На правах рукописи

Баранчикова Екатерина Александровна

МОДЕЛИ, АЛГОРИТМЫ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ФИЛЬТРАЦИИ ЭЛЕКТРОННОЙ КОРРЕСПОНДЕНЦИИ ДЛЯ ИНФОРМАЦИОННОЙ СИСТЕМЫ С ОГРАНИЧЕННЫМИ РЕСУРСАМИ

Специальность: 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Рязань 2011

Работа выполнена в ФГБОУ ВПО «Рязанский государственный радиотехнический университет».

Научный руководитель: доктор технических наук, профессор, заслуженный деятель науки и техники Российской Федерации **Корячко Вячеслав Петрович**

Официальные оппоненты: доктор технических наук, профессор **Белов Владимир Викторович**
кандидат технических наук, доцент **Буланкин Валерий Борисович**

Ведущая организация: Государственный научно-исследовательский институт информационных технологий и телекоммуникаций (г. Москва)

Защита диссертации состоится «25» января 2012 года в 12 часов на заседании диссертационного совета Д 212.211.01 в ФГБОУ ВПО «Рязанский государственный радиотехнический университет» по адресу: 390005, г. Рязань, ул. Гагарина, 59/1.

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВПО «Рязанский государственный радиотехнический университет».

Автореферат разослан «09» декабря 2011 года.

Ученый секретарь
диссертационного совета
кандидат технических наук, доцент  В.Н. Пржегорлинский

Общая характеристика работы

Актуальность работы. Стабильное и надежное функционирование современных информационных систем, спроектированных с учетом ограниченных ресурсов, является одним из необходимых условий успешного развития малых предприятий. При этом возникает проблема безопасности информационных систем, которая заставляет уделять особое внимание их защите от различного рода воздействий, способных привести к нарушениям конфиденциальности, целостности или доступности информации, а также работы самой ИС или к финансовым потерям предприятия. Данное обстоятельство наиболее актуально для малых предприятий, так как даже самые небольшие финансовые потери в условиях жесткого ограничения денежных ресурсов могут привести к негативным последствиям.

Во многих современных ИС, как правило, отдельно выделяется служба обмена почтовыми сообщениями как внутри организации, так и с внешними почтовыми серверами. При использовании электронной почты возникают дополнительные угрозы безопасности и стабильной работе ИС, связанные с различными внешними воздействиями, такими как спам, вирусы и другие, которые могут привести к нарушению работоспособности системы, а также к уменьшению производительности труда людей, непосредственно использующих данную ИС.

В настоящее время одной из актуальных угроз информационной безопасности и финансовой стабильности предприятия является спам, ставший существенной проблемой службы обмена почтовыми сообщениями. За последние десять лет спам превратился из легкого раздражающего фактора в одну из самых серьезных угроз информационной безопасности.

В работе рассматриваются алгоритмы классификации почтовых сообщений на легальную почту и спам с применением регулярных выражений, описываются математические модели различных подходов к фильтрации почты. Рассматриваемые алгоритмы реализованы с помощью средств реляционной СУБД PostgreSQL и независимых программных модулей и могут быть в дальнейшем использованы при реализации спам-фильтров для различных почтовых серверов.

Степень разработанности темы. Задача фильтрации входящей электронной корреспонденции активно разрабатывается последние десять лет. Наибольший вклад в нее внесли Грехэм П., Джиарски Дж., Йеразунис У., Чхабра Ш. В их работах рассмотрены базовые

принципы фильтрации электронной почты на основе классификации текстовой информации, содержащейся в письме.

В литературе уделяется мало внимания построению формальных математических моделей фильтрации входящей электронной корреспонденции. Поэтому в данной диссертационной работе рассматриваются наиболее распространенные и востребованные модели фильтрации электронной корреспонденции. Это позволит реализовывать алгоритмы фильтрации, наиболее полно отвечающие поставленным задачам. Самые современные разработки в этой области опубликованы в ACM.

Основное содержание настоящей диссертации составляют разработка математических моделей и алгоритмов реализации фильтрации входящей электронной корреспонденции с использованием регулярных выражений, а также решение проблем, возникающих при их применении.

