Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Оренбургский государственный университет»

Кафедра программного обеспечения вычислительной техники
и автоматизированных систем

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ЗАЩИТЫ ПОЧТОВЫХ СЕРВИСОВ ОТ НЕСАНКЦИОНИРОВАННЫХ РАССЫЛОК НА ОСНОВЕ КОНТЕНТНОЙ ФИЛЬТРАЦИИ ЭЛЕКТРОННЫХ СООБЩЕНИЙ

Монография

Рекомендовано к изданию Ученым советом федерального государственного бюджетного образовательного учреждения высшего образования «Оренбургский государственный университет»

Оренбург
2017

УДК 004.49
ББК 32.973.26-018.2
П 60

Рецензент – доктор технических наук, профессор А.М. Пищухин

Авторы: Н. А. Соловьев, Е. Н. Чернопрудова, Н. А. Тишина,

Л. А. Юркевская

П 60

Программное обеспечение защиты почтовых сервисов от несанкционированных рассылок на основе контентной фильтрации электронных сообщений: монография/ Н.А. Соловьев, Е.Н. Чернопрудова, Н.А. Тишина, Л.А. Юркевская; Оренбургский гос. ун-т – Оренбург: ОГУ, 2016. – 128.
ISBN 978-5-7410-1724-1

В монографии изложены результаты научных исследований соответствующих приоритетному направлению развития науки и техники РФ – информационно-телекоммуникационным технологиям и критическим технологиям федерального уровня – информационно-телекоммуникационным системам.

Методология научных исследований проблем построения средств защиты почтовых сервисов информационно-телекоммуникационных систем корпоративных предприятий с территориально-распределенной структурой может быть использована в научных исследованиях магистрантов, аспирантов по направлениям обучения: 09.04.01 – Информатика и вычислительная техника, 09.04.04 – Программная инженерия, 09.04.02– Информационные системы и технологии.

УДК 004.49
ББК 32.973-018.2

ISBN 978-5-7410-1724-1

© Соловьев Н.А.,
Чернопрудова Е.Н.,
Тишина Н.А.
Юркевская Л.А., 2017
© ОГУ, 2017

Содержание

Введение

Служба рассылки электронной почты информационно-телекоммуникационных систем, являющаяся средством документооборота, личной и служебной переписки корпоративных предприятий территориально-распределенной структуры, становится важнейшим информационным каналом реализации бизнес-процессов. Одной из проблем использования электронной почты на современном этапе является массовая рассылка несанкционированных электронных сообщений (НЭС) адресатами коммерческой или иной информации.

Специалисты информационной безопасности (ИБ) выделяют НЭС (спам) как один из видов угроз, требующих особого внимания не только в связи с мешающим технологическим эффектом, но и наносимым экономическим ущербом. По материалам департамента стратегического анализа аудиторской финансовой компании от спам-рассылок «экономика России ежегодно теряет 47,2 миллиарда рублей, или 1,9 миллиарда долларов».

Отсюда, противодействие НЭС становится актуальной задачей обеспечения ИБ информационно-телекоммуникационных систем (ИТКС) корпоративных предприятий с территориально-распределенной структурой.

Проблемам обеспечения ИБ электронной почты посвящены работы Валеева С.С., Семеновой М.А., Шварца А.А. и зарубежных исследователей B. Pfahringer, K. Junejo, D. Zhou и других. Обобщая результаты исследований, можно сделать вывод, что в настоящее время сложилась система методов, моделей и средств спам-фильтрации электронных сообщений (ЭС), позволяющая решать широкий спектр задач ИБ. Вместе с тем, анализ современного состояния электронной переписки корпоративных предприятий, имеющих территориально-распределенную структуру, выявил лавинообразный рост числа НЭС при относительно высокой ложной классификации сообщений. Кроме того, известные методы спам-фильтрации требуют значительных временных и ресурсных затрат. Поэтому развитие методов защиты электронной

почты остаётся актуальной тематикой научных исследований, **объектом** которых становится ИТКС корпоративных предприятий с территориально-распределенной структурой.

ИТ-рынок предлагает различные средства фильтрации содержимого информационного обмена по каналам Интернет. В настоящее время условно выделяют три типа средств, обеспечивающих контроль использования Интернет-ресурсов на корпоративном уровне:

– маршрутизаторы, межсетевые экраны, системы обнаружения вторжений, прокси-сервер и т.п.;

– антивирусное программное обеспечение, обладающие базовыми возможностями контентной фильтрации;

– специализированные средства, разработанные непосредственно для контроля использования интернет – ресурсов, такие как системы мониторинга электронной почты, средства контроля веб - трафика, антиспам - фильтры и т.п.

Исследования соответствуют **приоритетному направлению развития науки и техники РФ** – информационно-телекоммуникационным технологиям и **критическим технологиям федерального уровня** – информационно-телекоммуникационным системам.

Эти обстоятельства определяют **цель исследования**: *повышение достоверности идентификации легитимной почтовой корреспонденции на основе семантической подготовки электронных сообщений к интеллектуальной фильтрации и нейросетевой классификации в условиях изменяющегося контента служебной переписки.*

Для достижения сформулированной цели необходимо решение ряда задач научного характера, вызванных противоречиями между состоянием теории и требованиями практики обеспечения ИБ ИТКС:

1) Системный анализ защиты почтовых сервисов ИТКС предприятий с территориально-распределенной структурой.

2) Разработка модели электронного почтового сообщения, учитывающей семантику контента почтовой корреспонденции.

3) Разработка методики и алгоритмов фильтрации легитимных электронных сообщений почтовых сервисов в условиях изменяющихся интересов адресатов служебной корреспонденции.

4) Разработка средств фильтрации легитимной корреспонденции почтовых сервисов ИТКС корпоративных предприятий.

5) Проведение экспериментальной проверки почтовых сервисов со средствами фильтрации служебной корреспонденции и оценка их эффективности.

Научной основой для решения поставленной задачи являются: методы системного анализа и исследования операций; теоретические основы информатики; теория принятия решении; методы и средства защиты информации; интеллектуальные методы, методы теория статистических решений, объектно-ориентированное программирование; теория эксперимента.

Методология научных исследований проблем построения средств защиты почтовых сервисов информационно-телекоммуникационных систем корпоративных предприятий с территориально-распределенной структурой может быть использована в научных исследованиях магистрантов, аспирантов по направлениям обучения: 09.04.01 – Информатика и вычислительная техника, 09.04.04 – Программная инженерия.

1 Системный анализ проблем защиты почтовых сервисов информационно-телекоммуникационных систем корпоративных предприятий

Основа любой действующей организации это бизнес-процессы. Эффективность которых во многом определяется информационно- телекоммуникационными системами (ИТКС), в задачи которых входит создание единого информационного пространства в интересах корпоративных предприятий. Любая организация работает со значительным объемом различной корреспонденции, такой как входящие и исходящие письма, соответствующие приказы и распоряжения, нормативные акты и внутренняя документация, что становиться причиной повышения нагрузки на ИТКС организации. Что негативно сказывается на оперативности и достоверности принимаемых решений по управлению информационными процессами в ИТКС администраторами сети [10,19,21,75]

Анализ проблем связанный с информационными потоками ИТКС позволит более полно определить объект исследований.

1.1 Почтовые сервисы информационно-телекоммуникационных систем корпоративных предприятий

В организациях имеющих территориально удаленные подразделения, пользователям разных подразделений необходимо работать с общими документами. Участники деловых процессов в таких организациях могут находиться в разных городах, но при этом они должны взаимодействовать в едином информационном пространстве. Другими словами в организации с территориально-распределенной структурой должна действовать единая система с общим данными. Решение задач связанных с бизнес ориентированным механизмом Workflow [21,42,75,80,81] обеспечивают модули системы, представленные на рисунке 1.1.

На сегодняшний день на рынке программного обеспечения существует широкий спектр продуктов ИТКС [21,41,42]. Выбор прототипа для анализа особенностей информационных процессов электронной почты определялся внедрением и опытной эксплуатации разработанного прототипа системы фильтрации электронной почтовой корреспонденции на предприятии ОАО «Оренбургнефть» корпорации «ТБинфром», использующее для осуществления бизнес процессов корпорации систему электронного документооборота и управления взаимодействием (СЭДУВ) DIRECTUM. Следовательно, анализ особенностей информационных процессов обмена корреспонденцией как внутри организации так и с внешними корреспондентами целесообразно рассмотреть именно на системе DIRECTUM.

Рисунок 1.1 – Функциональность прототипа ИТКС

Решение задач бизнес-процессов организации обеспечивают модули системы DIRECTUM, представленные на рисунке 1.2 [81]:

Рисунок 1.2 – Модули системы DIRECTUM

Управление электронными документами. Создание и хранение различных неструктурированных документов (тексты Microsoft Word, таблицы Microsoft Excel, рисунки Visio, CorelDraw, видео и пр.); поддержка версий документов и электронной цифровой подписи (ЭЦП); структурирование документов по папкам; назначение прав доступа на документы; история работы с документами; полнотекстовый и атрибутивный поиск документов.

– *Управление деловыми процессами.* Поддержка процессов согласования и обработки документов, выдача заданий и контроль их исполнения, обеспечение взаимодействия между сотрудниками в ходе бизнес-процессов, поддержка свободных и жестких маршрутов (workflow).

– *Канцелярия.* Полное соответствие ГСДОУ, ведение номенклатуры дел с гибкими правилами нумерации, рассылка и контроль местонахождения бумажных документов, организация обмена электронными документами с ЭЦП между организациями.

– *Управление договорами*. Регистрация и согласование договоров и сопутствующих документов, ведение реестров документов, обеспечение оперативной работы с ними (поиск, подписание, анализ и т.д.).

– *Управление совещаниями и заседаниями.* Подготовка и проведение совещаний и заседаний (согласование места и времени, состава

участников, в т.ч. внешних, повестки), формирование и рассылка протокола, контроль исполнения решений совещания.

– *Управление взаимодействием с клиентами.* Ведение единой базы организаций и контактных лиц, истории встреч, звонков и переписки с клиентами, сопровождение процесса продаж в соответствии с регламентированными стадиями; планирование маркетинговых мероприятий, анализ их эффективности.

– *Обращения граждан и организаций.* Организация работы с обращениями граждан в соответствии с Федеральным законом № 59-ФЗ "О порядке рассмотрения обращений граждан Российской Федерации" в государственных организациях и на крупных предприятиях.

– *Управление показателями эффективности.* Оперативный контроль и анализ бизнес-процессов предприятия по ключевым показателям эффективности с поддержкой сбалансированной системы показателей (BSC).

– *Интеграция с системами обмена*. Организация взаимодействия ECM-системы с сервисами операторов межкорпоративного обмена электронными документами.

Система DIRECTUM так же предоставляет средства для обмена данными в территориально-распределенных организациях. Для чего в данной системе разработан механизм обмена электронными документами между системами различных организаций или системами структурных единиц одной организации. Для обмена документами между структурными единицами крупного предприятия или между независимыми организациями используется встроенный модуль осуществляющий механизм обмена электронными документами между системами.

Тем не менее, полностью исключить другие каналы передачи данных невозможно, так как значительная часть информации поступает в компанию через электронную почту или другие системы (СЭД). Сведения о корреспонденции поступившей через электронную почту регистрируются и пересыла-

ются ответственному за их обработку, который по каждому поступившему письму принимает решение принять документ в систему или нет.

Архитектура (СЭДУВ) DIRECTUM (рисунок 1.3) предусмотрена таким образом, что для любого пользователя данной системы не имеет значение через какую систему идет обмен сообщениями (рисунок 1.4).

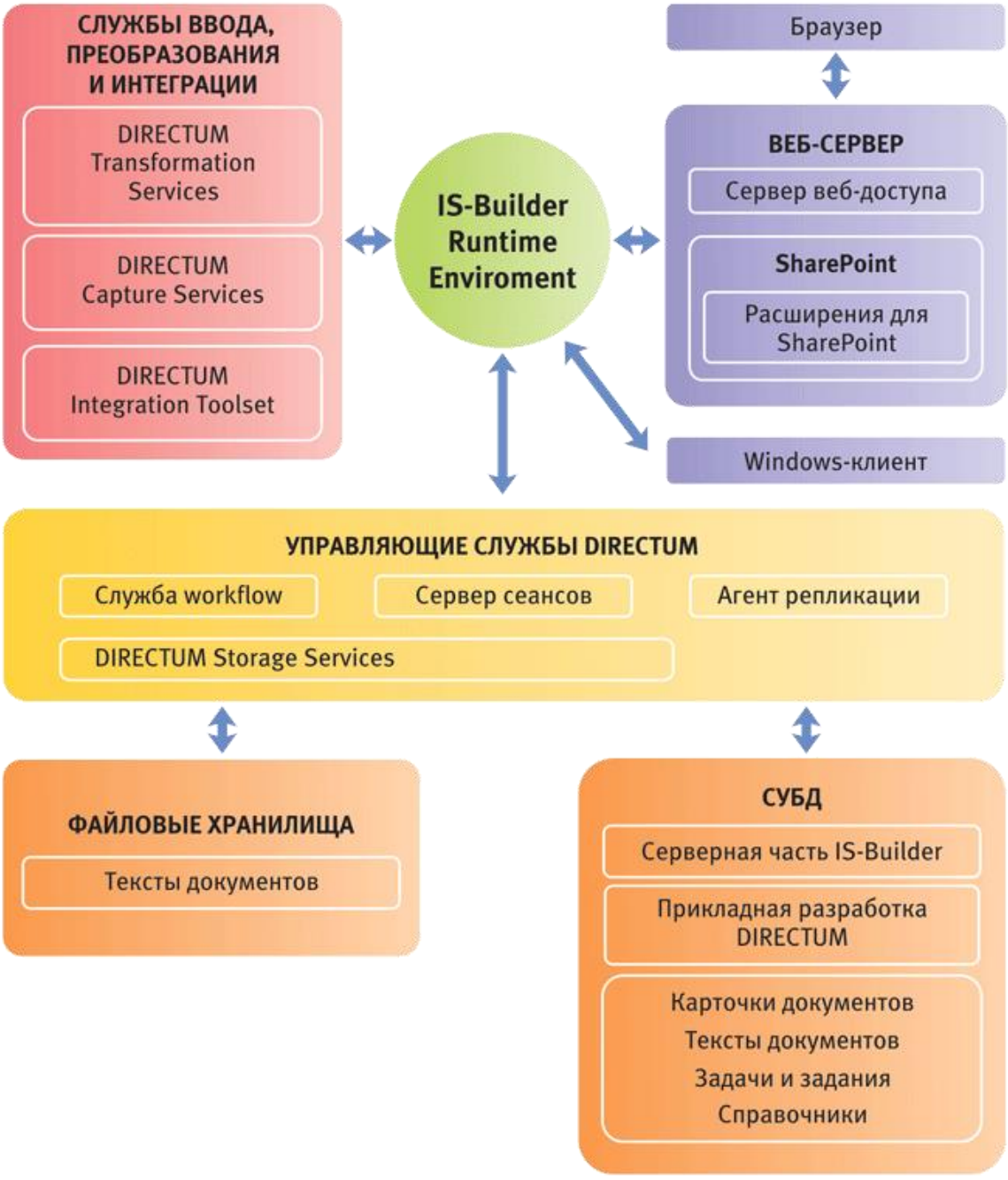


Рисунок 1.3 – Программная архитектура система СЭДУВ DIRECTUM

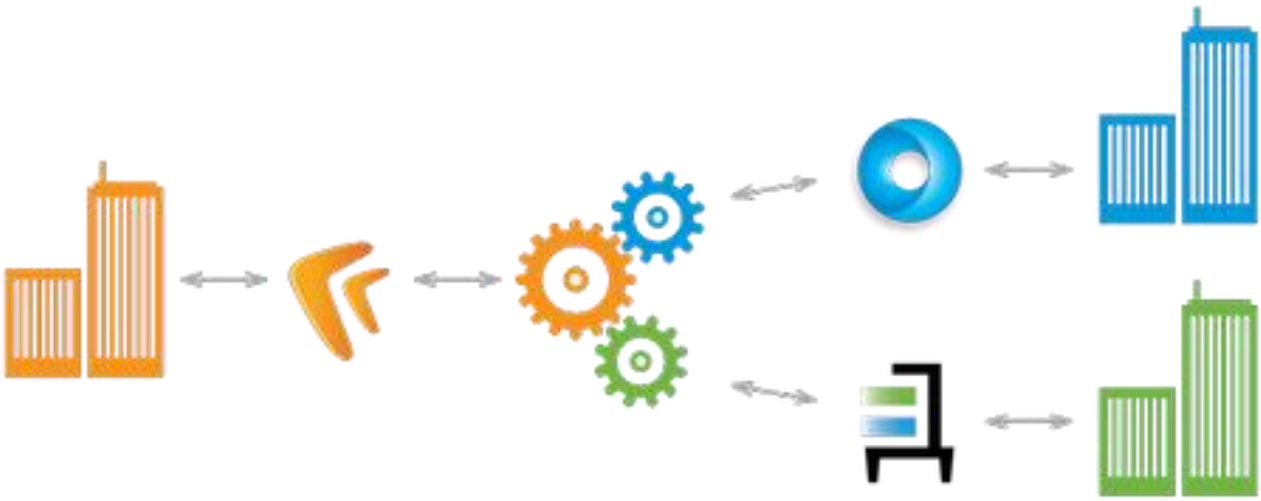


Рисунок 1.4 – Процесс обмена сообщениями

Система DIRECTUM предназначена для осуществления документооборота на предприятии, причем процесс передачи документов строго регламентирован политикой применения самой системы. Каждый электронный документ системы состоит из следующего набора информации: *карточки*, содержащей набор атрибутов, описывающий документ (автор, тип документа, дата создания, корреспондент) которые в дальнейшем используются в процессе работы и *содержимого самого документа*. Следовательно, все документы должны соответствовать требованиям данной системы. Всё это приводит к увеличению трудозатрат, связанных с первичной обработкой документов. Такая обработка требует проведения с документами типовых операций (оцифровка, классификация, регистрация). Для автоматизации данного процесса в СУЭД DIRECTUM существует набор компонент DIRECTUM Capture and Transformation Services (DCTS), позволяющий осуществлять массовый ввод документов в DIRECTUM с различных источников (электронная почта, файловая система, МФУ, факсы сканеры). Общая схема работы данной службы представлена на рисунке 1.5.

Служба ввода с факса

Служба принимает факсы на факс-модем, получая образы входящих многостраничных документов. Отправляет образы в DIRECTUM, предостав-

ляя информацию о количестве страниц, телефонном номере отправителя, телефонном номере получателя. Служба интегрирована с Microsoft Fax Service.

Служба ввода из файловой системы

Служба принимает документ из папки файловой системы и ее подпапок в заданном порядке (по дате создания, по имени файла и т.д.) и группирует отдельные образы страниц в документы. Отправляет тела документов в DIRECTUM, предоставляя информацию о полном пути к файлам документа. Служба поддерживает не только графические образы, но и изначально электронные документы и может работать с сетевыми сканерами (децентрализованное сканирование) – через папки файловой системы и FTP.

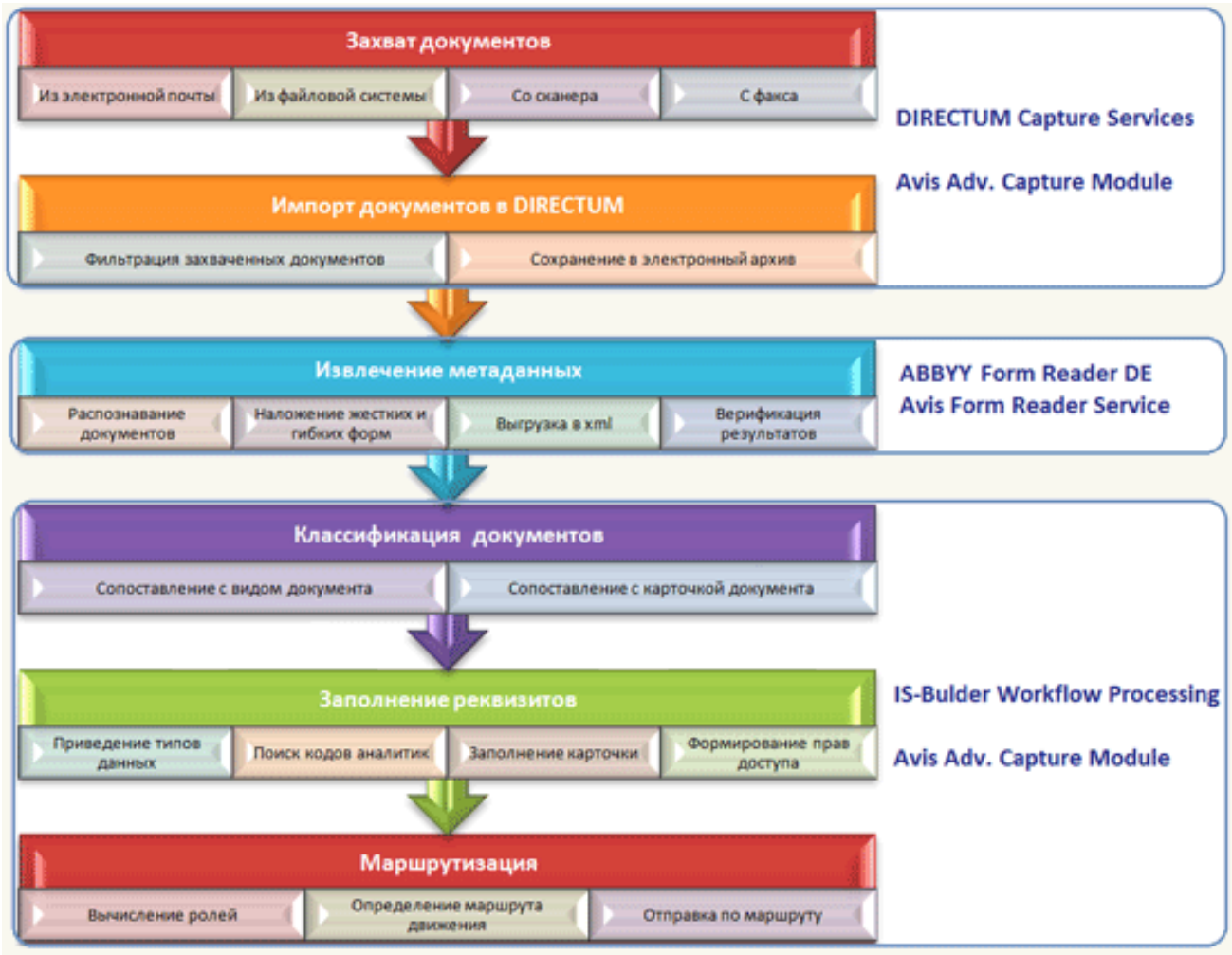


Рисунок 1.5 – Общая схема Capture and Transformation Services

Служба ввода с почтовых серверов

Служба загружает почтовые сообщения (текст письма и его вложения), фильтрует письма только по теме и адресату. Обработанные письма могут быть удалены из почтового ящика немедленно или спустя заданный промежуток времени.

В качестве недостатка данной службы можно отметить следующее: данное программное средство позволяет только захватывать документы с почтового сервера и передавать их в систему DIRECTUM без осуществления детектирования данных документов на нежелательные и легитимные письма. Т.е. письма проходят фильтрацию только средствами почтового сервера, что не предоставляет возможности анализа содержимого письма для исключения ложной идентификации легитимных сообщений. Так как любой входящий документ состоит из двух основных частей:

– служебная информация об отправителе и маршруте движения электронного документа (формальные признаки письма);

– текстовое содержимое письма (контент).

Для лучшего принятия решения о фильтрации входящей корреспонденции системе защиты необходимо проводить анализ обеих частей потому что служебная часть может быть подделана отправителем

Что особенно критично при условии мгновенного удаления письма из почтового сервера (без возможности его восстановления).

Принцип работы СУЭД DIRECTUM DCTS (Capture and Transformation Services) показан на рисунке 1.6.

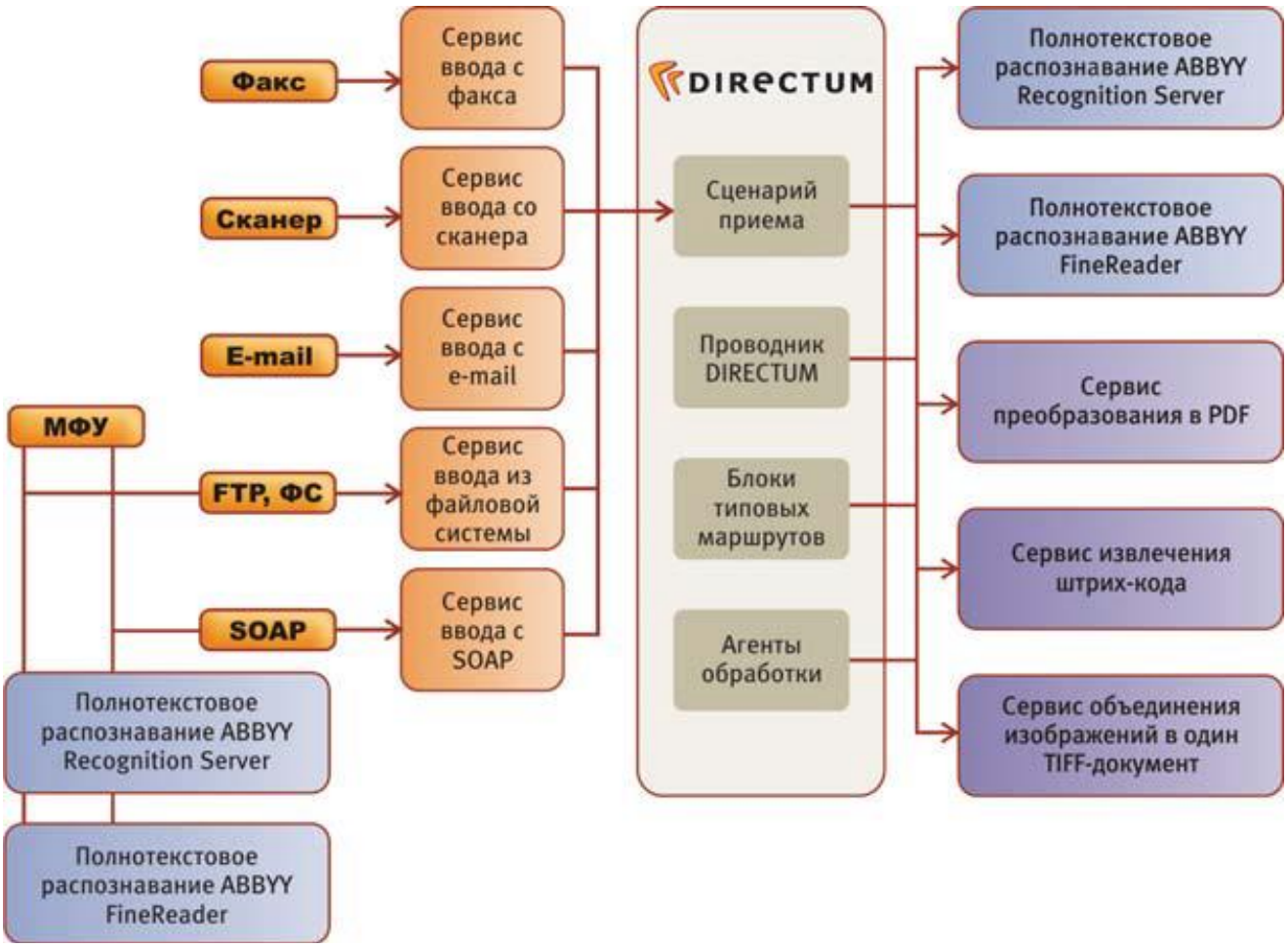


Рисунок 1.6 - Принцип работы Capture and Transformation Services

Таким образом, архитектура и функциональные возможности СЭДУВ DIRECTUM соответствуют современным требованиям к организации ИТКС, что позволяется обоснованно провести анализ особенностей информационных процессов в телекоммуникационных сетях с точки зрения информационной безопасности и обобщить его результаты.

1.2 Анализ проблем эксплуатации почтовых сервисов информационно телекоммуникационных систем корпоративных предприятий

Практически все бизнес-процессы в организации регламентированы нормативными документами (внутренними стандартами предприятия, положениями, уставами, приказами и т.д.), прохождение этих процессов должно быть в соответствии с требованиями Государственной системы документа-

ционного обеспечения управления (ГСДОУ), являющихся основой традиционной Российской технологии делопроизводства. В настоящее время организации используют системы электронного документооборота, которые позволяют автоматизировать данные информационные процессы. Современная ИТКС соответствует концепции ECM (Enterprise Content Management – расширенное управление предприятием) [80,81].

Архитектура и функциональные возможности СЭДУВ DIRECTUM соответствуют современным требованиям к организации ИТКС, что позволяется обоснованно провести анализ особенностей информационных процессов в телекоммуникационных сетях с точки зрения информационной безопасности и обобщить его результаты.

Анализ особенностей ИП систем электронной почты показал стремительный рост числа пользователей ЭП. Кроме того каждый из рассмотренных элементов, участвующих в процессе передачи ЭС может являться каналом для осуществления массовых несанкционированных рассылок (НР) сообщений электронной почты. Также анализ информационных потоков ЭП и мест расположения систем фильтрации показал необходимость разработки программного средства с учетом индивидуальных особенностей пользователя ЭП и специфики его деятельности, что позволит сделать процесс информационной безопасности ЭП более гибким а фильтрацию ЭПС персонифицированной.

Подводя итог анализу организации рассылки почтовых сообщений, а также мест расположения систем фильтрации ЭС, рассмотрение достоинств и недостатков существующих аналогов можно выделить следующие противоречия обеспечения информационной безопасности фильтрации почтовых сообщений в практике:

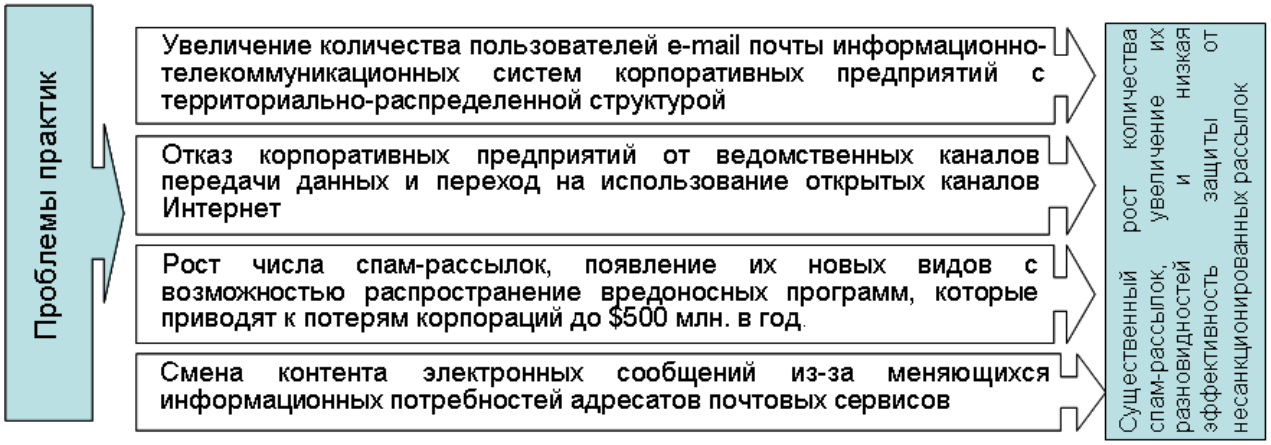


Рисунок 1.7 – Противоречия обеспечения защиты почтовых сообщений

Следовательно, основным противоречием в практике защиты электронной почты является существенно возросшее число спам рассылок и низкая эффективность обнаружений несанкционированных рассылок при относительно высокой ложной идентификации легитимных сообщений.

Указанные особенности фильтрации НЭС сообщений позволили определить **объект исследования** – защита почтовых сервисов информационно-телекоммуникационных систем от несанкционированных электронных почтовых сообщений.

1.3 Анализ аналогов систем защиты почтовых сервисов информационно телекоммуникационных систем корпоративных предприятий

Существует множество программных средств, как платных так и бесплатных или условно бесплатных, осуществляющих защиту от несанкционированных рассылок почтовых сервисов.

Одной из таких систем является система «Дозор - Джет» [33,62]. Система имеет следующую структуру:

- подсистема фильтрации;
- подсистема архивирования, реализованная на основе реляционной СУБД;

– модули, расширяющие возможности системы (в том числе модуль категоризации почтовых сообщений, который обеспечивает фильтрацию спама).

В состав системы «Дозор - Джет» входят утилиты при необходимости которые могут быть отключены администратором:

– проверка по RBL-спискам. В системе «Дозор-Джет» возможны «мягкие» настройки, учитывающие возможность того, что письмо, может быть занесено в черный список по ошибке;

– anti-spoofing – проверка подлинности адресов путем поиска соответствующей записи в DNS (или проверки существования такого домена в DNS);

– anti-relay – запрет вхождения и отправки писем, адреса которых отличны от внутренних.

Система «Дозор-Джет» подходит к фильтрации спама комплексно, используя статистические алгоритмы и технологию фильтрации на основе признаков спама. При этом фильтрация осуществляется двумя основными способами: по формальным признакам и по содержанию текстовой составляющей писем, то есть с помощью лингвистического метода.

