УДК 519.688

Е.Н. Чернопрудова, Н.А. Соловьев, Л.А. Юркевская

ФИЛЬТРАЦИЯ НЕСАНЦИОНИРОВАННЫХ СООБЩЕНИЙ В ПОЧТОВЫХ ЭЛЕКТРОННЫХ СЕРВИСАХ

Оренбургский государственный университет, Оренбург, Россия

В статье предложено решение задачи фильтрации электронной почтовой корреспонденции на основе предварительной интеллектуальной обработки электронных сообщений с использованием нейросетевого классификатора. Обработка электронных сообщений включает в себя автоматическую обработку текста на основе лингвистического подхода. В работе рассмотрена векторная модель отображения признаков электронного сообщения. Предложено в качестве меры значимости термов использовать Ltc-меру взвешивания. Также обоснован комбинированный подход сокращения признакового пространства путем расчета величины характеризующей значимость терма для определенного класса k и формированием коллоката сообщения с использованием показателей силы смысловой (синаптической) связи между качественными признаками (термами) словосочетаний. Обоснованно использование меры тесноты взаимосвязи двух качественных признаков словосочетаний: коэффициенты ассоциации K_a и контингенции K_k. Для решения задачи фильтрации несанкционированных электронных сообщений выбрана адаптивная нейронная сеть ART преимуществом которой является способность самообучаться (создавать образы) для адаптации к изменяющимся потребностям адресата корреспонденции исследована и подтверждена эффективность предложенной модели электронного сообщения, интегрированной с методом нейросетевой классификации для интеллектуальной фильтрации электронной корреспонденции.

Ключевые слова: электронная почта, интеллектуальная обработка текста, нейросетевой классификатор.

Введение. Служба рассылки электронной почты в Internet, 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 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» [1].

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

структурой.

Система защиты электронной почты. Проблемам обеспечения ИБ электронной почты посвящены работы Ашманова И.С., Валеева С.С., Никитина А.П., 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.

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 [2]:

– 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.

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 [3]. Поэтому развитие методов защиты электронной почты остаётся актуальной тематикой научных исследований, объектом которых становится ИТКС корпоративных предприятий с территориально-распределенной структурой.

Постановка задачи классификации электронных сообщений. Системный анализ защиты электронной почты от НЭС выявил ряд противоречий между требованиями практики и состоянием теории спам-фильтрации, основным из которых становится противоречие между существенно возросшей интенсивностью спам-рассылок при недопустимо высоком уровне ложной классификации и отсутствием методов адаптации модели описания ЭС к изменяющимся потребностям адресатов корреспонденции в алгоритмах спам-фильтрации, работающих в реальном масштабе времени [4]. Отсюда, предметом исследования становятся методы, модели и средства адаптивной фильтрации НЭС электронной почты.

Эти обстоятельства определяют цель исследования: повышение достоверности обнаружения спам-рассылок на основе интеллектуальной фильтрации электронных сообщений в реальном масштабе времени.

Для достижения поставленной цели необходимо решить ряд задач научного характера:

– разработать модель электронного сообщения, учитывающую семантические свойства текста;

– разработать методику фильтрации НЭС на основе интеллектуальных методов классификации;

– реализовать систему спам-фильтрации, адаптивную к изменению содержания легитимных сообщений, работающую в реальном масштабе времени;

– оценить эффективность предложенных методов, моделей и средств.

