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

УДК 004.93

Заграй
Владимир, Юрьевич

# Алгоритмы фильтрации спам-писем с помощью нейронных сетей

**АВТОРЕФЕРАТ**
на соискание степени магистра технических наук
по специальности 1-40 80 02 «Системный анализ, управление и обработка информации»

_____

Научный руководитель
Гуринович Алевтина Борисовна
Кандидат физико-математических наук

_____

Минск 2019

# КРАТКОЕ ВВЕДЕНИЕ

В настоящее время, век интернета, компьютерная техника используется для доступа к различной информации. Данная техника применяется для обмена различной информацией, в том числе в виде электронной почты и мгновенных сообщений. Но использование указанных средств передачи и получения сообщений влечёт за собой несколько проблем, одной из которых является проблема фильтрации спам-писем. В то же время количество пользователей всемирной паутины увеличивается и, следовательно, растут нагрузка на сеть, а также объём и количество отправляемых ежесекундно писем, среди которых присутсвуют и спам-письма.

Спам-письма – это нежелательные письма, которые могут приходить на адрес электронной почты или адреса некоторых систем.

Фильтр спам-писем – это алгоритм или система, релизованная на его основе, призванная защитить пользователя от спам-писем, которая используется для фильтрации входящей электронной почты и других видов сообщений.

Как упомянуто ранее, с появлением интернета объем информации увеличился во много раз и ежедневно этот объем информации увеличивается. Это утверждение влечёт собой требование к использованию высокопроизводительных алгоритмов фильтрации писем, рационально использующих вычислительные ресурсы.

Целью диссертации является сравнение и анализ существущих алгоритмов фильтрации спам-писем и определение оптимального и эффективного алгоритма.

В данной работе рассматривается проблема фильтрации спам-писем на почтовых серверах и оконечных ЭВМ пользователей, рассматриваются некоторые нейросетевых алгоритмы, связанные с обработкой писем, с целью оптимизации и ускорения процесса фильтрации писем в целом.

Нахождение эффективного алгоритма фильтрации спам-писем позволит уменьшить вероятность ложных срабатываний и уменьшить нагрузку на системы обмена сообщениями.

Задачами исследования являются:

– описание и исследование различных алгоритмов фильтрации на основе нейронных сетей, а также других типы алгоритмов фильтрации;

– сравнение их эффективности, времени калибровки, настройки и количества требуемого времени для этого обучения и классификации.

# ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

1 Цель исследования.

Целью настоящей работы является исследования и разработки алгоритмов фильтрации спам-писем, которые позволят защитить пользователй от спам-писем, которые могут быть использованы злоумышленниками для кражи такой личной информации, как данные банковских счетов и карт, логины и пароли интернет-служб и так далее..

2 Задачи исследования.

– описание и исследовать различные алгоритмы фильтрации на основе нейронных сетей, а также других типов алгоритмов фильтрации;

– сравнение их эффективности, времени калибровки, настройки и количества требуемого времени для этого обучения и классификации.

3 Личный вклад соискателя.

Соискателем выполнены все изложенные в работе разработки и исследования. Постановка задач и обсуждение результатов проводились совместно с научным руководителем Белорусского государственного университета информатики и радиоэлектроники. Обработка, интерпретация данных, а также выводы сделаны автором самостоятельно.

4 Апробация результатов диссертации

Основные положения диссертационной работы докладывались на следующих научных конференциях:

– 54-ая научно-техническая конференция магистрантов, аспирантов и студентов (Минск 2018);

– Международная научная конференция «Информационные технологии и системы» (Минск 2018);

– 55-ая научно-техническая конференция магистрантов, аспирантов и студентов (Минск 2019);

Опубликованность результатов диссертации

Основные результаты диссертации опубликованы в третьей статье в сборнике материалов научных конференций.

# 1 АНАЛИЗ СОСТОЯНИЯ ПРОБЛЕМЫ И ПОСТАНОВКА ЗАДАЧ ИССЛЕДОВАНИЯ

## 1.1 Общая характеристика почты, писем и их фильтрации

Электронная почта (или сокращенно почта) – это быстрый, эффективный и недорогой способ обмена сообщениями через сеть Интернет. Будь то частные сообщения от семьи, товарищей, сообщение для сотрудников всей компании от начальства, обмен исследователями со всего мира своими научными открытиями или астронавты, поддерживающие связь со своей семьей (по почте или по IP-телефонии), электронная почта является предпочтительным универсальным средством для общения.