Система «Дозор-Джет» функционирует на UNIX-платформе и включает в свой состав подсистему архивирования, реализованную на основе СУБД Oracle. Система «Дозор-Джет» используется в организациях, объем почтового трафика которых достигает 5 гигабайт в день, а количество почтовых адресов превышает 5000. Что требует применения аппаратных средств, которые способны обеспечить высокую производительность и отказоустойчивость системы.

AntiSPAMer 4.11 [62].

Программа AntiSPAMer 4.11 не требует установки, имеет возможность запуска выполняемого файла, и автоматической загрузки приложения при загрузке операционной системы. Для осуществления фильтрации элек-

тронных почтовых сообщений предусмотрена возможность создания «чёрных» и «белых» списков для сообщений.

Проверка сообщений происходит только по заголовкам, что позволяет удалять спам-рассылку на сервере. Система фильтрации AntiSPAMer 4.11 способна самостоятельно обучаться в соответствии с заданными правилами. Сообщения, прошедшие фильтрацию помещаются в один общий список, в котором сообщения помеченные как спам выделяются цветной подсветкой. Что является неудобным для пользователя.

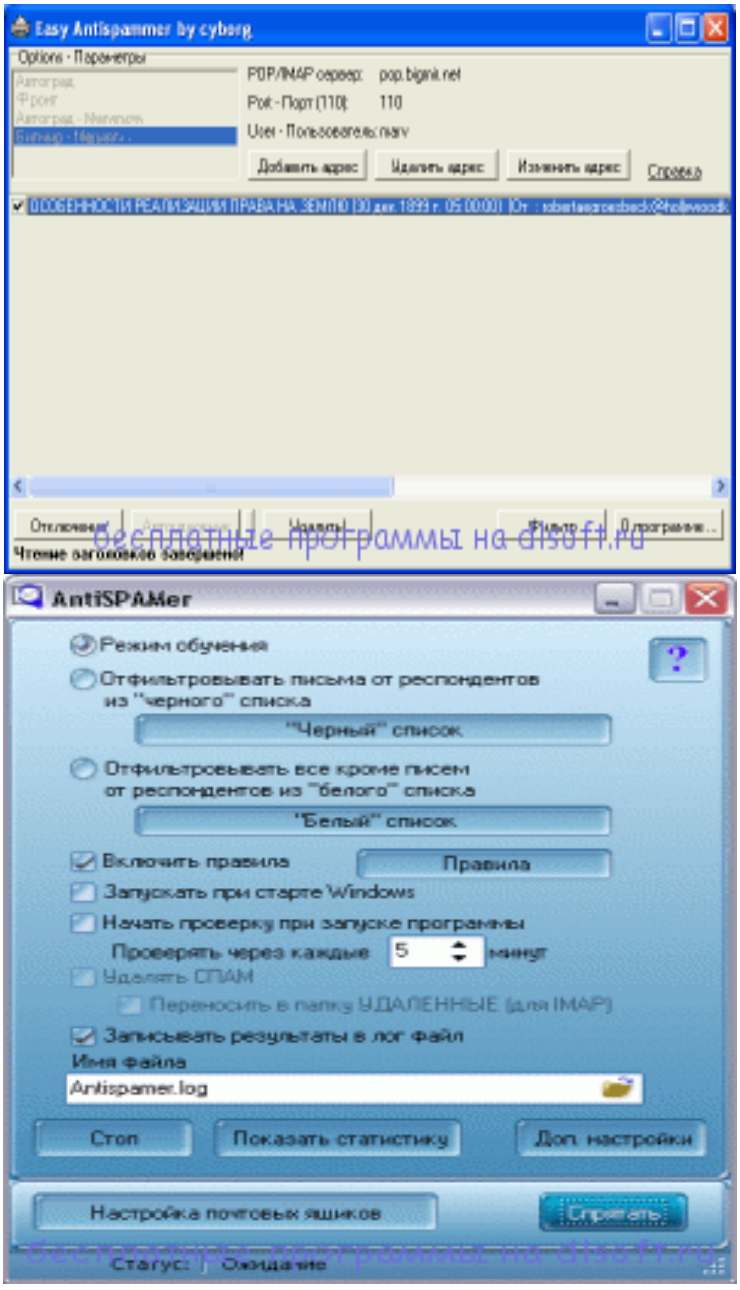


Рисунок 1.8 – Пример экранных форм программы AntiSPAMer 4.11

Недостатком данной системы фильтрации является необходимость длительного обучения фильтра для повышения эффективности. Кроме того проверка эффективности данной системы после первоначальной установки показала что только 25% из 400 спам рассылок были идентифицированы правильно [62]. Анализ правильно классифицируемых спам сообщений показал что в тематику «спам» попали сообщения имеющие кириллические заголовки, что говорит о требовательности и избирательности данной системы фильтрации. Однако, необходимо отметить что возможно, при дальнейшей работе и более тонкой настройке критериев отбора эффективность фильтрации ЭПС может быть выше.

Mailbox cleaner 1.5 [37,62].

Mailbox cleaner осуществляет проверку и блокирование несанкционированных электронных сообщений на почтовом сервере. Проверка ЭС осуществляется по заголовкам сообщений. Классифицированные сообщения распределяются по разным категориям, что позволяет быстро сориентироваться при большом объёме проверки.

Настройка системы фильтрации предусматривает возможность указания заголовков писем которые система фильтрации должна считать как спам сообщение. Но при таких настройках системы заблокированными могут быть легитимные письма от пользователей некорректно заполняющие заголовочные поля сообщений.

Недостатком системы Mailbox cleaner 1.5 является то что использование данной системы фильтрации требует постоянного контроля администратора безопасности для настройки системы в связи с тенденциями изменения спам сообщений.

Тестирование данной системы позволило осуществить идентификацию 55% -65% спам рассылок[62].

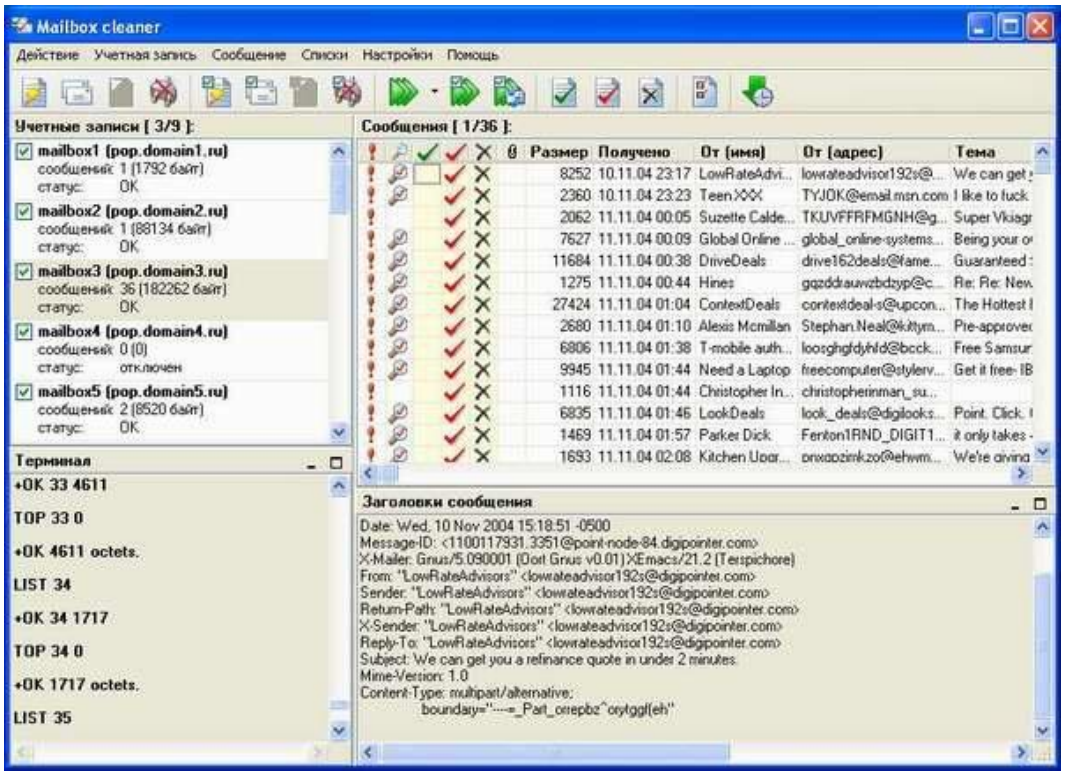


Рисунок 1.9 – Пример экранных форм программы Mailbox cleaner 1.5

Недостатком существующих систем фильтрации является не только низкое качество распознавания спама, но и блокировка легитимных сообщений[32,62,68].

Таким образом, анализ аналогов программных средств защиты от несанкционированных рассылок показал актуальность разработки программного средства защиты от НЭС, позволяющего уменьшить процент ложных срабатываний тем самым повысить его эффективность.

1.4 Анализ электронной почтовой корреспонденции

В настоящее время строго определения понятия несанкционированных рассылок электронных сообщений не существует. Известны следующие определения. Спам – незаконно распространяемая путем массовых рассылок информация рекламного характера, получение которой не согласованно с пользователем [47,32]. Лаборатория «Спамтест» понимает под спамом несанкционированные коммерческие рекламные рассылки, отвечающие требо-

ваниям массовости и анонимности [6]. В дальнейшем исследователи расширили границы этого понятия, приравнивая к спаму все виды неинформативных и нежелательных сообщений – автоответы почтовых роботов, письма с вирусами и т.п.

Но не все сообщения относятся к категории спам рассылки. Например, сообщения о всеобщей эвакуации, мобилизации граждан или надвигающихся стихийных бедствиях являются массовыми и незапрашиваемыми электронными сообщениями, но не являются спам-рассылкой.

В данной работе принята классификация входящей электронной почтовой корреспонденции показанная на рисунке 1.10

Рисунок 1.10 – Пример экранных форм программы Mailbox cleaner 1.5

Для выделения признаков несанкционированных рассылок целесообразно провести анализ существующих видов НЭС.

По данным Лаборатории Касперского в феврале 2013 года почтовый спам в интернете распределился по тематике следующим образом: 18,9 % — образование, 15,7 % — отдых и путешествия, 15,5 % — медикаменты, товары/услуги для здоровья, 9,2 % — компьютерное мошенничество, 6,5 % — компьютеры и интернет, 5,2 % — реплики элитных товаров, 4,1 % — реклама спамерских услуг, 2,7 % — для взрослых, 2,2 % — недвижимость, 2,2 % — юридические услуги, 1,9 % — личные финансы, 1,4 % — полиграфия.

Тематическое распределение спама показано на диаграмме рисунка 1.8:

Рисунок 1.8 – Тематическое распределение спам сообщений

Реклама

Некоторые легальные компании используют спам для рекламы своих товаров и услуг. Привлекательность данного метода рекламы состоит в его низкой стоимости и оповещении (предположительно) большого числа потенциальных клиентов. Впрочем, возможен обратный эффект, вызванный настороженностью получателей.

Реклама незаконной продукции

Спам так же используется для рекламы запрещенной продукции: порнографии, контрафактных товаров, программного обеспечения, лекарственных средств с ограниченной реализацией, неправомерно полученной закрытой информации (баз данных), а так же реклама услуг самих спам-рассылок.

Антиреклама

С помощью спама может распространяться информация, запрещенная на законодательном уровне, например, порочащая деятельность и продукцию конкурентов.

«Нигерийские письма»

Вид «Нигерийские письма» определяет сообщения, которые написаны от имени граждан государств, обладающих нестабильной экономикой. Автор

письма сообщает, что является обладателем миллионов долларов, хранящихся в обход закона, и поэтому не может разместить их в банке. Ему срочно необходим счет, на который можно перечислить незаконные деньги. Примером может служить письмо с информацией об аресте Михаила Ходорковского и нестабильной обстановке в компании «Юкос», в котором говорилось о конфиденциальном переводе десяти миллионов долларов и вознаграждении за оказанное содействие. После предоставления доверчивым пользователем автору письма доступа к счету деньги, хранящиеся на нем, исчезают.

Фишинг

«Фишинг» - вид спам сообщений предназначенный для получения личных данных пользователя ЭП, таких как номер кредитных карточек, номер счета в банке, паролей для доступа к счетами к системам онлайновых платежей. Данный вид спама обычно похоже на официальный запрос (письмо) либо от руководителя или администратора банка, либо администратора какой-нибудь другой коммерческой организации в которой у пользователя находятся или могут находиться счета. Сообщение несет в себе просьбу подтвердить сведения о себе(личные данные, номера счетов и пароли). Оформление письма-запроса почти полностью схоже со страницами и сообщениями от официального банка.

Так же существуют другие виды НЭС к которым относятся DDos-атаки, письма «счастья», Рассылка писем с информацией, за пересылку которой якобы осуществляется выплата семье пострадавшего некоторой суммы денег «на лечение». Цель подобной рассылки – сбор e-mail адресов, которые часто содержатся в тексте таких писем после многочисленных пересылок. При этом в числе очередных получателей может оказаться сам спамер.

Существуют типы массовых рассылок, не относящихся к категории спама, поскольку осуществляются неумышленно. Используя электронную почту, распространяют себя некоторые почтовые черви. После заражения очередного компьютера такой червь сканирует его в целях нахождения e-mail

адресов, которые будут использованы для дальнейшей рассылки почтового червя. Отвергнув письмо, почтовый сервер может уведомить отправителя о недоставке. Так как в спаме подделан адрес отправителя, пользователь, который не имеет отношения к рассылке, может получить множество уведомлений о недоставке (backscatter). Аналогично ведут себя некоторые спам-фильтры и антивирусы. Впрочем, подобные случаи очень редки и встречаются только у старых программ.

Результаты проведенного анализа можно обобщить и представить в виде таблицы 1.1.

Таблица 1.1 – Характеристики легитимной почты и спам сообщений

Спам сообщения	**Легитимная почта**
Получатель, как правило, не может отписаться от рассылки;	Существует возможность отписаться от рассылки;
Высокая периодичность сообщений;	Периодичность рассылки довольно низкая;
	Письма персонализированные (есть обращение либо к конкретному сотруднику отдела, либо обращение с указанием названия отдела). В легитимных письмах (особенно деловых) отправитель всегда подписываться, с указанием либо личности (Ф.И.О.) либо названия отдела или организации.
Информация, как правило, не полезна для получателя;	Полезная, актуальная и качественная информация;
Сообщения имеют массовый характер (списки получателей рассылок могут насчитывать миллионы получателей);	Сообщения тоже могут иметь массовый характер, но число получателей всё равно ниже;
Часто подделывается адрес отправителя спам-письма;	
Определенный набор слов и словосочетаний	

Таким образом, анализ видов несанкционированных рассылок позволил выделить признаки электронных почтовых сообщений, необходимые для идентификации несанкционированных электронных рассылок

1.5 Концептуальная постановка задачи исследований и её формализация

1.5.1 Анализ методов борьбы с несанкционированными рассылками

В рамках концепции ИБ систем почтового сервиса от несанкционированных воздействий можно выделить следующие классы мер противодействия спаму [25,29,47,48,63,85,86], показанные на рисунке 1.11

1) Нормативно – правовые или юридические.
2) Организационные мероприятия.
3) Программно-технические.

Рисунок 1.11 – Методы борьбы со спамом

Юридические меры

Решением проблемы несанкционированной рассылки могли бы стать соответствующие поправки в законодательство, которые наказывали бы провайдеров, способствующих распространению нежелательной рекламы. Надо отметить, что спам является достаточно доходным бизнесом. По данным содержащимся в отчете UNESCO IFAP Russia в рамках проекта "АнтиСпам российские спамеры зарабатывают порядка 55 млн руб. (2 млн долл.) в год [55,66,74]..

Организационные меры

К организационным мерам относятся разработка и внедрение в организации политики использования электронной почты [29]. Политика использования электронной почты это закрепленные в письменном виде и доведенные до сотрудников инструкции и другие документы, регламентирующие их

деятельность и процессы, связанные с использованием системы электронной почты.

Политика использования электронной почты должна соответствовать критериям:

– доступно и понятно изложенной для *всех* сотрудников организации или компании. При этом понятность и доступность изложения не должно привести к потери юридического статуса документа;

– как документ, должна быть одобрена и подписана соответствующими должностными лицами организации, т.е. иметь законную силу;

– не противоречить федеральным и местным законам, а также внутри корпоративным законам;

– регламентировать меры воздействий на сотрудников, не придерживающихся установленной политики;

– соблюдать баланс между степенью защищенности информации и продуктивностью деятельности организации;

– детально определять мероприятия по обеспечению политики использования электронной почты в компании.

– исходить из необходимости защиты информации в процессе экономической деятельности компании.

Технические методы борьбы с несанкционированными рассылками [32,36,38,47,48,59,63,68,88] в электронной почте представлены на следующем рисунке.

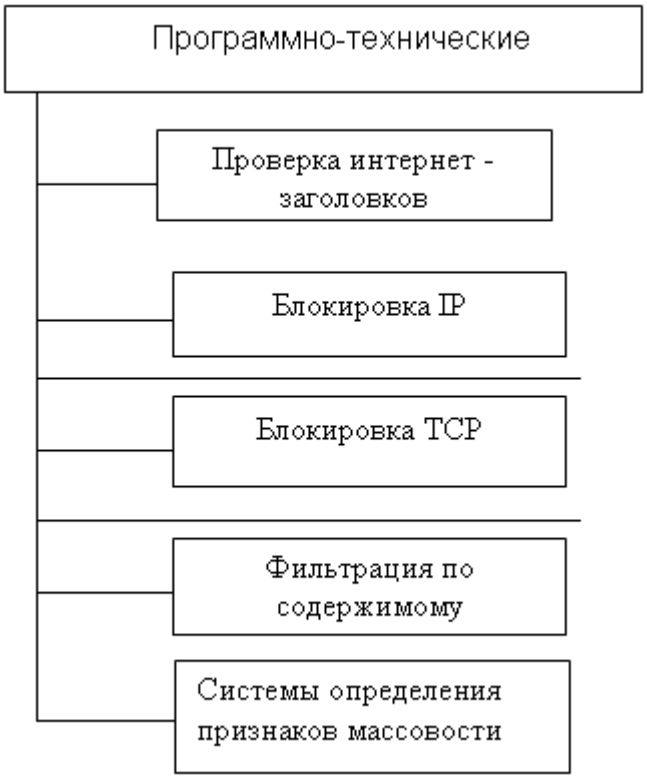


Рисунок 1.12 - Классификация технических методов

Блокировка IP

Блокировка IP осуществляется в соответствии с серыми и черными списками. Серые списки – временный отказ в приеме сообщений. Принимающий письмо сервер может сообщить о временной ошибке, т.е. в данный момент он не может принять письмо. Причины – недостаток дискового пространства, чрезмерная нагрузка и т.п. Но при изменении ситуации письмо будет принято.

При попытке послать это же письмо еще раз (что свойственно протоколу SMTP) использующий серые списки сервер обнаружит в своей базе данных соответствующую запись, и письмо будет принято. Увеличение надежности метода достигается наложением дополнительного ограничения: наличие определенного временного промежутка после первой попытки. В дальнейшем, письма, посланные тем же отправителем тому же получателю через тот же сервер, будут приниматься без задержки, поскольку в базе данных уже существует необходимая запись.

К достоинствам данного метода относятся как высокая надежность и малое количество ложных срабатываний, так и легкость реализации, почти не использующей ресурсы сервера. Недостатком является задержка доставки почты. Кроме того, злоумышленники могут повторно отправить спам-сообщение. В таком случае эффективность данного метода низкая.

Черные списки – антиспам-технология, являющаяся одной из самых старых. Она подразумевает блокировку почты, идущей с IP-адресов, которые перечислены в списке. Данные списки строятся на основе DNS. Данная технология позволяет осуществлять однозначную блокировку, либо использовать взвешенный подход. При однозначной блокировке отклоняются все сообщения, пришедшие с IP-адресов, находящихся в DNSBL. При взвешенном подходе такие сообщения не блокируются, но данный факт учитывается в процессе классификации «спамности» письма.

При использовании первого метода однозначно отклоняются все письма, которые пришли с IP-адресов, попавших в DNSBL, без учета ошибочности занесения IP в черный список. Второй подход хорошо иллюстрируется свободным спам-фильтром spamassassin. Когда классификация сообщения производится на основе анализа по множеству критериев. В этом случае наличие IP-адреса в черном списке не является единственным определяющим фактором определения класса сообщения. Что снижает количество ложных срабатываний фильтра при случайном попадании IP-адреса в черный список.

Применение технологии DNSBL в соответствии с первым принципом на данный момент не приносит больших проблем. Но использование большого числа блокирующих серверов все же нежелательно, поскольку это может сказаться на доставке чистых писем и производительности почтового сервера.

Достоинством является полное исключение почты из подозрительного источника на основе черного списка. Основной недостаток: такие списки приводят к высокому уровню ложных срабатываний, из-за чего применять их необходимо с осторожностью.

Контроль массовости

Контроль массовости позволяет выявлять в потоке почты массовые сообщения, абсолютно идентичные или незначительно различающиеся. Построение работоспособного «массового» анализатора предполагает наличие огромных потоков почты, поэтому данная технология предлагается крупными производителями, которые способны подвергнуть анализу значительные объемы почты. Достоинства: гарантированность определения массовой рассылки при срабатывании технологии. Недостатки: отнесение легитимной «большой» рассылки к категории спама, возможность взлома подобной защиты с помощью интеллектуальных технологий: генерации разного контента (текста, графики и т.п.) в каждом спамерском письме.

Проверка интернет-заголовков сообщений

Для генерации и мгновенного распространения спамерских сообщений существуют специальные программы. Впрочем, такой спам не всегда соответствует требованиям почтового стандарта RFC, поскольку содержит ошибки в заголовках, что позволяет классифицировать подобные сообщения как спам. Достоинства: надежность и прозрачность процесса распознавания спама и его фильтрации. Недостатки: данная технология позволяет обнаружить не более трети всего спама. Кроме того, анализ спамерами существующих заголовков приводит к уменьшению ошибок в заголовках спама.

Фильтрация

Методы фильтрации в общем случае можно разделить на две группы:

– фильтрация по формальным признакам основанная на фильтрации по признакам спама, и фильтрации по спискам (почтовых адресов, IP-адресов);

– фильтрация по лингвистическим признакам включает в себя распознавание по содержанию письма (словосочетания, эвристики, статистика) и распознавание по образцам писем (распознавание по сигнатурам, с голосованием и пр.) [11].

Существуют фильтры осуществляющие централизованную и распределенную фильтрацию. Можно выделить следующие достоинства фильтров осуществляющие централизованную фильтрацию: это возможность быстрого обнаружения и отсечения массовых рассылок одинаковых сообщений, а так же возможность оперативной адаптации к вновь разработанным методам обхода известных способов фильтрации, используемых в спам – фильтрах. К недостаткам таких фильтров можно отнести невозможность учета интересов конкретного пользователя, что приводит к принятию ошибочного решения о пропуске спама или блокировании полезного письма. В случае распределенной фильтрации учитывается область интересов пользователя, но не используются интересы других пользователей организации.

В общем виде можно выделить два вида фильтров

– фильтры работающие на основе поиска в письме признаков «спама»

– фильтры на основе статистических (вероятностные) методов детектирования несанкционированных рассылок.

Эти средства осуществляют контекстную фильтрацию общений, т.е. по содержимому письма. Все традиционные фильтры обладают следующим недостатком – они учитывают признаки «спама» в письме, но не берут во внимание признаки «легитимности» сообщения характерные для деловой переписки.

Системы с запросом на подтверждение

На e-mail отправляется запрос, позволяющий убедиться в том, что ххх использование данного способа защиты от НЭС потребует от отпарвителя выполнение действий связанных с подтверждением того что он является человеком, а не программой для массовой рассылки сообщений. Однако данные системы не могут отличить подобные программы от пользователя осуществляющего например рассылку приглашений на научно практическую конференцию. Кроме того, подобные подтверждения по тем или иным причинам не всегда возможны и удобны для пользователей (отпарвителей).

Способ временных адресов заключается в том что бы использовать временные адреса для почтовых ящиков. В том случае если на данный потовый ящик приходит большое количество несанкционированных сообщений, то работа с данным e-mail прекращается, что может являться не совсем удобным для пользователя данного почтового ящика, и недопустимо для осуществления бизнес процессов в организациях и корпорациях.

Голосование пользователей

Данный способ заключается в том, что любой пользователь обнаружив в почтовом ящике спам сообщение информирует систему ЭП тем самым накапливая статистику для соответствующего сообщения. При набранном количестве уведомлений от пользователей превышающий установленный порог образец данного сообщения помещается в базу для последующего сравнения с входящими сообщениями. Точность данного метода зависит от голосования пользователей что можно выделить в качестве недостатка данного метода. Указанная особенность делает данный метод не применимым для фильтрации на корпоративном уровне.

Распознавание на основе сигнатур

Метод, позволяющий по каждому спам сообщению создавать сигнатуры (это могут быть лексические, семантические или графические сигнатуры) позволяющие распознавать спам сообщения. Использование сигнатур позволяет идентифицировать сообщение с небольшими изменениями в содержимом. Данные методы не способны к детектированию новых сообщений, сигнатуры которых не были представлены базе, кроме того средства на основе данного метода требуют постоянного обновления. Часто измененные сообщения дают ложные срабатывания.

Таким образом, определены достоинства и недостатки существующих методов борьбы с несанкционированными рассылками.

1.5.2. Анализ методов классификации

Принадлежность электронного сообщения к тому или иному классу определяется используемым для фильтрации методом классификации. Методы можно классифицировать по схеме принятия решения о наличии факта нарушения [9,16,46,58,72,85,86] классификация рассматриваемых методов представлена на рисунке 1.13.

Рисунок 1.13 – Классификации методов по схеме принятия решения

Структурные методы распознавания формируют строгую модель либо заведомо корректного состояния или воздействия, либо заведомо злоумышленного воздействия. К преимуществам методов данного класса относится полное отсутствие ложных срабатываний в области, описываемой моделью корректного состояния или злоумышленного воздействия.

Недостатком данных методов является принципиальная невозможность описания новых, неизвестных ранее, либо не укладывающихся в разработанную модель злоумышленных воздействий, а несанкционированные рассылки и легитимные сообщения постоянно изменяются.

Корреляционные методы вводят метрики отличия наблюдаемого вектора признаков, либо более сложной (например, поведенческой) характеристики от заведомо корректного, либо заведомо злоумышленного состояния. Преимуществом корреляционных методов является покрытие всего множества допустимых воздействий, что гипотетически позволяет принимать корректные решения и в отношении неизвестных ранее атак.

Корреляционные методы подразделяются на алгоритмы «без памяти» и алгоритмы «с памятью».

Алгоритмы "с памятью" анализируют события с учетом некоторой предыстории, а также, возможно, истинного или предполагаемого состояния системы.

Детерминированные алгоритмы контроля поведения генерируют события по любому факту отклонения поведения системы от профиля, созданного на этапе обучения.

Нечеткие алгоритмы контроля поведения вычисляют в ходе анализа последовательности событий тем или иным образом вектор вероятностных характеристик и генерируют событие только по превышению им некоторых пороговых значений.

Алгоритмы "без памяти" рассматривают каждое событие как отдельный элемент множества, в отношении которого необходимо принять решение. Данный класс также называют *методами пространства признаков.*

Существуют методы с одномерным вектором признаков, которые включают в себя пороговые алгоритмы, генерирующие информационное событие о факте обнаружения аномалии по превышению наблюдаемого значения некоторой граничной величины.

Методы с многомерным вектором признаков делятся на алгоритмы линейной классификации (уступившим позиции алгоритмам кластерного и нейросетевого анализа как более гибким), кластерный анализ, нейросетевые методы и иммунные методы.

Кластерный анализ как зарекомендовавший себя метод классификации позволяет производить обнаружения компьютерных вторжений как в направлении обнаружения «без учителя», так и кластеризации с предварительным обучением на размеченных входных данных.

Иммунные методы строятся по аналогии с иммунной системой живого организма и предпринимают попытку распространить принципы обнаружения противодействия иммунной системы живых существ чужеродным вирусам.

Нейросетевые методы используют для принятия решения о наличии либо отсутствии злоумышленного воздействия на базе нейронной сети.

Наиболее важное преимущество нейросетей заключается в их способности изучать характеристики умышленных атак и идентифицировать элементы, которые не похожи на те, что наблюдались в сети раньше.

Анализ методов показал, что наиболее перспективным направлением исследований в области фильтрации НЭС являются нейросетевые методы, основными достоинствами которых являются: возможность анализа данных в условиях неполноты, искаженности и неточности информации; работа в режиме реального времени; независимость объема вычислений от числа объектов обнаружения и идентификации.

Подводя итог анализу методов защиты от несанкционированных рассылок почтовых сообщений можно выделить следующие проблемы обеспечения информационной безопасности фильтрации почтовых сообщений в теории:

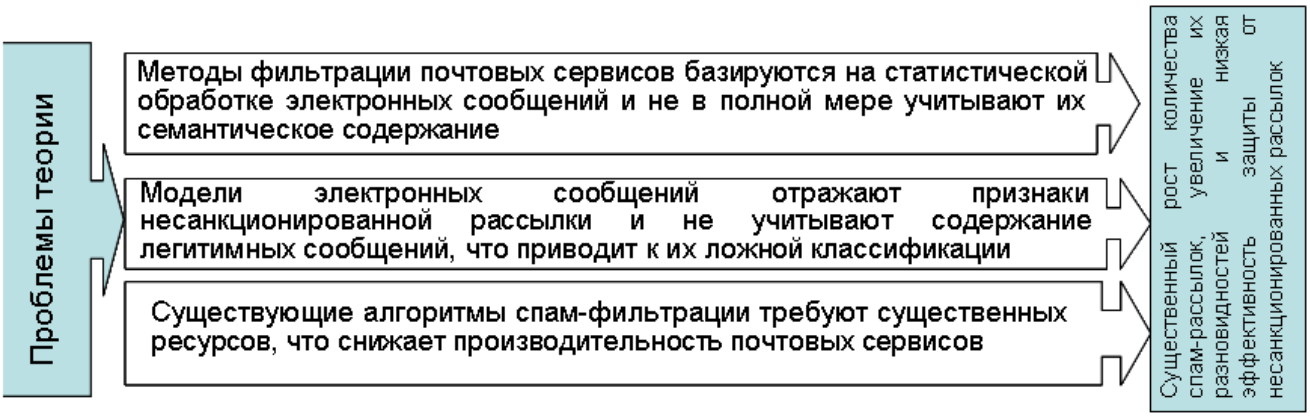


Рисунок 1.14 - Проблемы обеспечения ИБ почтовых сообщений

Таким образом, системный анализ проблем обеспечения ИБ почтовых сервисов ИТКС позволил выявить основное противоречие между существенно возросшей интенсивностью спам-рассылок и высоким уровнем ложной классификации легитимной почтовой корреспонденции в силу отсутствия методов, учитывающих признаки легитимности сообщений при изменяющихся информационных потребностях адресатов.

Снизить процент ложных срабатываний предлагается и за счет разработки двухуровневой системы фильтрации, представленной на рисунке 1.15, и состоящей из формального и интеллектуального фильтров.

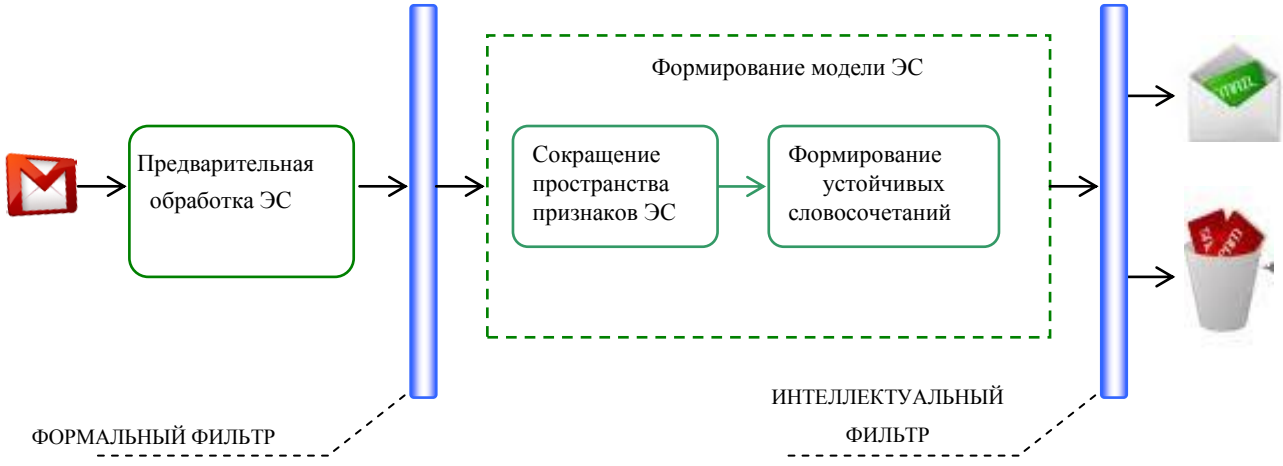


Рисунок 1.15 – Предложенная технология фильтрации ЭПС

Предварительная обработка ЭПС заключается в приведение к стандартному типу кодировки, удаление стоп-слов, гиперссылок и знаков пунктуации.