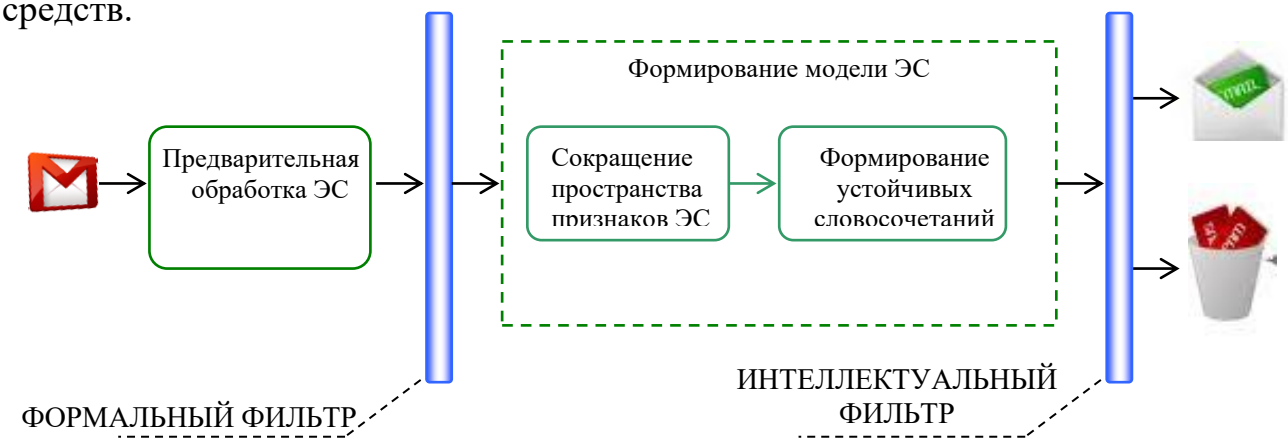


Рисунок 1 – Технология фильтрации электронных сообщений

Повысить достоверность разрабатываемой системы предлагается за счет использования двухуровневой системы фильтрации, состоящей из формального и интеллектуального фильтров.

На Рисунке 1 предложена технология реализации предлагаемой системы фильтрации электронных сообщений.

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

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

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

Пусть, для формальной постановки задачи, $L \in \{Let_i\}$ множество ЭС (писем), 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.

Тогда процедуру автоматической фильтрации P_f электронных сообщений L на множестве классов K можно представить в следующем виде

$$R(L(p_i), A(k_j)) \xrightarrow{P_f} \min \qquad (1)$$

Задача интеллектуальной фильтрации требует предварительной обработки сообщений

Модель электронного сообщения для интеллектуальной классификации. Содержание сообщения l описывается с помощью термов t, множество которых образует тезаурус $m^k\{t_1,....t_j\}$, $j=\overline{1,n}$ определенного класса k. В качестве термов выступают слова, отражающие содержание сообщения. В [5-7] предложено в качестве термов использовать не отдельные слова, а словосочетания или n-граммы. Кроме того, для повышения адекватности модели описания сообщения в [8] предложено учитывать информацию о структуре документа путём присваивания веса словам в зависимости от его месторасположения в документе (заголовок, тело сообщения, подписи). Однако тексты почтовых эс сообщений не всегда структурированы.

Результатом исследования существующих мер взвешивания термов текста является формирование пространства признаков, определяющих значимость соответствующего терма j в i-ом ЭС, состоящего из весового коэффициента w_{ij} и частоты термов f_{ij} в сообщении. Для устранения эффекта больших различий в частотах термов сообщения предложено в качестве меры значимости термов использовать Ltc-меру взвешивания, расчет которой определяется зависимостью вида [9]:

$$Ltc_{ij} = \frac{\log(f_{ij}+1)\log\left(\frac{M}{M_j}\right)}{\sqrt{\sum_{t_j=1}^{N}\left[\log(f_{ij}+1)\log\left(\frac{M}{M_j}\right)\right]^2}}, \quad (2)$$

где M – общее число сообщений в выборке;

N – число термов в выборке после удаления стоп-слов;

M_j – общее число сообщений, содержащих терм t_j.

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

$$L_k = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{j1} \\ w_{12} & w_{22} & \cdots & w_{j2} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1i} & w_{2i} & \cdots & w_{ji} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1N} & w_{2N} & \cdots & w_{MN} \end{bmatrix} \quad (3)$$

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

Однако, результирующая матрица признаков ЭС (3) будет имеет размерность, обработка которой потребует недопустимо больших вычислительных ресурсов и времени.

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.

Для сокращения признакового пространства задачи классификации известны следующие способы:

1) сокращение пространства признаков непосредственно для каждого класса,

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

Для реализации указанных способов известны методы многомерного статистического анализа, ориентированные на работу с текстовыми данными, такие как подсчет взаимной значимости термов [10], кластеризация термов относительно введенной метрики [11], выделение только тех термов, вес которых является максимальным [12].

В работе предложен комбинированный метод, основанный на том, что для каждого терма в сообщениях определенного класса вычисляется

величина $RF_{t_j}^k$, характеризующая значимость терма для определенного класса k [13]:

$$RF_{t_j}^k = \log_2(2 + \frac{a_i}{\max(1, b_i)}), \quad (4)$$

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.

В результате использования (4) пространство анализируемых термов сокращается. Термы, значимость которых $RF_{t_q}^k \le 0{,}5$, исключаются в данном классе k.

Модель ЭС будет более информативна, т.е. размерность матрицы определенного класса уменьшится, если помимо последовательности термов и их значимости учесть связи между термами.

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

В качестве меры смысловой близости между термами d в сообщении предложено использовать меру Дайса [14], которая рассчитывается по зависимости вида:

$$d(t_1, t_2) = \log_2(\frac{2*(f(t_1, t_2))}{f(t_1) + f(t_2)}) \quad (5)$$

где $f(t_1)$, $f(t_2)$ – частоты встречаемости термов t_1 и t_2 в сообщении;
$f(t_1,t_2)$ – частота совместной встречаемости термов t_1 и t_2.

