычной человеку уже не хватает, поэтому необходимо создавать технические решения. Пример такого канала поступления информации является электронная почта. Пользователь сталкивается с большим потоком электронных сообщений, которые они не в состоянии обработать самостоятельно. Более того, большинство сообщений - это реклама или спам рассылка, в то время как некоторые из сообщений могут быть очень важны для пользователя и нельзя допустить чтобы они были оставлены без внимания.

Сегодня, пользователи вручную создают папки и группируют свои сообщения с их помощью. Но ручная группировка может быть длительным процессом, если пользователь получает их в большом количестве. Для борьбы со спамом на почтовые сервера устанавливаются брандмауэры. Данный подход позволяет заблокировать только IP и DNS адреса. Для того чтобы обойти защиту брандмауэра, достаточно изменить IP или DNS адрес, с которого отправляется спам рассылка.

В качестве решения проблемы будут рассмотрены методы распределения по заданному условию и автоматического распределения сообщений по папкам, то есть их классификации. Также будут рассмотрен процесс автоматической обработки сообщений. Под автоматической обработкой понимается выполнение какого действия над сообщений в зависимости от его классификации. Данные решения значительно уменьшат время обработки электронных сообщений пользователей.

I. Метод распределения сообщений по условию

Рассмотрим простейший способ классификации электронных сообщений с технической точки зрения. Метод распределения сообщений по условию работает следующим образом: пользователь создает папку в электронном ящике и создает набор условий для данной папки. Если новое сообщение удовлетворяет всем условиями данной папки, то сообщение автоматически перемещается в данную папку.

Примерами условий могут быть:

- адрес отправителя;
- дата отправки сообщения;
- размер сообщения;
- наличие прикрепленных файлов в сообщении;
- наличие тега в письме.

На рисунке 1 вы можете увидеть пример работы метода распределения сообщений по условию.

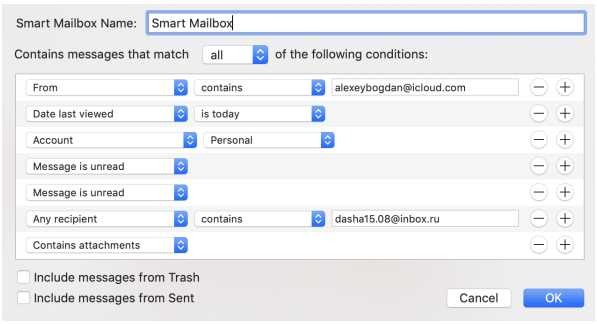

Рис. 1 – Распределение сообщений по условию в приложении Apple Mail

II. Автоматическое распределение писем

В случае автоматического группирования писем, алгоритм будет основан на классификации. В качестве классов будут выступать существующие папки. При синхронизации нового сообщения, сначала будет применяться метод распределения сообщений по условию. Если новое сообщение подойдет одной из папок по выражению, то оно будет перемещено в папку. Если нет, то будет применятся алгоритм автоматического группирования писем. При автоматическом распределении необходим алгоритм, который сам определит, что некоторые сообщения являются похожими и их необходимо переместить в одну папку.

В качестве метода автоматического распределения сообщений можно использовать метод максимина(Рис. 2). Он предназначен для разделения объектов на кластеры, причем количество кластеров заранее неизвестно. Оно определяется автоматически в процессе разбиения объектов. Принцип работы метода следующий. Выбирается один из объектов и назначается прототипом первого кластера. Находится объект, наиболее удаленный от выбранного и назначается прототипом второго кластера. Все объекты распределяются по двум кластерам. Каждый объект относится к кластеру, представленному ближайшим прототипом. Затем в каждом из кластеров находится объект, наиболее удаленный от своего прототипа. Если расстояние между этим объектом и прототипом кластера оказывается значительным (превышающим некоторую предельную величину), то объект становится новым прототипом, т.е. образуется новый кластер. После этого распределение объектов по кластерам выполняется заново. Процесс продолжается, пока не будет получено такое разбиение на кластеры, при котором расстояние от каждого объекта до прототипа кластера не будет превышать заданную предельную величину.