Электронная почта появилась и стала использоваться достаточно давно. Помимо обычной почты стали приходить так называемые спам-письма. За пару десятилетий объем спама в электронной почте увеличивался в геометрической прогрессии и стал не просто раздражающим, а также и угрозой безопасности, поскольку его потенциал продолжают развивать, чтобы использовать накопленный потнециал для нанесения серьезного ущерба людям, бизнесу и экономике для получения собственной выгоды.

Невозможно точно сказать, кто первым пришел к идее распространения спама к такой достаточно простой идее, как если рекламная почта отправляется миллионам людей, то по крайней мере один человек отреагирует на нее независимо от того, является ли оно предложением чего-либо или нет. С помощью электронной почты предоставляется возможность бесплатно отправлять миллионы рекламных объявлений отправителям, и этот факт в настоящее время широко используется рекламными и не только организациями.

В результате электронные почтовые ящики миллионов людей заполняются со всей этой так называемой «непрошенной» массовой электронной почтой, также известной как «спам» или «нежелательная почта». Будучи невероятно дешевым для отправки, спам вызывает много проблем для интернет-сообщества: большие объемы спам-трафика между серверами вызывают задержки при доставке нормальных сообщений, люди с поминутным доступом к Интернету должны тратить трафик и время, загружая нежелательную почту. Сортировка нежелательных сообщений требует времени и вводит риск ошибочного удаления обычных сообщений. Наконец, существует довольно количество спама взрослого содержания, которые не должны попадаться детям и не должны влиять на них.

Было предложено много способов борьбы со спамом. Существуют «социальные» методы, такие как юридические меры и простые инструкции о

том, что не желательно делать человеку, если он не хочет получать спам-письма. Существуют такие «технологические» способы, как блокирование IP-адресов спамеров, которые занимаются рассылкой спама, и, наконец, фильтрация электронной почты. К сожалению, пока нет универсального и идеального способа устранения спама, поэтому количество нежелательной почты продолжает расти. Например, около 50% сообщений, поступающих на мой почтовый ящик среднестатистического пользователя, являются спамом.

Автоматическая фильтрация электронной почты и сообщений является наиболее эффективным методом противодействия спаму на данный момент, и продолжается напряженная борьба между спамерами и методами и алгоритмами фильтрации спама: чем более эффективны методы защиты от спама, тем более эффективны и методы спамеров. Около 10 лет назад большую часть спама можно было надежно разрешить, блокируя электронные письма, исходящие от определенных адресов, или отфильтровывая сообщения с определенными темами. Чтобы преодолеть эти методы, спамеры начали отправлять спам со случайных адресов и добавлять случайные символы в тему сообщения. Алгоритмы фильтрации спама, скорректированные с учетом отдельных слов в сообщениях, могут справиться с этим, но затем появилась нежелательная почта со специально написанными словами или просто с ошибками. Чтобы обмануть более продвинутые фильтры, которые полагаются на частоты слов, спамеры добавляют большое количество «обычных слов» в конец сообщения. Кроме того, есть спам-письма, которые вообще не содержат текста, например, *HTML*-сообщения с одним или несколькими изображениями, загружаемыми из Интернета при открытии сообщения, и есть даже саморасшифровывающееся спам-письма, например, зашифрованное *HTML*-сообщение, содержащее код *JavaScript* который расшифровывает его содержимое при открытии.

Существует два общих подхода к фильтрации почты: инженерия знаний и машинное обучение.

Инженерия знаний – это область наук об искусственном интеллекте, связанная с разработкой экспертных систем и баз знаний. Изучает методы и средства извлечения, представления, структурирования и использования знаний. В этом случае создается набор правил, согласно которым сообщения классифицируются как спам или законная почта. Типичное правило такого типа может выглядеть так: «если тема сообщения содержит текст «купить сейчас», тогда сообщение является спам-письмом». Набор таких правил должен быть создан либо пользователем фильтра, либо каким-либо другим лицом. Основным недостатком этого метода является то, что набор правил должен постоянно обновляться, а его поддержка в актуальном состоянии не подходит большинству

пользователей, не разбирающимся в нюансах спам-фильтров. Разумеется, правила могли бы централизованно обновляться разработчиком средства фильтрации спама, и даже существует решение для одноранговой базы данных, но, когда правила общедоступны, спамер имеет возможность корректировать текст его сообщения, чтобы он проходил через фильтр. Следовательно, лучше, когда правила хранятся не в открытом виде, но также хранятся глобально.