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

Алгоритм формирования коллоката:

1) выделение значимых термов с учетом (4) для соответствующего класса k (spam/legitim);
2) расчет близости термов (5) и принятие решения о формировании коллоката. Решение о формировании коллоката для каждой пары термов принимается, если значение коэффициента Дайеса (5) выше, чем в соседних парах термов (левой и правой);
3) подтверждение смысловой значимости коллоката.

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

Наиболее распространенными мерами ассоциации являются MI-score, t-score и log-likelihood [15,16], которые признаны показателями силы смысловой (синаптической) связи между качественными признаками (термами) словосочетаний.

Мерой тесноты взаимосвязи двух качественных признаков словосочетаний являются коэффициенты ассоциации K_a и контингенции K_k, которые рассчитываются по следующим зависимостям [17]:

$$K_a = \frac{ad - bc}{ad + bc} \tag{6}$$

$$K_k = \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+c)(c+d)}} \tag{7}$$

где a – число сообщений, имеющих терм t_1, встречающихся в классе k;

b – число сообщений, в которых терм t_1 встречается в других классах;

c – число сообщений, имеющих терм t_2, встречающихся в классе k;

d – число сообщений, в которых терм t_2 встречается в других классах.

Экспериментально установлено [18,19], 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:

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

где T^k – терм устойчивых словосочетаний в сообщении;

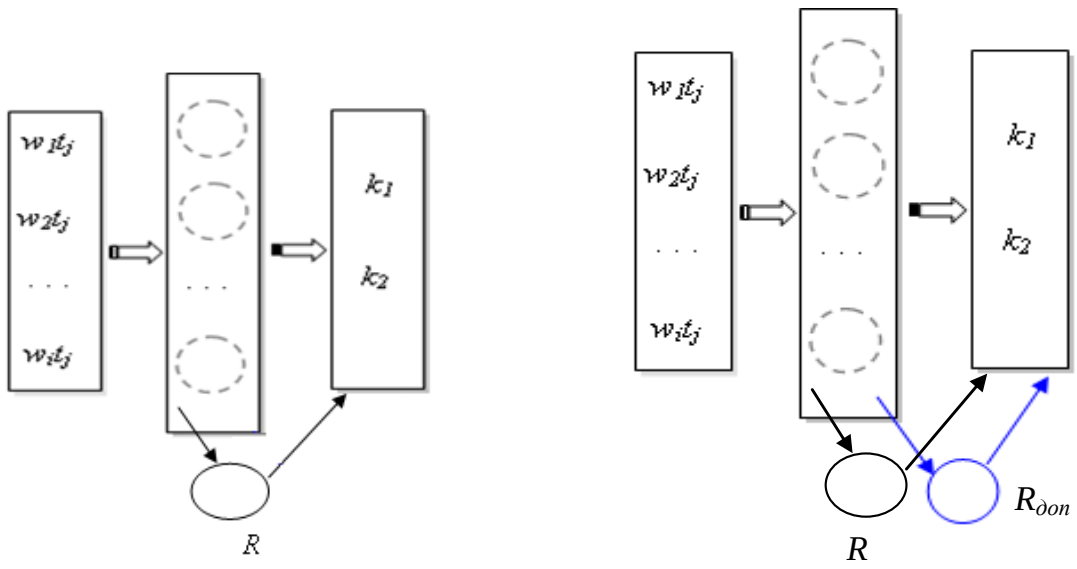
$w^*(t_j)$ – вес терма в сообщении после сокращения матрицы признаков (3).

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

Интеллектуальная классификация электронной корреспонденции. Исследования методов классификации текстов показали [9,12,14,15], что наиболее перспективным направлением в области классификации ЭС являются нейросетевые методы, преимуществами которых становятся: способность самообучаться (создавать образы) для адаптации к изменяющимся потребностям адресата корреспонденции, возможность распараллеливания процессов обработки информации для классификации ЭС в реальном масштабе времени в условиях неполноты, искаженности и неточности информации.

Для решения задачи фильтрации НЭС выбрана адаптивная нейронная сеть ART [19,20], структура которой представлена на рисунке 2а. Входной слой нейронной сети содержит столько нейронов сколько термов в словаре обучающей выборки (тезаурусе), элементами которого являются значения весов термов $w^*(t_j)$ анализируемого ЭС. Слой распознавания представляет собой набор нейронов, каждый из которых отвечает за один экземпляр класса.

