УДК 681.3:621.391

Е.А. Баранчикова

СПОСОБ ФИЛЬТРАЦИИ ЭЛЕКТРОННЫХ ПОЧТОВЫХ СООБЩЕНИЙ

Рассматриваются аспекты выборки данных для анализа почтовых сообщений на предмет их принадлежности к спаму. Проведен анализ способов выборки данных и возможных методов классификации входящей корреспонденции на легальную и спам. Предложен подход к фильтрации почтовых сообщений, позволяющий снизить финансовые потери предприятий, связанные с получением спама.

Ключевые слова: *спам, выборка данных, информационная система, формат почтовых сообщений, фильтрация почты, лексический анализ.*

Введение. В настоящее время многие предприятия различных форм собственности используют при работе современные информационные системы (ИС). Во многих современных ИС 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.

Постановка задачи. Спам (любое сообщение, пришедшее по электронной почте, без явного на него запроса со стороны пользователя ИС) в наше время стал существенной проблемой службы обмена почтовыми сообщениями. За последние несколько лет спам превратился из легкого раздражающего фактора в одну из самых серьезных угроз информационной безопасности. Основными негативными последствиями спама являются [1]:

– получение корреспонденции, содержащей вирусы;

– увеличение трафика, вследствие чего может быть парализована работа почтового сервера, увеличивается время доставки легальной почты (под легальной корреспонденцией в данной статье будем понимать корреспонденцию, запрашиваемую самим пользователем в явном виде), увеличиваются расходы предприятия на оплату услуг провайдера;

– постоянный рост времени, которое сотрудники вынуждены тратить на разбор и чтение писем электронной почты, а следовательно, и финансовых потерь предприятий (например, в 2007 году, по подсчетам исследовательской фирмы Nucleus Research, потери в производительности одного рядового сотрудника американской компании, вызванные спамом, в денежном эквиваленте составили 712 долларов [2]).

В статье рассматриваются методы выборки данных для анализа почтовых сообщений электронной почты с целью выявления их принадлежности к спаму, а также принцип построения антиспамовского фильтра на основе лексического и семантического анализа информации, содержащейся в письме.

Методика. Служебная информация зависит от протокола, по которому было получено письмо, и определяется стандартом работы протокола. Наиболее известными, являются следующие протоколы.

1. SMTP (Simple Mail Transfer Protocol) – протокол обмена почтовыми сообщениями в сетях, работающих по стеку протоколов TCP/IP, наиболее распространенному в сети Internet.

2. UUCP (Unix to Unix Copy Protocol) – протокол взаимодействия между удаленными узлами сети, в настоящее время используется крайне редко.

3. Протокол обмена сообщениями в сети Fidonet, которая является международной некоммерческой компьютерной сетью, построенной по технологии «из точки в точку» и «коммутация с запоминанием».

В настоящее время адреса протокола SMTP являются стандартными почтовыми адресами в сети Internet и фактически являются стандартом де-факто.

Для передачи почты по протоколу SMTP в TCP/IP сетях разработан ряд специальных программ - почтовых серверов, одной из разновидностей которых являются агенты пересылки сообщений (англ. mail transfer agent, MTA). Наибо-

лее популярными свободно распространяемыми MTA в настоящее время являются Sendmail, Postfix, Exim и Qmail. По различным данным, среди открытых серверов электронной почты наблюдается примерно такое распределение Sendmail – 24 %, Postfix – 17 %, Exim – 9 %, Qmail – 4 %.

Наиболее распространенным MTA является **Sendmail,** имеющий ряд преимуществ по сравнению с другими вышеперечисленными почтовыми серверами (неограниченные возможности описания конфигурации; подключение внешних почтовых фильтров через специальный интерфейс; простота в настройке при работе с использованием стандартных решений). В стандартный дистрибутив Sendmail версии 1.8.12 и выше включен *Sendmail Content Management API (Milter)*, предназначенный для взаимодействия MTA Sendmail и внешних программ, обеспечивающих фильтрацию почты. Поскольку данный MTA наиболее гибко решает задачу выборки и фильтрации входящей корреспонденции с использованием сторонних программ, целесообразным является его дальнейшее использование для решения поставленной задачи.

Интерфейс Milter API, входящий в Sendmail, обеспечивает взаимодействие MTA со сторонними программами (рисунок 1), которые осуществляют проверку сообщения на его принадлежность либо к спаму либо к легальной корреспонденции во время его прохождения через службу доставки почты и в зависимости от полученных результатов могут вносить изменения в само сообщение (например, изменив тему письма и добавив туда пометку, что данное письмо было отнесено к спаму при обработке внешней программой фильтрации). Далее письма, которые имеют такую пометку, можно либо удалять, либо помещать в специальную папку, для последующей фильтрации пользователем [3].

