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

УДК [004.773 + 004.724.3] - 048.445

Богдан
Алексей Анатольевич

Методы обработки и классификации электронных сообщений

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности 1-40 80 05 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Научный руководитель
Лапицкая Н.В.
к.т.н., доцент

Минск 2020

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является разработка алгоритмов классификации электронных сообщений и создание на их основе программного сервиса автоматического распределения электронных сообщений по папкам. Для достижения поставленной цели необходимо решить следующие задачи:

1. Проанализировать существующие алгоритмы классификации текстовой информации.
2. Проанализировать технические решения для решения задачи электронных сообщений.
3. Разработать архитектуру программного средства для решения задачи, а также функционал и компоненты, которые позволят разрабатываемому программному средству быть полноценным продуктом.
4. Определить методы и подходы для решения поставленной задачи.
5. Разработать программное средство для обработки и классификации электронных сообщений.
6. Провести экспериментальные исследования разработанного программного средства.

Объектом исследования являются классификационные признаки, определяющие информационное наполнение электронных сообщений.

Предметом исследования является возможность автоматической классификации электронных сообщений исходя из их информационного наполнения.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность поиска классификационных признаков в информационном наполнении электронных сообщений с целью их автоматической классификации.

Связь работы с запросами реального сектора экономики

С экономической точки зрения, данная проблема является актуальной, так как в 2020 году невозможно представить пользователей компьютеров и телефонов, у которых нет электронной почты. При постоянном росте информации, время на обработку электронных сообщений неуклонно растет. Решение поставленной проблемы может позволить пользователям автоматизировать ежедневные рутинные задачи по работе с почтой.

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя Н. В. Лапицкой, заключается в формулировке проблемы исследования, целей и задач исследования.

Опубликованность результатов диссертации

По теме диссертации опубликовано 2 печатные работы в сборниках трудов и материалов конференции БГУИР. Одна из работ косвенно относится к диссертации, так как является описанием механизмом по работе с фоновыми процессами, которые используются в разрабатываемом программном средстве.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, четырех глав, заключения, списка использованных источников, списка публикаций автора и приложений. В первой главе представлен анализ предметной области, выявлены основные существующие проблемы в рамках тематики исследования. Вторая глава посвящена постановке функциональных требований, выбор алгоритмов и технологий для решения задачи. В третьей главе рассмотрены алгоритмы и подходы для решения поставленных задач. Также описаны реализации компонентов, которые являются важными для построения программного средства. В четвертой главе предложена представлена практическая реализация программного средства с тестированием и руководством пользователя.

Общий объем работы составляет 75 страниц, из которых основного текста – 40 страниц, 20 рисунков на 10 страницах, 13 таблиц на 10 страницах, список использованных источников из 18 наименований на 2 страницах и 1 приложение на 13 страницах.

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 он 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.

В конечном итоге будет разработано программное средство, которое будет решать поставленные задачи и также решать простейшие задачи почтового клиента.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе проанализированы существующие решения для работы с почтой, такие как Gmail, Яндекс.Почта, Spark, Apple Mail, Windows Mail. Для данных решений был проведен сравнительный анализ и были составлены рейтинги по следующим критериям:

- наличие функционала и доступность использования;
- поиск сообщений;
- классификация сообщений по условию;
- автоматическая классификация сообщений;
- автоматическая обработка сообщений.

Был сделан вывод, что они частично решают решают проблемы обработки и классификации сообщений. В большинстве случаев это спам фильтры, которые помогут избавиться от нежелательных сообщений. Сегодня рассмотренные приложения делают акцент на удобство представления и поиска нужных сообщений, а также большое количество интеграции с другими приложениями. Хотелось бы отметить почтовый клиент Apple Mail, который поддерживает добавление условию по различным параметрам, что делает его на голову выше конкурентов с точки зрения автоматизации классификации сообщений. С точки зрения удобства использования, я бы выделил приложение Spark, у которого есть возможность перемещать сообщения по папкам без привязки к электронному адресу а также поддержка базовой классификации на обычные сообщения, нотификации и новости.