Формальный фильтр использует адреса (IP, e-mail), разделяющие ЭПС на разрешенные и запрещенные, и представляет собой базу данных признаков ЭПС, формируемую администратором.

Интеллектуальный фильтр осуществляет семантическую классификацию ЭПС конкретного адресата электронной почты, что требует предварительного обучения классификатора.

Определена и формализована задача фильтрации почтовой корреспонденции корпоративного предприятия.

Пусть $L \in \{Let_i\}$ множество писем (ЭПС), предназначенных для обучения классификатора. Модель L характеризуется пространством признаков $P=(p_1,p_2,p_3,....p_l)$, где p_l – значение l-го признака ЭС. А – алгоритм классификации, относящий L к одному из классов $K \in \{k_1,k_2\}$, (spam/legitim).

Задача фильтрации заключается в построении такого решающего правила, при котором классификация проводится с минимальным числом ошибок $R_{I,II}$ в реальном масштабе времени.

Тогда процедура автоматической фильтрации P_f L ЭПС на множестве классов K примет вид целевой функции

$$R(L(p_i),A(k_j)) \xrightarrow{P_f} \min$$

Таким образом, определена и формализована задача фильтрации почтовой корреспонденции корпоративного предприятия.

2 Моделирование контента электронной почты

Определение контент-ана́лиза «контент-ана́лиз (от англ.: contents - *содержание, содержимое*) или анализ содержания, - стандартная методика исследования в области общественных наук, предметом анализа которой является содержание текстовых массивов и продуктов коммуникативной корреспонденции». Согласно определению выделяют количественный и качественный контент-анализ текстов для следующей их обработки, анализа и определения числовых закономерностей [1,35,88].

2.1 Исследование моделей описания текстового контента электронных сообщений

Модель представления текстового документа является важной составляющей в процессе его компьютерной обработки. Выбор модели определяет эффективность выделения смыслового содержания и структуры электронного сообщения. В рамках данных исследований под моделью текста понимается его приближённое описание, выраженное с помощью математической символики. Модель всегда проще самого текста, отражает лишь некоторые его свойства, стремиться выделить главное, не отвлекаясь на детали. Наиболее распространенной моделью является представление текста в виде набора числовых признаков, т. к. подавляющее большинство алгоритмов обработки текста является числовыми [9,17,28], т. е. использующими для работы не непосредственно слова и фразы, составляющие текст, а числовые характеристики документов.

Количественный контент-анализ (содержательный) предназначен для анализа отдельных слов, словосочетаний, предложений сообщения подвергая анализу и обработке содержание сообщения.

Качественный контент-анализ (структурный), позволяет определять не ***что*** говориться в тексте, т.е. его содержание, а ***как*** отражается этот объект в тексте, не уделяя внимания анализу самого содержания.

2.1.1 Векторная модель

Текст электронного почтового сообщения можно представить в виде термов (слов), что позволяет представить каждый документ в виде вектора в пространстве признаков. Графическое отображение документа D состоящее из трех термов t_1 t_2 t_3 представлено на рисунке 2.1.

Трехмерное отображение документа, состоящее из трех термов может быть распространенно и на N-мерные документы, где N – количество термов в документе.

В рамках векторной модели ЭПС описывается вектором в некотором пространстве признаков, в котором каждому используемому в сообщении терму ставится в соответствие его вес (значимость)

$$d_j = (w_{1j}, \ldots, w_{|T|j}), \qquad (2.1)$$

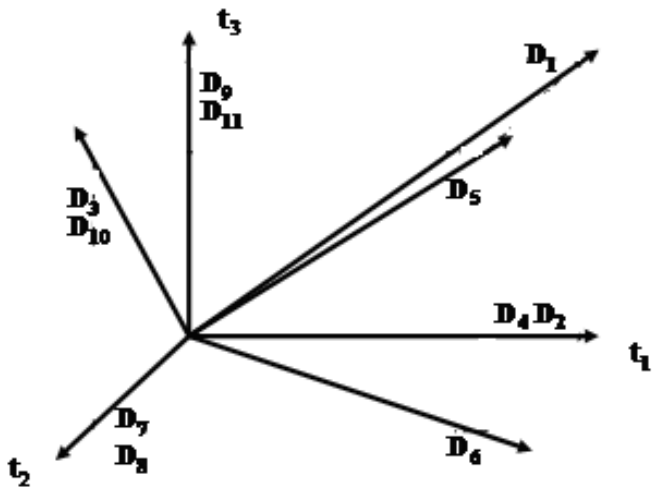


Рисунок 2.1– Пример графического отображения документа

В матричном виде ЭПС можно представить в виде матрицы столбец S_i элементами которой, являются веса соответствующих термов в сообщении.

$$S_i = \begin{bmatrix} w_{1j} \\ \vdots \\ w_{ij} \\ \vdots \\ w_{Mj} \end{bmatrix} \tag{2.2}$$

где w_{ij} – вес терма j в сообщении i

M – число термов(слов) в сообщении

Если одно ЭС можно представить в виде 2.2, то вся коллекция сообщений примет вид матрицы столбцами которой будут являться письма, а строками термы содержащиеся в данных письмах.

$$Let = \begin{bmatrix} w_{11} & \cdots & w_{1j} & \cdots & w_{1N} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ w_{i1} & \cdots & w_{ij} & \cdots & w_{iN} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ w_{M1} & \cdots & w_{Mj} & \cdots & w_{MN} \end{bmatrix},$$

где

M – количество писем в выборке,

N – число термов(слов) в коллекции после удаления стоп-слов,

ω_{ij} – вес терма i в сообщении j;

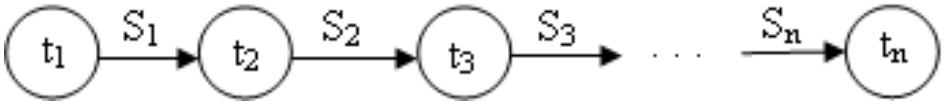
$i = 1,...,M$, $j = 1,...,N$.

Основным достоинством векторной модели является возможность использования алгоритмов классификации основанных на анализе статистических характеристик, а так же возможность сравнивать вектора в векторном пространстве признаков. Задача преобразования текста в вектор в пространстве признаков требует определения координаты признаков. Самым распространенным методом определения значимости терма является логическое взвешивание, заключающееся в том, что терму присваивается значение «1», если он встречается в сообщении и «0» - в случае, если терм в сообщении не встречается.

Основным недостатком данной меры является то, что такой подход не учитывает частоту встречаемости отдельного терма во всей коллекции документов. Таким образом, при оценке близости векторов с использованием данной мерой значимости термов результаты могут быть не всегда удовлетворительными[104,107].

2.1.2 Модель представления текста на основе графа

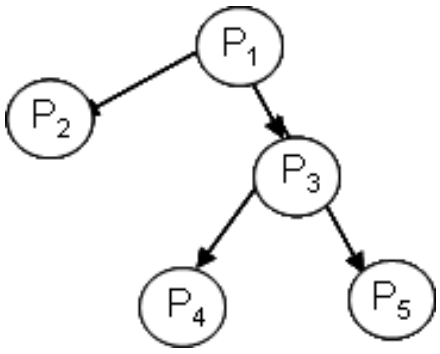
Графовая модель представления текста заключается в выделении объектов которые являются вершинами графа, и (связи) отношений между данными объектами, т.е. ребрами графа. В качестве объектов могут выступать как отдельные слова так и понятия. Графическое отображение модели представления текста показано на рисунке 2.2



t_1, t_n – слова в тексте (вершины графа)
S_1, S_n – расстояние между словами (дуги графа)

Рисунок 2.2 – Представление текста на основе графа

Использование понятий в качестве вершин графа требует ведения понятийных словарей. Вид графа в таком случае принимает вид показанный на рисунке 2.3



P_n – понятия в тексте
O_n – отношение между понятиями

Рисунок 2.3 – Представление текста на основе графа (дерево зависимостей)

На основе графовой модели часто используют семантические сети представление текста имеющие вид ориентированного графа вершинами которого являются объекты предметной области, а ребра определяют отношение между ними[51,67,83]. Таким образом, семантическая сеть отражает семантику предметной области в виде понятий и отношений. Семантический анализ позволяет построить семантическую структуру как одного предложения, так и всего текста.

В качестве вершин в семантических сетях обычно используют понятия базы знаний, отношение между этими понятиями отображаются ребрами. В более сложном представлении текста используют грамматические связи между словами или понятиям [21]. В результате чего получается граф грамматических связей. Обычно такое представление используют при синтаксическом анализе текста.

Выбор конкретной формы графа определяется методом, применяемым для последующей обработки, основной целью, и требуемым временем на обработку графа [51].

В рамках исследований был проведен сравнительный анализ векторного представления текста и описания на основе графа. Для проведения исследований было взято 100 сообщений (исходя из того что 50 спам сообщений и 50 легитимных сообщений учавствует в обучающей выборке системы фильтрации для одного пользователя). Анализировались сообщения количество термов которых составляет от 0 до 20, от 20 до 50, от 50 до 70, от 70 до 100, и свыше 100 слов в сообщении. Первый параметр который определялся это затраченное время на обработку – для векторной и графовой модели в качестве веса термов принято логическое взвешивание которое заключается в присвоении весу значения 1, если слово встречается в документе и 0 в противном случае:

$$w_{ji} = \begin{cases} 1, & f_{ji} > 0 \\ 0 & f_{ji} = 0 \end{cases}, \qquad (2.4)$$

Так как графовая модель представления текста дополнительно требует определение связей между термами, то для определение связей использовалась мера Дайса[44].

2.1.3 Оценка вычислительных ресурсов основных моделей

Результаты сравнения моделей (ограничения рассматриваемых моделей по времени обработки (сек), по размерности матрицы (терм*документ), по размерности матрицы (терм*документ)) представлены на рисунке 2.3 – 2.5.

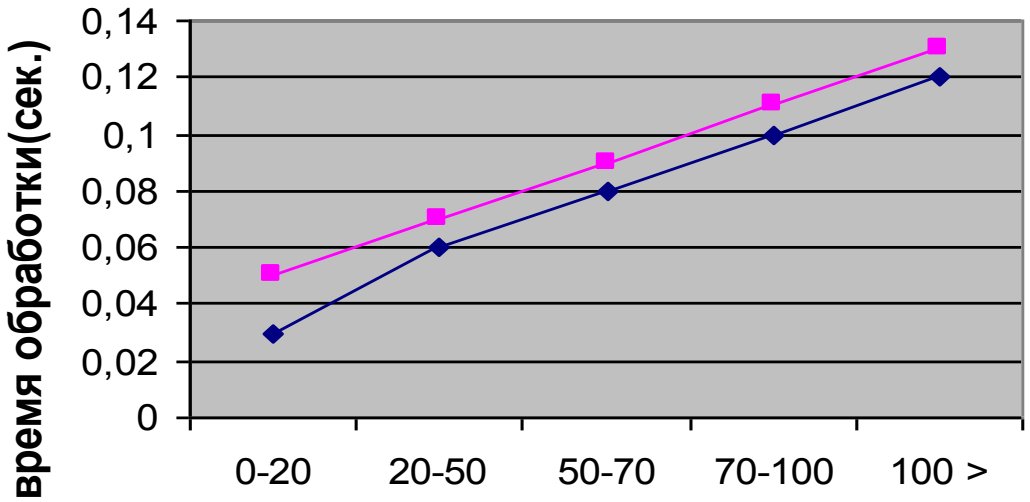


Рисунок 2.3 – Ограничения рассматриваемых моделей по времени обработки (сек)

Векторная модель позволяет определить значимость коллекции сообщений в диапазоне от 0,03 до 0,1секунды. Графовая модель на определение веса терма и нахождение связей между ними затрачивает от 0,05 до 0,13 секунды.

Второй параметр по которому происходило сравнение это размерность полученной матрицы. При таком сравнении векторная модель и модель на основе графа показали практически идентичные результаты. При векторном представлении в сообщениях, в которых количество термов порядка двадцати размерность составляет 3250 термов на 100 сообщений, и при

увеличении числа термов в сообщениях она достигает 12000 термов для исследуемых сообщений. Результат сравнения представлен на рисунке 2.4.

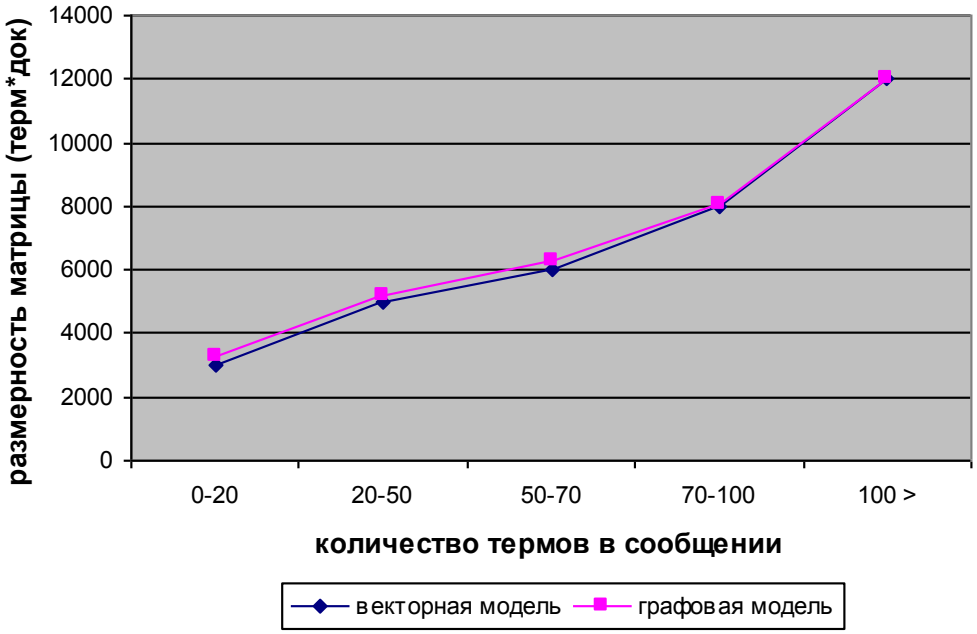


Рисунок 2.4 – Ограничения рассматриваемых моделей по размерности матрицы (терм*документ)

Анализ размерности базы термов занимаемой на жестком диске показал (рисунок 2.5), что векторная модель занимает 124 Кб при количестве слов в сообщении до 20, и возрастает до 385Кб при увеличении числа термов до 100. Графовая модель при количестве слов в сообщении до 20 занимает 155 Кб, и возрастает до 475Кб при увеличении числа термов до 100.

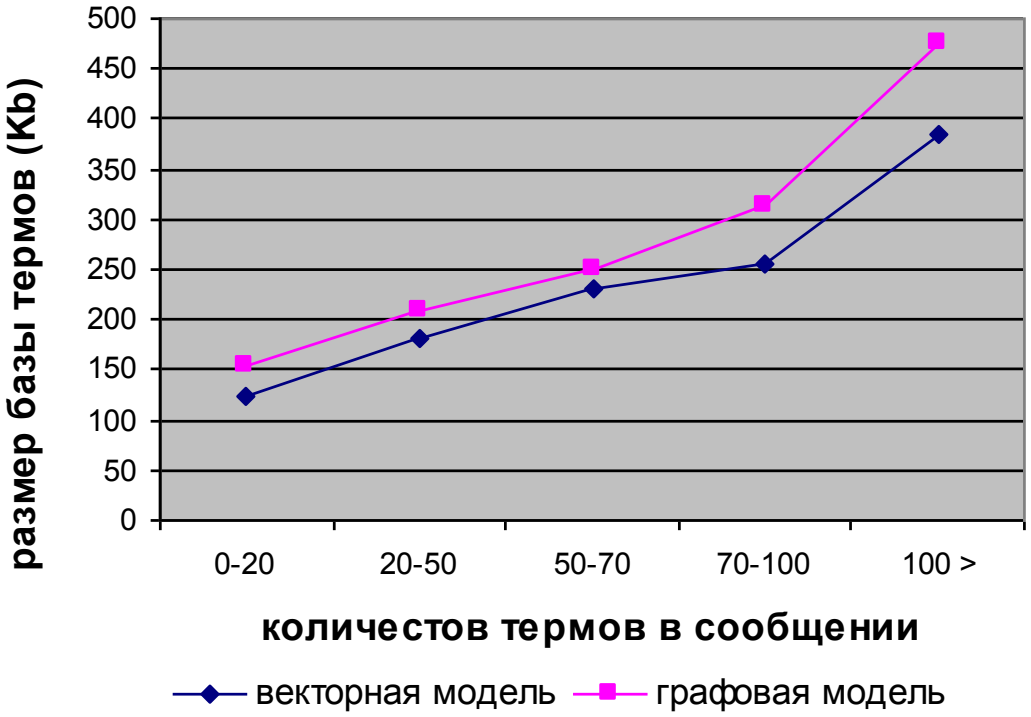


Рисунок 2.5 – Ограничения рассматриваемых моделей по размерности матрицы (терм*документ)

Таким образом, в результате проведенного анализа ограничений рассмотренных моделей определенно что модель на основе графового описания текста требует дополнительных ресурсных и временных затрат на установление связей между термами. Кроме того в случае более глубокого анализа текста при использовании графовой модели в том случае если в качестве вершин используются понятия слов, или при необходимости выделения частей речи, падежей временные и ресурсные затраты возрастают.

2.2 Развитие векторной модели электронного сообщения для задачи классификации

При использовании векторной модели содержание ЭПС можно описать с помощью термов t, множество которых будет образовывать тезаурус $T\{t_1,....t_q\}$ определенного класса k. В качестве термов используются слова, составляющие содержание сообщения. Анализ пространства признаков $S(p_i)$

представленных на рисунке 2.6, позволил выбрать в качестве признака вес терма.

Тогда модель ЭПС можно представить в виде 2.3

$$S(p_i)=<t_j,w(t_j)> \quad (2.3)$$

где t – j-ый терм в сообщении;

p_i – пространство признаков, определяющих сообщение;

$w(t_j)$ – вес терма в сообщении после удаления стоп-слов.

Если одно электронное сообщение можно представить в виде 2.3, то всю коллекция сообщений определенного класса L_k примет вид 2.4

$$L_k=<T_k ,w(t_j)> \quad (2.4)$$

где T_k – k-ый тезаурус сообщения класса k;

$w(t_j)$ – вес терма в сообщении.

Рисунок 2.6 – Пространство признаков ЭС

Базовый подход к построению векторного представления текста при использовании логической меры значимости термов обладает следующим недостатком: при оценке близости векторов с использованием данной меры значимости термов результаты могут быть не всегда однозначными[99,102,104,107]. Кроме того на результат классификации также может влиять изменение количества термов в сообщении. Преодоление данного недостатка возможно при изменении способа определения значимости терма. Следовательно, одной из основных задач при работе с текстовым содержимом ЭПС становится вычисление весового

коэффициента w_{iq}, определяющего значимость соответствующего терма t_j в i-ом документе.

Существуют несколько различных мер определения значимости термов [92,99,100,102,104] (рисунок 2.5).

Рисунок 2.5 – Существующие меры взвешивания термов

Примем за f_{ij} - частоту терма t_j в сообщении S_i, N – число сообщений в выборке определенного класса, M – число термов в сообщениях класса k после удаления стоп-слов, n_j - общее количество сообщений, содержащих терм t_j.

Логическое взвешивание.

Данная мера взвешивания основана на то что присвоить терму занчение 1 в том случае если он встречается в документе и 0 в противном случае.

$$w_{ji} = \begin{cases} 1, & f_{ji} > 0 \\ 0 & f_{ji} = 0 \end{cases}, \tag{2.5}$$

Основным достоинством данного метода является простота реализации и использования. Однако, в качестве недостатка можно отметить что при таком подходе никаким образом не учитывается важная информация о частоте встречаемости термина и отсутствует выделение информативных терминов.

TF – взвешивание (term freguency)

Другим простым подходом является использование частоты слова в документе

$$w_{ij} = f_{ij}, \tag{2.6}$$

TF – взвешивание выделяет в качестве информативных часто встречающиеся термины.

Использование частоты слова дает примерно 25% увеличение эффективности классификации по сравнению с логическим взвешиванием.

TF - IDF взвешивание. Предыдущие два метода не используют частоту встречаемости термина во всех документах коллекции. TF-IDF- взвешивание присваивает вес слову j в документе *i* пропорционально числу вхождений слова в документ, и обратно пропорционально числу документов в коллекции, в которые слово входит, по крайней мере, однажды. Логарифм в формуле используется для уменьшения веса часто встречающихся терминов и увеличения веса терминов которые встречаются редко, Таким образом в TE-IDF мере взвешивания происходит выделение средне и низкочастотных терминов.

$$w_{ij} = f_{ij} \quad \log\left(\frac{N}{n_j}\right) \quad , \tag{2.7}$$

TFC-взвешивание. В TF-IDF методе взвешивания не учитывается, что документы могут быть различной длины, что существенным образом влияет на качество классификации TFC - взвешивание подобно TF-IDF - взвешиванию за исключением того, что используется нормализация длины документа.

$$w_{ij} = \frac{f_{ji} \quad \log\left(\frac{N}{n_j}\right)}{\sqrt{\sum_{l=1}^{M}\left[f_{li} \quad \log\left(\frac{N}{n_l}\right)\right]^2}} \quad , \tag{2.8}$$

где n_l - общее количество документов исходной выборки, содержащих слово l.TFC мера взвешивания, как и TF – IDF мера, выделяет средне и низкочастотные термины.

LTC – взвешивание. Данный подход заключается в использовании логарифма частоты слова вместо просто частоты слова, таким образом сокращается эффект больших различий в частотах.

$$w_{ij} = \frac{\log\left(f_{ji}+1\right)\ \log\left(\frac{N}{n_j}\right)}{\sqrt{\sum_{l=1}^{M}\left[\log\left(f_{li}+1\right)\ \log\left(\frac{N}{n_l}\right)\right]^2}}, \qquad (2.9)$$

Также как и предыдущие две меры взвешивания выделяет средне и низкочастотные термины.

Энтропия - взвешивания. Энтропия – взвешивание основывается на идеях теории информации и является наиболее сложной схемой взвешивания.

$$x_{ji} = \log\left(f_{ji}+1\right)\cdot\left(1+\frac{1}{\log(N)}\sum_{l=1}^{N}\left[\frac{f_{ji}}{n_j}\log\left(\frac{f_{jl}}{n_j}\right)\right]\right), \qquad (2.10)$$

где $\frac{1}{\log(N)}\sum_{l=1}^{N}\left[\frac{f_{ji}}{n_j}\log\left(\frac{f_{jl}}{n_j}\right)\right]$ является энтропией слова j (равно 1, если слово одинаково распределено по всем документам, и 0, если слово встречается только в одном документе).

Таким образом, использование LTC меры позволяет сократить эффект больших различий в частотах, что делает использование данной меры взвешивания наиболее приемлемой.

Отсюда, сообщения, формирующие обучающую выборку, можно представить в виде матрицы, столбцами которой будут письма, а строками термы, содержащиеся в письмах:

$$L_k = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{j1} \\ w_{12} & w_{22} & \cdots & w_{j2} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1\mathrm{i}} & w_{2\mathrm{i}} & \cdots & w_{ji} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1\mathrm{N}} & w_{2\mathrm{N}} & \cdots & w_{MN} \end{bmatrix}$$

где $w_{it_j} = Ltc_{it_j}$, $j = 1,\ldots,M$, $i = 1,\ldots,N$.

Получаемая матрица признаков ЭС имеет размерность, обработка которой требует больших вычислительных ресурсов и времени.

Кроме того, согласно законам Ципфа, слова, встречающиеся в тексте обучающей выборки чаще других, являются малоинформативными, не имеющими решающего смыслового значения, что становится основой снижения размерности матрицы признаков за счет избавления от малоинформативных термов без потери смыслового содержания ЭС.

Так же проведенный анализ частотности термов в сообщениях класса легитим и класса спам (рисунок 2.6) показал что:

- всегда существуют термы, встречающиеся в одном классе, но не встречающиеся в другом классе. И термы, которые чаще всего встречаются в определенном классе, что говорит о возможности данных термов характеризовать тот или иной класс;

- существуют термы, по которым трудно определить принадлежность к тому или иному классу, так как частота встречаемости данных термов в обоих классах различается незначительно.

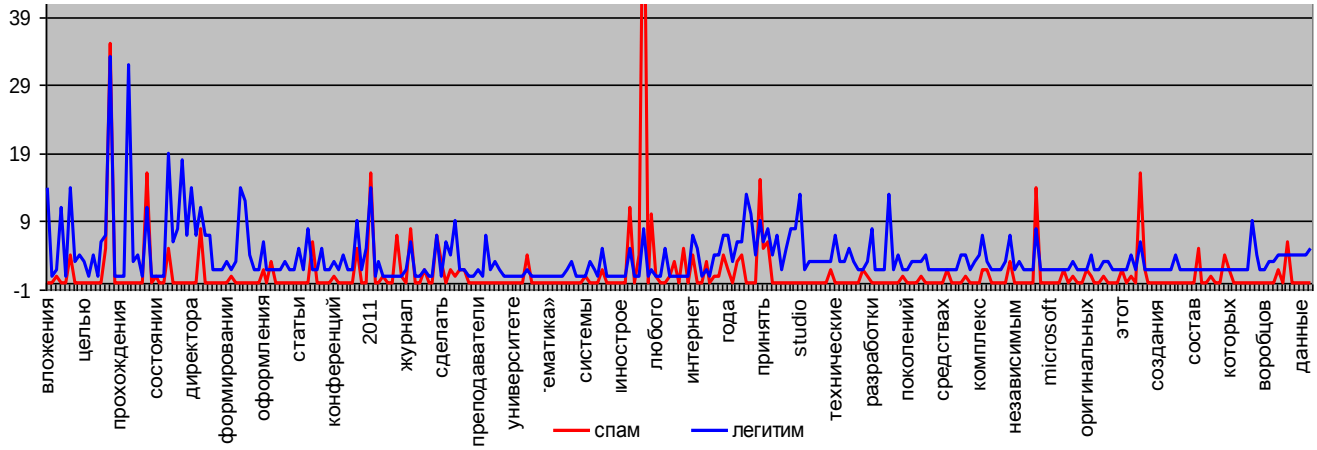


Рисунок 2.6 – Число термов по классам *(spam/legitim)* типового почтового электронного сообщения

Полученные выводы свидетельствуют о необходимости сокращения признакового пространства с целью избавления от малоинформативных термов и выделения термов способных характеризовать тот или иной класс.

Для сокращения признакового пространства задачи классификации известны следующие подходы [1,30,61,87]:

– сокращение пространства признаков непосредственно для *каждого* класса;

– сокращение пространства признаков для *всех* писем обучающей выборки без учета принадлежности к тому или иному классу.

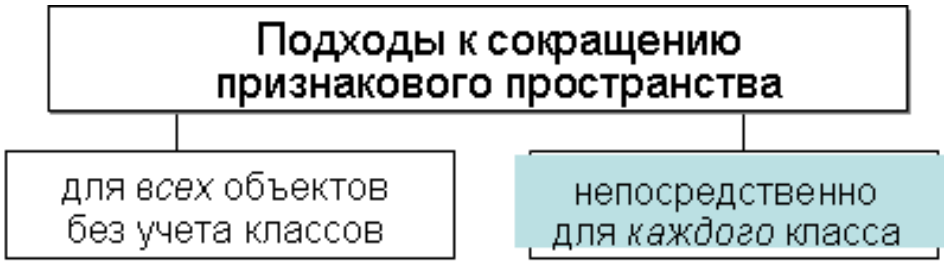


Рисунок 2.7 – Подходы к сокращению признакового пространства

Для реализации указанных способов известны методы многомерного статистического анализа, ориентированные на работу с текстовыми данными, такие как подсчет взаимной значимости термов [34], кластеризация термов относительно введенной метрики [98], выделение только тех термов, вес которых является максимальным [99].

В данной работе исследовались и сравнивались методы:

Метод взаимной значимости термов

Метод взаимной значимости термов позволят вычислить значимость термов в коллекции документов.

Данная мера рассчитывается по формуле вида:

$$MI(t_j,k) = \sum P(t_j,k) * \log \frac{P(t_{j,}k)}{P(t_j) * P(k)} \qquad (2.11)$$

где t_j – терм сообщения;

k – класс сообщения.

Данная мера основана на определении вероятности появления терма в сообщении. С точки зрения теории вероятности данная мера позволяет определить степень независимости термов. Сокращение числа термов происходит по установленному порогу. В качестве недостатка данной

методики можно выделить зависимость значения меры *MI* от количества сообщений в классе, а также свойство данной меры завышать значимость термов, частота которых ниже по сравнению с другими термами сообщениях класса [94,96,99].

Метод Хи-квадрат

Статистический метод Хи-квадрат широко применяется как метод выбора признаков для работы с текстом. Критерий Хи-квадрат позволяет ранжировать признаки только по степени их полезности(важности) и не позволяет сделать вывод о статистической зависимости или независимости признаков. Критерий Хи-квадрат ориентированный для работы с текстом [99], рассчитывается по зависимости вида:

$$\chi^2(t,k) = \frac{N*(AD-BC)^2}{(A+C)*(B+D)*(A+B)*(C+D)}, \qquad (2.12)$$

где A – количество сообщений, в которых *t* и *k* появились совместно;

B – количество сообщений, при которых слово *t* встречается с другим классом;

C – количество сообщений относящихся к классу *k*, но в которых не встречается *t*;

D – количество сообщений не относящихся к классу *k*, в которых нет слова *t*;

Метод главных компонент

Метод главных компонент, предназначенный для сокращения размерности данных, при этом позволяет осуществить отбор наиболее информативных термов. В [Айвазян] детально описаны основные определения, вычисления и числовые характеристики главных компонент. Формально задача снижения размерности признакового пространства для некоторой выборки объектов $X = \{X_n\}$ размерности $n=1,\infty$ состоит в получении представления этой выборки в пространстве меньшей размерности, т.е. в приведении X к виду $X' = \{X_m\},\ m < n$.

Однако метод не всегда эффективно снижает размерность при заданных ограничениях на точность. Прямые и плоскости не всегда обеспечивают хорошую аппроксимацию. Например, данные могут с хорошей точностью следовать какой-нибудь кривой, а эта кривая может быть сложно расположена в пространстве данных. В этом случае метод главных компонент для приемлемой точности потребует нескольких компонент (вместо одной), или вообще не даст снижения размерности при приемлемой точности.

Глобальный вес $RF_{t_q}^k$

В данной работе для сокращения признакового пространства предложен комбинированный подход, основанный на том, что для каждого терма в сообщениях определенного класса вычисляется величина $RF_{t_j}^k$, характеризующая значимость терма для определенного класса *k*:

$$RF_{t_j}^k = \log_2(2+\frac{a_i}{\max(1,b_i)}), \quad (2.13)$$

где a_i – количество ЭС, содержащих t_j-ый терм и относящихся к классу *k*;

b_i – количество ЭС, содержащих t_j-ый терм и не относящихся к классу *k*.

Термы, значимость которых $RF_{t_q}^k \le 1{,}5$, исключаются в данном сообщении класса *k*.

В рамках исследований был проведен сравнительный анализ методов сокращения признакового пространства рассмотренных выше.

Для проведения исследований было использовано 100 сообщений (50 сообщений спам тематики и 50 легитимных сообщений). Для описания текстового содержимого электронных почтовых сообщений использовалась векторная модель с использованием меры определения значимости термов LTC. Общее количество термов в сообщениях составляет 8723.

При использовании метода определения информационной значимости (IG) происходит сокращение числа термов на 2523 и составляет 6200. Компонентный анализ позволил сократить число термов на 1923, в результате чего количество термов стало 6800. Метод Хи-квадрат позволил сократить число термов всего на 1611, таким образом, количество термов стало составлять 7112. Метод на основе определения глобального веса $RF_{t_q}^{k}$ сократил число термов на 2566, и количество термов стало составлять соответственно 6157. Еще один параметр по которому происходило сравнение рассматриваемых методов – время, затраченное на сокращение размерности матрицы. Метод $RF_{t_q}^{k}$ позволяет сократить размерность термов за 17 секунд, компонентный анализ за 23 секунды, метод Хи-квадрат за 15 секунд и метод на основе определения информационной значимости за 16 секунд. Результаты анализа сведены в таблицу 2.1 и представлены в виде диаграммы на рисунке 2.7.