Целью диссертационного исследования является:

- минимизация доли нелегальной корреспонденции в общем объёме электронной почты, получаемой предприятием;
- повышение эффективности работы малых предприятий за счет сокращения трудовых затрат, связанных с необходимостью обработки электронной корреспонденции.

Задачи. Для достижения поставленной цели решаются следующие задачи:

1) анализ существующих научных исследований в области фильтрации входящей электронной корреспонденции;

2) построение математических моделей, описывающих процесс фильтрации входящей электронной корреспонденции;

3) разработка методики фильтрации входящей электронной корреспонденции на основе модели статистической фильтрации с применением регулярных выражений;

4) реализация предложенных алгоритмов, оценка их производительности, выбор рекомендаций при их реализации.

Научная новизна работы состоит в следующем:

- предложен способ, позволяющий реализовать ускоренный доступ по заданному входному слову к соответствующим ему регулярным выражениям, хранимым в базе данных, с помощью индексной таблицы;

– предложен способ фильтрации входящей электронной корреспонденции, основанный на использовании регулярных выражений для распознавания слов, входящих в электронное сообщение, и последующей их обработки статистическим спам-фильтром.

На защиту выносятся следующие научные результаты:

1) формулировка и решение задачи классификации входящей электронной корреспонденции на легальную и спам;

2) модель нарушителя в информационной системе в случае борьбы с нелегальной корреспонденцией;

3) математические модели основных подходов к фильтрации электронной корреспонденции, обобщенная модель классификации текстовой информации;

4) алгоритм, реализующий быстрый доступ к соответствующим регулярным выражениям, хранимым в базе данных;

5) алгоритм фильтрации входящей электронной корреспонденции на основе статистической модели фильтрации с применением регулярных выражений.

Практическая ценность работы состоит в том, что применение предложенных алгоритмов позволяет реализовать контекстно зависимый статистический спам-фильтр электронной корреспонденции, осуществляющий сортировку почты с помощью регулярных выражений, хранимых в базе знаний. Предложенные алгоритмы для работы с регулярными выражениями в составе программы фильтрации почты могут быть реализованы в любой реляционной СУБД.

Соответствие паспорту специальности. Согласно паспорту специальности 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» проблематика, рассмотренная в диссертации, соответствует следующим областям исследований:

– модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем;

– системы управления базами данных и знаний.

Реализация и внедрение. Диссертация выполнена в Рязанском государственном радиотехническом университете.

Результаты диссертационной работы внедрены:

– в ООО «Алеатис» (г. Москва) в виде разработки почтового фильтра, работающего в составе почтового сервера Sendmail;

– в ООО «Русофт-Ритейл» (г. Рязань) в виде разработки почтового фильтра, работающего в составе почтового сервера Sendmail;
– в ООО «Арго-тур» (г. Рязань) в виде разработки почтового фильтра, работающего в составе почтового сервера Sendmail;
– в учебном процессе Рязанского государственного радиотехнического университета.

Апробация работы. Основные положения диссертационной работы докладывались на следующих конференциях:

1) Электронное обучение и управление знаниями высшего учебного заведения, МЭСИ, г. Рязань, 2007 г.;

2) 13-й Всероссийской научно-технической конференции студентов, молодых ученых и специалистов, г. Рязань, 2008 г.;

3) Информационные и телекоммуникационные технологии, г. Рязань, 2009 г.;

4) 14-й Всероссийской научно-технической конференции студентов, молодых ученых и специалистов, г. Рязань, 2009 г.;

5) XXX VI ГАГАРИНСКИЕ ЧТЕНИЯ, г. Москва, 2010 г.

Публикации. По теме диссертации опубликовано 11 работ: 5 статей, в том числе в сборниках рекомендованных ВАК РФ - 3, и 5 тезисов докладов на международных и всероссийских конференциях, 1 свидетельство о государственной регистрации программы для ЭВМ.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, списка используемых источников и приложения. Основной текст работы содержит 135 с., 32 рисунка и 3 таблицы. Список используемых источников на 12 с. включает 112 наименований. В приложении на 130-й с. приведены документы о внедрении и практическом использовании результатов диссертации и свидетельство о регистрации программного продукта в Федеральной службе по интеллектуальной собственности, патентам и товарным знакам.