Во второй главе были рассмотрены алгоритмы для автоматической классификации сообщений. В результате был выбран метод максимина для решения данной задачи, так как он позволяет классифицировать данные для разных наборов данных, не прибегая к обучению, а также создает необходимое количество классов для разных пользователей. Были предложены функциональные и нефункциональные требования для разрабатываемого прототипа. Было проведено обоснование технологий, которые использовались для решения задачи.

В третьей главе был описан процесс разработки следующих методов:

- классификация сообщений по выражениям;
- автоматическая классификация сообщений;
- метод событий.

Для указанных методов была предложена реализация с детальным описанием шагов. Также были приведены потенциальные проблемы, с которыми может столкнуться пользователь при использовании данных методов для упрощения своей работы с корреспонденцией.

В четвертой главе был проведен эксперимент для разработанных методов обработки и классификации почтовых сообщений. Для этого был использован личный почтовый аккаунт соискателя с достаточно большим количеством сообщений. Что касается метода классификации сообщения по выражению, результаты эксперимента показали ожидаемые результаты, так как были созданы папки с выражениями, которые могут достаточно точно классифицировать сообщения. Как и говорилось, данный механизм является довольно простым для реализации и использования. Эксперимент классификации сообщений, используя метод автоматической классификации, пока удовлетворительные результаты. Было создано пять папок и их классификация имеет смысл. Недостатком является то, что достаточно сложно определить имена для таких папок, так как их имена создаются по специальному шаблону. С высокой вероятностью, пользователь будет переименовывать автоматически созданные папки. Механизм событий показал свою работоспособность и позволяет пользователей избавиться от рутинных операций.

ЗАКЛЮЧЕНИЕ

В рамках работы была сделана попытка решения задачи автоматизации работы с корреспонденцией, поступающей посредством электронной почты. Были проанализированы самые популярные приложения для работы с электронной почтой. Большинство существующих решений делают акцент на доступность использования, интеграцию с другими приложениями, нежели ускорения процесса работы с сообщениями. Результаты показали, что одно из существующих приложений предоставляет механизм классификации по выражениями, а другое приложение реализует частичную автоматическую классификацию сообщений.

В работе предложены три метода: метод классификации сообщений по выражениям, автоматическая классификация сообщений и метод событий. Метод классификации сообщений по выражениям позволяет группировать сообщения в определенную папку, устанавливая выражения для папки. При выполнении заданных условий, новые сообщения перемещаются в папку. Достоинством данного метода является высокая точность работы, а также очень эффективен для группировки специфических сообщений. Недостатком является некорректная группировка сообщений в случае добавление слишком обобщенных выражений к папкам. Автоматический метод классификации сообщений основан на частоте встречания самых популярных слов в теме и теле сообщения всех сообщений почтового аккаунта. Результаты эксперимента показали удовлетворительные результаты, где было создано восемь папок в каждой из которых находилось более двадцати сообщений. Сообщения были сгруппированы корректно с точки зрения их тематики. Достоинством метода является полная автоматизация. Недостатком является не всегда корректное определение имен для папок, что требует дополнительной активности конечного пользователя. Метод событий позволяет делать автоматические действия для сообщений, который были перемещены в папку. Данный механизм позволяет добавить обработку над сообщениями без участия пользователя. Достоинством метода является возможность избегать ненужных сообщений и получать мгновенные уведомления. Недостатком является возможная потеря важной корреспонденции при добавлении событий на автоматически созданные папки. По результатам проведенных исследований был создан прототип программного средства для работы с электронной почты, реализующий указанные функции.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Богдан А.А. Использование бессерверных вычислений для решения задач обработки информации / Богдан А.А. // Информационные технологии и системы. Системы обработки информации. - 2019. - №4(55). - с. 190 - 191.

2. Богдан А.А. Классификация и обработка электронных сообщений / Богдан А.А., Лапицкая Н.В. // Информационные технологии и системы. Системы обработки информации. - 2019. - №5(55). - с. 192-193.