Таблица 2.1 – Результаты работы методов по количеству сокращения термов в коллекции сообщений и времени на их обработку

Вид проведенной работы	Название методов			
	RF	IG	Хи-квадрат	Компонентый анализ
размерность матрицы термов	6157	6200	7112	6800
время обработки (сек)	17	16	15	23

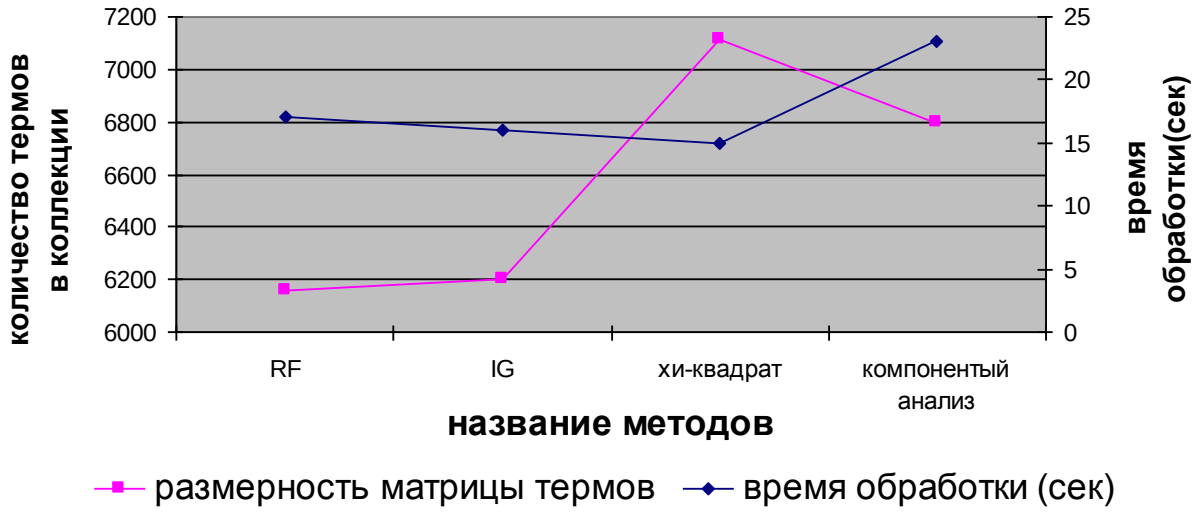


Рисунок 2.7– Сравнительная характеристика методов

Таким образом, проведенный анализ показал что больше всего на сокращение числа термов «затратил» компонентный анализ, при этом сократив количество термов с 8723 до 6800. Скорее всего это связанно с наибольшими вычислительными затратами данного метода. Метод RF и IG показали близкие результаты, однако использование метода на основе определения глобального веса оправданно его результативностью.

2.3 Разработка методики выявления устойчивых словосочетаний

Апробация программного средства и его дальнейшее использование осуществлялись на базе предприятия ОАО «Оренбургнефть» корпорации «ТБинфром». В соответствии с Федеральным законом "О коммерческой тайне" [78] обрабатываемые письма не могут быть приведены в настоящей работе. В связи с этим, в качестве примера будет рассмотрен пример легитимного ЭПС сообщения кафедры программного обеспечения вычислительной техники и автоматизированных систем на базе которой была выполнена диссертационная работа.

В соответствии с законом о персональных данных [79] адрес отправителя и получателя, а так же номера телефонов будут скрыты, а все личные данные (Ф.И.О.) заменены на Иванов Иван Иванович.

От: xxxxx@mail.osu.ru
Отправлено: 14 февраля 2012 г. 17:04
Кому: xxxx@unpk.osu.ru

Вложения: Информационные ресурсы научной библиотеки.doc

Уважаемые заведующие кафедрами!
С целью информированности об информационных ресурсах научной библиотеки, а также для успешного прохождения аккредитации в 2012 г. образовательных программ, рассылаю Вам справку о состоянии фонда библиотеки вуза.

С уважением,
Иванов Иван Иванович,
зам. директора Научной библиотеки ОГУ
тел.: (xxxx) xx-xx-xx; внутр. xx-xx

Рисунок 2.8 – Пример легитимного ЭПС

В таблице 2.8 приведены термы и рассчитанное значение глобального веса RF. Согласно принятому порогу термы $RF_{t_q}^{k} \le 1{,}5$, исключаются в данном

сообщении. Следовательно, такие термы, как *аккредитации, информационные, рассылаю, справку, состоянии, вуза, информированности, ресурсах, успешного, прохождения* исключаются из данного сообщения.

Таблица 2.2 – Рассчитанные значения RF термов в сообщении

Термы в сообщении	RF
внутр	3
аккредитации	1
вложения	3,90689
2012	4,45943
информационные	1
образовательных	2
ресурсы	1,32193
программ	2,32193
научной	3,32193
рассылаю	1
библиотеки	3,32193
справку	1
уважаемые	2,39232
состоянии	1
заведующие	2
фонда	1
кафедрами	2,32193
вуза	1
целью	2
уважением	2,48543
информированности	1
Иванов	2,80735
информационных	2
Иван	3,16993
ресурсах	1
Иванович	4,24793
также	1,80735
директора	3,58496
успешного	1
3532	3
прохождения	1

Анализ значимости термов в сообщении после сокращения признакового пространства (рисунок 2.10) показал что существуют термы которые об-

ладают большим классификационным свойством и следовательно могут в большей степени повлиять на результат фильтрации, а некоторые меньшим свойством что говорит о необходимости выделения термов, наиболее важных для детектирования сообщений. Данное обстоятельство говорит о необходимости применения методики выделения ключевых термов из текстового содержимого сообщения, которые отражают специфику ЭПС и позволяют выделить термы имеющие наиболее выраженные классификационные свойства.

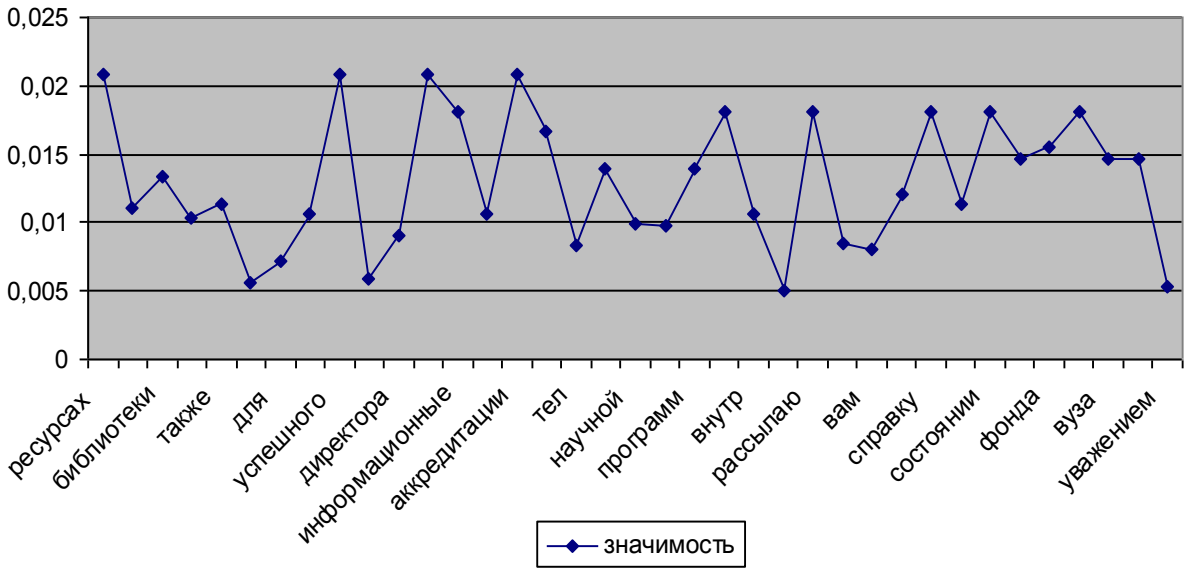


Рисунок 2.10 - Значимость термов сообщения после сокращения пространства признаков статистическими методами

В то же время, не все термы, содержащиеся в сообщении, отражают его тематику. Служебные слова не несут смысловой нагрузки, а используются для связи слов в предложениях (предлоги, союзы, частицы). Самостоятельные слова, частота употребления которых в тексте невелика, также не отражают тематику текста[56,81,100]. В связи с этим необходимо выделить термы способные отражать содержание сообщения. Извлечение таких термов может быть смоделировано через процедуры выделения ключевых слов текста.

Тогда задачу фильтрации входящей почтовой корреспонденции можно представить в виде этапов показанных на рисунке 2.11

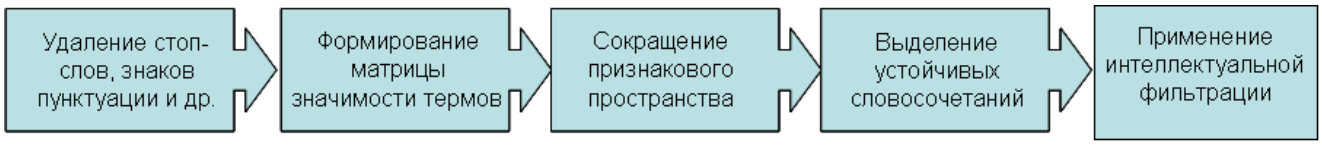


Рисунок 2.11 – Технология решения задачи фильтрации корреспонденции

В данной работе под устойчивым словосочетанием понимается комбинация двух и более термов, имеющих тенденцию к совместной встречаемости [60,83,89], т.е. речь идет об осмысленных последовательностях слов в тексте. Такое рассмотрение устойчивых словосочетаний позволяет применять к ним различные статистические меры, с помощью которых можно определить связь между элементами словосочетаний и силу связи между ними.

В настоящее время в лингвистике существует несколько способов для вычисления степени связанности частей устойчивых словосочетаний[83]. В качестве таких статистических мер были выбраны меры ассоциации, которые чаще всего используются при вычислении степени близости между компонентами словосочетаний в корпусе [58].

Существуют два основных подхода к автоматическому выделению терминов [44,45,93,96,97].

Первый подход относится к области статической обработки естественного языка - вычисление различных мер ассоциативной связи, которые оценивают, является ли взаимное появление лексических единиц случайным, или оно статически значимо.

Второй подход опирается на семантическую близость термов. Предполагающую определение семантической связности термов в сообщениях.

Пусть $D_i=\{d_{jq}\}, _{j=1,\ldots,N}$ характеристика связи между термами в i-ом сообщении, где d_{jq} – степень смысловой близости j-го и q-го термов.

В качестве меры близости между термами в сообщении предложено использовать расстояние Дайса. Данная статистическая мера позволит объединить термы в устойчивые (ключевые) словосочетания, характеризующие семантическое содержание сообщений.

Близость D и частота $f(t_1,t_2)$ совместной встречаемости термов становятся предпосылкой для нахождения устойчивых словосочетаний.

Алгоритм формирования устойчивого словосочетания:

1) выделение значимых термов с учетом (4) для соответствующего класса k (spam/legitim);

2) расчет близости термов и принятие решения о формирования устойчивого словосочетания;

3) подтверждение смысловой значимости устойчивого словосочетания.

Мера Дайса D рассчитывается по зависимости вида:

$$D(t_1,t_2)=\log_2\left(\frac{2*(f(t_1,t_2))}{f(t_1)+f(t_2)}\right) \qquad (2.14)$$

где $f(t_1)$ и $f(t_2)$ – частота встречаемости термов t_1 и t_2 в сообщении;

$f(t_1,t_2)$ – частота совместной встречаемости термов t_1 и t_2.

Для задачи фильтрации электронных почтовых сообщение в данной работе предлагается формировать устойчивое словосочетание, если значение D_{jq} равное или выше, чем в соседних парах термов.

Для подтверждения смысловой значимости полученных устойчивых словосочетаний предлагается оценить тесноту взаимосвязи между термами в словосочетании, метрикой которой могут выступать меры ассоциации или контингенции.

Наиболее распространенными мерами ассоциации являются *MI-score*, *t-score* и *log-likelihood* [84,89,90], которые признаны показателями силы смысловой (синаптической) связи между качественными признаками (термами) словосочетаний. В данной работе в качестве меры тесноты взаимосвязи двух качественных признаков словосочетаний предложено использовать коэффициенты ассоциации K_a и контингенции K_k, которые рассчитываются по следующим зависимостям:

$$K_a=\frac{ad-bc}{ad+bc} \qquad (2.15)$$

$$K_k = \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+c)(c+d)}} \qquad (2.16)$$

где a – количество сообщений, имеющих терм t_1, который встречается в классе k;

b – количество сообщений, в которых терм t_1 встречается с другим классом;

c – количество сообщений, имеющих терм t_2, который встречается в классе k;

d – количество сообщений, в которых терм t_2 встречается с другим классом.

Экспериментально установлено, что связь между элементами словосочетания считается подтвержденной, если $K_a \geq 0{,}5$ или $K_к \geq 0{,}3$.

Тогда модель почтового ЭС можно представить в виде:

$$L(p_i) = <T^k, w^*(t_j)>,$$

где T^k – терм устойчивых словосочетаний в сообщении;

$w^*(t_j)$ – вес терма в сообщении после сокращения матрицы признаков (3).

Таким образом, модель ЭС в форме устойчивых словосочетаний позволяет без потери смыслового содержания обеспечить интеллектуальную классификацию почтовой электронной корреспонденции.

3 Информационное обеспечение системы контентной фильтрации электронной корреспонденции

В результате анализа предметной области было получено формализованное описание предметной области, определена иерархия функций приведённая в приложении, получено формализованное описание предметной области, выявлен состав пользователей и описаны уровни доступа пользователей к проектируемой БД (приложение А). Построена концептуальная инфологическая модель предметной области.

3.1 Проектирование базы данных

В процессе разработки информационного обеспечения системы контентной фильтрации ЭС разработана информационно-логической и даталогическая модели предметной области**.** Исходными данными для построения информационно-логической модели предметной области (ИЛМ) являются результаты анализа предметной области, представленные в виде описания классов объектов и связей между ними. Чаще всего ИЛМ предметной области представляют в терминах семантической модели данных, в виде ER-диаграммы предметной области [28].

В настоящее время существуют разнообразные нотации построения ER-модели. Подробно рассмотрим самую распространённую из них – методологию Ричарда Баркера.

Методология Ричарда Баркера построена на базе следующих элементов: класс объектов, свойство класса объектов, уникальные идентификаторы, опциональность свойств, мощность (тип), опциональность и переносимость связей, уникальность объектов из связей, супертипы, подтипы, арки.

В методологии используются следующие соглашения:

- класс объектов отображается в виде четырехугольника с закругленными углами, а имя класса объектов указывается внутри четырехугольника,

это имя существительное в единственном числе, отображенное заглавными буквами;

- свойства записываются внутри четырехугольника, отображающего класс объектов строчными буквами, это имя существительное в единственном числе;

- четырехугольник, отображающий класс объектов, можно увеличивать до любых размеров, четырехугольники могут быть разных размеров;

- опциональность свойств помечается: обязательное свойство – звездочкой (*), необязательное – кружочком (о);

- уникальный идентификатор помечается #, если уникальных идентификаторов несколько, тогда каждый помечается номером, указанным в скобках, например, # (1), #(2);

- обязательная связь помечается сплошной линией, необязательная связь пунктирной линией;

- тип (мощность) связи «один» помечается линией, «много» — «вороньей лапой».

Каждый объект обладает определённым набором свойств. Для объектов одного класса набор этих свойств одинаков, а эти значения могут различаться.

При описании предметной области необходимо отразить связь между объектами разных классов. Различают связи типа «один к одному» (1:1), «один ко многим» (1:M), «многие ко многим» (M:M).

Концептуальная инфологическая модель предметной области, построенная по методологии Ричарда Баркера, представлена на рисунке 3.1.

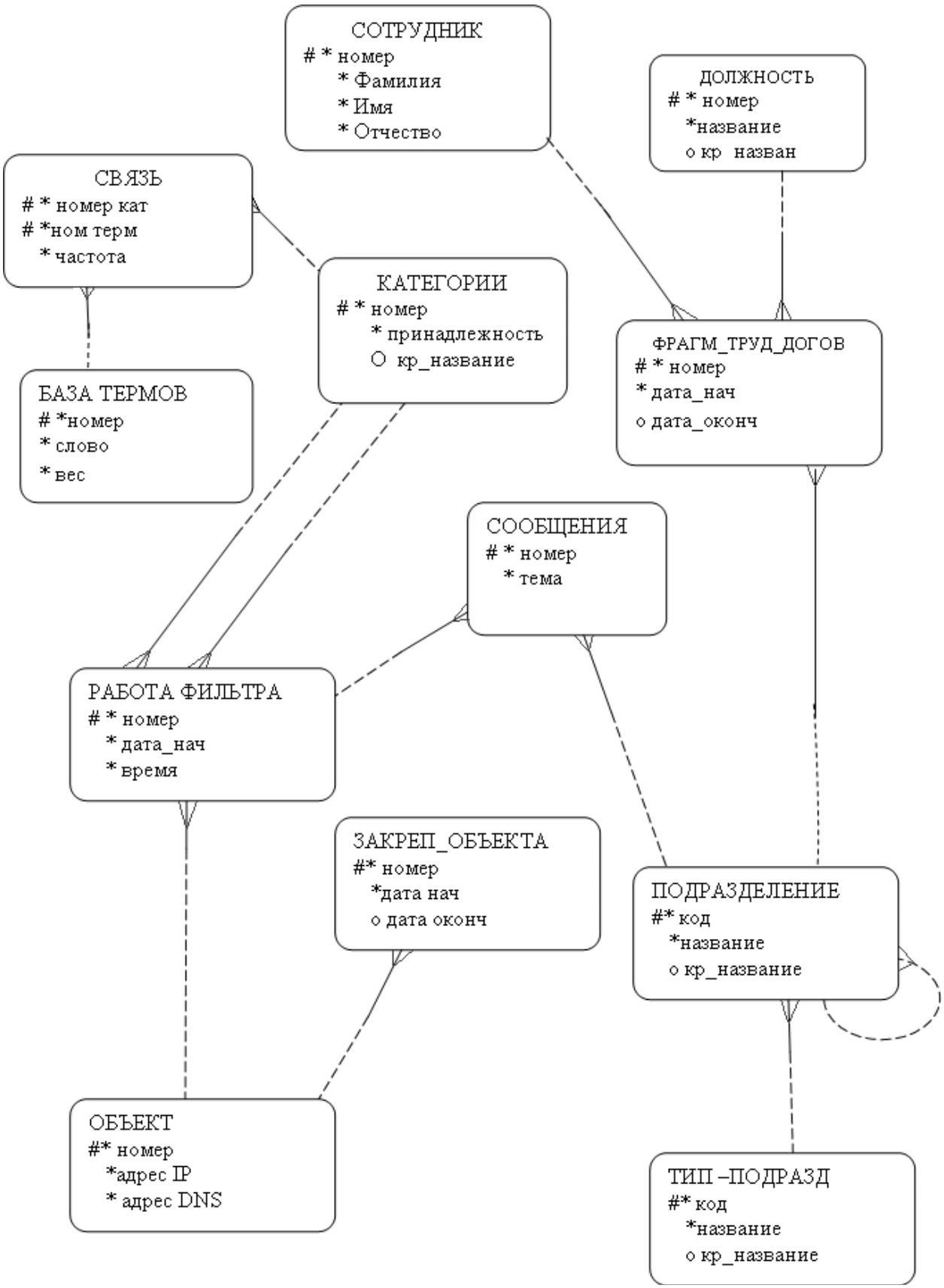


Рисунок 3.1 – Инфологическая модель предметной области, построенная по методологии Ричарда Баркера

После построения инфологической модели предметной области необходима проверка полученной модели предметной области – поддерживает ли она выполнение функций разрабатываемого программного продукта. Проверка приводится в формализованном виде, в форме таблицы 3.1.

Таблица 3.1 – Перекрестная проверка

Элементарные функции	Классы объектов			
	Категории	База термов	Сообщения	Конфигурация
Ф1	U			
Ф2	R			
Ф3		I, U		
Ф4		R		
Ф5			I, U	
Ф6			R	
Ф7				U
Ф8				R
Ф9	R	R	I,U	U,R
Ф10	U	I,U		

Анализируя таблицу 3.1, можно отметить, что каждой функции соответствует хотя бы один класс объектов. С другой стороны, каждый класс объектов, отображенный в модели, необходим для реализации хотя бы одной функции.

Следующим этапом проектирования базы данных методом «нисходящего» проектирования было построение даталогической модели базы данных. Исходными данными для даталогического проектирования является информационно-логическая модель предметной области. В результате даталогического моделирования получена логическая структура базы данных, описанная в терминах реляционной модели данных на основе физических записей.

При переходе к даталогической модели следует помнить, что инфологическая модель включает в себя всю информацию о предметной области, необходимую и достаточную для проектирования баз данных.

Даталогическая модель базы данных представлена на рисунке 3.2.

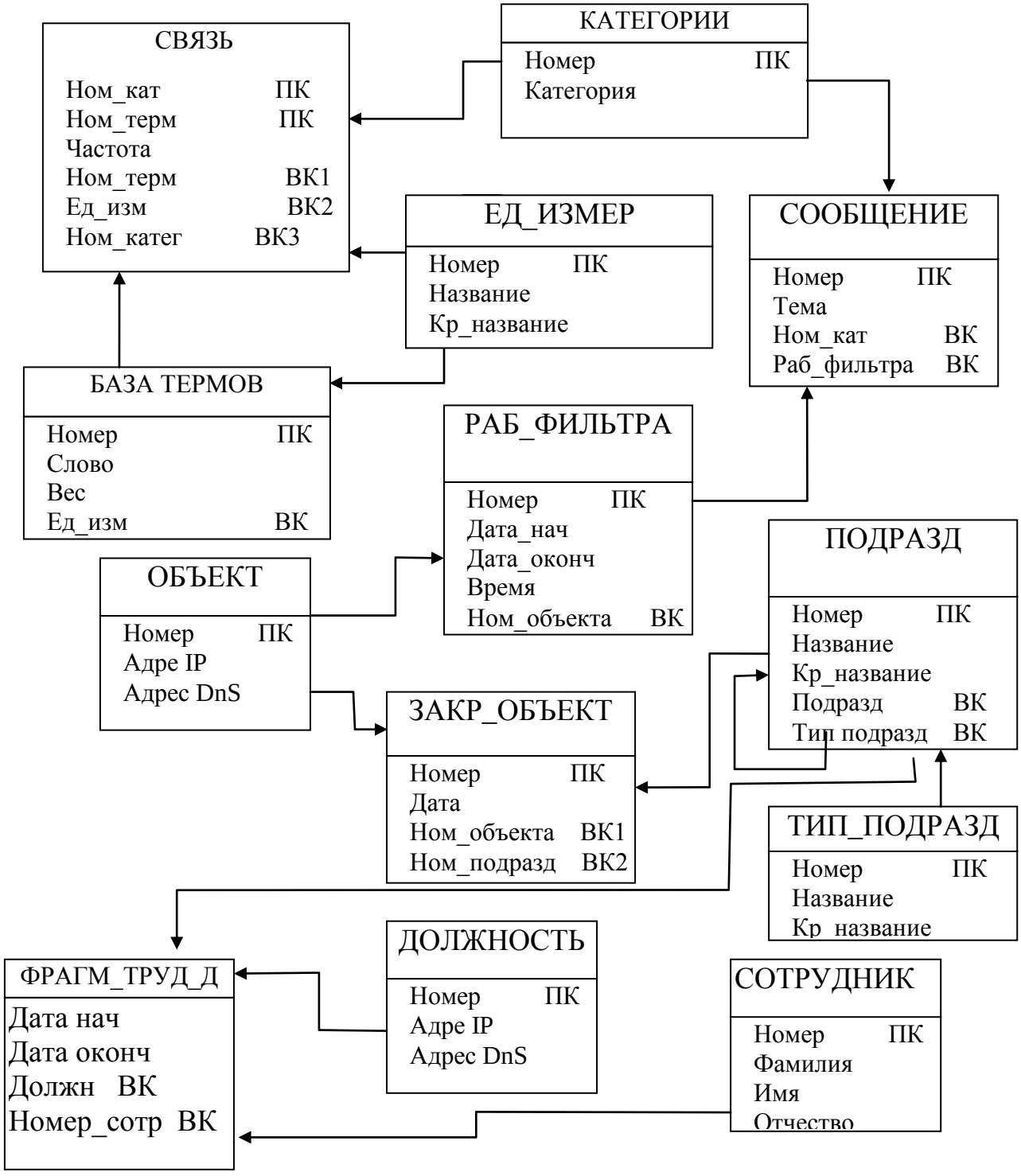


Рисунок 3.2 – Даталогическая модель базы данных

3.2 Физическая модель базы данных

Физическое проектирование БД заключается в преобразовании даталогической модели данных в такую форму, которая позволит реализовать проект в среде СУБД MySQL

Техническое описание реляционных таблиц на языке определения данных представлено в таблицах 3.8-3.18

Таблица 3.2 – Реляционная таблица «Classes»

Имя поля	Clas_id	Type
Ключ	Primary Key	
Тип, длина	Integer(10)	Integer(2)
Обязательность значения	Not Null	Not Null
Логическое ограничение на поле	Check (value>0)	
Примеры данных	1	1

Таблица 3.3 – Реляционная таблица «Link»

Имя поля	Clas_id	Term_id	Freq
Ключ	Primary Key	Primary Key	
Тип, длина	Integer(10)	Integer(10)	double
Обязательность значения	Not Null	Not Null	
Логическое ограничение на поле	Check (value>0)		
Примеры данных	1	1	

Таблица 3.4 – Реляционная таблица «Term»

Имя поля	Term_id	Name	Weight
Ключ	Primary Key		
Тип, длина	Integer(10)	varchar (255)	double
Обязательность значения	Not Null	Not Null	
Логическое ограничение на поле	Check (value>0)		
Примеры данных	1	терм	

Таблица 3.5 – Реляционная таблица «Сообщение»

Имя поля	Mess_id	Classes	Tema
Ключ	Primary Key		
Тип, длина	Integer(10)	Integer(10)	Integer(25)
Обязательность значения	Not Null	Not Null	
Логическое ограничение на поле	Check (value>0)		
Примеры данных	1	spam	

Таблица 3.6– Реляционная таблица «Объект»

Имя поля	Mess_id	Classes	Tema
Ключ	Primary Key		
Тип, длина	Integer(10)	Integer(10)	Integer(25)
Обязательность значения	Not Null	Not Null	
Логическое ограничение на поле	Check (value>0)		
Примеры данных	1	spam	

Таблица 3.7 – Реляционная таблица «Закрепл_объект»

Имя поля	Mess_id	Classes	Tema
Ключ	Primary Key		
Тип, длина	Integer(10)	Integer(10)	Integer(25)
Обязательность значения	Not Null	Not Null	
Логическое ограничение на поле	Check (value>0)		
Примеры данных	1	spam	

Таблица 3.8 – Реляционная таблица «Труд_догов»

Имя поля	Mess_id	Classes	Tema
Ключ	Primary Key		
Тип, длина	Integer(10)	Integer(10)	Integer(25)
Обязательность значения	Not Null	Not Null	
Логическое ограничение на поле	Check (value>0)		
Примеры данных	1	spam	

Таблица 3.9 – Реляционная таблица «Подразделение»

Имя поля	Mess_id	Classes	Tema
Ключ	Primary Key		
Тип, длина	Integer(10)	Integer(10)	Integer(25)
Обязательность значения	Not Null	Not Null	
Логическое ограничение на поле	Check (value>0)		
Примеры данных	1	spam	

Таблица 3.10 – Реляционная таблица «Тип_подразделения»

Имя поля	Mess_id	Classes	Tema
Ключ	Primary Key		
Тип, длина	Integer(10)	Integer(10)	Integer(25)
Обязательность значения	Not Null	Not Null	
Логическое ограничение на поле	Check (value>0)		
Примеры данных	1	spam	

Таблица 3.11 – Реляционная таблица «Сотрудник»

Имя поля	Mess_id	Classes	Tema
Ключ	Primary Key		
Тип, длина	Integer(10)	Integer(10)	Integer(25)
Обязательность значения	Not Null	Not Null	
Логическое ограничение на поле	Check (value>0)		
Примеры данных	1	spam	

Таблица 3.12 – Реляционная таблица «Должность»

Имя поля	Mess_id	Classes	Tema
Ключ	Primary Key		
Тип, длина	Integer(10)	Integer(10)	Integer(25)
Обязательность значения	Not Null	Not Null	
Логическое ограничение на поле	Check (value>0)		
Примеры данных	1	spam	

Реализация ПС, имеющая следующие основные экранные формы, представлена в приложении Б.

4 Программное обеспечение системы контентной фильтрации электронной корреспонденции

4.1 Разработка архитектуры контентной фильтрации

Для выбора архитектуры разрабатываемого фильтра проанализированы основные достоинства и недостатки фильтрации на стороне клиента и на стороне сервера. Взаимодействие модулей в процессе работы фильтра показано на рисунке 4.1

Модуль обучения запрашивает каталог со спамом и каталог с легитимной почтой, список адресов к которым относится эта почта. Обучающая база отдела закреплена за списком адресов сотрудников данного отдела, что позволяет при идентификации почтовой корреспонденции по e-mail адресу получателя обращаться к базе характеризующей интересы данного пользователя.

Классификация сообщений происходит на стороне почтового сервера посредством модуля формальной фильтрации и модуля контент-фильтрации (модуля фильтрации по содержимому).

Модуль предварительной обработки ЭС предназначен для приведения текста сообщения к единому регистру и кодировке, удаления стоп-слов, знаков пунктуации, разбиения сообщений на термы.

Модуль расчета значимости термов позволяет в предварительно подготовленном сообщении осуществить расчет частот термов в сообщении и в классе и определить значимость терма в каждом сообщении.

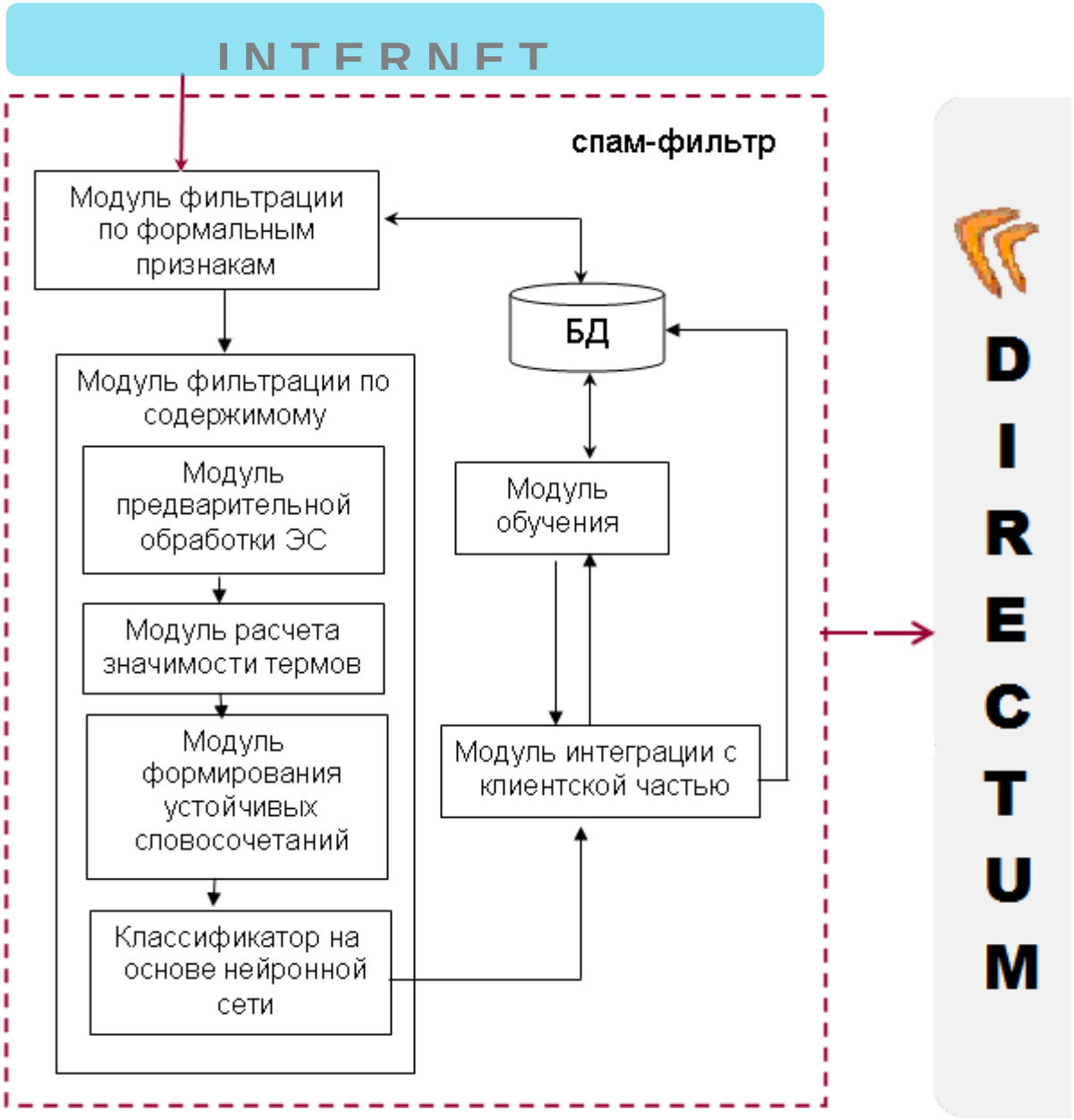


Рисунок 4.1 – Взаимодействие модулей

Модуль формирования устойчивых словосочетаний позволяет определить частоту совместной встречаемости термов, подтвердить смысловую значимость термов, тем самым выделить термы характеризующие содержание электронного почтового сообщения.

Классификатор на основе нейронной сети осуществляет детектирование ЭПС.

Модуль интеграции с клиентской частью отвечает за взаимодействие системы фильтрации с пользователем.