Содержание работы

Во введении дается обоснование актуальности темы работы, формулируются цели исследования, кратко излагается содержание диссертации.

Первая глава посвящена обоснованию темы диссертации. В главе определяются основные цели и задачи проектирования средств

защиты информации, приводится обзор работ по теме диссертации, вводятся основные понятия и определения. Особое внимание уделено вопросам проектирования компонент программного обеспечения фильтрации входящей электронной корреспонденции для информационных систем с ограниченными ресурсами.

Описана специфика работы малых предприятий, обозначены основные угрозы и риски для информационных систем, функционирующих на таких предприятиях и имеющих ограниченные ресурсы. Показано, что основной угрозой информационной безопасности для малых предприятий являются вредоносные программы и спам и в случае их игнорирования предприятие может понести значительные потери.

Для определения соответствующих мер защиты ИС следует использовать системный подход. Решение о том, как обеспечить защиту, где реализовать защиту в ИС, какими должны быть типы и качество мер и средств защиты, требует проведения соответствующего анализа уязвимости и рисков.

Произведен расчет рисков по информационной системе с ограниченными ресурсами на основе модели угроз и уязвимостей.

В результате работы алгоритма было получено, что риск по информационной системе на основе ОС MS Windows по одной суммарной угрозе без организации защиты стремится к 90-95 %.

Показано, что при создании системы защиты информации следует учитывать, что она требует реализации не только программных и технических, но также и организационных мер по защите информации, которые являются необходимым начальным условием для успешной реализации всех остальных методов.

В результате применения стандартных технических и организационных мер риск по информационной системе на основе рабочих станций ОС MS Windows и серверов под управлением ОС класса Unix по одной суммарной угрозе стремится к 25-30 %.

Таким образом, предложенный подход к повышению уровня защиты малых ИС позволил при жестких финансовых ограничениях добиться значительного результата.

Подробно рассмотрена угроза информационной безопасности — спам — любая информация, полученная предприятием без явного на то запроса и приводящая к увеличению трафика и уменьшению пропускной способности каналов, а также увеличивающая время обработки корреспонденции и поиска нужной информации.

Предложена методика расчета потерь предприятия при реализации данной угрозы. Показана необходимость дальнейшей работы над уменьшением вероятности ее реализации в рамках малых информационных систем. Исходя из экспериментально полученных данных, на примере работы туристической фирмы, имеем число писем спама n_1=600 шт. без применения какой-либо фильтрации. В итоге ежемесячные потери предприятия, связанные только с получением спама, составляют $P(n_1)$=21540 руб.

Рассмотрены основные подходы в борьбе со спамом. Показано, что существующих средств недостаточно для борьбы с данной угрозой либо существующие средства сложно применимы в контексте малого предприятия.

Показана необходимость дальнейшей работы в направлении сокращения доли нелегальной корреспонденции в общем потоке электронной почты. Для достижения поставленной цели на данном этапе требуется сформулировать задачу, решаемую при защите информационной системы от спама, построить математические модели, описывающие процесс фильтрации входящей электронной корреспонденции.

Вторая глава

Рассмотрены основные угрозы информационной системе со стороны спама.

Для определения необходимых средств противодействия реализации угрозы проникновения в систему спама, построена модель нарушителя в информационной системе. Для этого предложена содержательная модель нарушителя, включающая сценарии и математическую модель воздействия нарушителей. Проведено статистическое исследование рассмотренных воздействий.

Проведенные исследования показывают, что нагрузка на информационную систему со стороны спамеров составляет до 90 % от общего потока входящей электронной корреспонденции, а штатные средства рассмотренной информационной системы не дают требуемого уровня фильтрации.

Сформулирована задача классификации входящей электронной корреспонденции на легальную и спам. Функция, описывающая идеальный спам-фильтр, представлена в следующем виде:

$$Flt(Q)=Flt(H\cup S)\rightarrow(H,S).$$

Реальный спам-фильтр описывается функцией:

$$Flt(Q)\rightarrow(H\setminus H_S\cup S_H, S\setminus S_H\cup H_S),$$

где H — легальное письмо, S — спам, H_s — легальное письмо, ошибочно распознанное как спам, S_H — спам, ошибочно принятый за легальное письмо.