Для исключения ложного распознавания 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 [16,17,18] (Рисунок 2б).



а б

Рисунок 2 – Компоненты нейросетевого классификатора ЭС
(а – компоненты нейронной сети ART 2а;
б – структура модифицированного классификатора ART 2а)

Модифицированный алгоритм нейросетевой классификации примет вид, представленный на Рисунке 3.

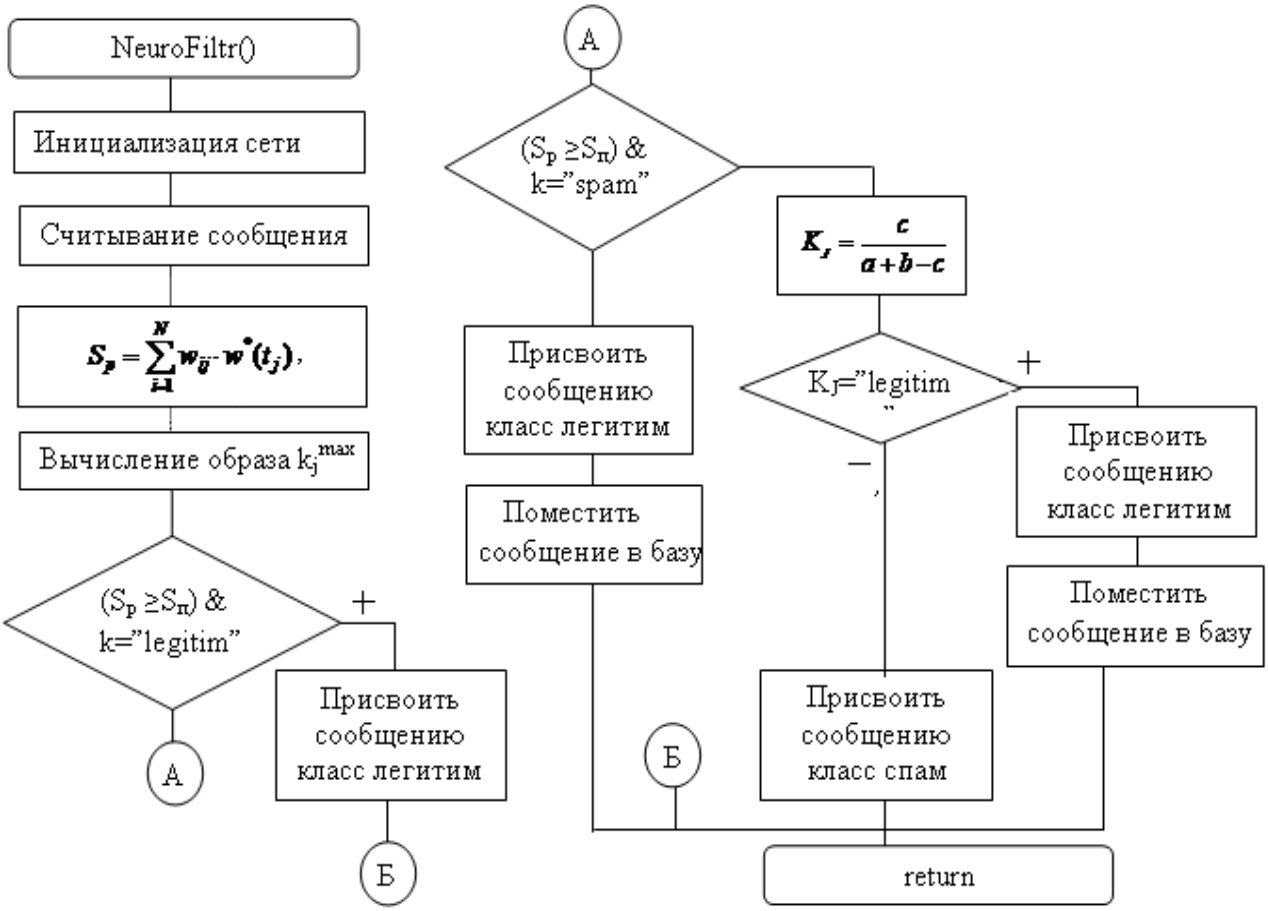


Рисунок 3 – Укрупненный алгоритм нейросетевой классификации нейронной сети

Программный проект прототипа системы спам-фильтрации реализован в соответствии с моделью IDEF0, представленной на рисунке 4[20,21].

Интерфейс программной системы спам-фильтрации в процессе обучения и тестирования представлен в виде экранных форм на Рисунке 5.

Для осуществления тестирования была разработана имитационная модель. Обучающая выборка представляла собой набор данных состоящих из электронных сообщений, представляющих собой спам-рассылку и легитимных писем, полученных из общедоступных ресурсов от определенных пользователей.