Модуль обучения производит обучение системы фильтрации и позволяет пользователю выбрать и подготовить письма для обучения.

Архитектура программного средства – это его строение, т.е. представление программной системы состоящей из некоторой совокупности взаимодействующих подсистем. В качестве подсистем будут выступать программные модули, так как программный комплекс имеет модульную структуру. При разработке был использован метод нисходящего проектирования, который состоит в том, что сначала строится модульная структура программы в виде дерева и модули проектируются поочередно, начиная с модуля самого верхнего (головного) уровня, а переход к программированию какого-либо другого модуля осуществляется только в том случае, если уже запрограммирован модуль, который к нему обращается. После того, как все модули программы запрограммированы, производится их поочередное тестирование и отладка в таком же (нисходящем) порядке. Метод нисходящего проектирования иногда называют функциональной декомпозицией.

Прототип разрабатываемой системы фильтрации состоящий из проекта e-mail_filtering в который входит 9 модулей представлен на рисунке 4.2

Модули, реализующие основные функции, вызываются из основного модуля menu при выборе определенного пункта меню. Вспомогательные модули используются в процессе работы основных модулей.

Спецификация модулей приведена в таблице 4.1.

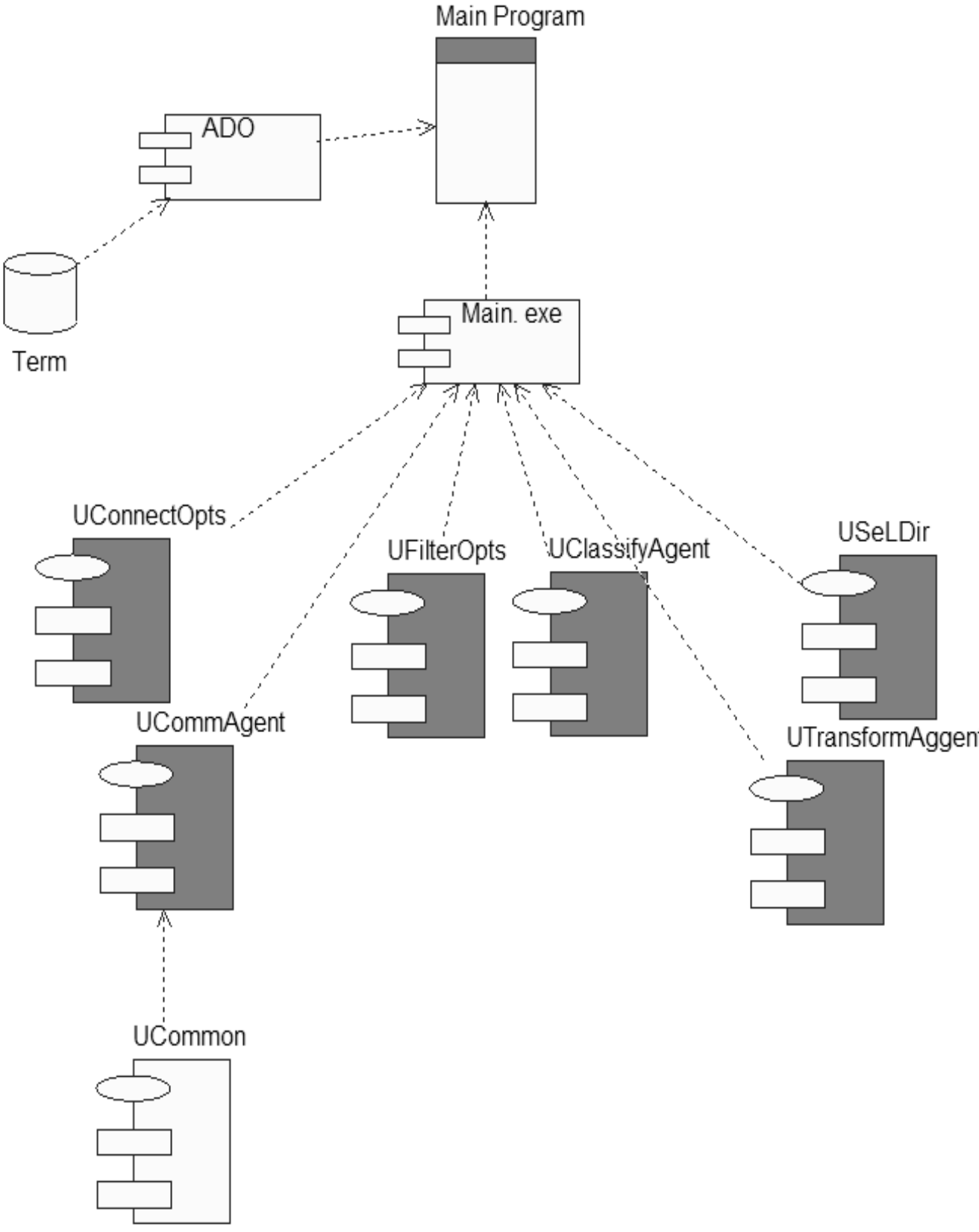


Рисунок 4.2 – Диаграмма компонентов программного проекта системы контентной фильтрации

Таблица 4.1- Спецификация модулей

Название	Назначение
UMain	Модуль служит для реализации главного окна программы, создания и анимации значка в системном дереве. Хранит компонент прокси-сервера, который обеспечивает взаимодействие между POP-сервером электронной почты и клиентской программой доставки почты.
UClassifyAgent	Определяет принадлежность письма к соответствующей категории. Реализованы методы обучения системы.
UCommon	Модуль, в котором храниться описания типов, константы, процедуры и функции, общие для всей системы. В нем реализована загрузка и сохранение настроек.
UTransformAgent	Модуль реализует трансформирующего агента, который преобразует входное сообщение в вид, удобный для анализа классифицирующим агентом
UCommAgent	Модуль реализует взаимодействие приложения и базы данных, где хранится информация о нейронах системы
UConnectOpts	Предоставляет пользовательский интерфейс для настройки подключения к серверу электронной почты и управления базой данных MySQL.
UFilterOpts	Модуль реализует диалоговое окно, предоставляющее пользовательский интерфейс для настройки самого приложения. В нем задаётся порог чувствительности системы, особенности обработки сообщений.
USelDir	Модуль, реализующий диалог для выбора каталога
UTransformAgent	Модуль реализует трансформирующего агента, который преобразует входное сообщение в вид, удобный для анализа классифицирующим агентом

4.2 Разработка алгоритмов нейросетевого классификатора

Так как задача фильтрации входящих электронных почтовых сообщений является задачей классификации, то для решения задачи классификации были проанализированы методы классификации и сделан вывод что наиболее эффективным для фильтрации электронных сообщений являются нейронные сети. В результате чего проведен анализ нескольких архитектур нейронных сетей предназначенных для классификации [14,27,31,37,54,64,82,84,95].

Сеть Кохонена

Сеть Кохонена обычно используется для решения задач классификации. Принцип работы данной сети состоит в нахождении центров объектов классов и отнесение входного вектора к одному из классов используя меру близости двух векторов (обычно применяется евклидово расстояние). Данная нейронная сеть обучается без учителя на основе самоорганизации. В процессе обучения вектора веса нейронов становятся прототипами классов – групп векторов обучающей выборки. Структура нейронной сети представлен на рисунке 4.3.

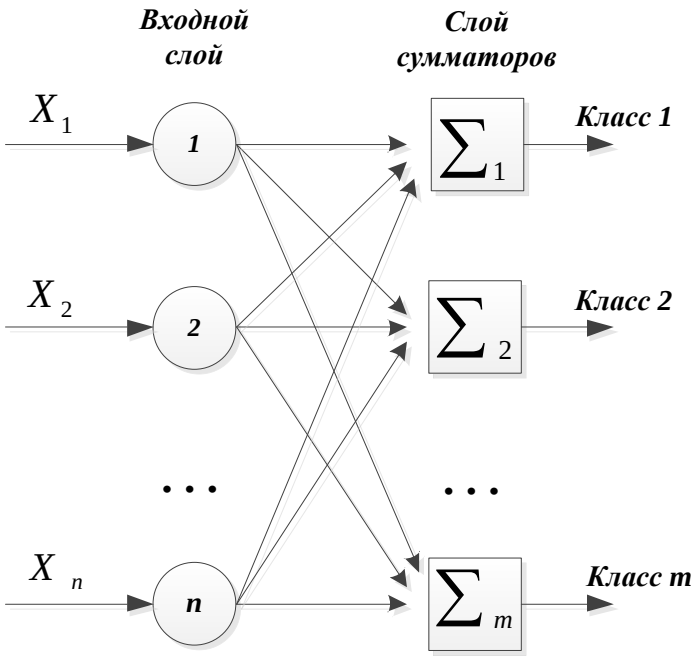


Рисунок 4.3 – Структура нейронной сети Кохонена

Сеть Кохонена состоит из одного слоя нейронов. Число входов каждого нейрона равно размерности вектора параметров объекта. Количество нейронов соответствует количеству классов, на которые нужно разбить объекты.

В общем виде обучение начинается с задания небольших случайных значений элементам весовой матрицы W, после чего происходит процесс так называемой самоорганизации, состоящий в модификации весов при предъявлении на вход векторов обучающей выборки. Алгоритм функционирования представлен на рисунке 4.4.

Рисунок 4.4 – Алгоритм функционирования нейронной сети

Каждый столбец весовой матрицы представляет собой параметры соответствующего нейрона-классификатора. Для каждого *j*-го нейрона определяется расстояние от него до входного вектора. После чего выбирается нейрон для которого это расстояние минимально. При дальнейшем обучении будет происходить модификация весов только тех нейронов, которые ближе всего находятся к выбранному нейрону .

Работа с данной нейронной сетью при решении задачи классификации показала высокий процент ложных срабатываний (22%-25%). Возможно, это связано с принципом работы данной сети (определение центра классов) не учитывающей тематический «разброс» электронных сообщений.

Многослойный персептрон

Нейронная сеть данного типа содержит один или несколько скрытых слоев нейронов, не являющихся частью входа или выхода нейронной сети. Сеть обладает высокой связностью синаптических соединений нейронов. Изменение уровня связанности нейронной сети требует изменения синаптических весов или коэффициентов.

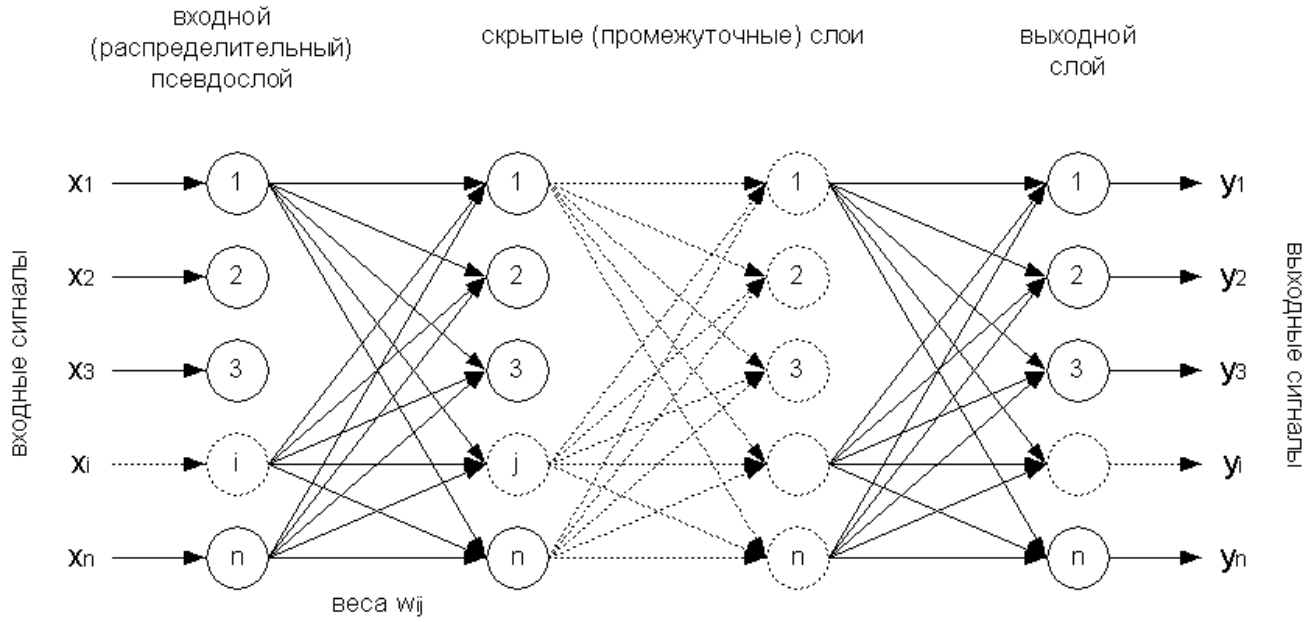


Рисунок 4.5 – Пример архитектуры многослойного персептрона

Определение количества слоев и типа функции активации зависит от той или иной задачи, которую приходится решать нейронной сети. В качестве недостатка данной нейронной сети можно отнести неспособность к самообучению и неустойчивость к помехам во входных образах, так же наличие скрытых нейронов делает процесс обучения более трудным для визуализации. Именно в процессе обучения необходимо определить, какие признаки входного сигнала следует представлять скрытыми нейронами [14,27,37].

Алгоритм функционирования данной нейронной сети представлен на рисунке 4.6

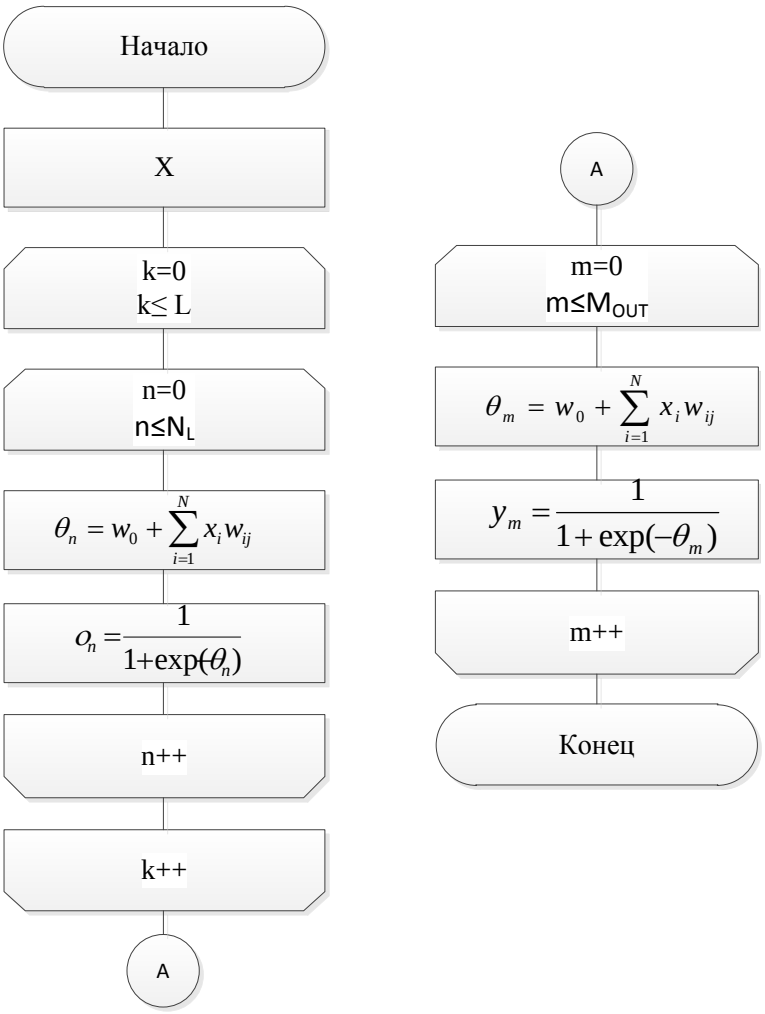


Рисунок 4.6 – Алгоритм работы многослойного персептрона

Вероятностные нейронные сети

Вероятностная нейронная сеть (PNN – Probabilistic Neural Network) основана на статистических метода Байеса и представляет собой параллель-

ную реализацию этих методов. В PNN образцы классифицируются на основе оценок их близости к соседним образцам.

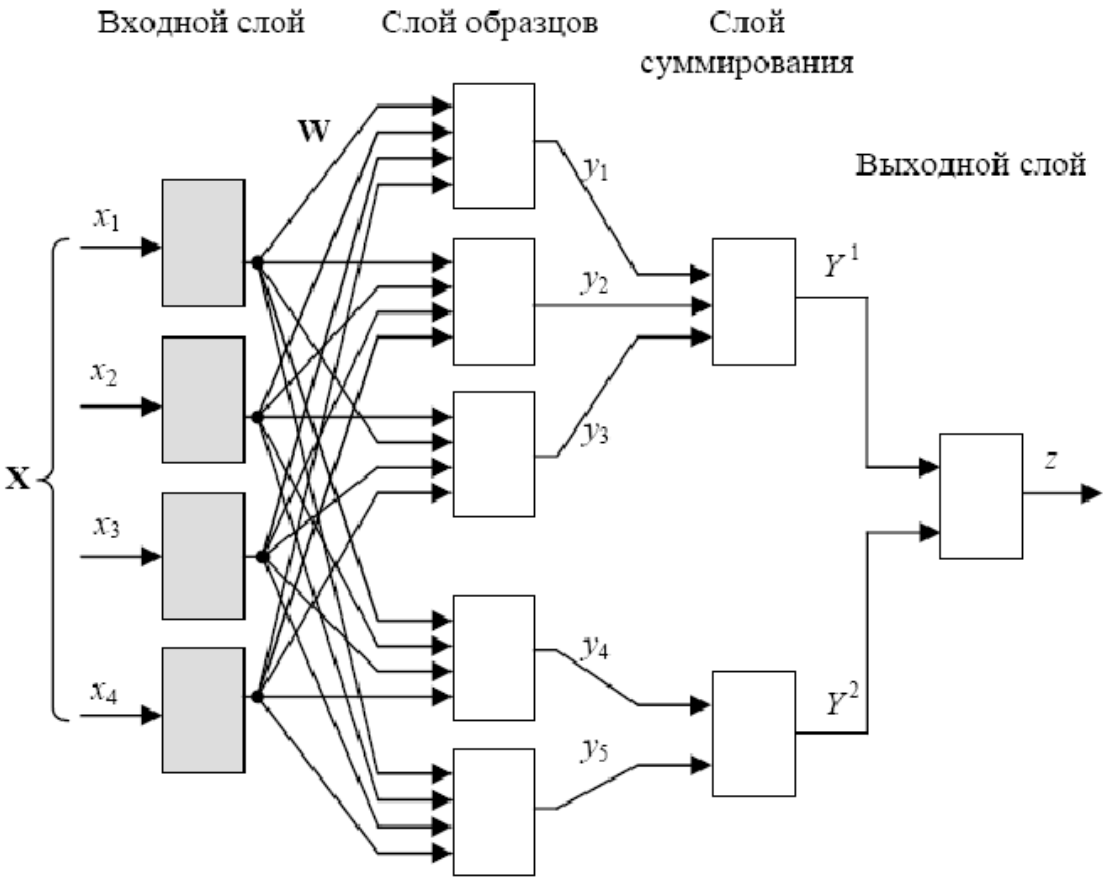


Рисунок 4.7 – Пример структуры вероятностной нейронной сети для разделения 4-компонентныхвходных векторов на два класса

При классификации векторов данной нейронной сетью используется ряд критериев статистических методов, на основе которых принимается решение о принадлежности к тому или иному классу. Формальным правилом при классификации является то, что класс с наиболее плотным распределением в области неизвестного образца, а также с более высокой априорной вероятностью и с более высокой ценой ошибки классификации, будет иметь преимущество по сравнению с другими классами, т.е. считают что вектор X принадлежит классу A (из двух рассматриваемых классов A и B) если :

$$hA \cdot cA \cdot fA(x) > hB \cdot cB \cdot fB(x),$$

где h – априорная вероятность; c – цена ошибки классификации; $f(x)$ – функция плотности вероятностей.

Для оценки функции плотности распределения вероятностей применяют метод Парцена, в соответствии с которым для каждого учебного образца рассматривается некоторая весовая функция, называемая *ядром*. Чаще всего в качестве ядра используется *функция Гаусса.*

Алгоритм работы данной нейронной сети представлен на рисунке 4.8

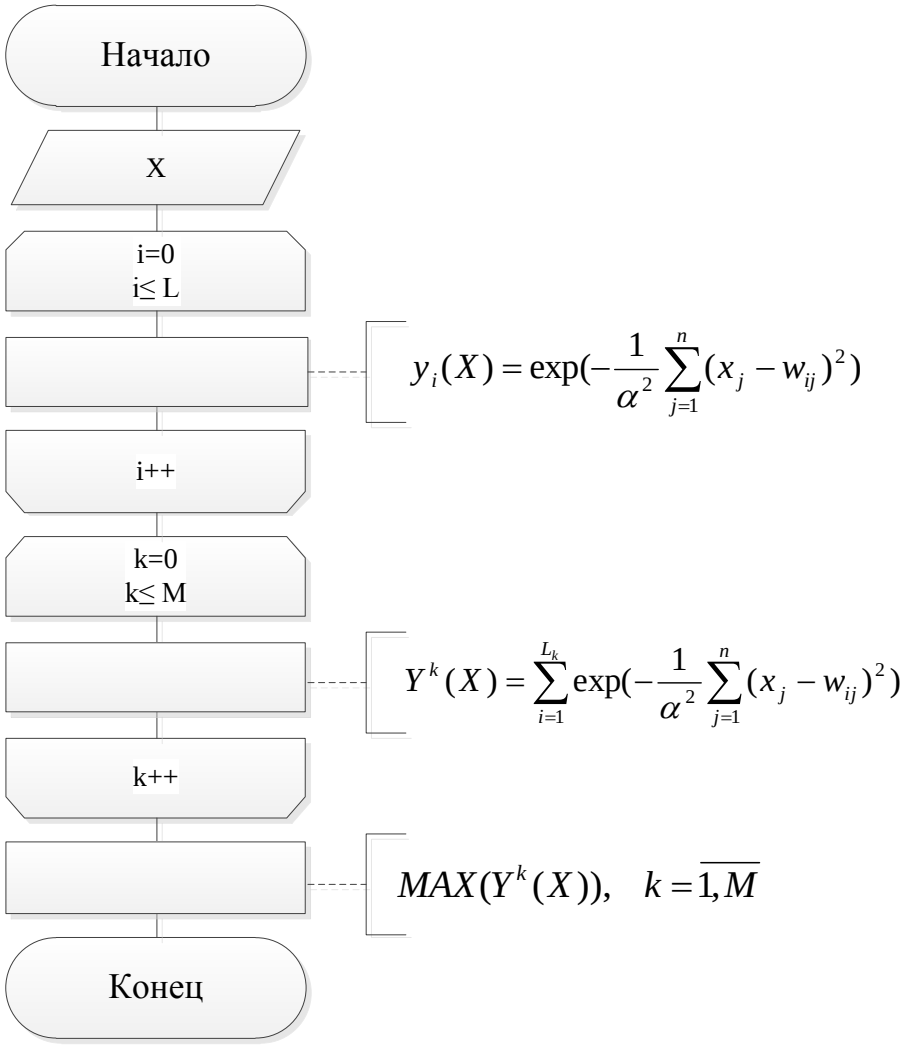


Рисунок 4.8 – Алгоритм работы вероятностной нейронной сети

В процессе работы с данной нейронной сетью и проведения эксперимента было выявлено сильное влияние изменения степени сглаживания ядерной функции на адекватность результатов функционирования нейронной сети. В связи с этим, данная нейронная сеть была исключена из рассмотрения.

Рассматриваемы нейронные сети при обучение новому образу уничтожают или изменяют результаты предшествующего обучения. В некоторых

случаях это не существенно. Если имеется только фиксированный набор обучающих векторов, они могут предъявляться при обучении циклически. Например, многослойный персептрон, обучающийся по методу обратного распространения, запоминает весь пакет обучающей информации, при этом образы обучающей выборки предъявляются в процессе обучения многократно . Обучение персептрона новому образу модифицирует синаптические связи с неконтролируемым разрушением структуры памяти о предыдущих образах[37]. Таким образом, персептрон не способен к запоминанию новой информации, и необходимо полное переобучение сети [64].

Аналогичная ситуация имеет место и в сетях Кохонена и Хемминга, обучающихся на основе самоорганизации [37,54]. Данные нейронные сети не могут отделить новые образы от искаженных или зашумленных версий старых образов.

В процессе решения задачи фильтрации электронных почтовых сообщений определенно, что не всегда по содержанию электронного сообщения можно определить является ли текст сообщения «новым» или данное сообщение является модифицированным вариантом «старого» сообщения. Такая ситуация (определение является ли предъявленный образ «новым» или модифицированным старым) нашла свое развитие в теории адаптивного резонанса нейронных сетей типа ART, и получила название проблемы *стабильности-пластичности,* т.е. когда восприятие нового образа *пластично*, адаптировано к новой информации и при этом *стабильно* – не разрушает память о старых образах. В настоящее время существуют следующие виды нейронных сетей типа ART:

1) ART 1 – нейронная сеть, предназначенная для обработки двоичных векторов;

2) ART 2 и ART 2a – нейронные сети позволяющие работать как с двоичными, так и с аналоговыми векторами;

3) ART 3 – нейронная сеть, предназначенная для моделирования временных, химических и биологических процессов;

4) ARTMAP – комбинация двух нейронных сетей;

5) FuzzyART – гибридная сеть созданная на основе нечеткой логики и ART сетей.

Базовая архитектура сетей ART состоит из трех видов нейронов: входных нейронов (слоя F_0 и F_1), распознающего слоя F_2 и управляющего нейрона R.

Слой F_0 – входной слой принимающий входные вектора для последующей обработке.

Слой F_1 – интерфейсный слой производит дополнительную обработку входного образа (соответствующие формулы в зависимости от вида нейронной сети) и передает входной вектора для классификации слою F_2.

Слой F_2 – распознающий слой, предназначенный для формирования нейрона (в процессе обучения) отвечающего за определенный образец соответствующего класса и определения нейрона имеющего максимальный резонанс.

Управляющий нейрон – нейрон, отвечающий за принятие решения о результатах классификации на основе определения меры сходства входного вектора и образа, хранящегося в памяти нейронной сети.

Базовая архитектура нейронных сетей ART представлена на рисунке 4.9.

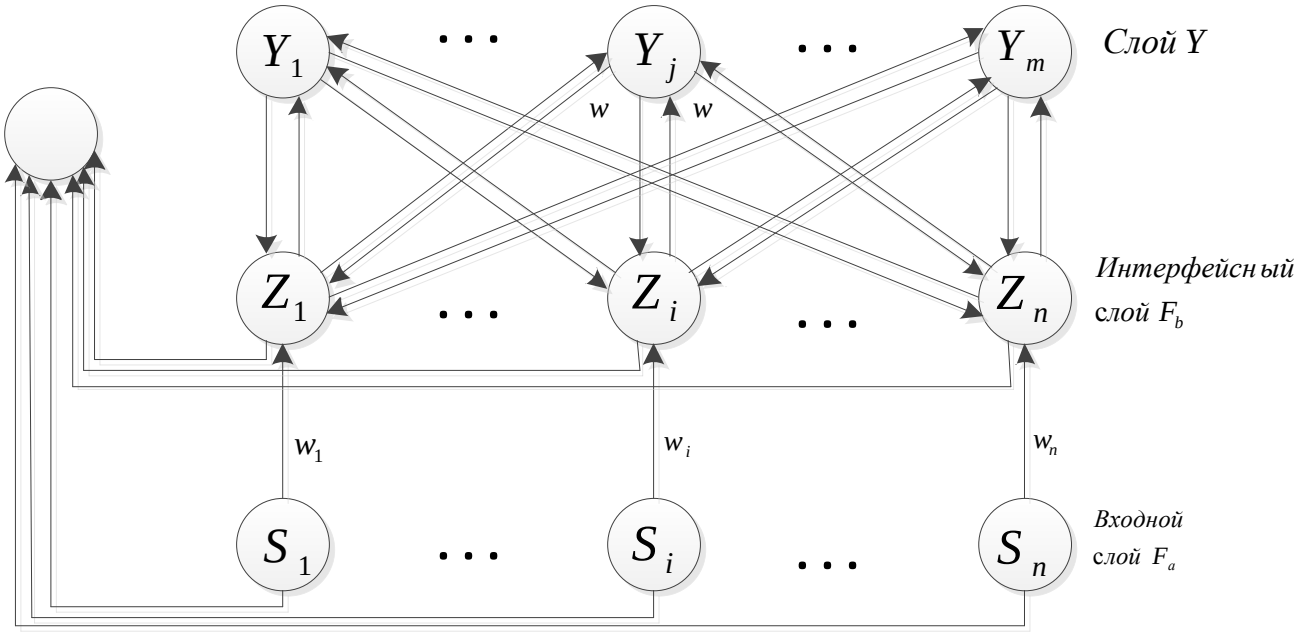


Рисунок 4.9 – Основные компоненты нейросетевого классификатора ЭС

Каждый нейрон слоя F_1 связан с каждым нейроном слоя F_2 и управляющим нейроном R.

В результате анализа различных архитектур нейронных сетей для фильтрации ЭПС предложен методический аппарат на основе адаптивной нейронной сети ART2a, которая была разработана Карпентер и Гроссбергом в 1986 г.

Нейронная сеть ART2a совмещает слой F_0 и F_1 за счет чего скорость обучения у данной нейронной сети выше чем у ART2, также обработка входного вектора происходит в два-три раза быстрее и за счет изменения в латеральных тормозящих связях нейронов. Таким образом, ART2 может использоваться для тех задач при решении которых применение ART2a не даёт желаемые результаты, но нейронная сеть ART2a может эффективно заменять ART2 в большинстве решаемых задач.

Алгоритм нейронной сети, в общем виде состоит из следующих шагов:

Стадия предварительной обработки входного вектора (входной слой) - входной вектор S_i проходит стадию предварительной обработки и подается на вход нейронной сети.

Для каждого вектора S_i из обучающей выборки вычисляется:

$$I = X' \cdot X''$$

где

$$X' = \frac{x}{\|x\|}$$

$$X'' = \begin{cases} x_i, \text{ если } x_i > \theta \\ 0, \text{в других случаях} \end{cases}$$

причем Θ принимает значения $0 < \theta \le 1/\sqrt{M}$, где M количество элементов в векторе.

Слой активации F_2

Для каждого *j-го* нейрона слоя F_2 выполняется проверка

$$T_j = \begin{cases} \alpha \cdot \sum I, \text{ еслс } j \text{ нейрон не задействован} \\ I \cdot z_j, \text{ если } j \text{ нейронзадействован} \end{cases}$$

Изначально все нейроны слоя F_2 незадействованные. Параметр α является константой, причем $\alpha \le 1/\sqrt{M}$. z_j – вес связи от нейрона входного слоя к нейрону слоя F_2. Кроме того согласно методологии нейронной сети ART2a значение α должно быть достаточно малым чтобы в том случае если $z_j=I$ для некоторых векторов, то при предъявлении этого вектора нейронной сети j нейрон должен быть выбран.

Выбор образца с наибольшим значением соответствия

Находится реакция каждого нейрона слоя F_2 с входным вектором. Затем выбирается нейрон с максимальной реакцией.

$$T_j = \max(T_j)$$

Если нейронов с максимальной реакцией больше одного то выбирается нейрон случайным образом.

Сравнение с порогом (управляющий нейрон)

Порог p показывает насколько должен входной сигнал совпадать с одним из запомненных образцов чтобы они считались похожими.

$$T_j^{\max} \ge p$$

Близкое к единице значение порога требует почти полного соответствия входного образа и образа хранящегося в памяти нейронной сети. Если значение отношения меньше установленного порога, то считается, что входной вектор отличается от образа, хранящегося в нейронной сети, и происходит поиск другого нейрона.

Если входной вектор отличается от всех образов, то он рассматривается как новый образец.

Обучение

После представления нейронной сети входного вектора z_j изменяется, так что

$$z_j^* = \begin{cases} I, \ \textit{если нейрон не задействован} \\ \beta\psi + (1-\beta)z_j, \ \textit{если нейрон задействован} \end{cases}$$

где $\psi = \begin{cases} I, \textit{если } z_j > \theta \\ 0, \textit{в противном случае} \end{cases}$

Параметр β должен принимать значение от 0 до 1.

Для адаптации нейронной сети ART2a к решению задачи идентификации электронных почтовых сообщений в её алгоритм работы были внесены следующие изменения:

1) Изменена предварительная обработка входного вектора (стадия предварительной обработки разработана для того чтобы входной вектор соответствовал основному требованию нейронной сети предъявляемому к входным сигналам: входной вектор должен оставаться неизменным без возможности сброса своих параметров до того как внутренний слой распознавания не станет активным и не начнет с ним работать).

Согласно методологии нейронной сети ART2a стадия предварительной обработки предназначена для уменьшения шума во входном векторе, т.е. выделения малоинформативных элементов и соответственно сокращения числа этих элементов. Для решения задачи фильтрации электронных сообщений данная стадия была заменена предложенной в главе 2 методикой формирования устойчивых словосочетаний.

2) Изменена структура нейронной сети ART2a добавлением дополнительного управляющего нейрона.

