УДК 8133:004.056

Соловьев Н.А., Чернопрудова Е.Н.
Оренбургский государственный университет
E-mail: povt@unpk.osu.ru

ФОРМИРОВАНИЕ УСТОЙЧИВЫХ СЛОВОСОЧЕТАНИЙ В ЗАДАЧЕ КОНТЕНТНОЙ ФИЛЬТРАЦИИ ЭЛЕКТРОННЫХ СООБЩЕНИЙ

Рассмотрены проблемы формирования устойчивых словосочетаний в задаче контентной фильтрации электронных почтовых сообщений. Предложено решение задачи на основе предварительной семантической обработки текста сообщений для использования нейросетевого классификатора. Разработана методика формирования устойчивых словосочетаний, в основу которой положен контент-анализ для формирования тезауруса системы защиты почтовых сервисов служебной переписки.

Ключевые слова: электронные почтовые сообщения, семантика текста, контент-анализ, устойчивые словосочетания, интеллектуальная обработка, спам, контентная фильтрация

Почтовые сервисы информационно-телекоммуникационных систем (ИТКС) корпоративных предприятий с территориально-распределенной структурой являются средством документооборота и служебной переписки, важнейшим информационным каналом реализации бизнес-процессов. Одной из проблем использования электронной почты становится массовая рассылка несанкционированных электронных сообщений (НЭС) (спам) субъектами коммерческой или иной информации. Отсюда, противодействие НЭС становится актуальной задачей обеспечения информационной безопасности (ИБ) ИТКС.

Специалисты информационной безопасности (ИБ) выделяют НЭС как один из видов угроз, требующих особого внимания не только в связи с неблагоприятным 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].

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

Лавинообразный рост интенсивности НЭС и изменение способов их доставки приводят к ложной классификации контента и, что особенно важно, к частичной потере легитимных сообщений [2], [3]. Кроме того, известные методы фильтрации НЭС идентифицируют спам-рассылки и не учитывают изменяющиеся потребности адресатов служебной корреспонденции. Поэтому развитие методов защиты электронной почты остаётся актуальной задачей научных исследований в области ИБ, объектом которых становится защита почтовых сервисов ИТКС от НЭС, предмет – методы, модели и средства контентной фильтрации легитимной корреспонденции электронной почты; границы исследований – почтовые сервисы ИТКС корпоративных предприятий с территориально-распределенной структурой.

Системный анализ ИБ 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 наличии ложной классификации и отсутствием методов идентификации легитимной почтовой корреспонденции с учетом изменяющихся потребностей адресатов, работающих в реальном масштабе времени. Отсюда, целью исследования является повышение достоверности идентификации легитимной почтовой корреспонденции на основе семантической подготовки электронных сообщений к интеллектуальной фильтрации и нейросетевой классификации в условиях изменяющегося контента служебной переписки.

Повысить достоверность разрабатываемой системы возможно за счет использования двухуровневой системы фильтрации, состоящей из формального и интеллектуального фильтров. На рисунке 1 представлена технология реализации системы фильтрации электронных сообщений.

Предварительная обработка ЭС заключается в приведение текста к стандартному типу

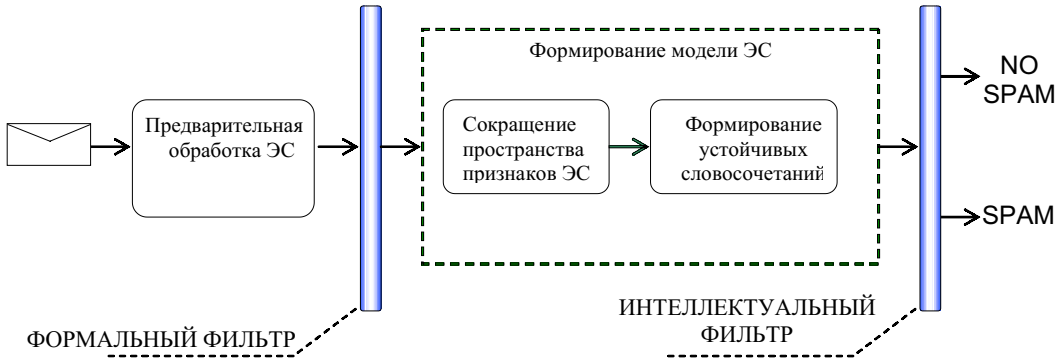

Рисунок 1. Технология фильтрации ЭПС

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 писем (ЭС), 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. A – 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

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

Решение задачи автоматической фильтрации электронных сообщений предлагается решать на основе нейросетового классификатора с предварительной обработкой текстового содержимого сообщения.

МЕТОДИКА ФОРМИРОВАНИЯ УСТОЙЧИВЫХ СЛОВОСОЧЕТАНИЙ

Содержание сообщения L описывается с помощью термов t, множество которых образует тезаурус $T^k\{t_1, ..., t_2\}$, $j = \overline{1,n}$ 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 исследований 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]:

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

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 классификатора, 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 $w_{it_j} = Ltc_{it_j}$, $j = 1,...,M$, $i = 1,...,N$.

Однако, получаемая матрица признаков ЭС имеет размерность, обработка которой потребует недопустимо больших вычислительных ресурсов и времени.

Согласно законам Ципфа, слова, встречающиеся в тексте обучающей выборки чаще других, являются малоинформативными, что становится основой уменьшения размерности матрицы за счет избавления от малоинформативных термов без потери смыслового содержания ЭС.

Для сокращения признакового пространства задачи классификации используются 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].

Нами 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 предложенного метода пространство анализируемых термов сокращается. Термы, значимость которых $RF_{t_q}^k \le 1{,}5$, исключаются в данном классе k.