Рассмотрены два вида возможных ошибок при фильтрации спама: первого и второго рода:

$$Err_1=\frac{|S_H|}{|S|},\quad Err_2=\frac{|H_S|}{|H|}.$$

Под ошибкой первого рода будем понимать возможность прохождения спама через фильтр данной категории, под ошибкой второго рода будем понимать возможность классификации легальной корреспонденции как спам. Ошибки второго рода более критичны с точки зрения надежности ИС, чем ошибки первого рода в связи с тем, что при ошибке второго рода может возникнуть потеря важной деловой информации, а как следствие - финансовые потери для предприятия. В случае ошибки первого рода вместе с легальной корреспонденцией конечному пользователю также будет доставлен спам, и тогда потери будут не столь значительны.

При этом задача фильтрации входящей электронной корреспонденции принимает вид:

$$\begin{cases} Err_1\to min, \\ Err_2\to 0. \end{cases}$$

Разработаны математические модели основных подходов к фильтрации спама (белые и черные списки, эвристическая фильтрация, статистическая фильтрация). Описаны преимущества и недостатки этих моделей.

Определены функции принадлежности:

1. для белых списков:

$$\mu(q_i)=\begin{cases} 1 & при\ \ q_i\in WL, \\ 0 & при\ \ q_i\notin WL; \end{cases}$$

2. для черных списков:

$$\mu(q_i)=\begin{cases} 1 & при\ \ q_i\notin BL, \\ 0 & при\ \ q_i\in BL; \end{cases}$$

где q_i — письмо, WL — белый список, BL — черный список.

Показано, что при пересечении методов черных и белых

списков возможно возникновение критических ошибок второго рода, а при объединении этих методов возникает неопределенность, когда одно и то же письмо может быть классифицировано и как спам, и как легальная корреспонденция в зависимости от порядка применения правил классификации при реализации объединения.

Описана математическая модель эвристической фильтрации. Наиболее существенным отличием между списками и обычной эвристической фильтрацией является то, что во втором методе письмо классифицируется как спам на основе его содержимого, тогда как белые и черные списки идентифицируют отправителей письма и на основе этого принимают решение. Функция принадлежности в этом случае имеет вид:

$$\mu(q_i)=\begin{cases}1 & при \quad k\cdot f_H(q_i)\geq f_S(q_i),\\ 0 & при \quad k\cdot f_H(q_i)<f_S(q_i).\end{cases}$$

где $fs(q_i)$ и $f_H(q_i)$ — функции, характеризующие принадлежность письма к спаму и легальной корреспонденции соответственно.

Статистическая фильтрация реализуется с помощью вероятностного подхода. В этом заключается ее основное преимущество по сравнению с другими методами фильтрации, связанное с тем, что реальные данные накапливаются с течением времени и не зависят от субъективных пользовательских оценок.

Математическая модель статистической фильтрации имеет вид:

$$score(x_j)=max\left(0.1, min\left(0.99, \left(\frac{min\left(1,\frac{S_j}{Count_S}\right)}{min\left(1,\frac{2\cdot H_j}{Count_H}\right)+min\left(1,\frac{S_j}{Count_S}\right)}\right)\right)\right),$$

где *score* — вероятностная оценка, x_j — некоторая лексема, содержащаяся в обучающей выборке, S_j — количество писем спама, в которых встречается x_j, H_j — количество легальных писем, в которых встречается x_j, $Count_H$ — общее количество легальных писем в обучающей выборке, $Count_S$ — общее количество писем спама в обучающей выборке.

$$S(q_i)=\frac{\prod score_{i,j}}{\prod score_{i,j}+\prod (1-score_{i,j})},$$

где $S(q_i)$ — итоговая вероятностная оценка письма;

$$score_{i,j}=\begin{cases} score(x_{i,j}) & при\ \ x_{i,j}\in X, \\ 0{,}4 & при\ \ x_{i,j}\notin X. \end{cases}$$

В итоге функция принадлежности примет вид:

$$\mu(q_i)=\begin{cases} 1 & при\ \ S(q_i)>0{,}7, \\ 0 & при\ \ S(q_i)<0{,}4, \\ 0{,}5 & при\ \ 0{,}4\leqslant S(q_i)\leqslant 0{,}7. \end{cases}$$

На основе выявленных общих закономерностей при построении математических моделей была предложена обобщенная модель классификации текстовой информации, позволяющая взглянуть на процесс фильтрации в более общем виде.