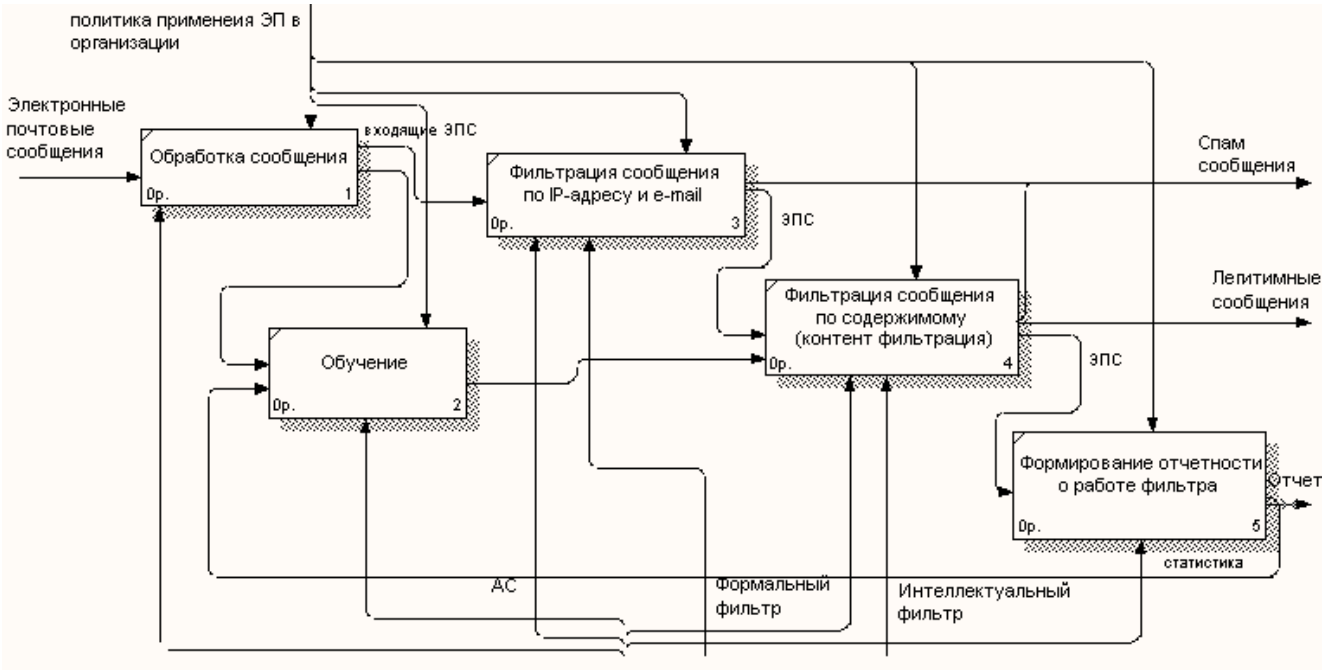
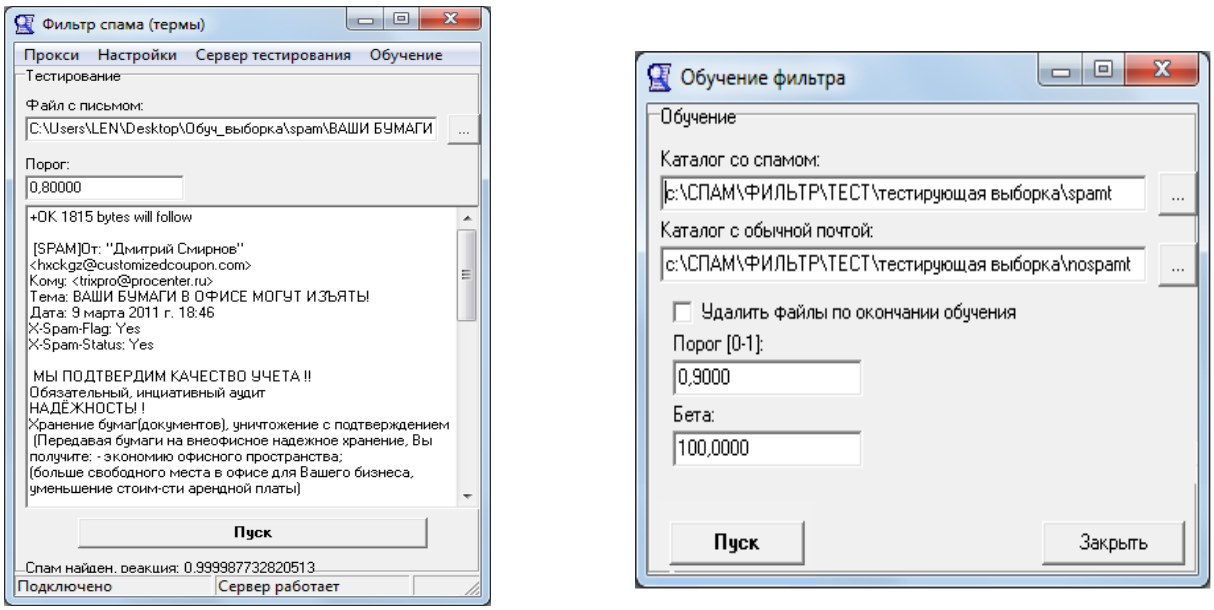