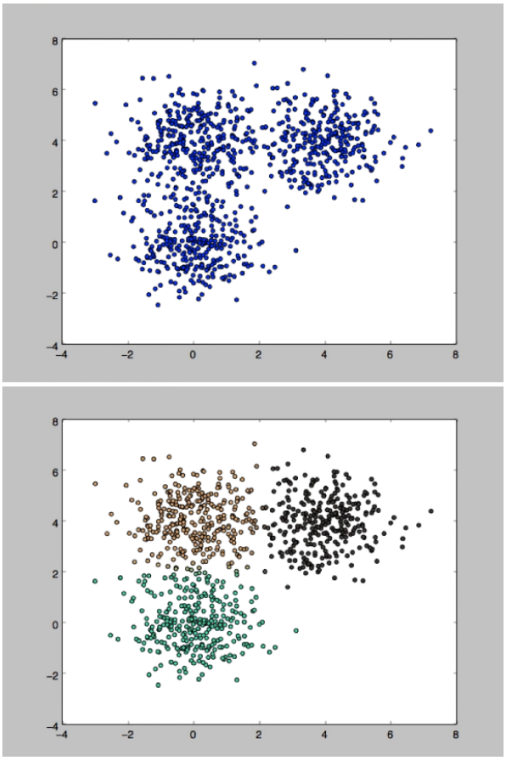

Рис. 2 – Пример работы метода максимина

III. Автоматическая обработка сообщений

Под автоматической обработкой сообщений понимается механизм, который будет выполнять определенные действия на сообщением автоматически. Данный метод работает следующим образом: пользователь добавляет обработчики для определенной папки и при попадании нового сообщения в папку автоматические обработчики начинают выполняться. Автоматические обработки могут быть следующими:

- удаление сообщения;
- отправка push нотификации на смартфон;
- sms оповещение;
- ответ на письмо с предопределенным текстом письма.

Рассмотрим следующий пример. Вам, как пользователю, хотелось бы избавиться от постоянных писем со спамом и рекламой. Тогда вы можете сделать следующее: создать папку, в которую будут попадать письма такого рода, и добавить действие удалить письмо. Таким образом вы избавитесь от таких писем. Рассмотрим следующий вариант. Вы, как сотрудник компании, должны в срочном порядке отвечать на все письма своего начальника. В таком случае вы можете создать создать папку, в которую будут сохраняться все письма вашего начальника, и добавить sms оповещение. В таком случае вы не упустите из виду важные сообщения.

IV. Заключение

В работе представлены три подхода к решению задачи повышения скорости обработки входных электронных сообщений. Предложены подходы обработки сообщений без участия пользователя. В дальнейшем будет проведен сравнительный анализ эффективности предложенных методов в зависимости от персональных характеристик пользователя.

Список литературы

1. An Introduction to Data Science [Electronic resource] / Saint Petersburg State University. – Coursera, 2019. – Mode of access: https://www.coursera.org/lecture/vvedeniye-v-nauku-o-dannykh/alghoritm-k-means-hOVUY. – Date of access: 25.08.2019.
2. Types of classification algorithms in Machine Learning [Electronic resource] / Mandy Sinada. – Medium, 2017. – Mode of access: https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14. – Date of access: 28.02.2017.
3. How to Organize Your Email with Smart Mailboxes in Apple Mail [Electronic resource] / Matt Clain. – HowToGeek, 2016. – Mode of access: https://www.howtogeek.com/252635/how-to-organize-your-email-with-smart-mailboxes-in-apple-mail. – Date of access: 07.03.2017.
4. How do email servers detect spam? [Electronic resource] / Joy Larkin. – Quora, 2016. – Mode of access: https://www.quora.com/How-do-email-servers-detect-spam. – Date of access: 10.01.2016.
5. Обучение без учителя: 4 метода кластеризации данных на Python [Электронный ресурс] / Библиотека программиста. – Proglib, 2019. – Метод доступа: https://proglib.io/p/unsupervised-ml-with-python/. – Дата доступа: 24.05.2018.