На основе предложенной унифицированной модели предполагается дальнейшая реализация спам-фильтра как классификатора входящей текстовой информации, отвечающего заданным входным параметрам. Поскольку предполагается, что модель является унифицированной, она позволяет комбинировать описанные выше подходы для решения поставленных задач.

В процессе фильтрации полученной корреспонденции необходимо выполнить следующие шаги для классификации входящего текста:

1) начальные преобразования;

2) разбиение на лексемы, или токенизация;

3) определение отличительных характеристик текста;

4) оценка характеристик;

5) вычисление суммарной оценки;

6) принятие решения об отнесении письма к той или иной классификационной категории;

7) коррекция ошибок с целью получения обратной связи.

Полученная модель представлена на рисунке 1. Она включает не только вышеперечисленные этапы, в ее описание также входят обучающие алгоритмы, база знаний, заданные параметры фильтрации и схемы вычисления итоговой функции принадлежности входящего письма к различным классификационным подмножествам.

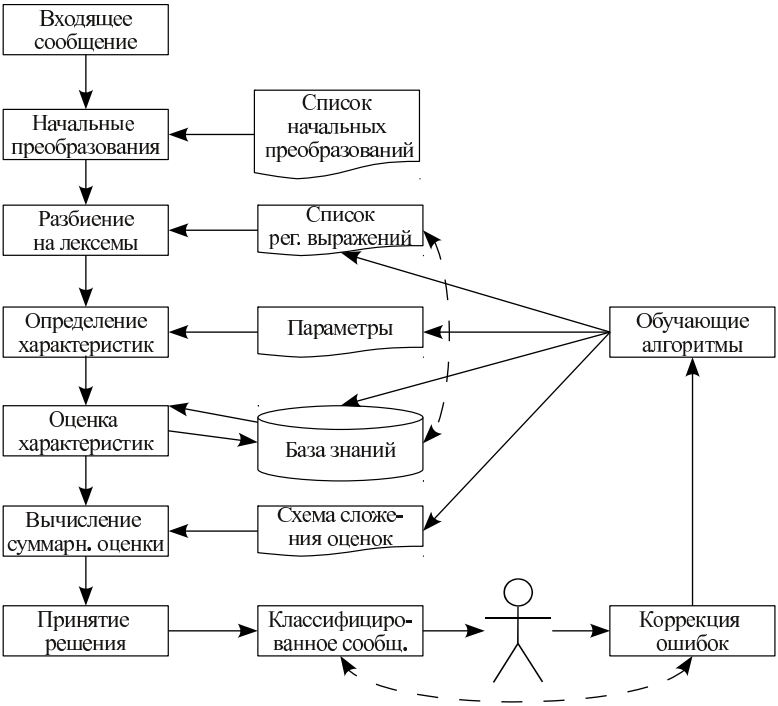

Рисунок 1 - Унифицированная модель фильтрации электронной корреспонденции

Необходимо отметить, что структура базы знаний, обучающие алгоритмы и параметры системы существенно влияют на конечный результат при классификации текстовой информации и формируются в соответствии с выбранным подходом. В данной работе для разработки фильтра на основе предложенной модели был выбран статистический метод фильтрации.

Третья глава посвящена вопросам разработки алгоритмов программного обеспечения, входящего в состав спам-фильтра.

Представленные в третьей главе алгоритмы позволяют реализовать программное обеспечение для фильтрации почты, основанное на теореме Байеса, но в отличие от классического байесовского фильтра, предложенного Полом Грехэмом, для расчета вероятностей используются не лексические последовательности, а регулярные выражения. Для оптимизации работы данного спам-фильтра были предло-

жены вспомогательные алгоритмы, облегчающие работу фильтра с регулярными выражениями и повышающие его быстродействие.

При построении современных почтовых фильтров требуется возможность тонкой настройки правил фильтрации почты с целью уменьшения доли спама в корреспонденции при получении почты, при этом необходимо стремиться и к уменьшению времени, затрачиваемого на настройку программы фильтрации, и к автоматизации реализации правил обработки почты.