Для исключения ложного распознавания легитимного ЭПС в нейронную сеть в дополнение к управляющему нейрону *R*, обеспечивающему вычисление скалярного произведения векторов, введен дополнительный управляющий нейрон $R_{доп}$ определяющий меру сходства по коэффициенту Жаккара вида:

$$K_J = \frac{c}{a+b-c},$$

где *a* – количество термов во входном сообщении;

b – количество термов эталона, хранящегося в базе,

c – количество общих термов, встречаемых в 1-ом и 2-ом сообщении.

Структура нейронной сети с введенным дополнительным управляющим нейроном примет вид, представленный на рисунке 4.10

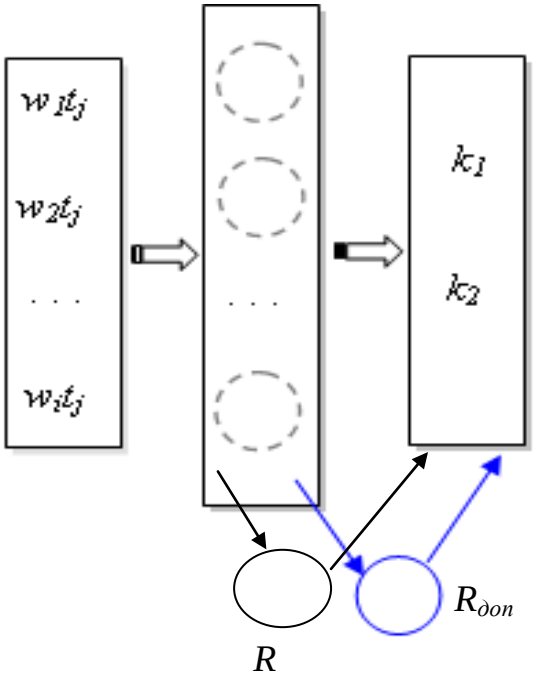


Рисунок 4.10 – Структура модифицированного классификатор ART 2a

Внесенные изменения не противоречат основным теоремам нейронной сети АРТ:

1 По достижении стабильного состояния обучения предъявление одного из обучающих векторов будет сразу приводить к правильной классификации без фазы поиска, на основе прямого доступа.

2 Процесс поиска устойчив.

3 Процесс обучения устойчив.

4 Процесс обучения конечен. Обученное состояние для заданного набора образов будет достигнуто за конечное число итерации, при этом дальнейшее предъявление этих образов не вызывает циклических изменений значений весов.

Модифицированный алгоритм нейросетевой классификации примет вид, представленный на рисунке 4.10

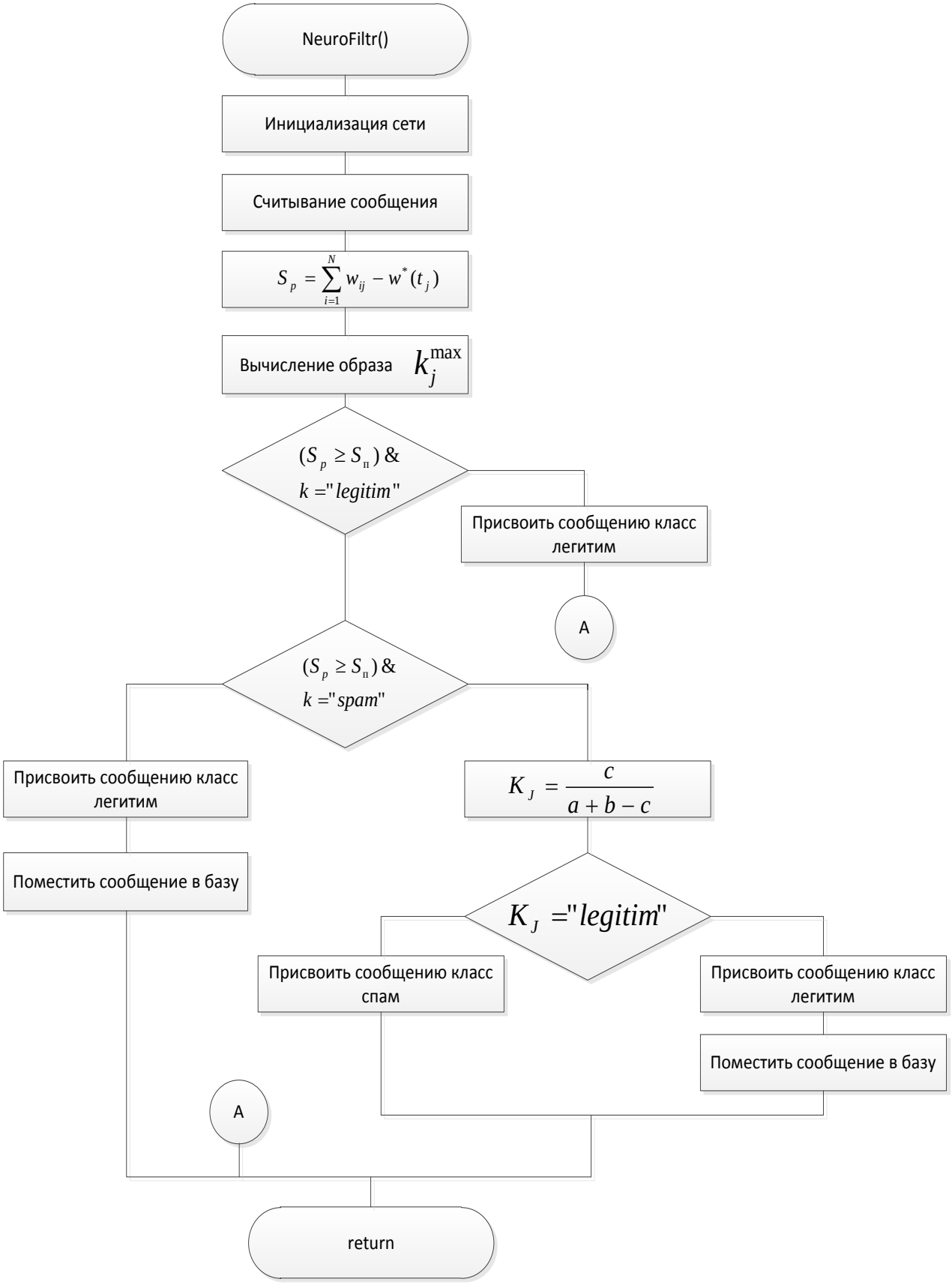


Рисунок 4.10 – Укрупненный алгоритм нейросетевой классификации

Таким образом, получил развитие метод основанный на нейронной сети ART2a для задачи интеллектуальной фильтрации ЭПС .

4.3 Разработка алгоритмов классификации электронной корреспонденции

Анализ объекта и предмета исследования позволили выявить информационные потоки представленные в виде контекстной диаграммы, нотация DFD на рисунке 4.11.

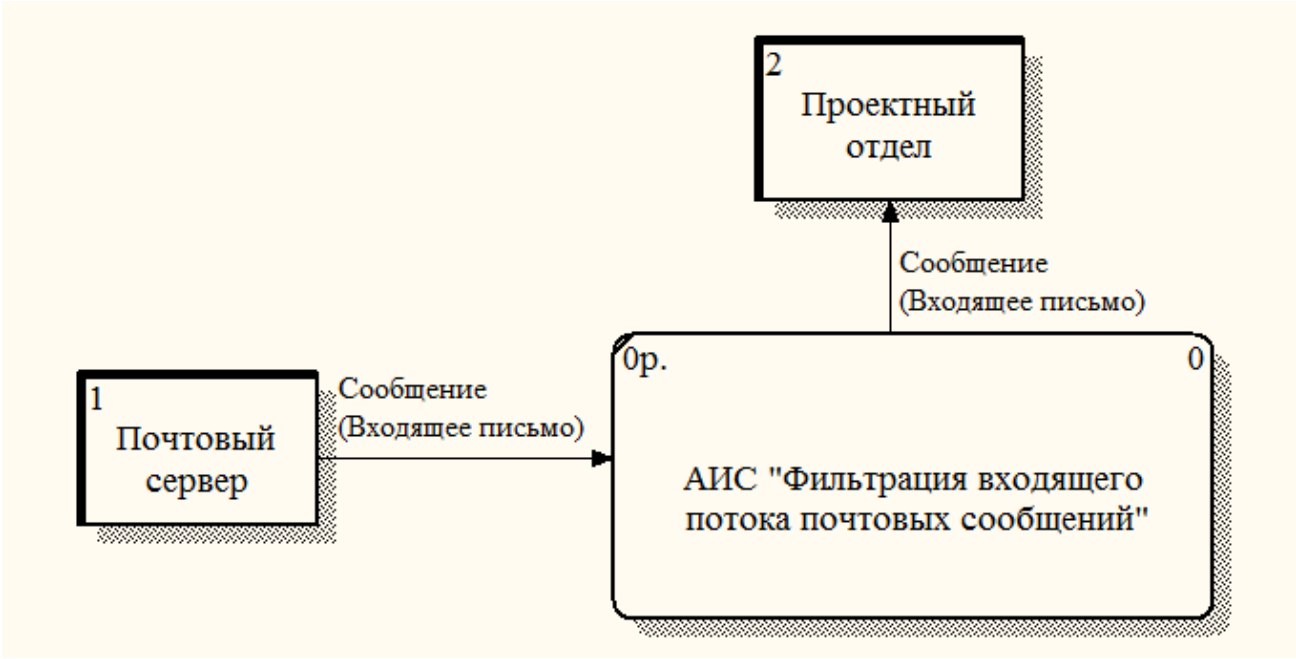


Рисунок 4.11 - Информационный поток.

Программный проект прототипа спам-фильтра реализован в соответствии с моделью IDEF0, представленной на рисунке 4.12 – 4.13.

Важной стороной проектирования является описание функционального назначения программной системы, которое позволяет определить масштаб разработки.

Проект программной системы (ПС) позволяет решать следующие задачи: фильтрация входного потока электронных сообщений, настройка фильтрации электронных сообщений, настройка подключений фильтра, обучение фильтра

на спамных и легитимных сообщениях, тестирование настроенного и обученного фильтра, ведение базы термов и легитимных сообщений, ввод и обновление информации в базе данных, формирование выходных отчётных документов.

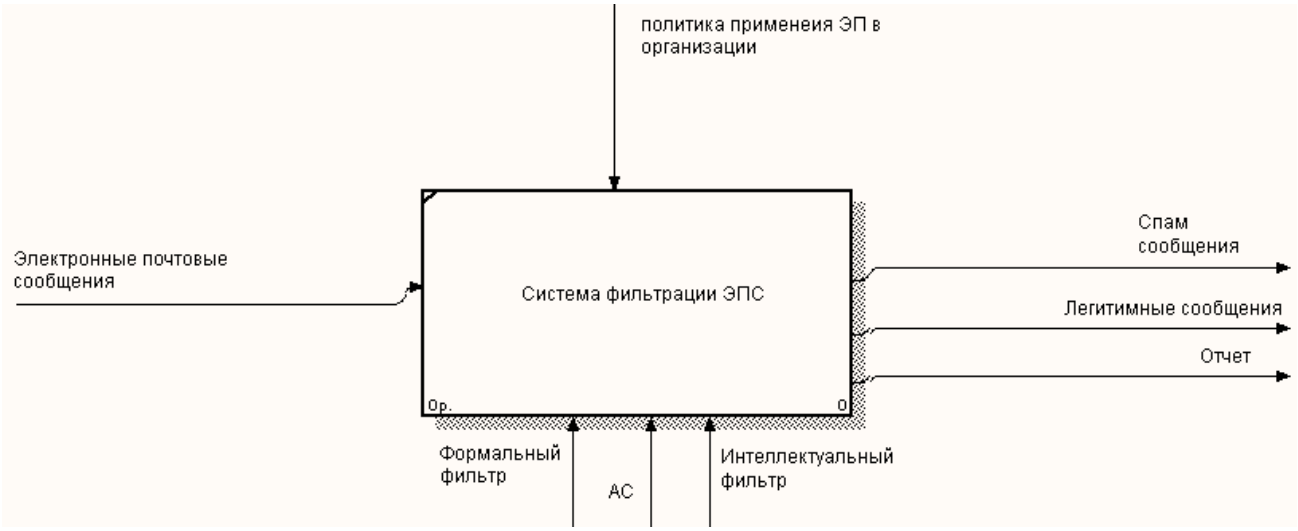


Рисунок 4.12 - Контекстная диаграмма в нотации IDEF0

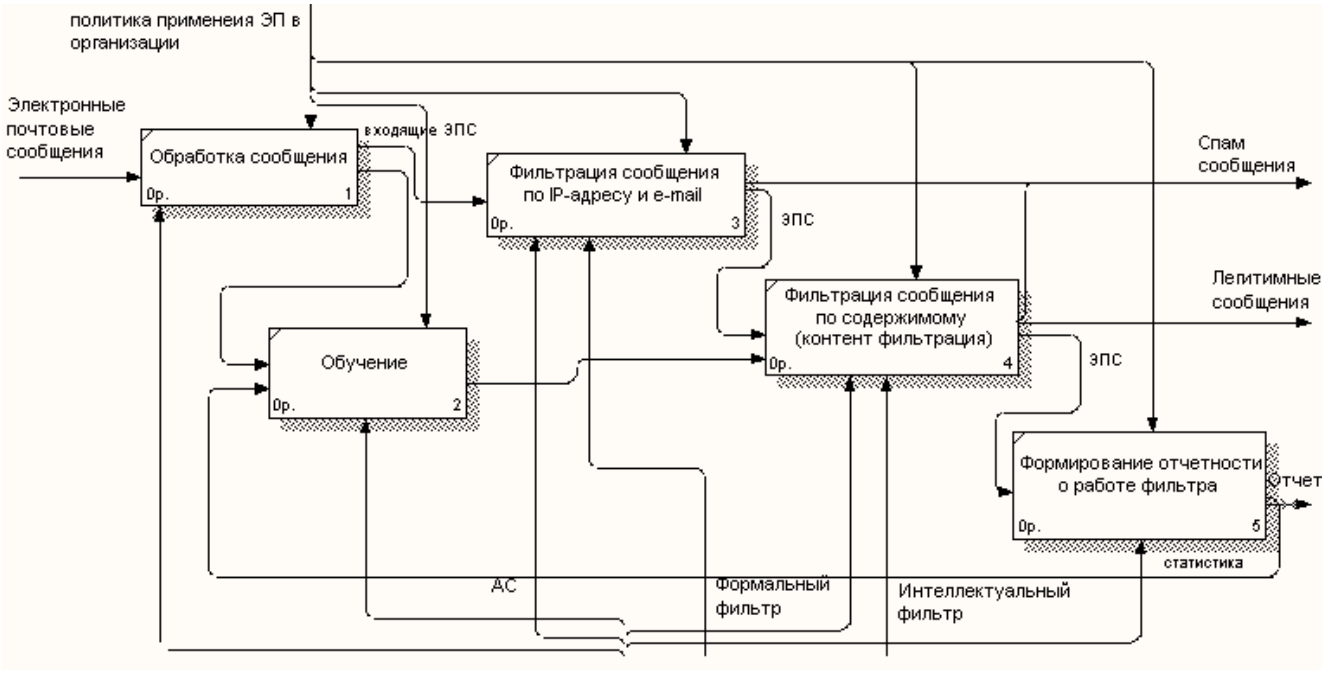


Рисунок 4.13 - Функциональная модель прототипа системы фильтрации по методологии IDEF0

Функциональная схема строится с целью однозначного понимания всех функций, выполняемых данной системой. В большинстве случаев функциональная спецификация формулируется на естественном языке при помощи специальных объектов и утверждений, конкретно описывающих функции программного средства.

Разработанная функциональная схема ПС представлена на рисунке 4.14. ПС поддерживает режим администрирования, позволяющий проводить обучение и тестирование системы фильтрации.

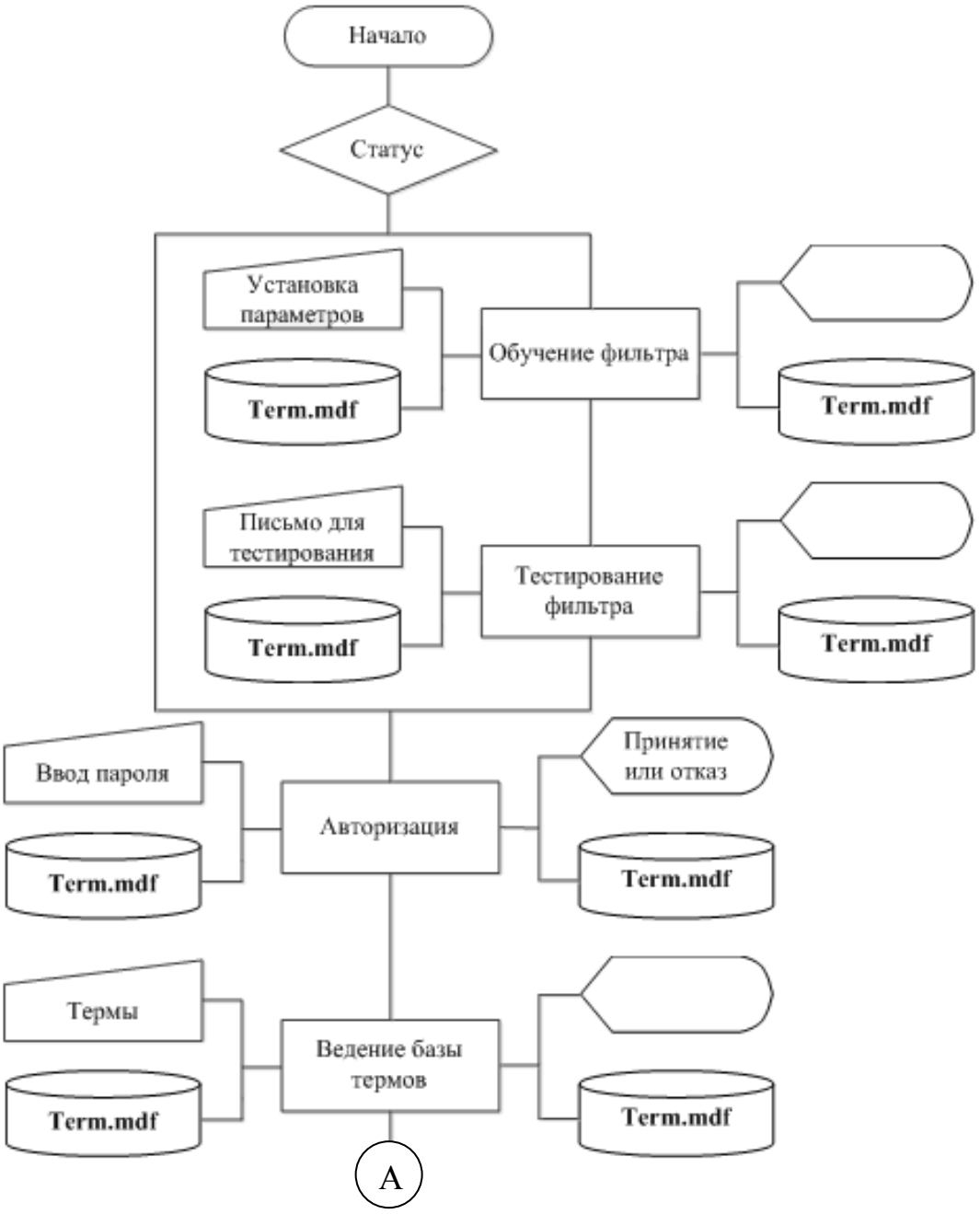


Рисунок 4.14 – Функциональная схема ПС

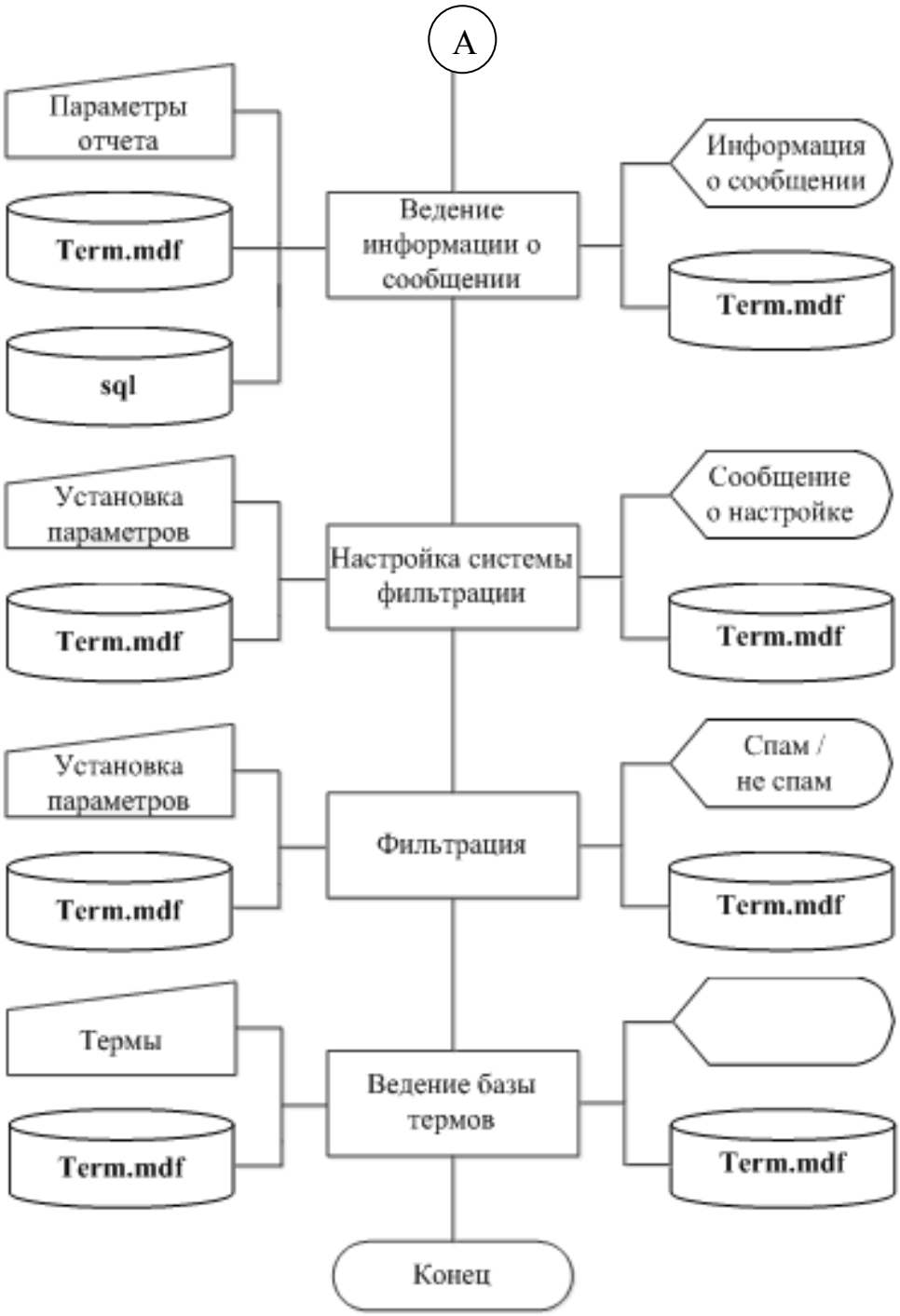


Рисунок 4.14 – Функциональная схема ПС

Реализация ПС, имеющая следующие основные экранные формы, представлена в приложении Б.

5 Исследование достоверности системы контентной фильтрации почтовых серверов информационно-телекоммуникационных систем корпоративных предприятий

Предмет исследования позволяет выполнить эксперимент на основе реальных электронных сообщений, часть из которых можно отнести к несанкционированным рассылкам. Проведение эксперимента включает две стадии: составление плана эксперимента и проведение исследования [3].

План эксперимента должен отражать последовательность работы с интеллектуальной системой фильтрации. В плане должен быть предусмотрен эксперимент удовлетворяющий целям исследования. Эксперимент должен сопровождаться осмыслением итогов, что служит основой анализа результатов и принятия решений.

5.1 Методика оценки эффективности фильтрации электронной почты

Методика проведения эксперимента включает выполнение следующих этапов [87]:

– формулировка целей исследования;

– выбор существенных факторов (параметров);

– разработка и реализация плана эксперимента;

– обработка результатов эксперимента с помощью метода проверки статистических гипотез;

– формирование рекомендаций для дальнейших исследований.

Целью экспериментального исследования являлась проверка эффективности использования модели представления ЭС и метода классификации на основе нейронной сети.

Задачи экспериментального исследования:

1) проверка опытно-экспериментальным путем эффективности использования предложенной модели и метода нейросетевой фильтрации спама;

2) разработка рекомендаций для дальнейших исследований.

В ходе исследования эффективности использования предложенных модели и метода фильтрации спама возникла необходимость отслеживания динамики изменения степени корректной фильтрации входящих сообщений в зависимости от различных параметров нейронной сети и порога оценки ЭС при работе с системой фильтрации спама.

Экспериментальный набор данных состоит из электронных сообщений представляющих собой спам-рассылку и легитимных писем. Выборка спам тематики состоит из писем, предоставленных для тестирования лабораторией Касперского. Множество легитимных писем получено из общедоступных ресурсов и писем полученных от определенных пользователей.

В основу обоснования порога соответствия входящего ЭПС определенному классу положен метод теории статистических решений для задачи проверки двухальтернативной гипотезы $\boldsymbol{H_0}$ и $\boldsymbol{H_1}$, выражающие предположения о легитимности ЭПС или наличии спам-вторжения.

Для того, чтобы данная задача обрела математическую содержательность в качестве показателя эффективности системы фильтрации приняты ошибки о правильной классификации легитимных ЭПС.

Существуют несколько метрик оценки качества работы алгоритмов классификации [61,85,87,95]. Базовыми характеристиками качества классификации приняты уровни ошибок первого и второго рода.

Используемые переменные:

N_{sp} – количество объектов, относящихся к классу спам;

N_l – количество объектов, относящихся к классу легитимных сообщений;

FN_{sp} – количество спам-рассылок, классифицированных как легитимное письмо;

FP_l – количество легитимных писем, классифицированных как спам-рассылка.

Количество объектов относящихся к классу спам и количество объектов относящихся к классу легитимных сообщений в сумме должны соответствовать общему количеству объектов в экспериментальной выборке, т.е. $N = N_l + N_{sp}$.

Пусть выдвинута гипотеза H_0 о том, что программная система правильно классифицирует легитимные рассылки или распознает спам.

Тогда, количество ложных пропусков FN_{sp} и количество ложных обнаружений FP_l определят количество правильно классифицированных легитимных ЭС TP_l и верных обнаружений спама TN_{sp} вида:

$$TP_l = N_l - FP_l;$$

$$TN_{sp} = N_{sp} - FN_{sp};$$

Отсюда ошибка первого $\boldsymbol{\alpha}$ (вероятность отвергнуть нулевую гипотезу, когда она ложна, т.е. вероятность принять решение о легитимности сообщения, когда оно спам) и второго рода $\boldsymbol{\beta}$ (вероятность отвергнуть нулевую гипотезу, когда она справедлива, т.е. вероятность отвергнуть решение о легитимности сообщения, когда оно легитимно) будут определяться зависимостями:

$$\alpha = FN_{sp} / N_{sp} ,$$

$$\beta = FP_l / N_l ,$$

Данные величины характеризуют качество распознавания, т.к. не зависят от количества объектов в тестовом наборе.

На основе характеристик TP и TN можно рассчитать меру полноты и точности. Для наиболее наглядного представления при исследовании чаще оперируют не абсолютными показателями, а относительными (долями) выраженными в процентах. Мера полноты ($precision$) оценивает долю верного распознавания относительно всех объектов определенного класса. Мера точ-

ности ($recall$) оценивает долю верных обнаружений относительно всех объектов. Данные меры рассчитывают по следующим формулам:

$$precision = \frac{TP}{TP + FP_l} * 100\%,$$

$$recall = \frac{TP}{TP + FN_{sp}} * 100\%,$$

Используя следующую зависимость можно определить долю ложно классифицированных объектов FPR соответствующих классов:

$$FPR = \frac{FP_l}{TN + FP_l} * 100\%,$$

Сводная оценка качества классифицирования (F-мера), зависящая от полноты и точности, определяется зависимостью:

$$F = \frac{2}{1/\ precision + 1/ recall},$$

5.2 Технология проведения имитационного эксперимента

Проведение имитационного эксперимента позволило решить следующие задачи [87]:

1 Оценить настройку и обучение классификатора.

2 Определить объекты классификации наиболее вероятно вызывающие ошибки работы классификатора с целью корректировки или расширения обучающей выборки.

3 Обучение классификатора с разными параметрами и оценка результатов с целью определения наилучших параметров.

4 Выявление признаков лучше всего характеризующих определенные классы.

Метод тестовой выборки [54,61,87]

Является самым простым способом тестирования. Данный метод заключается в том, что для осуществления тестирования из обучающей выборки выбирается 10-20% образцов. Необходимо отметить, что такая выборка должна быть сбалансирована, т.е. должна состоять из одинакового количества объектов, предназначенных для тестирования каждого класса.

Для оценки результата тестирования необходимо сделать всего лишь одно обучение и одну проверку тестирования для каждого объекта из тестовой выборки. В качестве недостатка данного способа можно отметить сильную зависимость результатов тестирования от того какие объекты попали в тестовую выборку. Так, например, если в тестовую выборку попали объекты, большинство которых находилось в обучающей выборке - результаты тестирования будут хорошими, в противном случае, если объекты в тестовой выборке окажутся специфическими, то результаты тестирования покажут низкий уровень правильно классифицируемых объектов и высокое количество ложно классифицируемых тестовых примеров. Такой метод не позволит решить задачу настройки параметров классификатора, определение и подбора обучающей выборки.

Другой более сложный и трудоемкий *метод k-подмножеств* (*k-folds*)[54,85,87].

Сущность подхода заключается в разбиении экспериментальной выборки на *k* равных частей. Причем распределение сообщений по частям осуществлялось равномерно. Далее производится *k* запусков работы классификатора (обучение и тестирование). В ходе каждого запуска работы классификатора (*k-1)* часть участвует в обучении и одна в тестировании, при этом тестовая часть постоянно меняется.

В результате каждого запуска системы фильтрации фиксировались: значения двух вероятностных характеристик – ошибки I рода $\boldsymbol{\alpha}$ и II рода $\boldsymbol{\beta}$, сводная оценка качества классификации (F-мера), меры полноты и точности. По результатам всех тестов вычислялись средние значения всех величин. Ре-

зультаты классификации зависят от разбиения на подмножества тестовых данных, но определение средних величин позволит достаточно точно оценить качество работы классификатора. Следует отметить, что такая методика требует больших временных затрат что послужило целью разработки имитационной модели схема которой представлена на рисунке 5.2, алгоритм метода k–подмножеств (k-foldes) представлен на рисунке 5.1.