Рисунок 4 – Функциональная модель прототипа системы фильтрации по методологии IDEF0



а б

Рисунок 5 – Экранная форма спам – фильтра в режиме тестирования и обучения
(а – экранная форма спам – фильтра в режиме тестирования;
б – экранная форма спам – фильтра в режиме обучения)

В тестовых данных принимали участие письма, составляющие менее 10% от обучающей выборки.

На Рисунке 6 представлены сравнительные результаты тестирования предложенного спам-фильтра и фильтра, разработанного на основе байесовской методики классификации.

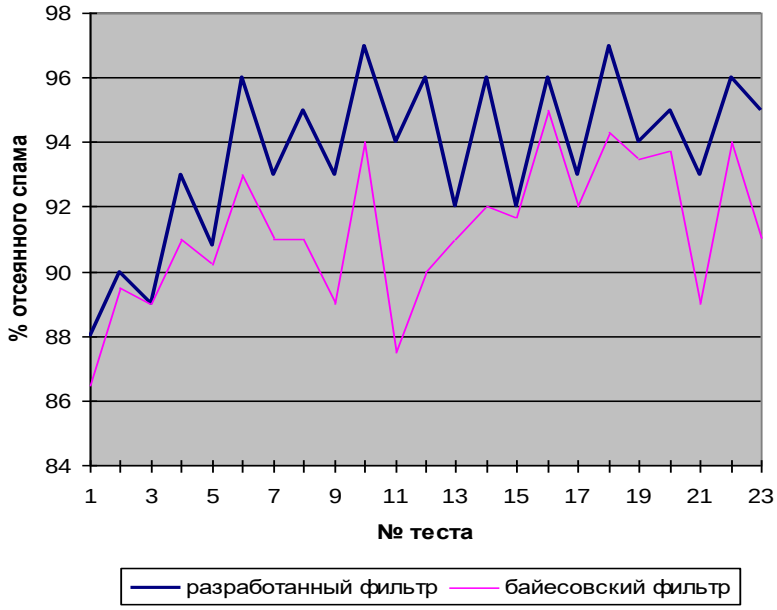


Рисунок 6 – Диаграмма результатов тестирования

Как следует из диаграммы, доля спама, отсеянного предложенной системой фильтрации, выше, чем у байесовского фильтра, при вероятности ложного срабатывания, не выходящего за пределы допустимой величины 0,05.

Таким образом, эффективность предложенной модели электронного сообщения, интегрированной с методом нейросетевой классификации для интеллектуальной фильтрации электронной корреспонденции подтверждается.

Заключение. В работе предложен прототип системы фильтрации, позволяющий в процессе обучения выделять ключевые словосочетания, характеризующие тематику корреспонденции конкретного пользователя электронной почты, что делает фильтрацию персонифицированной. В этом случае точность автоматической фильтрации электронных сообщений становится регулируемой относительно меняющихся интересов пользователя.

ЛИТЕРАТУРА

1. Николаев, И.А. Спам: экономические потери: Аналитический доклад /[Электронный ресурс] (http://www.fbk.ru/news/5419/83743/).
2. Слепов, О. Контентная фильтрация /О. Слепов // JetInfo № 10 (149)/2005 [Электронный ресурс] (http://www.jetinfo.ru/Sites/new/Uploads/ 2005_10.pdf).