Для решения этой задачи с учетом заданной предметной области предлагается использование регулярных выражений как системы синтаксического разбора текстовых фрагментов по формализованным шаблонам, основанной на системе записи образцов для поиска. Как показали исследования, проведенные в данной работе, регулярные выражения являются гибким инструментом для создания почтовых фильтров, ориентированных именно на конкретную предметную область.

Предложенный алгоритм автоматической генерации регулярных выражений на основе словарей русского языка и обучающей выборки позволяет добиться большей надежности при работе фильтра, использующего созданный список регулярных выражений, и предоставляет следующие возможности:

- автоматическое объединение разных вариантов написания одного слова в одно регулярное выражение повышает надежность почтового фильтра за счет генерации регулярных выражений, описывающих ранее не встречавшиеся формы слова;
- распознавание намеренно искаженных слов. Реализация данной возможности вручную трудоемка, а при автоматической генерации регулярных выражений — нет;
- результативный поиск встреченных слов в служебных словарях и словарях русского языка для исключения бессмысленных фраз. В словарях содержится перечень слов в их основной форме. Регулярные выражения позволяют осуществлять поиск слов по базе знаний без учета их окончаний.

При обработке больших объемов данных часто бывает необходимо реализовать поиск регулярных выражений, которым соответствует заданная строка. Предложен алгоритм оптимизации поиска регулярных выражений в базе данных с помощью индексной таблицы.

Общий алгоритм статистической фильтрации с применением

регулярных выражений представлен на рисунке 2.

При использовании регулярных выражений в процесс статистической фильтрации вносятся изменения, адаптирующие подход под использование именно регулярных выражений вместо лексем. Во-первых, использование регулярных выражений влияет на структуру базы знаний, в которой становится необходимым хранить не только лексе-

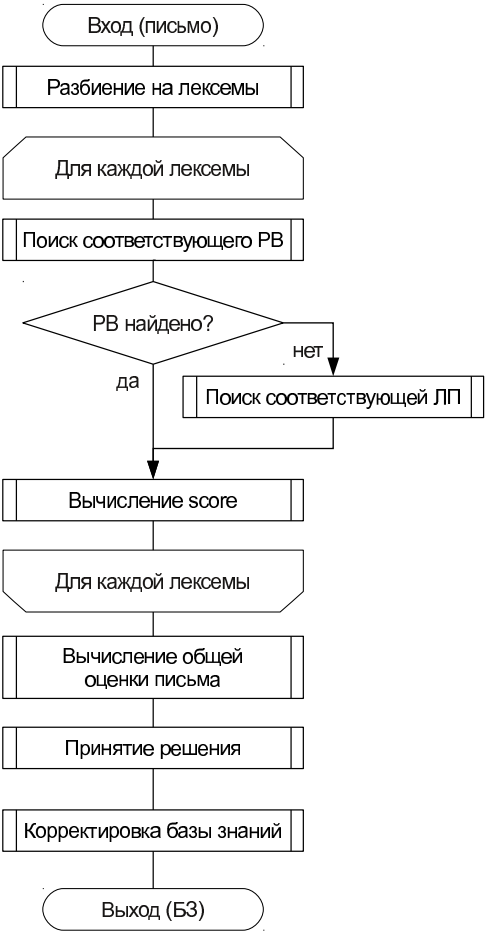

Рисунок 2 -Общий алгоритм статистической фильтрации с применением регулярных выражений

мы, но и соответствующие им регулярные выражения, а также индексные таблицы, позволяющие ускорить поиск по регулярным выражени-

ям. Во-вторых, использование регулярных выражений влияет на процесс первоначальной настройки фильтра на основе обучающей выборки, а также вносит изменения непосредственно в сам процесс фильтрации и принятия окончательного решения по классификации заданного входящего сообщения.

Четвертая глава посвящена практической реализации предложенных в предыдущих главах подходов к фильтрации входящей электронной корреспонденции.

Для разработки программного обеспечения были выбраны язык программирования C++, СУБД PostgreSQL, библиотека для работы с текстовой информацией Qt.