Для замены такого набора правил возможно использовать сериализованные модели обученных систем фильтрации. Такие модели возможно создать при использовании алгоритмов на базе нейронных сетей, так как в основу их работы входит отдельный этап обучения. Данные алгоритмы представляют собой часть алгоритмов машинного обучения.

Спам подразумевает собой широкое понятие, феномен которого до сих пор не полностью изучен. В целом, спам имеют множество форм – спам-сообщения в чаты, спам-объявления в интернет-блогах, спам-результаты в поисковых системах, которы часто вводят в заблуждение в то время, как социальные сети страдают от социального спама и так далее.

В диссертации рассматриваются некоторые из наиболее популярных алгоритмов фильтрации спама в области машинного обучения, а также классические алгоритмы и способ их совмещения с целью уменьшения вероятности ложного срабатывания и получение более стойкого к ложным срабатываниям и времени на выполнение алгоритма фильтрации.

## 1.2 Задачи фильтрации спам-писем

Целями или задачами фильтрации писем и сообщений является защита пользователей от спам-писем, которые могут быть использованы злоумышленниками для кражи такой личной информации, как данные банковских счетов и карт, логины и пароли интернет-служб, форумов, социальных сетей, служб обмена сообщениями и так далее. Это возможно осуществить с помощью подготовки программного обеспечения лоя фильтрации спам писем.

Фильтр спама – это функция $f(m)$, которая позволяет выяснить, является ли письмо $m$ спамом $(S)$ или обычным письмом $(L)$. Если обозначить множество всех писем через $M$, то ясно, что ищется функция $f : M \to \{S, L\}$. Данная функция ищется путём обучения одного из алгоритмов машинного обучения набором предварительно классифицированных писем $\{(m_1, c_1), (m_2, c_2), \ldots, (m_n, c_n)\}, m_i \in M, c_i \in \{S, L\}$. Это утверждение является практически общим утверждением задачи машинного обучения. Однако, на эту задачу есть две точки зрения: необходимо извлекать признаки из текстовых

строк, а также присутствуют очень строгие требования к точности классификатора, что рассматривается в данной работе.

### 1.3 Виды алгоритмов фильтрации

Одной из главных задач при выборе алгоритма фильтрации является обеспечение допустимой производительности фильтра спам-писем. Требования к такому фильтру отличаются от требований к обычному классификатору, то есть, если фильтр неправильно распознаёт спам-письмо как обычное, это не является особой проблемой для рядового пользователя. Однако ошибки другого типа – ошибочной классификации обычной почты как спама – являются совершенно неприемлемыми, так как нет особого смысла в фильтре спам-писем, который раз в несколько писем помечает обычную почту как спам, потому что в этом случае пользователь должен регулярно просматривать сообщения в папке со спам-письмами и это подвергает сомнению использование такого классфикатора для фильтрации спам-писем. Фильтр, совершающий небольшое число ложных срабатываний при классификации, ненамного лучше, потому что в этом случае пользователь системы будет склонен доверять фильтру и, скорее всего, не просматривать отфильтрованные сообщения, поэтому, если фильтр делает ошибку, важное электронное письмо может быть потеряно.

К сожалению, в большинстве случаев невозможно гарантировать, что фильтр не будет иметь ложных срабатываний. В большинстве алгоритмов обучения есть параметр, который возможно настроить для повышения важности правильной классификации обычной почты, но с другой стороны, если присвоить слишком высокую важность обычной почте, алгоритмы будут склонны классифицировать все письма как не спам, не принимая опасных решений, но это не имеет практического смысла.

Некоторые меры безопасности могут компенсировать ошибки классификатора. Например, если сообщение классифицируется как спам, отправителю этого письма может быть отправлен ответ с предложением повторно отправить это письмо на другой адрес или включить некоторые конкретные слова в заголовок письма. Другая идея заключается в использовании фильтра для оценки достоверности того, является ли письмо спамом, и сортировки списка писем в почтовом ящике пользователя в порядке возрастания почситанной оценки достоверности письма.

Существует немалое количество алгоритмов фильтрации спам-писем.

Среди них присутствуют:

– наивный байесовский классификатор, который называют «наивным» за то, что он основан на дополнительном предположении, что объекты описываются независимыми признаками;
– метод «усиление»;
– метод случайных деревьев;
– метод опорных векторов;
– генетические алгоритмы;
– оценочные алгоритмы, например, решающие деревья.