3. Соловьев, Н.А. Развитие концепции обнаружения вторжений /Е.Н. Чернопрудова, Н.А. Соловьев //Современные информационные технологии в науке, образовании и практике: материалы VIII Всерос. науч.-практ. конф., /Оренбург. гос. ун-т. – Оренбург: ГОУ, 2009. – С. 66-67. - ISBN 978-5-7410-0975-8
4. Чернопрудова, Е.Н. Нейросетевая модель интеллектуальной фильтрации несанкционированных рассылок /Е.Н Чернопрудова // Информации Материалы IX всероссийской научно-технической конференции – Оренбург: ОГУ, 2010, с. 44-47.
5. Чернопрудова, Е.Н Интеллектуальная фильтрация несанкционированных рассылок на основе нейронной сети /Е.Н. Чернопрудова, Н.А. Соловьев // Академический журнал «Интеллект. Инновации. Инвестиции». Спец. выпуск, 2011. с.106-107.
6. McCallum, A. A comparison of Event Models for Naïve Bayes Classifiection / A. McCallum, K. Nigam; // In AAAI-98 Workshop on Learning for Text Categorization. – 1998 – 8 с.
7. Fuernkranz, J. A study using n-gram features for Text Categoriza-tion / J. Fuernkranz // Tech report OEFAI-TR-98-30 – 1998.
8. 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.
9. Li, Y. Classification of Text Documents /Y.H. Li, A.K. Jain //The Computer Journal, Vol. 41, No. 8, 1998
10. Mingyong, L. An improvement of TFIDF weighting in text categorization / L. Mingyong, Y. Jiangang [Электронный ресурс] http:// www.ipcsit.com/vol47/009-ICCTS2012-T049.pdf)
11. Cover, T. Elements of Information theory / T. Cover, J. Thomas [Электронный ресурс] (https:// web.cse.msu.edu/cse842/Papers/CoverThomas-Ch2.pdf)
12. Кондратьев, М.Е Двухуровневая иерархическая кластеризация новостного потока в РОМИП 2006 / М.Е. Кондратьев // Российский семинар, по оценке методов информационного поиска. Труды четвертого российского семинара РОМИП'-2006. Санкт-Петербург: НУ ЦСИ, 2006, 274 с. 126-138 [Электронный ресурс] (http://romip.narod.ru/ romip2006/index.html)
13. Hotho, A. Ontology-based Text Clustering / A. Hotho, S. Staab, A. Maedche [Электронный ресурс] (http://www.cs.cmu.edu/mccallum/textbeyond/papers/hotho.pdf)
14. Lan, M. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization / M. Lan, C. L. Tan, Senior Member [Электронный ресурс] (https://www-old.comp.nus.edu.sg/~tancl/publications/j2009/PAMI2007-v3.pdf)
15. Маннинг, К.Д. Введение в информационный поиск/К.Д. Маннинг, П. Рагхаван, Х. Шютце, Вильямс. М.: 2011, 528 с.

16. Ягунова, Е.В. От коллокаций к конструкциям /Е.В. Ягунова, Л.М Пивоварова // Русский язык: конструкционные и лексико-семантические подходы - СПб, 2011. – 43с.
17. Хохлова, М.В. Экспериментальная проверка методов выделения коллокаций /М.В. Хохлова [Электронный ресурс] (http://www.helsinki.fi/ slavicahelsingiensia/ preview/sh34/pdf/21.pdf).
18. 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.
19. Чернопрудова, Е. Н. Защита почтовых сервисов от несанкционированных рассылок на основе контентной фильтрации электронных сообщений: дис. … канд. тех. наук: 05.13.19 /Чернопрудова Елена Николаевна. – Уфа, 2013. – 131 с.
20. Хайкин, С. Нейронные сети: полный курс /С. Хайкин - М.: Вильямс, 2006, - 1104 с.
21. Свидетельство о государственной регистрации программы для ЭВМ № 2013617655 «Программа фильтрации спама на основе нейронной сети» / Е. Н. Чернопрудова, Н. А.Соловьев, В. А. Пучков. Заявка № 2013617655. Дата поступления 21 мая 2013 г. Зарегистрировано в Реестре программ для ЭВМ 21 августа 2013 г.
22. 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.

E.N. Chernoprudova, N.A. Soloviev, L.A. Yurkevskaya

FILTERING UNAUTHORIZED MESSAGES IN POSTAL ELECTRONIC SERVICES

Orenburg State University, Orenburg, Russia

In article the solution of the task of filtering electronic mail correspondence on the basis of preliminary intellectual processing of electronic messages with use of the neural network qualifier is proposed. Processing of electronic messages includes hands-off processing of the text on the basis of linguistic approach. In operation the vectorial model of display of signs of the electronic message is considered. It is offered to use a weighing Ltc-measure as a measure of the significance of terms. Combined approach of abbreviation of character space by calculation of the value characterizing the significance of a term for a certain class k and formation of a collocate of the message with use of indices of force of semantic (synoptic) communication between qualitative characters (terms) of phrases is also reasonable. Use of a measure of narrowness of correlation of two qualitative characters of phrases is reasonable: coefficients of association of KA and kontingention Kk. For the decision of the task of filtering unauthorized electronic messages the adaptive neural ART network by advantage of which is selected the ability to samoobuchatsya (to create images)

for adaptation to the changing needs of the addressee of correspondence is the efficiency of the offered model of the electronic message integrated with method of neural network classification for intellectual filtering electronic correspondence is probed and confirmed.