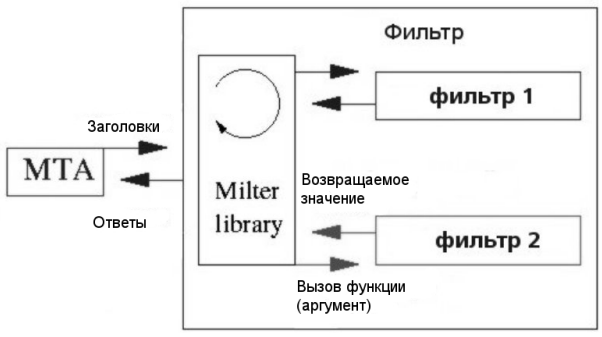

Рисунок 1 – Взаимодействие MTA с внешними почтовыми фильтрами

Рассмотрим более подробно процесс получения электронного письма, который условно можно разбить на несколько стадий:

1) получение заголовка сообщения почтовым сервером;

2) анализ полученной информации из заголовка сообщения;

3) получение/отказ от получения тела письма;

4) анализ тела письма, на основе которого принимается окончательное решения о принадлежности данного письма к спаму;

5) доставка письма конечному пользователю/удаление письма.

Следует отметить особую важность второго этапа, поскольку именно в этот момент времени принимается первичное решение о причислении данного письма к спаму или к легальной корреспонденции. На основе принятого решения происходит загрузка почтовым сервером или отказ от нее тела письма, которое может содержать как вредоносную, так и просто ненужную информацию в большом объеме, замедляющую работу ИС. Отсечение информации на данном этапе приводит к существенной экономии ресурсов ИС за счет уменьшения объема сетевого трафика, а также позволяет избежать дополнительного риска, связанного с возможностью получения вирусов вместе со спамом.

Для того чтобы понять, на основе какой информации производится анализ заголовков электронных писем, передаваемых по SMTP протоколу, рассмотрим более подробно их формат. SMTP протокол реализует передачу сообщений, формат которых описывается стандартом RFC 822 (рисунок 2), предусматривающим их разбиение на две части. Первая часть называется заголовком письма. В нее вносятся все данные, идентифицирующие сообщение. Вторая часть называется телом письма и непосредственно содержит текст письма. Заголовок состоит из полей данных, которые используются по мере необходимости внесения дополнительной информации в сообщение. Для полей заголовка не существует определенного порядка следования, т.е. поля заголовка могут располагаться в произвольном порядке. Кроме того, в одном сообщении поля заголовка могут повторяться. На рисунке 2 представлен общий вид почтового сообщения, соответствующего требованиям RFC 822. Обязательными в заголовке сообщения являются поля Received, Return-Path, To, Reply-To, From, Date и некоторые другие. На основе информации, содержащейся в этих полях, производится первичная проверка на принадлежность письма к спаму. Рассмотрим формат данных полей. На основе информации, содержащейся в них, может быть произведена фильтрация почты [4].

Рисунок 2 – Формат сообщения согласно RFC-822

Поле заголовка Received используется для идентификации SMTP-серверов, которые принимали участие в процессе доставки сообщения от отправителя получателю. Каждый сервер добавляет к почтовому сообщению свое поле Received, с указанием специфических сведений о себе. На основе информации, содержащейся в этом поле, проводится проверка сетевого узла, который передал почту, на соответствие MX-записям (от англ. mail exchanger) DNS (доменную систему имен, от англ. Domain Name System).

Поле Return-Path указывает, на какой адрес должно быть переслано почтовое сообщение в случае, если оно не может быть доставлено адресату.

Поле Reply-To содержит адрес электронной почты, на который будет послан ответ на сообщение.

Поле From — адрес отправителя почтового сообщения. На основе информации, содержащийся в этом поле, может быть реализована фильтрация почтовых сообщений с помощью белых и черных списков, если заведомо известные почтовые адреса относятся к той или иной категории .

Поле Date содержит информацию о дате и времени отправления письма.

Поле To содержит адрес получателя письма. В случае если адрес получателя отсутствует в этом поле, письмо может быть классифицировано как спам, но существует вероятность возникновения ошибки второго рода, когда адрес получателя в силу каких-либо причин может быть указан в полях CC или BC.

Поля CC и BC содержат адреса получателей, которые должны получить копию письма, видимые другим пользователям и скрытые соответственно.

В связи с возросшей активностью спамеров в настоящее время вышеперечисленные способы фильтрации почты не дают требуемых результатов и являются недостаточно эффективными в использовании.