Также существуют алгоритмы на основе нейронных сетей, которых рассматриваются в данной работе. К ним относятся многослойный перцептрон, рекуррентная нейронная сеть с различными типами «памяти».

Нейронные сети являются разновидностью машинного обучения. Машинное обучение – это класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Подход машинного обучения не требует четкого указания каких-либо правил. Вместо этого необходим набор предварительно классифицированных наборов признаков, то есть образцов для обучения. Затем конкретный алгоритм используется для «изучения» скрытых шаблонов в данных по этих обучающим данным. Предмет машинного обучения широко изучен и существует множество алгоритмов, подходящих для этой задачи.

Для улучшения общей производительности алгоритма возможно комбинировать различные виды нейронных сетей, а также и классические алгоритмы.

Несмотря на широкое разнообразие алгоритмов, ни один метод не является идеальным решением проблемы фильтрации спама, и каждый алгоритм имеет некоторые компромисы между неправильно отфильтроваными письмами и не отфильтровыванием всех спам-писем – и требуют на это время, усилия по подбору коэфициентов и цену за неправомерное воспрепятствование прохождения обычной почты через фильтр.

# 2 КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Магистерская диссертация представлена в виде пояснительной записки на 67 страницах, состоящей из введения, четырёх разделов и заключения.

В первом разделе приведена характеристика почты, писем и проблем, связанных с ней. В нём приведены описания электронной потчты как средства обмена собщениями, разновидностей спам-писем и алгоритмов фильтрации.

В разделе приведён перечень существующих классических алгоритмов фильтрации и упоминание применения нейронных сетей для фильтрации спама, которые рассматриваются во втором и третьем разделах.

Во втором разделе приводится описание понятия нейронной сети, её структуры, описанию структур перцептронной и рекуррентной нейронных сетей и нейронных сетей в целом. В подразделе про структуру сети описаны входной, скрытые и выходной слои. Приведёно описание проблемы переобучения и недообучения сети и метод её решения – метод отсеивания. Также приведены преимущества и недостатки нейронных сетей и их применимость для решения различных задач, а в частности – задачи фильтрации спам-писем.

В этом разделе также рассмотрены следующие виды нейронных сетей:

- многослойный перцептрон;
- свёрточные неёронные сети;
- рекурсивные нейронные сети;
- рекуррентные нейронные сети;
- сети с долгой краткосрочной памятью;
- ограниченные нейронные сети.

Описаны виды передаточных функций, а также методы обучения сетей, которые и применены в диссертации для обучения нейросетевых спам-фильтров.

Третий раздел содержит подготовку данные для классификации, описание классических и разработанных на базе нейронных сетей алгоритмы, а также приведены подробные алгоритмы их обучения.

В разделе приведён алгоритм для преобразования текста в вектор признаков, пригодный для классификации нейронными сетями и классическим алгоритмами. В качестве признаков выбраны колиечество слов, частоты символов, количество предложений, различные лингвистические меры текста, средняя длина слова, число коротких слов и так далее. Описаны методы обучения различных алгоритмов для спам-фильтрации, том числе рекурентная нейронная сеть и многослойный перцептрон.

В четвёртом разделе приведены результаты работы классических и нейросетевых алгоритмов, а также сравнение их эффективности. Проведена серия испытаний как классических, так и нейросетевых алгоритмов, в результате которых рассчитаны время обучения, время классификации и точность классификации и найдена зависимость данных результов от размеров выборок.

В результате работы над магистерской диссертацией разработаны два алгоритма на основе двух видов нейронных сетей для классификации писем и сообщений электронной почты и систем обмена сообщениями в целом.

Результаты, полученные в ходе исследования для магистерской диссертации, могут использоваться в таких множественных сферах, как фильтрация спам-писем, для систем поиска информации, для исключения из результатов вредноносных ссылок, для блокировки сетевых провокаторов и так далее.

# ЗАКЛЮЧЕНИЕ

В данной работе рассмотрены и описаны алгоритмы фильтрации спам-писем на основе нейронных сетей, предоставлена информация о нюансах применения данных алгоритмов, а также проведён анализ и сравнение их времён обучения и классификации, а также точности классификации предварительно заданных наборов сообщений разных размеров.

Нейронные сети, зарекомендовавшие себя, как мощный алгоритм для классификации изображений, в последнее время стали активно использоваться и для других задач машинного обучения.

Использование многослойных и рекуррентных нейронных сетей для фильтрации писем позволит меньше времени на обработку нежелательной и вредноностоной информации и больше на обычную почту.