Keywords: e-mail, intellectual text processing, neural network qualifier.

REFERENCES

1. Nikolaev, I.A. Spam: economic losses: Analytical report / [Electronic resource] (http://www.fbk.ru/news/5419/83743/).

2. Slepov, O. Content filtering / O. Slepov // JetInfo № 10 (149) / 2005 [Electronic resource] (http://www.jetinfo.ru/Sites/new/Uploads/ 2005_10.pdf).

3. Soloviev, N.A. The development of the concept of intrusion detection / EH. Chernoprudova, N.A. Solov'ev // Modern Information Technologies in Science, Education and Practice: Materials of the VIII All-Russia. scientific-practical. Conf., / Orenburg. state. un-t. - Orenburg: GOU, 2009. - P. 66-67. - ISBN 978-5-7410-0975-8

4. Chernoprudova, E.N. Neural network model of intellectual filtration of unauthorized mailings / E. Chernoprudova // Information Materials of the IX All-Russian Scientific and Technical Conference - Orenburg: OSU, 2010, p. 44-47.

5. Chernoprudova, E.N. Intellectual filtering of unauthorized mailings based on a neural network / E.H. Chernoprudova, N.A. Solovyov // Academic Journal of Intellect. Innovation. Investments". Specialist. issue, 2011. p.106-107.

6. McCallum, A. A comparison of Event Models for Naïve Bayes Classifiection / A. McCallum, K. Nigam; // In AAAI-98 Workshop on Learning for Text Categorization. – 1998 – 8 c.

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

8. 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.

9. Li, Y. Classification of Text Documents /Y.H. Li, A.K. Jain //The Computer Journal, Vol. 41, No. 8, 1998

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

11. Cover, T. Elements of Information theory / T. Cover, J. Thomas [Электронный ресурс] (https:// web.cse.msu.edu/cse842/Papers/CoverThomas-Ch2.pdf)

12. Kondratiev, M.E. Two-level hierarchical clustering of the news flow in ROMIP 2006 / M.E. Kondratiev // Russian seminar on the evaluation of methods of information retrieval. Proceedings of the fourth Russian seminar

ROMIP'-2006. St. Petersburg: NU CSI, 2006, 274 p. 126-138 [Electronic resource] (http://romip.narod.ru/ romip2006 / index.html)

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

14.14.Lan, M. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization / M. Lan, C. L. Tan, Senior Member [Электронный ресурс] (https://www-old.comp.nus.edu.sg/~tancl/pub_lications/j2009/PAMI2007-v3.pdf)

15.Manning, K.D. Introduction to Information Search / K.D. Manning, P. Raghavan, H. Schutze, Williams. Moscow: 2011, 528 p.

16. Yagunova, E.V. From collocations to constructions / E.V. Yagunova, LM Pivovarova // Russian language: constructional and lexical-semantic approaches - St. Petersburg, 2011. - 43c.

17. Khokhlova, M.V. Experimental verification of methods for isolating collocations. Khokhlova [Electronic resource] (http://www.helsinki.fi/ slavicahelsingiensia / preview / sh34 / pdf / 21.pdf).

18. Minashkin, V.G. Theory of statistics / V.G. Minashkin, RA Shmoilova, N.A. Sadovnikova, L.G. Mosesikina, E.S. Rybakova, Moscow: ed. the EAOI Center. 2008, 296 pp.

19. Chernoprudova, E. N. Protection of postal services from unauthorized mailings based on content filtering of electronic messages: dis. ... cand. those. Sciences: 05.13.19 / Chernoprudova Elena Nikolaevna. - Ufa, 2013. - 131 p.

20. Khaikin, S. Neural networks: full course / C. Khaikin - M: Williams, 2006, - 1104 p.

21.Certificate of state registration of the computer program No. 2013617655 "A program for spam filtering based on a neural network" / E. N. Chernoprudova, N. A. Soloviev, V. A. Puchkov. Application No. 2013617655. Date of receipt May 21, 2013 Registered in the Register of Computer Programs August 21, 2013

22. Valeev, S.S. Multilevel system of spam filtration based on artificial intelligence technologies / SS. Valeev, A.P. Nikitin // Bulletin of the USATU, 2008, vol. 11, No. 1 (28). Pp. 215-219.