Для организации взаимодействия разрабатываемого фильтра с агентом передачи почты (MTA) был выбран Sendmail как самый гибкий и широко распространенный агент передачи почты. Он предоставляет API для подключаемых модулей фильтрации (milter), позволяющее наращивать и без того гибкий функционал Sendmail.

Предложены подходы и конкретная реализация разработанных алгоритмов, проведен анализ производительности различных реализаций, предложенного программного обеспечения, с выбором оптимального по времени при равной точности.

Разработанное программное обеспечение состоит из следующих модулей, отображенных на рисунке 3.

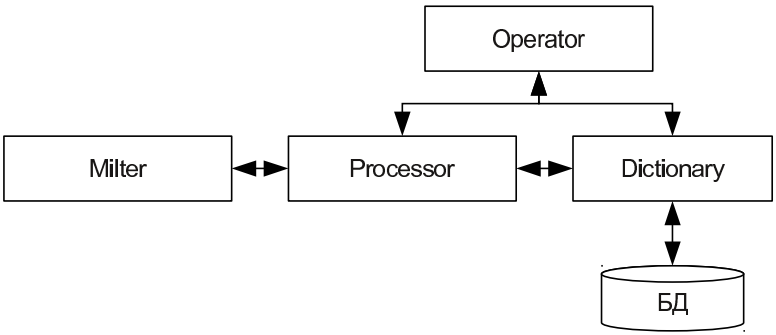

Рисунок 3 - Общая структура разработанного программного обеспечения.

1. Dictionary — модуль работы со словарем регулярных выражений и базой знаний, содержащей данные об уже обработанной корреспонденции. Фактически в этот модуль заложено довольно мало бизнес-логики и он является объектом доступа к данным (DAO — data

access object).

2. Processor — модуль, отвечающий за сборку сообщения в единое целое, предварительную обработку текстов писем, перекодирование письма в UTF8. Модуль Processor использует Qt Messaging Framework для разбора писем по частям.

3. Milter — модуль интеграции с SendMail через Milter API.

4. Operator — модуль вызова различных функций с командной строки, в том числе обучение фильтра, тестирование производительности и перестройка индексов.

Проведены статистические исследования различных конфигураций обучающей выборки и тестирования точности на контрольной выборке. Контрольная выборка составляла 10 % от объема обучающей. Разработанный фильтр показывает, что ошибка ложной идентификации легальных писем в качестве спама стремится к нулю, что является необходимым требованием к такого рода фильтрам (рисунок 4).

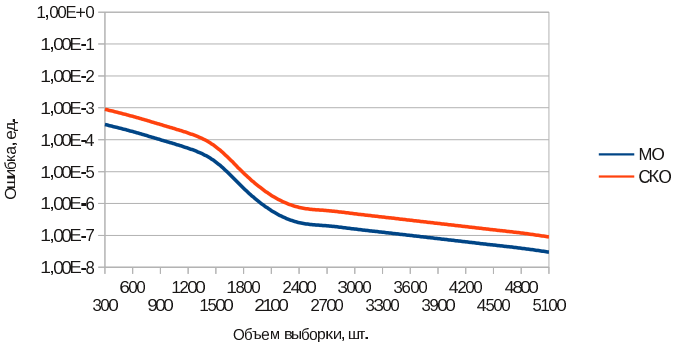

Рисунок 4 - Зависимость точности идентификации легального письма от объема обучающей выборки

В заключении приводится обобщение основных результатов диссертационной работы.

Основные результаты работы

В диссертационной работе разработаны теоретические основы и прикладные методы классификации входящей электронной почты на легальную корреспонденцию и спам. Решены следующие основные задачи:

1. Разработаны математические модели, описывающие рассмотренные методы классификации входящей корреспонденции.

2. Предложена обобщающая математическая модель фильтрации спама.

3. Предложен метод классификации электронной почты, основанный на использовании регулярных выражений в статистической модели фильтрации, позволяющий улучшить качество идентификации спама и легальной корреспонденции. Для реализации данного методы были предложены и разработаны сопутствующие алгоритмы, оптимизирующие работу с регулярными выражениями, хранимыми в базе данных.

4. Предложены алгоритмы индексации поиска регулярных выражений в базе данных, позволяющие оптимизировать скорость работы предложенных алгоритмов по классификации электронной корреспонденции.