Также в заголовке могут присутствовать и необязательные поля, которые более подробно идентифицируют сообщение для сервера SMTP, однако согласно RFC 822 могут и не присутствовать в сообщении. Тем не менее, эти поля в настоящее время широко распространены и содержат полезную информацию для борьбы со спамом. Наиболее часто используемым и полезным с этой точки зрения является поле Subject (Тема), это связано с тем, что данное поле является одним из самых слабых мест спама и обязательно должно содержать краткую информацию о теме письма, которая дает конечному получателю представление о содержании самого сообщения, и без которой спам как коммерческая рассылка не имеет смысла. Это объясняется тем, что в случае отсутствия темы письма или ее несоответствия содержанию письма получатель может не обратить на сообщение внимания, т.е. на основе данных, содержащихся в этом поле, мы имеем возможность провести лексический и семантический анализ полученной информации с целью выявления принадлежности данного сообщения к спаму. Этот метод позволяет получить приемлемые результаты и более подробно будет рассмотрен ниже.

На основе данных, содержащихся в заголовке письма, почтовый сервер проводит анализ почтового сообщения на предмет его принадлежности к спаму как самостоятельно, так и с привлечением внешних программ фильтрации. Данный анализ может включать в себя следующие стадии.

1. Проверка домена, с которого направлена почта на существование. Почтовый сервер производит проверку через DNS, существует ли сервер, с которого пришло письмо или нет. Если такой сервер не существует, пришедшие с него письма классифицируются как спам.

2. Проверка сетевого узла, который передал почту на соответствие MX-записям DNS. Данная проверка основана на том, что при получении нового письма IP-адрес, с которого оно было доставлено, должен соответствовать имени сервера, передавшего данное сообщение.

3. Проверка адреса назначения электронного письма. В случае, если адрес получателя электронного письма отсутствует в соответствующем поле, письмо классифицируется как спам. В данном случае существует вероятность возникновения ошибки второго рода, т.к. адрес получателя электронного письма может быть указан в скрытом поле, что необходимо учитывать при

проведении данной проверки.

4. Анализ темы письма с помощью лексического анализатора является наиболее сложным и наиболее перспективным направлением в борьбе со спамом. Для реализации данного метода можно использовать как встроенные средства почтовых серверов, так и специально разработанные внешние программы фильтрации почты, адаптированные под конкретную предметную область. Процесс фильтрации почты с применением лексического и семантического анализа схематично представлен на рисунке 3.

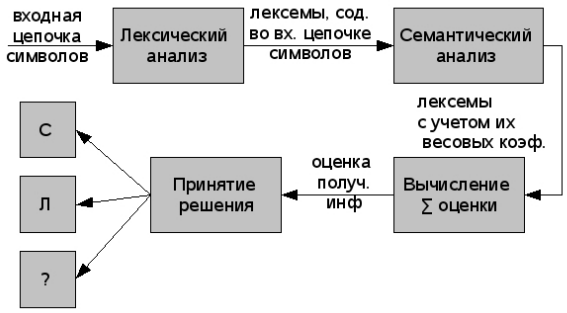

Рисунок 3 – Классификация почтовых сообщений с целью выявления их принадлежности к спаму с помощью лексического и семантического анализа

На первом этапе осуществляется лексический анализ входной цепочки символов, в нашем случае — либо темы, либо непосредственно текста письма. В информатике под лексическим анализом понимается процесс аналитического разбора входной последовательности символов с целью получения на выходе последовательности символов, называемых «токенами» [5,6]. При этом, группа символов входной последовательности, идентифицируемая на выходе процесса как токен, называется лексемой. Другими словами, в процессе лексического анализа производятся распознавание и выделение лексем из входной последовательности символов. Как правило, лексический анализ производится с точки зрения определенного языка или набора языков. Язык, а точнее его грамматика, определяет определенный набор лексем, которые могут встретиться на входе процесса. Распознавание лексем в контексте грамматики обычно производится путем их идентификации (или классификации) согласно идентификаторам (или классам) токенов, определяемых грамматикой языка:

$$q = delim \cup phrase\ , \qquad (1)$$

где q – множество всех лексем входной последовательности; $delim$ – множество делителей языка {«_», «;», «,»....}; $phrase$ – множество значащих лексем входной последовательности. Данное множество задается в следующем виде:

$$phrase = t_{id} \mid >< \mid t_{lex}\ , \qquad (2)$$

где t_{id} – таблица идентификаторов; t_{lex} – таблица лексем.

При этом значение каждой лексемы из t_{lex} должно принадлежать либо множеству всех возможных словоформ русского языка (или любого другого, на котором может быть получено сообщение), либо множеству служебных терминов для конкретной предметной области (3). Например, множество может включать термин «*Oracle*» для предметной области, связанной с разработкой баз данных с использованием СУБД Oracle, т.е.

$$lex \subset D_R \cup D_S\ , \qquad (3)$$

где lex – значение лексем из t_{lex} ; D_R – словарь русского языка; D_S – служебный список лексем, хранящийся в базе данных, с указанием их типов и весовых коэффициентов.