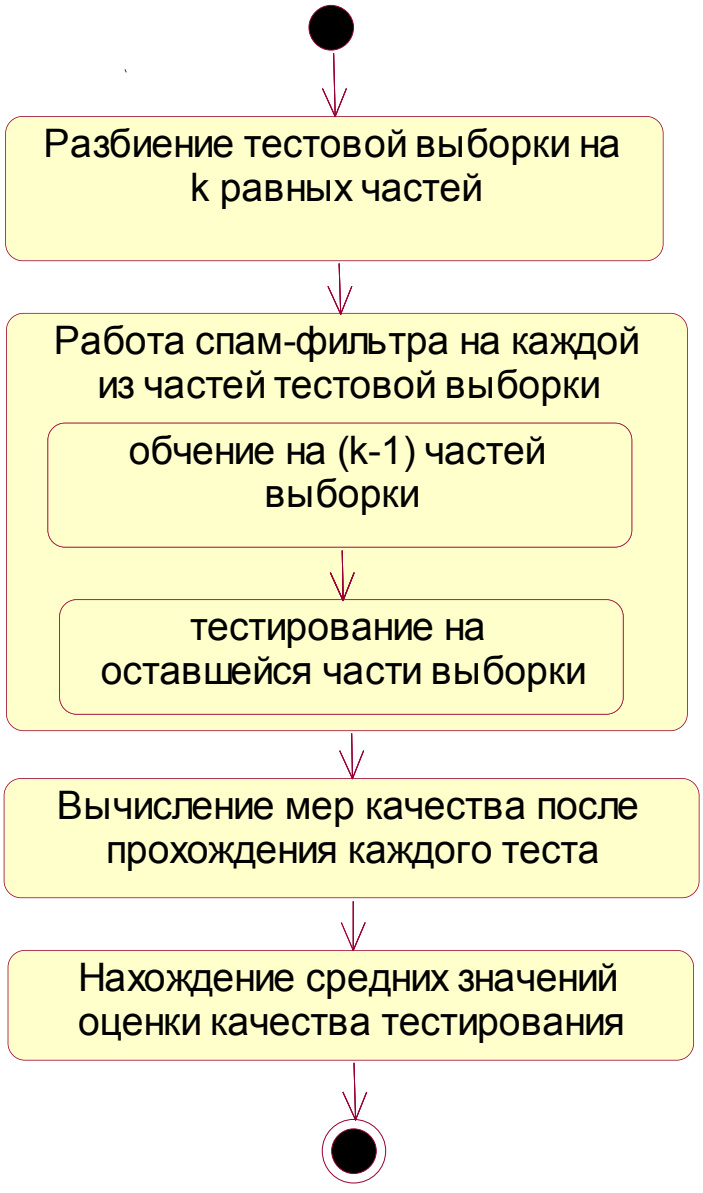


Рисунок 5.1 – Алгоритм оценки результатов методом k–подмножеств

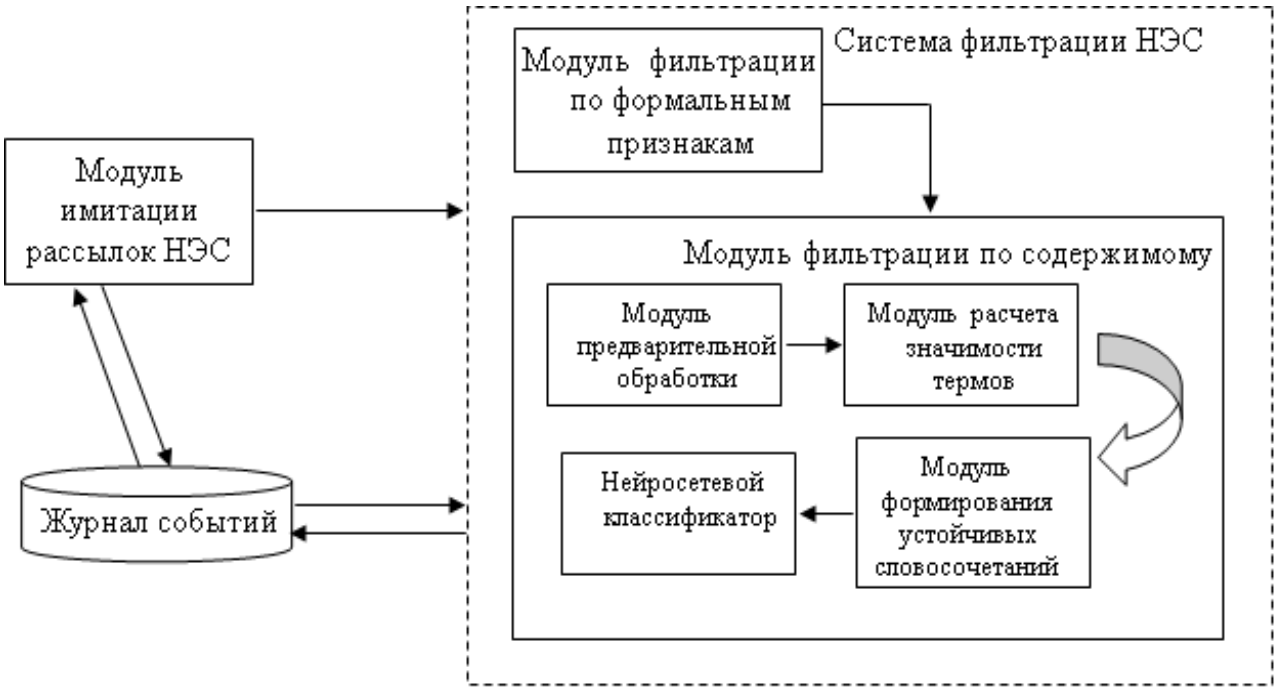


Рисунок 5.2 – Схема имитационного эксперимента

Модуль имитации рассылок электронных почтовых сообщений позволяет осуществлять рассылку спам-сообщений и легитимной почты.

Журнал событий осуществляет фиксацию результатов имитационного эксперимента, а так же сбор и хранение неправильно классифицированных сообщений для последующего анализа.

Модуль фильтрации по формальным признакам осуществляет прием письма от модуля рассылки сообщений и осуществляет фильтрацию по формальным признакам, введенным в модуле «установка формальных признаков».

Модуль фильтрации по содержимому осуществляет контентную фильтрацию электронных почтовых сообщений и включает в себя *модуль предварительной обработки, модуль расчета значимости термов, модуль формирования устойчивых словосочетаний, нейросетевой классификатор.*

Работа модуля предварительной обработки заключается в приведение электронного сообщения к стандартному типу кодировки, удаление стоп-слов, знаков пунктуации, нормализация орфографии.

Модуль расчета значимости термов позволяет произвести расчет весов термов, при этом структура модуля предусмотрена таким образом, что суще-

ствует возможность выбора различных мер значимости термов.

Модуль формирования устойчивых словосочетаний предназначен для выделения устойчивых словосочетаний, согласно предложенной методике.

В ходе проведения эксперимента исследованы несколько версий системы фильтрации, представленные в таблице 5.1. Для каждой из версии систем фильтрации изменялся порог на соответствие классу S_n. Экспериментальная выборка ЭПС для оценки эффективности прототипа системы фильтрации состояла из легитимных сообщений документооборота и спам-рассылки. Тематика сообщений экспериментальной выборки представлена в таблице 2. Всего исследовано 908 ЭПС (424 легитимных сообщений и 484 спам-сообщений) и осуществлено 13 запусков прототипа предложенной системы фильтрации ЭПС. Порог соответствия S_n изменялся в диапазоне от 0,4 до 0,9.

Таблица 5.1 – Варианты построения классификатора

Название	Модель ЭС	Вес	Метод сокращения признакового пространства	Выделение устойчивых словосочетаний	Алгоритм классификации
Met1	векторная	Tf-idf	RF	+	нейрон. сеть Art
Met2	векторная	Ltc	IG	+	нейрон. сеть Art
Met3	векторная	Ltc	RF	-	нейрон. сеть Art
Met4	векторная	Ltc	IG	+	нейрон. сеть Art
Met5	векторная	Tf-idf	RF	+	нейрон. сеть Art

Таблица 5.2 – Тематика сообщений

Вид сообщения	Тематика сообщений
Спам сообщения	“Пустые” сообщения, содержащие только ссылки или вложения.
Спам сообщения	Реклама товаров
Спам сообщения	Реклама услуг (юридических, бухгалтерских, строительных, образовательных, туристических, медицинских и проч.)
Спам сообщения	Приглашения на курсы, предложения схем «отмывания» денег («нигерийские» письма)
Легитимные сообщения	Деловая переписка (свободная форма)
Легитимные сообщения	Деловая переписка (приказы, распоряжения, отчеты и т.п.)
Легитимные сообщения	Приглашения на участия в грантах, конференциях, выставках и т.п.

Таким образом, предложенная методика эксперимента позволяет оценить эффективность использования предложенной модели ЭС на основе устойчивых словосочетаний и метода классификации основанного на нейронной сети адаптивного резонанса.

5.3. Сравнительная оценка эффективности контентной фильтрации почтовых сервисов

Экспериментальная выборка ЭПС для оценки эффективности прототипа системы фильтрации состояла из легитимных сообщений документооборота и спам-рассылок. Тематика сообщений экспериментальной выборки представлена в таблице 2. Всего исследовано 908 ЭПС (424 легитимных сообщений и 484 спам-сообщений) и осуществлено 13 запусков прототипа предложенной системы фильтрации ЭПС. Порог соответствия S_n изменялся в диапазоне от 0,4 до 0,9.

В таблице 5.3 представлены показатели эффективности версий и сравнительные результаты оценки предложенного фильтра легитимных ЭПС при различных значениях порога (таблица 5.4).

Таблица 5.3 - Показатели эффективности версий системы

Показатели эффективности	Название версий				
	Met1	Met2	Met3	Met4	Met5
ошибка I рода (%)	13	17	19	16	15
ошибка II рода (%)	9	12	14	7	3
мера полноты (%)	96	88	85	91	98
мера точности (%)	90	87	83	93	96
F-мера (%)	80	79,9	80	81	83

Таблица 5.4 – Показатели эффективности при изменении значения порога

Показатели эффективности	Значение порога			
	Sп=0,4	Sп=0,7	Sп=0,8	Sп=0,9
ошибка I рода (%)	15	17	7	8,7
ошибка II рода (%)	7	3	0.1	5
мера полноты (%)	75	95,7	98	97
мера точности (%)	87	95	96	95
F-мера (%)	78,9	82	83	83,4

Результаты проведенного эксперимента представлены в виде диаграммы на рисунке 5.3-5.5

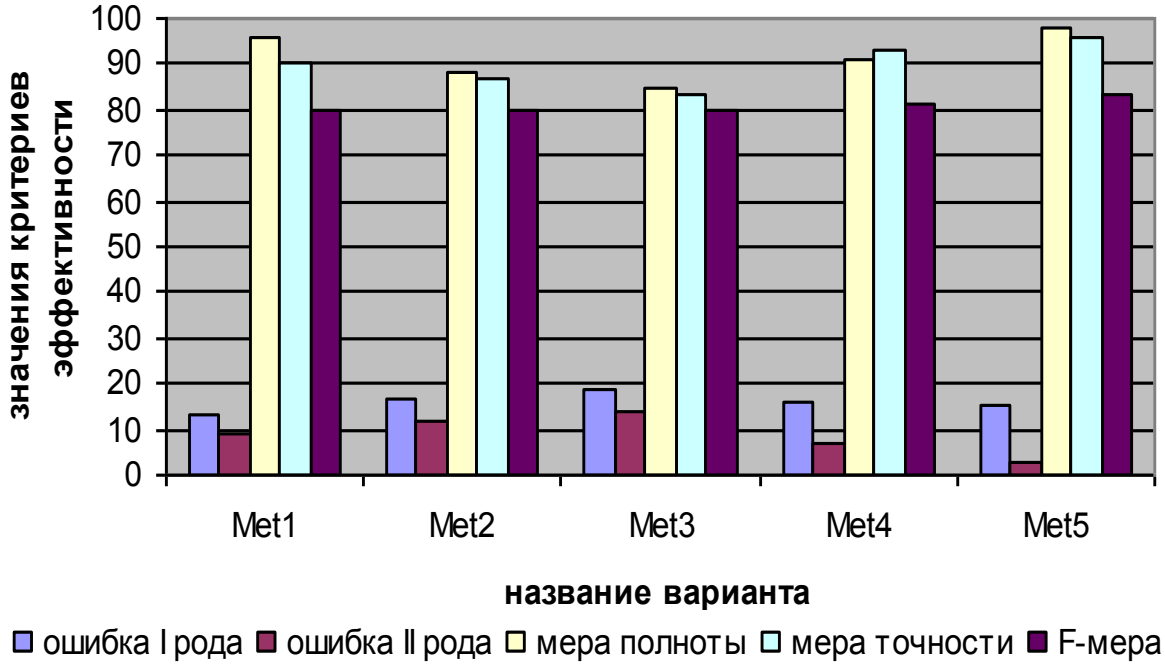


Рисунок 5.3 –Значения критериев эффективности разных версий

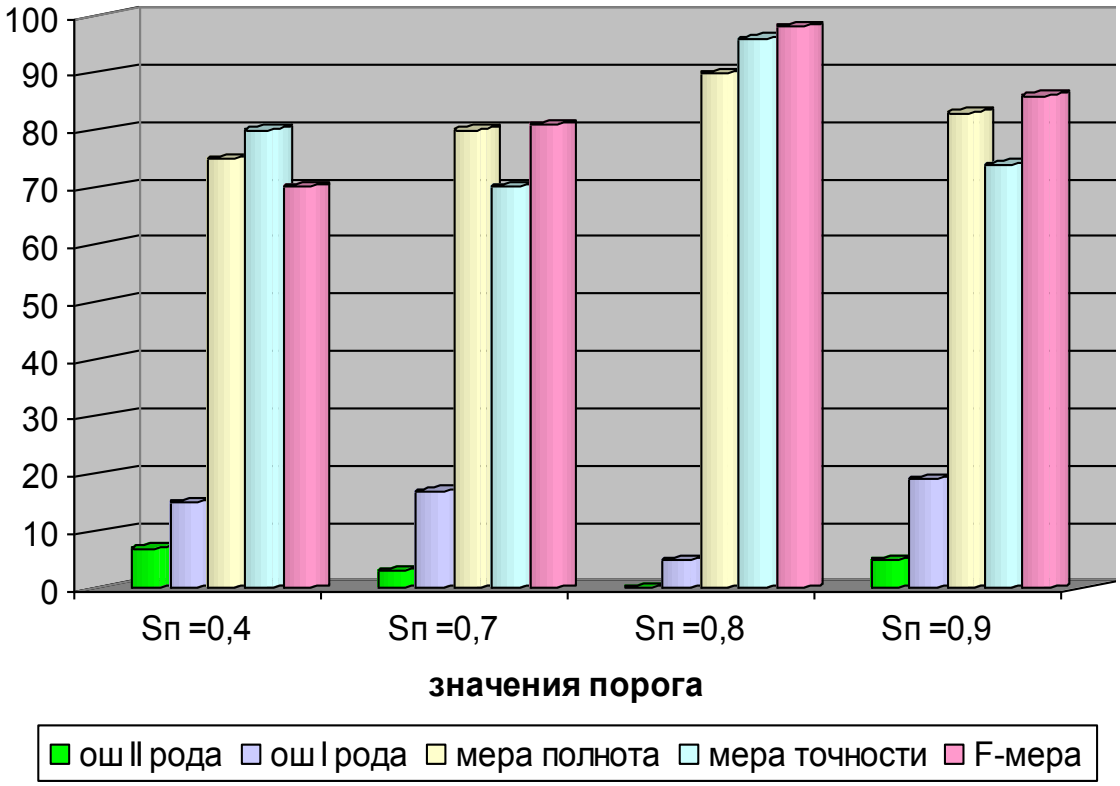


Рисунок 5.4 – Диаграмма результатов имитационного эксперимента

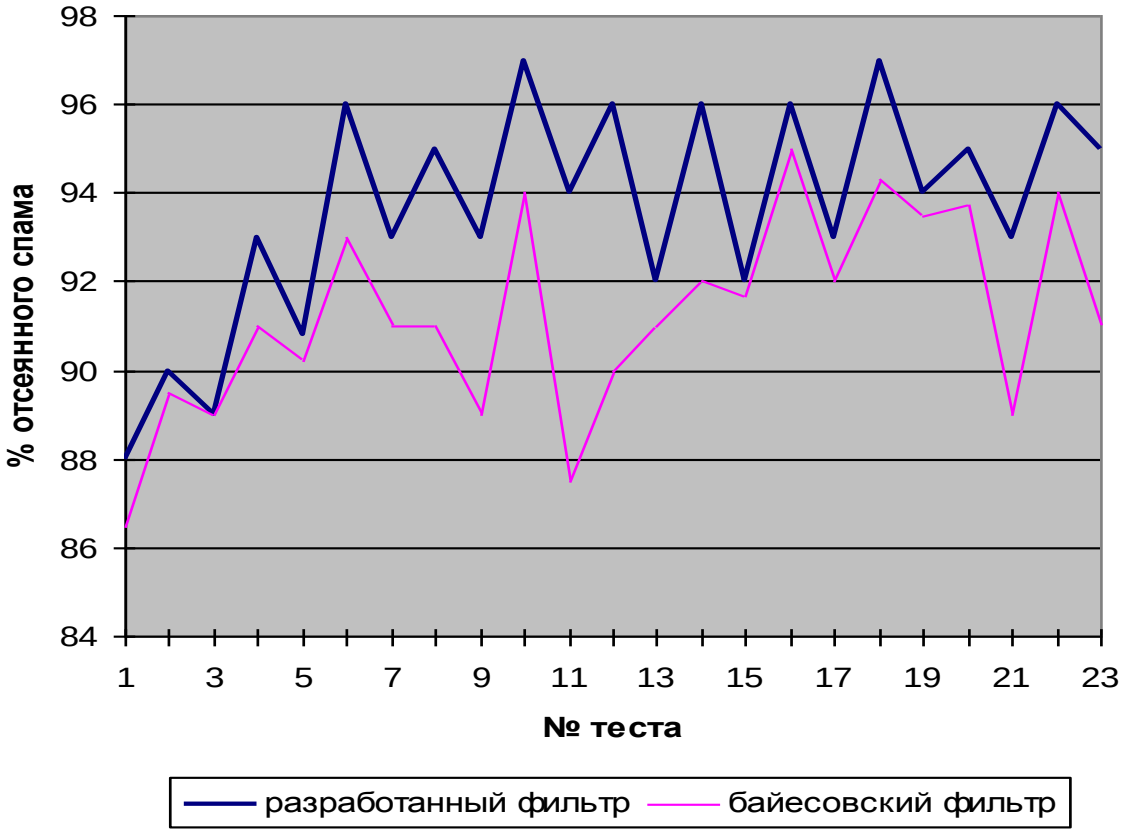


Рисунок 5.5 –Сравнительная оценка эффективности предложенных технических решений

Как видно из результатов имитационного эксперимента наиболее эффективна версия met5. Анализ результатов исследований met5 показал, что при изменении порога соответствия S_n изменяются показатели качества фильтрации ЭПС. При установке порога $S_п$ = 0,4 число легитимных сообщений принятых за спам составляет 7% , а число спам-сообщений принятых за легитимные составляет 15%. При увеличении порога $S_п$ до 0,7 снижается уровень ошибки 2 рода до 3%, однако уровень ошибки 1 рода составляет 17% и при дальнейшем увеличении порога S_n продолжает расти, что свидетельствует о высокой требовательности нейронной сети (при установленном пороге, близком к единице, нейронная сеть требует почти полного соответствия входного сообщения и прототипа хранящегося в базе). Установка порога $S_п$=0,8 показывает лучшие результаты: ошибка 2 рода стремиться к 0 и составляет 0,001, ошибка 1 рода – 0,07. Доля НЭС, выявленная предложенной системой фильтрации, выше, чем у байесовского фильтра, при вероятности ложного срабатывания не более 0,05.

Таким образом, результаты эксперимента подтверждают достижение поставленной цели исследования, и свидетельствуют об эффективности разработанного прототипа системы фильтрации ЭПС при пороге $S_п$=0,8.

5.4 Направления дальнейших исследований

Исследования выявили основные тенденции развития систем защиты электронной почты от несанкционированных рассылок:

– защита электронной почты приобретает черты фундаментального научного направления, имеющая свой методологический базис, математический аппарат, принципы, технологии и ряд общих закономерностей;

– средства массовых рассылок и защита от их несанкционированной части продолжают повышать уровень интеллектуализии, приобретая характер систем принятия решения;

– появляется идеология активной и адаптивной спам-защиты, мощность которой может изменяться, оставаясь адекватной угрозе;

– при оценке достаточности и эффективности спам-защиты могут быть использованы методы оценки риска как общий принцип определения минимальной достаточности.

Дальнейшие направления исследований в области системы защиты от несанкционированных рассылок, которые, по мнению автора, актуальны, приведены в таблице 5.5.

Таблица 5.5 – Дальнейшие направления исследований в области защиты от несанкционированных рассылок

Анализ среды Internet как предпосылки НСР	Исследования механизмов спам-рассылок	Разработка методов и средств обнаружения спам-рассылок
Анализ системного и прикладного ПО. Исследование сетевого оборудования и анализ протоколов. Анализ современных типовых технологии получения информации о спам-рассылках.	Систематизация и моделирование механизмов спам-рассылок и других аномальных событий. Модели спам-рассылок, интегрированных с информационными атаками. Создание механизмрв адаптивной защиты	Обнаружение и предотвращение спам-рассылок. Обнаружение и защита от сетевых вирусов Активное противоборство спам-воздействиям Анализ рисков возникновения спам-атак, их последствий и определение фактической степени необходимой защиты.

Таким образом, настоящая работа является очередным шагом в совершенствовании методов и средств защиты от несанкционированных рассылок в информационно-телекоммуникационных системах.

Заключение

Результаты и выводы проведенных исследований:

1 Системный анализ защиты почтовых сервисов ИТКС корпоративных предприятий с территориально-распределенной структурой свидетельствует о необходимости развития существующих методов фильтрации ЭПС на основе моделей, отражающих семантику легитимной почтовой корреспонденции и учитывающих изменяющиеся информационные потребности адресатов, для исключения ложной классификации легитимных ЭПС. Выявлены основные признаки электронных почтовых сообщений, необходимые для классификации электронных рассылок.

2 Разработана модель электронного сообщения в форме устойчивых словосочетаний, которая позволяет без потери смыслового содержания обеспечить классификацию легитимной электронной корреспонденции в реальном масштабе времени. Эффект достигается применением меры значимости термов для устранения больших различий в частотах фиксации термов, исключением термов с малой информативной нагрузкой, выделением устойчивых словосочетаний, позволяющих усилить смысловое содержание термов и сократить пространство признаков на 25% за счет использования дополнительных мер близости между термами в сообщении и тесноты взаимосвязи между ними.

3 Предложена методика и разработаны алгоритмы контентной фильтрации электронной корреспонденции почтовых сервисов на основе нейросетевого классификатора АРТ2а, отличающиеся использованием дополнительного нейрона для проверки сообщений, идентифицируемых как несанкционированные сообщения, мерой сходства векторов Жаккара.

4 Предложен прототип системы защиты почтовых сервисов, основанный на двухуровневой фильтрации электронных почтовых сообщений, отличающийся предварительной подготовкой сообщений к нейросетевой классификации и обеспечивающий контентную фильтрацию легитимной корреспонденции в реальном масштабе времени.

5 Результаты экспериментальных исследований предложенного прототипа системы защиты почтовых сервисов свидетельствуют о повышение достоверности идентификации почтовой корреспонденции по ошибке классификации легитимных сообщений до 0,1% , а по ошибке классификации спам-рассылок до 7%

Список использованных источников

1 Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Е. И. Большакова [и др.] ; Моск. гос. ин-т электроники и математики. — Москва : МИЭМ, 2011. — 272 с. - ISBN 978–5–94506–294–8.

2 Агеев, М. С. Выбор факторов для классификации веб-страниц и веб-сайтов [Электронный ресурс] / М. С. Агеев, Б. В. Добров, Н. В. Лукашевич // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2010 : семинар в рамках Всерос. науч. конф. RCDL'2010, 15 окт. 2010 г., Казань / Казан. ун-т. - Казань, 2010. - С. 28-39. – Режим доступа : http://www.cir.ru/docs/ips/publications/2010_romip_uis.pdf . - Дата обращения 08.11.2013.

3 Адлер, Ю. П. Планирование эксперимента при поиске оптимальных условий / Ю. П. Адлер, Е. В. Маркова, Ю. В. Грановский. - 2-е изд., доп. и перераб. – Москва : Наука, 1976. – 280 с.

4 Алгоритм обучения по Байесу [Электронный ресурс] / Учеб.-науч. комплекс «Института прикладного системного анализа» М-ва образования и науки Украины. - Режим доступа : http://iasa.org.ua/ lections/tpr/studying/bayes.htm . - Дата обращения 07.11.2013.

5 Арутюнова, Н. Д. Предложение и его смысл / Н. Д. Арутюнова. — М. : УРСС, 2005. – 122 с.

6 Афонин А.И. Что такое спам? [Электронный ресурс] / «Научно - технический журнал» №2, февраль 2013. - Режим доступа : http://technomag.bmstu.ru/doc/551869.html. - Дата обращения 12.07.2011.

7 Бакаткин, А. Спам помогает в создании ИИ [Электронный ресурс] / Александр Бакаткин // SubsCribe.RU : информационный канал. – Электрон. данные. – Режим доступа: http://digest.subscribe.ru/inet/inet/n70149181.html. - Дата обращения 08.11.2013.

8 Барабанов, А. Борьба со спамом как фактор, снижающий надежность почтовой доставки. //Системный администратор, №8, 2007 г., – С. 6-13.

9 Баранова, М. И. Информационные технологии: открытые системы, сети, безопасность в системах и сетях: учеб. пособие / М. И. Баранова, В. И. Кияев ; Санкт-Петербург. гос. ун-т экономики и финансов, Каф. информатики. – Санкт-Петербург : СПбГУЭФ, 2010.– 267 с. - ISBN 978-5-7310-2593-5.

10 Баранов, В. В. Реинжиниринг бизнес-процессов: этапы разработки и реализации [Электронный ресурс] / В. В Баранов. – Режим доступа: http://www.elitarium.ru/2012/11/14/reinzhiniring_biznes_processov_jetapy_razrabotki_realizacii.html.

11 Белоглазова, А.В. Спам и методы борьбы с ним [Электронный ресурс] // Молодежь и наука: сборник материалов VI-й Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых /отв. ред. О.А. Краев - Красноярск : Сиб. федер. ун-т, 2011. – Режим доступа: http://conf.sfu-kras.ru/sites/mn2010/pdf/8/16_8.pdf.

12 Борьба со спамом в 2005 году [Электронный ресурс] // КомпьютерПресс. – 2005. - № 7. – Режим доступа: http://www.compress.ru/article.aspx?id=11438&iid=449. – Дата обращения 07.11.2013.

13 Браславский, П. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста = Comparison of four methods for automatic two-word term extraction [Электронный ресурс] / П. Браславский, Е. Соколов // Компьютерная лингвистика и интеллектуальные технологии: тр. Междунар. конф. «Диалог 2006». – Москва, 2006. - С. 88-94. – Режим доступа : http://www.dialog-21.ru/digests/dialog2006/materials/html/ Braslavski.htm . - Дата обращения 08.11.2013.

14 Браун, М. Методы интеллектуального анализа данных [Электронный ресурс] / М. Браун // IBM : developerWorks : ресурс IBM для разработчиков и ИТ-специалистов. – Режим доступа: http://www.ibm.com/developerworks/ru/ library/ba-data-mining-techniques/ . - Дата обращения 07.11.2013.

15 Виды интернет-рекламы [Электронный ресурс] // Услуги в интернете – Нижний Новгород. - Нижний Новгород, 2013. – Режим доступа: http://internetusl-nn.ru/publ/2-1-0-10. - Дата обращения 07.11.2013.

16 Валеев, С.С. Многоуровневая система фильтрации спама на основе технологий искусственного интеллекта / С.С. Валеев, А.П. Никитин // Вестник УГАТУ, 2008, т.11, №1(28). С. 215-219.

17 Власова, А. Kaspersky Security Bulletin, январь-июнь 2007. Спам в первом полугодии 2007 года [Электронный ресурс] / А. Власова // SECURELIST. - Электрон. данные. - Москва : Лаборатория Касперского, 1997-2013. - Режим доступа: http://www.securelist.com/ru/analysis?pubid=204007568. - Дата обращения 08.11.2013.

18 Волкова, Т. В. Проектирование и создание БД : учеб. пособие для вузов / Т. В. Волкова ; М-во образования и науки Рос. Федерации, Федер. агентство по образованию, Гос. образоват. учреждение высш. проф. образования "Оренбург. гос. ун-т". - Оренбург : ОГУ, 2006. - 140 с. - ISBN 5-02-011452-9.

19 Гагарский, В.А. Бизнес-процессы: основные понятия [Электронный ресурс] // В.А. Гагарский. – Режим доступа: http://www.elitarium.ru/2013/02/08 /biznes_processy_osnovnye_ponjatija.html.

20 Габриелян, В. Проблемы и решения в области фильтрации спама в почтовой системе Mail.Ru - практика 2003-2004 гг. [Электронный ресурс] // В. Габриелян. – Режим доступа: http://www.securelist.com/ru/analysis/19210/Prolemy_i_resheniya_v_oblasti_filtratsii_spama_v_pochtovoy_sisteme_Mail_Ru_praktika_2003_2004_gg.

21 Глинских, А Современные системы электронного документооборота [Электронный ресурс] // А. Глинских. – Режим доступа: http://www.ci.ru/inform09_01/p223edoc.htm.

22 Гмурман, В. Е. Теория вероятностей и математическая статистика : учеб. пособие для вузов / В. Е. Гмурман. – 9-е изд., стер. – Москва : Высшая школа, 2003. – 479 с. : ил.

23 Губин, М. В. Модели и методы представления текстового документа в системах информационного поиска : дис. … канд. физ.-мат. наук : 05.13.11 / Губин Максим Вадимович. – Москва, 2005. – 95 с. – Библиогр. : с. 89-95.

24 Гудкова, Д. Спам в первом квартале 2011 года. Закрытие ботнетов и доля спама в почтовом трафике [Электронный ресурс] / Д. Гудкова, М. Наместникова // SECURELIST. - Электрон. данные. - Москва : Лаборатория Касперского, 1997-2013. – Режим доступа: http://www.securelist.com/ru/analysis/208050697/Spam_v_pervom_kvartale_2011. - Дата обращения 08.11.2013.

25 Доля, А. Тенденции развития спама и средства борьбы с ним / А. Доля // Компьютер Пресс. –2006. – № 10. – с. 4 –7.

26 Дубров, А. М. Многомерные статистические методы : учеб. / А. М. Дубров, В. С. Мхитарян, Л. Н. Трошин. – Москва : Финансы и статистика, 2002. – 352 с. : ил.

27 Евлоев, О. В. Нейронные сети. Гл. 2 : Теория нейронных сетей [Электронный ресурс] / О. В. Евлоев // Частная страничка Евлоева Олега. - Режим доступа : http://evloevoleg.narod.ru/indexrh.html . - Дата обращения 08.11.2013.

28 Зубкова, Т. М. Технология разработки программного обеспечения : учеб. пособие / Т. М. Зубкова ; Оренбург. гос. ун-т. - Оренбург : РИК ГОУ ОГУ, 2004. – 102 с. : ил.

29 Игнатьев, В. А. Информационная безопасность современного коммерческого предприятия : моногр. / В. А. Игнатьев. - Старый Оскол : ТНТ, 2005. – 366 с. - ISBN 5-94178070-2.

30 Калинина, В. Н. Введение в многомерный статистический анализ: учеб. пособие для студентов всех специальностей / В. Н. Калинина, В. И. Со-

ловьев ; Гос. ун-т управления ; Ин-т информ. систем управления. – Москва : [ГУУ], 2003. – 66 с. - ISBN 5-215-01514-7.

31 Каширина, И. Л. Нейросетевые технологии : учеб.-метод. пособие для вузов / И. Л. Каширина ; Воронеж. гос. ун-т. - Воронеж : Издат.-полиграф. центр ВГУ, 2008. – 72 с.

32 Касперски, К. Безопасность электронной почты / К Касперски [Электронный ресурс] // «Журнал сетевых решений/LAN» №05,2010 . – Режим доступа: http://www.osp.ru/lan/2001/05/135166/#top. – Дата обращения 05.07.2011.

33 Комплекс защиты от утечек информации «Дозор - Джет» [Электронный ресурс] // Дозор-Джет : центр информационной безопасности «Инфосистемы Джет». – Москва, 2002-2013. – Режим доступа: http://www.anti-malware.ru/files/Комплекс%20защиты%20от%20утечек%20информации%20Дозор-Джет.pdf. – Дата обращения 07.11.2011.

34 Кондратьев, М. Е. Двухуровневая иерархическая кластеризация новостного потока в РОМИП 2006 [Электронный ресурс] / М. Е. Кондратьев // Российский семинар по оценке методов информационного поиска : тр. четвертого рос. семинара РОМИП'2006. - Санкт-Петербург, 2006. - С. 126-138. – Режим доступа : http://romip.narod.ru/ romip2006/index.html.

35 Контент-анализ [Электронный ресурс] // ВикипедиЯ : свободная энциклопедия. – Режим доступа : http://ru.wikipedia.org/wiki/%CA%EE%ED%F2%E5%ED%F2-%E0%ED%E0%EB%E8%E7 . - Дата обращения 08.11.2013.

36 Корнеева, В. Виды угроз для информационных систем, внешние и внутренние угрозы [Электронный ресурс] / В. Корнеева // Частная страничка Корнеевой Виктории. - Режим доступа : https://sites.google.com/site/victoriakorneeva136/otvety-na-voprosy/vidy-ugroz-dla-informacionnyh-sistem-vnesnie-i-vnutrennie--ugrozy. - Дата обращения 08.01.2012.

37 Круг, П. Г. Нейронные сети и нейрокомпьютеры : учеб. пособие по курсу «Микропроцессоры» для студентов, обучающихся по направлению

«Информатика и вычислительная техника» / П. Г. Круг ; Моск. энергет. ин-т (Техн. ун-т). – Москва : МЭИ, 2002. – 176 с. - ISBN 5-7046-0832-9.

38 Левин, М. Антиспам без секретов: практ. рекомендации по борьбе с нелегальной рассылкой по электронной почте / М. Левин. – Москва : Новый издательский дом, 2005. – 320 с. : ил.

39 Магнус, Я. Р. Эконометрика. Начальный курс : учеб. для студентов / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий.- 4-е изд., изм. и доп. - Москва : Дело, 2004. - 576 с. - Библиогр.: с. 561-570. - Предм. указ.: с. 570. - ISBN 5-7749-0055-X.

40 Макаров, С. Workflow и Enterprise Content Management [Электронный ресурс] / С. Макаров. – Режим доступа : www.directum.ru/workflow. - Дата обращения 08.11.2011.

41 Макаров, С. Обмен электронными документами между разными системам [Электронный ресурс] / С. Макаров. – Режим доступа : www.directum.ru/interchange_systems_integration.aspx - Дата обращения 10.11.2011.

42 Макаров, С. Интеграция с сервисами обмена электронных документов [Электронный ресурс] / С. Макаров. – Режим доступа : www.directum.ru/directum_OverDoc - Дата обращения 10.11.2011.