5. Предложен алгоритм автоматической генерации регулярных выражений для почтовых фильтров, позволяющий уменьшить время первоначальной настройки программы фильтрации входящей электронной корреспонденции.

6. Разработан программный продукт, реализованный в виде подключаемого модуля для MTA Sendmail, подтверждающий достоверность полученных результатов.

Разработанные в диссертационной работе модели и алгоритмы можно применять при построении новых и доработке существующих программ фильтрации входящей электронной корреспонденции, например при разработке новых библиотек на основе Milter API для MTA Sendmail, Postfix и др.

При соответствующих модификациях предложенных алгоритмов их можно применять не только для фильтрации входящей корреспонденции, но и для классификации любой текстовой информации на несколько заранее определенных категорий.

Публикации по теме диссертации

1. Баранчиков А.И., Баранчиков П.А., Баранчикова Е.А. Формирование правил для создания антиспамовского фильтра в системах дистанционного обучения // Электронное обучение и управление знаниями учебного заведения. Рязань: РФ МЭСИ, 2007. С.201.

2. Баранчиков А.И., Баранчикова Е.А. Негативное влияние спама

на работу малых информационных систем и основные методы борьбы с ним //Вестник РГРТУ. 2007. № 21 с. 51-53.

3. Баранчикова Е.А. Выборка данных для анализа почтовых сообщений электронной почты на предмет их принадлежности к спаму //Новые информационные технологии в научных исследованиях и образовании: материалы 13 всероссийской научно-технической конференции студентов. Часть 1. Рязанский государственный радиотехнический университет. 2008. С. 145 — 147.

4. Баранчикова Е.А. Способ фильтрации электронных почтовых сообщений // Вестник РГРТУ. 2009. №2. — С. 56—60.

5. Баранчикова Е.А. Алгоритм автоматической генерации регулярных выражений (РВ) для спам-фильтра на основе обучающей выборки // Информационные и телекоммуникационные технологии: материалы 34-й всероссийской научно-технической конференции. Часть 1. Рязань: РВВКУС, 2009. С. 380-381

6. Баранчикова Е.А. Метод фильтрации электронной почты на основе теоремы Байеса с применением регулярных выражений//Новые информационные технологии в научных исследованиях и образовании: материалы 14 всероссийской научно-технической конференции студентов, молодых ученых и специалистов. Рязанский государственный радиотехнический университет. 2009. С. 154 — 156.

7. Баранчикова Е.А., Макаркина Л.Г. Построение автоматизированной информационной системы для туристического агентства с учетом ограниченности ресурсов предприятия// XXX VI ГАГАРИНСКИЕ ЧТЕНИЯ: научные труды международной молодежной научной конференции в 8 томах. Москва, 6 — 10 апреля 2010 г. М.: МАТИ, 2010. Т.4. С. 64 — 65.

8. Баранчиков А.И., Баранчикова Е.А. Анализ подходов к программной реализации статистического почтового фильтра // Информационные технологии в научных исследованиях. Рязань: РГРТУ, 2010. С.33-39.

9. Баранчикова Е.А. Особенности проектирования систем защиты малых информационных систем // Информационные технологии. Рязань: РГРТУ, 2011. С.30-36.

10. Баранчиков П.А., Баранчикова Е.А. Оптимизация поиска регулярных выражений в базе данных с помощью индексной таблицы //Вестник РГРТУ. 2011. № 37. С. 59-64.

Баранчикова Екатерина Александровна

МОДЕЛИ, АЛГОРИТМЫ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ФИЛЬТРАЦИИ ЭЛЕКТРОННОЙ КОРРЕСПОНДЕНЦИИ ДЛЯ ИНФОРМАЦИОННОЙ СИСТЕМЫ С ОГРАНИЧЕННЫМИ РЕСУРСАМИ

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Подписано в печать 02.12.2011. Формат бумаги 60× 80 1/16.
Бумага офисная. Печать трафаретная. Усл. печ. л. 1.
Уч.-изд. л. 1. Тираж 100 экз.

Редакционно-издательский центр
Рязанского государственного радиотехнического университета.
390005, г.Рязань, ул. Гагарина, 59/1.