Данное условие позволяет исключить из анализа сообщения, не несущие смысловой нагрузки или написанные на иностранном языке, когда достоверно известно, что такие сообщения причисляются к спаму по умолчанию. На начальном этапе внедрения данной системы это условие может быть не активно для того, чтобы эксперт имел возможность заполнить служебный словарь, поскольку легальная корреспонденция может содержать не только слова русского языка, но и слова другого языка, которые используются в качестве терминов в рассматриваемой предметной области. После его заполнения данная возможность может быть включена в рассматриваемую систему по согласованию с экспертом.

На втором этапе (рисунок 3) производится семантический анализ, результатом работы которого является перечень лексем, содержащихся во входной цепочке символов с указанием их принадлежности к определенному множеству и их весовых коэффициентов в данном множестве. На следующем этапе вычисляется суммарная оценка, на основе которой принимается решение о принадлежности данного сообщения либо к спаму, либо к легальной корреспонденции, либо выдается ответ о невозможности классифицировать данное письмо. Рассмотренный метод является эффективным инструментом, позволяющим отсеивать спам с большой долей вероятности, в то время как другие методы могут давать более низкие результаты.

Разработка структуры данных для реализации предложенной методики.

Для хранения данных, полученных в результате анализа сообщений, разрабатывается схема базы данных, в которой накапливаются данные, требующие экспертной оценки, данные, харак-

теризующие сообщения, которые должны быть обязательно доставлены адресатам и определенные как спам.

Списки лексем (рисунок 3), хранящихся в базе данных, могут быть представлены в виде следующих множеств:

$$D_S = D_W \cup D_B \cup D_G, \quad (4)$$

где D_W – белый список; D_B – черный список; D_G – серый список.

При добавлении новых лексем в служебный словарь они помечаются как неизвестные. В этом случае получаем следующее выражение:

$$D_S = D_W \cup D_B \cup D_G \cup D_U, \quad (5)$$

где D_U – список неизвестных лексем.

На основе информации в базе данных генерируется фильтр почтового сервера, время генерации которого позволяет поддерживать уровень спама не выше определенного уровня, с одной стороны, и не перегружать эксперта — с другой.

$$t_{гф} \le t_{опт} : t_{гф} \to t_{опт}, \text{ при котором } (K_c \to min) \wedge (t_э \to max), \quad (6)$$

где $t_{гф}$ – время между двумя последовательными генерациями фильтра; $t_{опт}$ – оптимальное время между двумя последовательными генерациями фильтра, при котором уровень спама не превышает заданных значений; K_C – количество нераспознанных писем спама в процентном отношении к общему числу полученных сообщений; $t_Э$ – промежуток времени, через который эксперт должен проводить обновление фильтра.

Выводы: широко эксплуатируемые на данный момент, обучаемые фильтры на основе теоремы Байеса дают точность идентификации письма около 90 %. Предполагается, что методы на основе лексического анализа могут снизить вероятность неверной идентификации почтовых сообщений примерно вдвое.

За неделю наблюдаемым сервером было обработано 64945 электронных писем. 47727 были отсеяны черными списками DNS (dnsbl), 640 были отвергнуты алгоритмом SPF, 15316 писем были отвергнуты алгоритмом autoSPF. Сервером было принято 1262 письма, среди которых доля «спама» составляет около 30 %, то есть – примерно 400 – 450 писем.

Предполагаемое количество «спама» за неделю при использовании фильтра на основе лексического анализа должно составить около 20 сообщений.

Рассмотренный подход к фильтрации почты на основе лексического анализа темы и тела письма позволяет получить динамический антиспамовский фильтр, адаптирующийся к конкретной предметной области и накапливающий информацию для анализа в течение некоторого промежутка времени. На основе предложенного подхода возможна более детальная проработка семантического анализа текстовой информации, содержащейся в письме, с целью минимизации ошибок первого и второго рода при принятии решения о принадлежности почтового сообщения либо к легальной корреспонденции, либо к спаму.

Библиографический список

1. *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.

2. Google расскажет, почем сегодня спам // статья с ресурса http:// www.compulenta.ru/

3. Filtering Mail with Sendmail // статья с ресурса https://www.milter.org/

4. *Блам Р.* Система электронной почты на основе LINUX. Руководство администратора. – М.: Вильямс, 2001. – 448 с.

5. *Молчанов А.Ю.* Системное программное обеспечение. – СПб.: Питер, 2003. 395 с.

6. *Ахо А., Сети Р., Ульман Дж.* Компиляторы: принципы, технологии, инструменты. – М.: Вильямс, 2003. – 767 с.