Представленные алгоритмы определения спама, основанные на классификации признаков писем с использованием нейронных сетей, являются современееыми подходами к классификации писем. При этом эффективность работы алгоритма на основе рекуррентной нейронной сети гораздо выше, чем при использовании многослойного пецептрона, но они оба выполняют классификацию гораздо эффективнее классических методов классификации, так как используют нейронные сети в своей основе. Это было достигнуто благодаря использованию нейронных сетей со всеми их преимуществами в плане переиспользования обученной модели. А также в алгоритме на основе рекуррентных нейронных сетей эффективность выше по сравнению с алгоритмом на основе многослойного перцептрона в связи с использованием в архитектуре первой сети узлов памяти и замыканий.

До написания данной работы, было твердое мнение, что достаточно хороший алгоритм фильтрации спам-писем невозможно создать, и единственным надежным способом фильтрации спам-писем является ручное создание набора правил, или периодическая корректировка коэффицинтов классификации, или полагание на проприетарное программное обеспечение, использование которого крайне ограничено в больших компаниях ввиду риска утечки ценной, секретной информации третьим лицам.

Практическая реализация рассмотренных алгоритмов значительно повысит эффективность фильтрации сообщений электронной почты и иных сообщений на оконечных устройствах, используя алгоритмы фильтрации на основе нейронных сетей, и предоставит дополнительную возможность для отбрасывания спам-писем в других приложениях и системах.

С точки зрения программиста использование алгоритмов фильтрации спама на основе нейронных сетей позволяет сократить время ожидания при подключении пользователя к почтовому серверу для получения сообщений. При использовании серверами классических алгоритмов фильтрации, на сервере рабочим персоналом, то есть программистами, вручную устанавливались параметры для фильтрации писем. Но со временем появляются новые типы спам-писем, которые перестают определяться установленными ранее параметрами для классических алгоритмов, которые необходимо постоянно перенастраивать. Именно поэтому использование новых алгоритмов фильтрации спама на основе нейронных сетей является более эффективным, чем использование классических алгоритмов.

Для классификации писем реализован алгоритм для извлечения признаков из писем для последующей их обработки. Данные лексикографиские признаки, представляющие собой числовой вектор, необходимы для обучения фильтров и минимизации данных для обучения.

Из результатов анализа алгоритмов определено, что на малых объёмах информации классические алгоритмы показали себя лучше в плане времени её обработки, но при увеличении объёма сообщений растёт сложность вычислений, а в частности – сравнение с образцами.

В свою очередь, нейронные сети на достаточно большом объёме сообщений имеет в несколько раз большее временя для обучения и более высокую точность классификации. При этом время классификации является сравнительно малым относительно классических аналогов.

Для устранения недостатков нейронных сетей и классических алгоритмов их следует комбинировать. Комбинация нейронных сетей с классическими алгоритмами позволила уменьшить количество спам-писем и вероятность их появления и увеличить надежность системы фильтрации при умеренном использовании вычислительных ресурсов за счёт увеличения точности разпознавания спам-писем.

Таким образом, сфера возможного применения результатов научного исследования весьма широка, что показывает высокую ценность проделанной работы.

# СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1–А. Заграй, В. Ю. Исследование алгоритмов фильтрации спам-писем / В. Ю. Заграй // Информационные технологии и управление : материалы 54-й научной конференции аспирантов, магистрантов и студентов. (Минск, 23 – 27 апреля 2018 года). – Минск : БГУИР, 2018. – С. 71

2–А. Заграй, В. Ю. Фильтрация спам-писем с помощью алгоритмов на основе нейронных сетей / В. Ю. Заграй, А. Б. Гуринович // Информационные технологии и системы 2018 (ИТС 2018) = Information Technologies and Systems 2018 (ITS 2018) : материалы международной научной конференции, Минск, 25 октября 2018 г. / Белорусский государственный университет информатики и радиоэлектроники ; редкол. : Л. Ю. Шилин [и др.]. – Минск, 2018. – С. 288 - 289.

3–А. Заграй, В. Ю. Сравнение и выбор оптимального нейросетевого алгоритма фильтрации спам-писем / В. Ю. Заграй, А. Б. Гуринович // Информационные технологии и управление : материалы 55-й научной конференции аспирантов, магистрантов и студентов. (Минск, 22 – 26 апреля 2019 года). – Минск : БГУИР, 2019. – С. 70.