43 Максаков, А. В. Исследование способов уменьшения набора характеристик в алгоритмах классификации текстов [Электронный ресурс] / А. В. Максаков. – Режим доступа : http://old.lvk.cs.msu.su/files/mco2003/maksakov.pdf . - Дата обращения 08.11.2013.

44 Маннинг, К.Д. Введение в информационный поиск/ К.Д. Маннинг, П. Рагхаван, Х. Шютце, Вильямс. М.: 2011, 528 с.

45 Машечкин, И. В. Методы вычисления релевантности фрагментов текста на основе тематических моделей в задаче автоматического аннотирования / И. В. Машечкин, М. И. Петровский, Д. В. Царев // Вычислительные методы и программирование. - 2013. - Т. 14. – С. 91-102.

46 Местецкий, Л. М. Математические методы распознавания образов [Электронный ресурс] : курс лекций / Л. М. Местецкий ; Моск. гос. ун-т, ВМиК, Каф. «Математические методы прогнозирования». – Москва, 2002-2004. – Режим доступа : http://www.ccas.ru/frc/papers/mestetskii04course.pdf . - Дата обращения 08.11.2013.

47 Методы борьбы со спамом [Электронный ресурс] // Securelist. – Москва : Лаборатория Касперского, 1997-2013. - Режим доступа: http://www.securelist.com/ru/threats/spam?chapter=157. – Дата обращения 07.11.2013.

48 Методы борьбы со спамом (mail spam filter sendmail) [Электронный ресурс] // OpenNet : портал открытого ПО. – 1996-2013. - Режим доступа: http://www.opennet.ru/base/net/spam_greylist.txt.html. - Дата обращения 07.11.2013.

49 Миркин, Б. Г. Методы кластер-анализа для поддержки принятия решений : обзор : препринт WP7/2011/03 / Б. Г. Миркин ; Нац. исслед. ун-т «Высш. шк. экономики». – Москва : Изд. дом Нац. исслед. ун-та «Высш. шк. экономики», 2011. – 88 с.

50 Минашкин, В.Г. Теория статистики / В.Г. Минашкин, Р.А. Шмойлова, Н.А. Садовникова, Л.Г. Моисейкина, Е.С. Рыбакова, М.: изд. центр ЕАОИ. 2008, 296 с.

51 Моченов, С. В Применение статистических методов для семантического анализа текста / С. В. Моченов

52 Насакин, Р. «Рыбалка» в Интернете [Электронный ресурс] / Р. Насакин // КомпьютерПресс. – 2004. - № 10. - Режим доступа: http://www.compress.ru/article.aspx?id=12156&iid=468. – Дата обращения 07.11.2013.

53 Неделько, В. М. Оценивание похожести текстов на основе канонического представления / В. М. Неделько, Ю. Д. Манузина, М. А. Назарьева // Сборник научных трудов НГТУ / Новосиб. гос. техн. ун-т. – Новосибирск, 2008. – № 3 (53). – С. 59–68.

54 Нейронная сеть АРТ-1 адаптивной резонансной теории [Электронный ресурс] // Языки программирования - Life-prog.ru. - Режим доступа : http://life-prog.ru/view_neyrocomputer.php?id=1 . - Дата обращения 08.11.2013.

55 Николаев, И. А. Спам: экономические потери: [Электронный ресурс] : аналит. доклад / И. А. Николаев, М. В. Титова ; Департамент стратегического анализа. – Режим доступа : http://www.ifap.ru/pr/2009/n090217a.pdf . - Дата обращения 08.11.2013.

56 Новикова, Д. С. Автоматическое выделение терминов из текстов предметных областей и установление связей между ними [Электронный ресурс] / Д. С. Новикова // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем-2012 : материалы Всерос. конф. (с междунар. участием), 23-27 апр./ Рос. ун-т дружбы народов, Фак. физ.-мат. и естеств. наук. – Москва, 2012. – Режим доступа : http://conf.sci.pfu.edu.ru/index.php/ittmm/2012/paper/view/245 . - Дата обращения 08.11.2013.

57 Обмен электронными документами между разными системами Directum [Электронный ресурс] // Derectum. – Режим доступа : http://www.directum.ru /1093814 .aspx - Дата обращения 07.11.2011.

58 Петренко С. А. Сравнительный анализ методов обнаружения компьютерных атак / С. А. Петренко, А. В. Беляев // Проблемы информационной безопасности. Компьютерные системы. – 2008. - № 2. – С. 44-53.

59 Петров, В. Н. Информационные системы : учеб. пособие / В. Н. Петров. - Санкт-Петербург : Питер, 2002. - 688 с. : ил. - ISBN 5-318-00561-6.

60 Пивоварова, Л. М. Извлечение и классификация терминологических коллокаций на материале лингвистических научных текстов (предварительные наблюдения) [Электронный ресурс] / Л. М. Пивоварова, Е. В. Ягунова // Материалы Симпозиума "Терминология и знание", 21-22 мая 2010 г., Москва. - Москва, 2010. - Режим доступа : http://www.webground.su/data/lit/pivovarova_yagunova/

Izvlechenie_i_klassifikatsiya_terminoligicheskih_kollokatsyi.pdf . - Дата обращения 08.11.2013.

61 Прикладная статистика: классификации и снижение размерности : справ. изд. / С. А. Айвазян [и др.] ; под ред. С. А. Айвазяна. – Москва : Финансы и статистика, 1989. - 607 с.

62 Программы для фильтрации спама [Электронный ресурс] // Компьютера Online. – Москва, 1997-2013. – Режим доступа: http://old.computerra.ru/gid/soft/38070. - Дата обращения 07.11.2013.

63 Прохоров, А. Вредоносные программы, и как их победить [Электронный ресурс] / А. Прохоров // КомпьютерПресс. – 2006. - № 3. – Режим доступа: http://www.compress.ru/article.aspx?id=16109&iid=736. – Дата обращения 07.11.2003.

64 Пятковский, О. И. Интеллектуальные информационные системы (Нейронные сети) : учеб. пособие / О. И. Пятковский ; Алтай. гос. техн. ун-т им. И. И. Ползунова. – Барнаул : АлтГТУ, 2010. – 125 с.

65 Рассел, С. Искусственный интеллект: современный подход : пер. с англ. / С. Рассел, П. Норвиг. - 2-е изд. – Москва : Вильямс, 2006. – 1408 с. : ил.

66 Российская экономика теряет из-за спама $1,9 млрд в год [Электронный ресурс] // Securitilab.ru – Режим доступа : http://www.securitilab.ru/news/spam. - Дата обращения 12.06.2011.

67 Селезнев, К. Обработка текстов на естественном языке / К. Селезнев // Открытые системы. – 2003. – № 12.

68 Слепов, О. Борьба со спамом [Электронный ресурс] / О. Слепов ; Компания «Инфосистемы Джет» // Jet Info : информ. бюл. - 2004. - № 9. – С. 13-19. - Режим доступа: http://www.jetinfo.ru/Sites/new/Uploads/2004_9.7BBAD6EFC6554E8791CCBF730A438BA8.pdf. - Дата обращения 07.11.2013.

69 Слепов, О. Контентная фильтрация [Электронный ресурс] / О. Слепов // Jet Info. - 2005. - № 10 (149). – Режим доступа :

http://www.jetinfo.ru/Sites/new/Uploads/2005_10.F4446FCDF75C49DEA68C395 1291A359D.pdf . - Дата обращения 08.11.2013.

70 Слепов, О. Безопасность систем электронной почты [Электронный ресурс] / О. Слепов, А. Таранов // Jet Info. - 2003. - № 6 (121). – Режим доступа : http://www.jetinfo.ru/Sites/new/Uploads/2003_6.3 19A4A356B684F33 A06E15C657633935.pdf. - Дата обращения 12.11.2012.

71 Соколовский, В. В. Исследование качества автоматической классификации текстовых документов с использованием семантического графа документа [Электронный ресурс] / В. В. Соколовский // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса : материалы Тринадцатой междунар. конф. "Крым 2006", 10-18 июня, 2006, Судак, Автономная Республика Крым, Украина. – Режим доступа: http://www.gpntb.ru/win/inter-events/crimea2006/disk2/037.pdf. - Дата обращения 08.07.2012

72 Соловьев, Н. А. Развитие концепции обнаружения вторжений / Н. А. Соловьев, Е. Н. Чернопрудова // Современные информационные технологии в науке, образовании и практике : материалы VIII Всерос. науч.-практ. конф., / Оренбург. гос. ун-т. - Оренбург, 2009. - С. 66-67. - ISBN 978-5-7410-0975-8.

73 Система электронного документооборота и управления взаимодействием Directum [Электронный ресурс] // Derectum. – Режим доступа : http://www.directum.ru /1093814 .aspx - Дата обращения 07.11.2011.

74 $71 млрд. в год составляют убытки от спама [Электронный ресурс] // Delaemreklamuru. – Режим доступа : http://www.delaemreklamu.ru /info/novosti/internet . - Дата обращения 07.11.2013.

75 Теория статистики : учеб.-метод. комплекс / В. Г. Минашкин [и др.] ; Междунар. консорциум «Электронный университет», Моск. гос. ун-т экономики, статистики и информатики, Евраз. открытый ин-т. – Москва : Изд. центр ЕАОИ, 2008. - 296 с. - ISBN 978-5-374-00041-2.

76 Уваров, А.С. Почтовый сервер. Структура и принцип работы [Электронный ресурс] / А.С Уваров. – Режим доступа : http://interface31.ru/tech_it . - Дата обращения 12.07.2012.

77 Федоровский, А. Н. Mail.ru на РОМИП-2005 / А. Н. Федоровский, М. Ю. Костин // Труды РОМИП'2005 : тр. третьего рос. семинара по оценке методов информационного поиска / под ред. И. С. Некрестьянова ; НИИ Химии СПбГУ. – Санкт-Петербург, 2005. – 17 с.

78 Федеральный закон «О коммерческой тайне» [Электронный ресурс] / Информационно-аналитический центр «Консультант». – Электрон. дан. - Режим доступа: http://www.consultant.ru/document/cons_doc_LAW_116684 Дата обращения 04.05.2012

79 Федеральный закон «О персональных данных» [Электронный ресурс] / Информационно-аналитический центр «Консультант». – Электрон. дан. - Режим доступа: http://www.consultant.ru/document/cons_doc_LAW_149747 Дата обращения 04.05.2012

80 «Фишинг»: спам с неприятными последствиями [Электронный ресурс] / Информационно-аналитический центр по параллельным вычислениям. – Электрон. дан. - Режим доступа: http://dw.de/p/95W9 Дата обращения 07.11.2013

81 Функциональный подход к выделению ключевых слов: методика и реализация / И. Е. Воронина [и др.] // Вестник ВГУ. Серия Системный анализ и информационные технологии. – 2009. - № 1. – С. 68-72.

82 Хайкин, С. Нейронные сети: полный курс / С. Хайкин. – Москва : Вильямс, 2006. - 1104 с.

83 Хохлова, М. В. Исследование лексико-синтаксической сочетаемости в русском языке с помощью статистических методов (на базе корпусов текстов) : автореф. дис. канд. филолог. наук : 10.02.21 / Хохлова Мария Владимировна. – Санкт-Петербург, 2010. – 26 с. – Библиогр. : с. 25-26.

84 Хохлова, М. В. Экспериментальная проверка методов выделения коллокаций / М. В. Хохлова // Инструментарий русистики: корпусные подходы / под ред. А. Мустайоки, М. В. Копотева, Л. А. Бирюлина, Е. Ю. Протасовой. - Хельсинки, 2008. – С. 343-356.

85 Чернопрудова, Е. Н. Интеллектуальная фильтрация несанкционированных рассылок на основе нейронной сети / Е. Н. Чернопрудова, Н. А. Соловьев // Интеллект. Инновации. Инвестиции. - 2011. - Спец. вып. - С.106-107.

86 Чернопрудова, Е. Н. Нейросетевая модель интеллектуальной фильтрации несанкционированных рассылок / Е. Н Чернопрудова // Материалы IX всерос. науч.-техн. конф. – Оренбург, 2010. - С. 44-47.

87 Шевелев, О.Г. Методы автоматической классификации текстов на естественном языке / О.Г. Шевелев //Учебное пособие. – Томск:ТМЛ-Пресс, 2007. – 144 с.

88 Яворский, В. Борьба со спамом: история и методы [Электронный ресурс] / В. Яворский // МФТИ : сайт Московского физико-технического института. – Долгопрудный, 2001-2013. - Режим доступа: http://bio. fiz-teh.ru/student/diff_articles/no_spam.html . - Дата обращения 07.11.2013.

89 Ягунова, Е. В. Коллокации и конструкции в исследовании структуры текста [Электронный ресурс] / Е. В. Ягунова, Л. М. Пивоварова. – Режим доступа : http://iling.spb.ru/confs/rusconstr2011/pdf/JagunovaPivovarova.pdf . - Дата обращения 08.11.2013.

90 Ягунова, Е. В. От коллокаций к конструкциям / Е. В. Ягунова, Л. М Пивоварова // Русский язык: конструкционные и лексико-семантические подходы / отв. ред. С. С. Сай. – Санкт-Петербург, 2011. – С. 27.

91 Яремчук, С. Технология борьбы со спамом DKIM //Системный администратор, №11(72), 2008 г., – С. 6-13.

92 Ananth Ullal Kini On the Effect of INQUERY Term-Weighting Scheme on Query-Sensitive Similarity Measures [Электронный ресурс] : a Thesis / Ananth Ullal Kini. – Режим доступа:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.9597&rep=rep1&type=pdf . - Дата обращения 08.11.2013.

93 Cover, T. Elements of Information Theory / T. M. Cover, J. A. Thomas. - John Wiley & Sons, 1991. – 542 p. - (Wiley Series in Telecommunications). - ISBN 0-471-20061-1.

94 Cover, T. Elements of Information theory [Электронный ресурс] / T. Cover, J. Thomas. – Режим доступа : https:// web.cse.msu.edu/cse842/Papers/CoverThomas-Ch2.pdf. - Дата обращения 08.11.2013.

95 Dasigi, V. Neural Net Learning Issues in Classification of Free Text Documents / V. Dasigi, R. Manu // AAAI spring symposium on Machine Learning in Information Access – 1996.

96 Daudaravicius, V. Automatic Identification of Lexical Units / V. Daudaravicius // Informatica. - 2010. - № 34. – P. 85-91.

97 Fuernkranz, J. A study using n-gram features for Text Categorization / J. Fuernkranz // Tech report OEFAI-TR-98-30 – 1998.

98 Hotho, A. Ontology-based Text Clustering [Электронный ресурс] / A. Hotho, S. Staab, A. Maedche. – Режим доступа : http://www.cs.cmu.edu/mccallum/ textbeyond/papers/hotho.pdf .

99 Lan, M. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization [Электронный ресурс] / M. Lan, C. L. Tan, J. Su // Journal of IEEE PAMI. – 2007. - Vol. 10, № 10, july. - Режим доступа : https://www-old.comp.nus.edu.sg/~tancl/publications/j2009/PAMI2007-v3.pdf . - Дата обращения 08.11.2013.

100 Li, Y. H. Classification of Text Documents /Y. H. Li, A. K. Jain // The Computer Journal. – 1998. - Vol. 41, № 8. - P. 537-546.

101 Lippmann, R. P. An introduction to computing with neural nets / R. P. Lippmann // IEEE ASSP Magazine. - 1987. - № 4. – P. 4-22.

102 Mammadov, M. A New Supervised Term Ranking Method for Text Categorization / Musa Mammadov, John Yearwood, Lei Zhao // AI 2010 : Ad-

vances in Artificial Intelligence : proceedings 23rd Australasian Joint Conference, 7-10 December 2010, Adelaide, Australia / ed. J. Li. – Springer Berlin Heidelberg, 2011. - P. 102-111. - (Lecture Notes in Computer Science, Vol. 6464). - ISBN 978-3-642-17432-2.

103 McCallum, A. A comparison of Event Models for Naïve Bayes Classifiection / A. McCallum, K. Nigam // AAAI-98 : Workshop on Learning for Text Categorization. – 1998. – 8 с.

104 Mingyong, L. An improvement of TFIDF weighting in text categorization / L. Mingyong, Y. Jiangang [Электронный ресурс] – Режим доступа : http:// www.ipcsit.com/vol47/009-ICCTS2012-T049. pdf

105 Microsoft Corporation Проектирование и реализация баз данных Microsoft SQL Server 2000. Учебный курс MCAD/MCSE : пер. с англ. – 2-е изд., испр. – Москва : Русская Редакция, 2003. – 512 с. : ил.

106 Neural Computing: NeuralWorks Professional II / Plus and NeuralWorks Explorer. – Pittsburgh : NeuralWare, Inc., 1991. - 355 p.

107 Salton, G. A Vector Space Model for Automatic Indexing [Электронный ресурс] / G. Salton, A. Wong, C. S. Yang. – Режим доступа : http://parnec.nuaa.edu.cn/xtan/IIR/readings/cacmSalton1975.pdf . - Дата обращения 08.11.2013.

Приложение А

(обязательное)

Таблица А.1 – Уровни доступа пользователей

Класс объектов/ Свойство	Пользователи системы			
	Прикладной программист	Администратор БД	Конечный пользователь	
			Админ. фильтра	Пользователь почты
1	2	3	4	5
Категирия				
Ном_катег	RIUD	RIUD	RUI	R
Принадлежность	RIUD	RIUD	RUI	R
Кр_название	RIUD	RIUD	RUI	R
Связь				
Ном_катег	RIUD	RIUD	RUI	R
Ном_терм	RIUD	RIUD	RUI	R
частота	RIUD	RIUD	RUI	R
Термы				
Ном_терм	RIUD	RIUD	RUI	R
Терм	RIUD	RIUD	RUI	R
Вес	RIUD	RIUD	RUI	R
Ед_изм				
Ном_ед	RIUD	RIUD	RUI	R
Название	RIUD	RIUD	RUI	R
Кр_название	RIUD	RIUD	RUI	R
Сообщение				
Номер	RIUD	RIUD	RUI	R
Тема	RIUD	RIUD	RUI	R
Работа фильтра				
Номер	RIUD	RIUD	RUI	R
Дата_нач	RIUD	RIUD	RUI	R
Дата_окон	RIUD	RIUD	RUI	R
Время	RIUD	RIUD	RUI	
Объект				
Номер	RIUD	RIUD	RUI	R
Адрес IP	RIUD	RIUD	RUI	R
Адрес DNS	RIUD	RIUD	RUI	R
Закреп_объекта				
Дата_нач	RIUD	RIUD	RUI	R
Дата_оконч	RIUD	RIUD	RUI	R
Подразделение				

Продолжение таблицы А.1

1	2	3	4	5
Номер	RIUD	RIUD	RUI	
Название	RIUD	RIUD	RUI	R
Кр_название	RIUD	RIUD	RUI	R
Тип_подраздел				
Номер	RIUD	RIUD	RUI	R
Название	RIUD	RIUD	RUI	R
Кр_название	RIUD	RIUD	RUI	R
Фрагм_труд_догов				
дата_нач	RIUD	RIUD	RUI	R
дата_оконч	RIUD	RIUD	RUI	R
Сотрудник				
Фамилия	RIUD	RIUD	RUI	R
Имя	RIUD	RIUD	RUI	R
Отчество	RIUD	RIUD	RUI	R
Должность				
номер	RIUD	RIUD	RUI	R
Название	RIUD	RIUD	RUI	R
Кр_название	RIUD	RIUD	RUI	R

В таблице использованы сокращения операций, производимых со свойствами классов объектов: R – read (чтение); I – insert (добавление); U – update (обновление); D – delete (удаление).

Таблица А.2 – Классы объектов, свойства

Класс объектов/ Свойство	Ключ (уникальный, первичный)	Физические характеристики (тип, длина)	Опциональность свойства	Логические ограничения	Процессы для значений свойств
1	2	3	4	5	6
Категория					
Номер_кат	УК,ПК	Число,10	Д.б.	>0	генерация, просмотр
принадлежн		Число,10	Д.б.		ввод, обновление, просмотр
Кр_название		Число,10	Д.б		ввод, обновление, просмотр
Связь					
Ном_катег	УК1,ПК	Число,10	Д.б.		генерация, просмотр
Ном_терм	УК2	Число,10	Д.б.		генерация, просмотр
частота		Число,10	Д.б.		ввод, обновление, просмотр
Термы					
Ном_терм	УК,ПК	Число,10	Д.б.	>0	генерация, просмотр
Терм		Varchar,255	Д.б		ввод, обновление, просмотр
Вес		Число,10	Д.б		ввод, обновление, просмотр
Ед_изм					
Ном_ед	УК,ПК	Число,10	Д.б.	>0	генерация, просмотр
Название		Число,10	Д.б.		ввод, обновление, просмотр
Кр_название		Число,25			ввод, обновление, просмотр
Сообщение					
Номер	У, П	число, 10	Д.б	>0	генерация
Тема		символ, 20	Д.б	Прописные, строчные	ввод, обновление, просмотр

Продолжение таблицы А.2

1	2	3	4	5	6
Работа фильтра					
Номер	У, П	число, 10	Д.б	>0	генерация
Дата_нач		дата, 8	Д.б	дд.мм.гггг	ввод, обновление, просмотр
Дата_окон		дата, 8	Д.б	дд.мм.гггг	ввод, обновление, просмотр
Время		дата, 8	Д.б		ввод, обновление, просмотр
Объект					
Номер	У, П	число, 10	Д.б	>0	генерация
Адрес IP	символ, 20	Д.б.	Прописные, строчные	ввод, обновление, просмотр	ввод, обновление, просмотр
Адрес DNS	символ, 20	Д.б.	Прописные, строчные	ввод, обновление, просмотр	ввод, обновление, просмотр
Закреп_объек					
Дата_нач		дата, 8	Д.б	дд.мм.гггг	ввод, обновление, просмотр
Дата_оконч		дата, 8	М.б	дд.мм.гггг	ввод, обновление, просмотр
Подразделение					
Номер	У, П	число, 10	Д.б	>0	генерация
Название		символ, 20	Д.б	Прописные, строчные	ввод, обновление, просмотр
Кр_название		символ, 10	Д.б	Прописные, строчные	ввод, обновление, просмотр
Тип_подразд					
Номер	У, П	число, 10	Д.б	>0	генерация
Название		символ, 20	Д.б	Прописные, строчные	ввод, обновление, просмотр
Кр_назв		символ, 10	Д.б	Прописные, строчные	ввод, обновление, просмотр

Продолжение таблицы А.2

1	2	3	4	5	6
Фрагм_труд_договов					
дата_нач		дата, 8	Д.б	дд.мм.гггг	ввод, обновление, просмотр
дата_оконч		дата, 8	М.б	дд.мм.гггг	ввод, обновление, просмотр
Сотрудник					
Номер	У, П	число, 10	Д.б	>0	генерация
Фамилия		символ, 20	Д.б	Прописные, строчные	ввод, обновление, просмотр
Имя		символ, 10	Д.б	Прописные, строчные	ввод, обновление, просмотр
Отчество		символ, 20	Д.б	Прописные, строчные	ввод, обновление, просмотр
Должность					
номер	У, П	число, 10	да	>0	генерация
Название		символ, 20	да	Прописные, строчные	ввод, обновление, просмотр
Кр_название		символ, 10	нет	Прописные, строчные	ввод, обновление, просмотр

В таблице использованы сокращения: У – уникальный ключ, П – первичный ключ (главный уникальный).

Приложение Б
(обязательное)

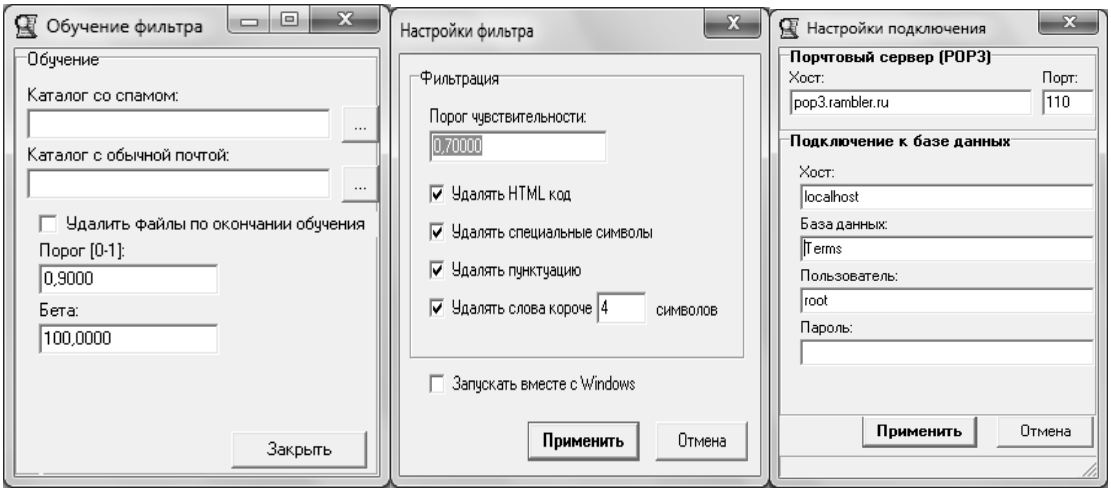


Рисунок Б.1 – Экранные формы

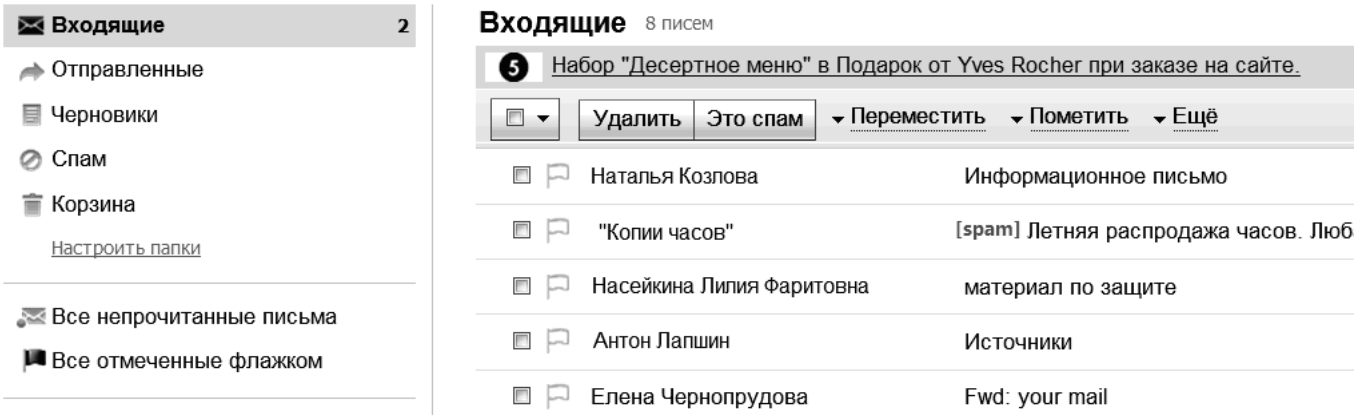
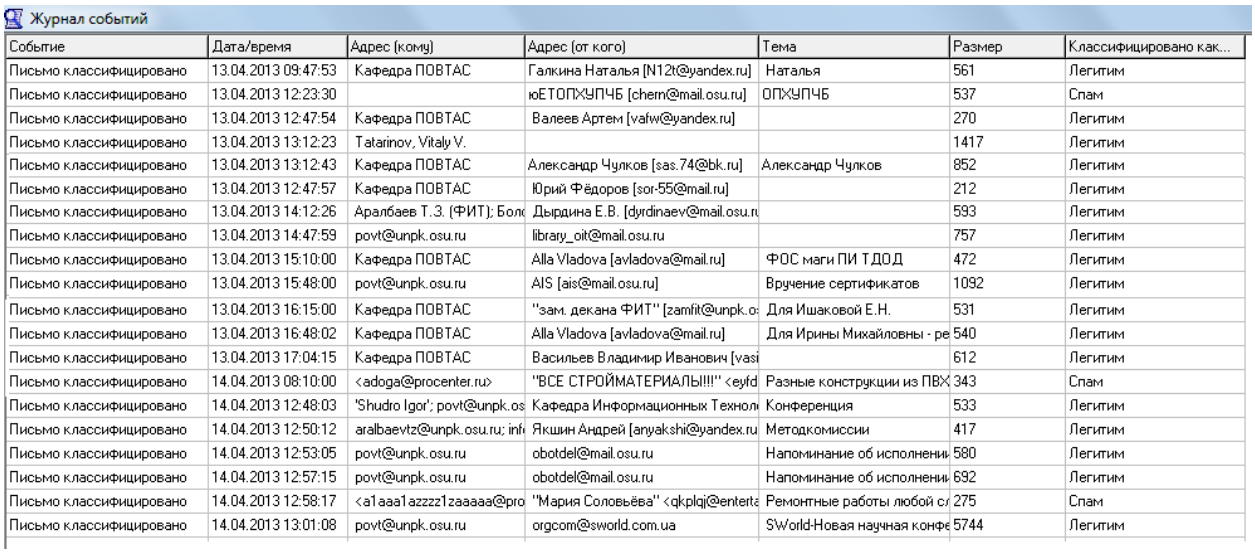


Рисунок Б.2 – Экранные формы системы фильтрации


Журнал событий

Событие	Дата/время	Адрес (кому)	Адрес (от кого)	Тема	Размер	Классифицировано как...
Письмо классифицировано	13.04.2013 09:47:53	Кафедра ПОВТАС	Галкина Наталья [N12t@yandex.ru]	Наталья	561	Легитим
Письмо классифицировано	13.04.2013 12:23:30		юЕТОПХУПЧБ [chern@mail.osu.ru]	ОПХУПЧБ	537	Спам
Письмо классифицировано	13.04.2013 12:47:54	Кафедра ПОВТАС	Валеев Артем [vafw@yandex.ru]		270	Легитим
Письмо классифицировано	13.04.2013 13:12:23	Tatarinov, Vitaly V.			1417	Легитим
Письмо классифицировано	13.04.2013 13:12:43	Кафедра ПОВТАС	Александр Чулков [sas.74@bk.ru]	Александр Чулков	852	Легитим
Письмо классифицировано	13.04.2013 12:47:57	Кафедра ПОВТАС	Юрий Фёдоров [sor-55@mail.ru]		212	Легитим
Письмо классифицировано	13.04.2013 14:12:26	Аралбаев Т.З. (ФИТ); Бол	Дырдина Е.В. [dyrdinaev@mail.osu.ru		593	Легитим
Письмо классифицировано	13.04.2013 14:47:59	povt@unpk.osu.ru	library_oit@mail.osu.ru		757	Легитим
Письмо классифицировано	13.04.2013 15:10:00	Кафедра ПОВТАС	Alla Vladova [avladova@mail.ru]	ФОС маги ПИ ТДОД	472	Легитим
Письмо классифицировано	13.04.2013 15:48:00	povt@unpk.osu.ru	AIS [ais@mail.osu.ru]	Вручение сертификатов	1092	Легитим
Письмо классифицировано	13.04.2013 16:15:00	Кафедра ПОВТАС	"зам. декана ФИТ" [zamfit@unpk.o	Для Ишаковой Е.Н.	531	Легитим
Письмо классифицировано	13.04.2013 16:48:02	Кафедра ПОВТАС	Alla Vladova [avladova@mail.ru]	Для Ирины Михайловны - ре	540	Легитим
Письмо классифицировано	13.04.2013 17:04:15	Кафедра ПОВТАС	Васильев Владимир Иванович [vasi		612	Легитим
Письмо классифицировано	14.04.2013 08:10:00	<adoga@procenter.ru>	"ВСЕ СТРОЙМАТЕРИАЛЫ!!!" <eyfd	Разные конструкции из ПВХ	343	Спам
Письмо классифицировано	14.04.2013 12:48:03	'Shudro Igor'; povt@unpk.os	Кафедра Информационных Технол	Конференция	533	Легитим
Письмо классифицировано	14.04.2013 12:50:12	aralbaevtz@unpk.osu.ru; inf	Якшин Андрей [anyakshi@yandex.ru	Методкомиссии	417	Легитим
Письмо классифицировано	14.04.2013 12:53:05	povt@unpk.osu.ru	obotdel@mail.osu.ru	Напоминание об исполнени	580	Легитим
Письмо классифицировано	14.04.2013 12:57:15	povt@unpk.osu.ru	obotdel@mail.osu.ru	Напоминание об исполнени	692	Легитим
Письмо классифицировано	14.04.2013 12:58:17	<a1aaa1azzzz1zaaaaa@pro	"Мария Соловьёва" <qkplqj@enterta	Ремонтные работы любой сл	275	Спам
Письмо классифицировано	14.04.2013 13:01:08	povt@unpk.osu.ru	orgcom@sworld.com.ua	SWorld-Новая научная конфе	5744	Легитим

Рисунок Б.3 – Окно журнала событий

Монография

Николай Алексеевич Соловьев

Елена Николаевна Чернопрудова

Наталья Александровна Тишина

Любовь Аркадьевна Юркевская

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ЗАЩИТЫ ПОЧТОВЫХ СЕРВИСОВ ОТ НЕСАНКЦИОНИРОВАННЫХ РАССЫЛОК НА ОСНОВЕ КОНТЕНТНОЙ ФИЛЬТРАЦИИ ЭЛЕКТРОННЫХ СООБЩЕНИЙ

ISBN 978-5-7410-1724-1

9 785741 017241