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

Пусть $D_i = \{d_{jq}\}$, $j = 1,...,N$ 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 d в сообщении возможно использовать расстояние Дайса [14]. Данная статистическая мера позволит объединить термы в устойчивые (ключевые) словосочетания, называемые коллокациями, характеризующими семантическое содержание сообщений.

Близость D и частота $f(t_1, t_2)$ совместной встречаемости термов становятся предпосылкой для нахождения устойчивых словосочетаний. Под устойчивыми словосочетаниями в данной работе понимается встречающиеся сочетания слов, появление которых рядом с друг другом можно определить

Мера Дайса D 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 $f(t_1)$ 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.

Алгоритм формирования устойчивых словосочетаний:

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

2) расчет близости термов D (5) и принятие решения о формирования устойчивого словосочетания;

3) подтверждение смысловой значимости словосочетания.

Решение о формировании устойчивого словосочетания для каждой пары термов принимается, если значение коэффициента Дайса (5)

выше, чем в соседних парах термов (левой и правой).

Для подтверждения смысловой значимости полученных устойчивых словосочетаний предлагается оценить тесноту взаимосвязи между термами всловосочетании, 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]:

$$K_a = \frac{ad - bc}{ad + bc}, \qquad (6)$$

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

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

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

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

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

Экспериментально установлено, что связь между элементами словосочетания считается подтвержденной (коллокат, устойчивое словосочетание), если $K_a \geq 0.5$ или $K_k \geq 0.3$.

Тогда модель текста почтового ЭС можно представить в виде:

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

где T^k – терм устойчивых словосочетаний в сообщении;

$w^*(t)$ – 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 ТЕКСТА ЭЛЕКТРОННОГО СООБЩЕНИЯ

Исследования методов классификации текстов показали [18], что наиболее перспективным направлением исследований 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 ART2a [18], структура которой представлена на рисунке 2.

Входной слой нейронной сети содержит столько нейронов, сколько термов в словаре обучающей выборки (тезаурусе), элементами которого являются значения весов термов $w^*(t)$ анализируемого ЭС. Слой распознавания представляет собой набор нейронов, каждый из которых отвечает за один экземпляр класса.

Алгоритм функционирования нейронной сети представлен на рисунке 3.

Отнесение к определенному классу осуществляется по наибольшему значению соответствия $k_j^{\max}$. Подтверждение решения реализуется на основе меры подобия векторов S_p, при котором расчетное значение больше установленного порога S_n.

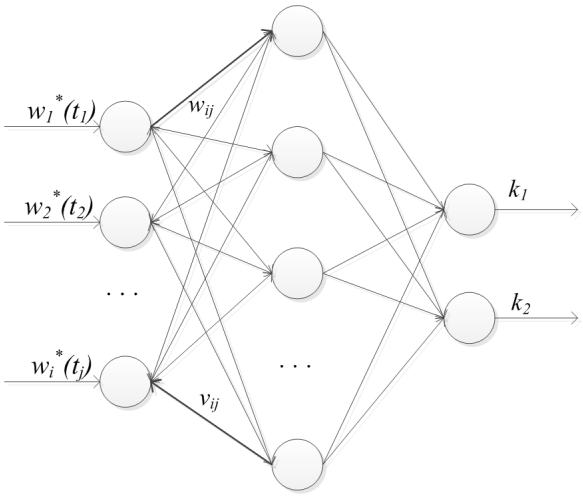

Рисунок 2. Основные компоненты нейросетевого классификатора ЭС

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

Для тестирования и оценки эффективности предложенных моделей методик, алгоритмов и средств защиты почтовых сервисов ИТКС разработана имитационная модель, схема которой представлена на рисунке 5.

В процессе эксперимента исследованы несколько версий средств фильтрации, представленные в таблице 1. Для каждой из версий изменялся порог на соответствие классу S_n. 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 3 представлены показатели эффективности версий и сравнительные результаты оценки предложенного фильтра легитимных ЭПС при различных значениях порога (таблица 4).

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 филь-

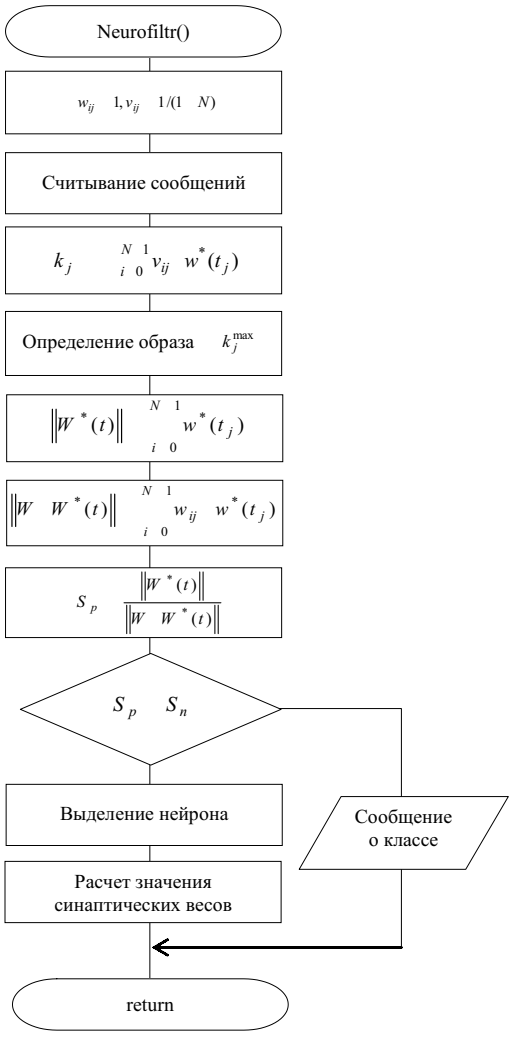

Рисунок 3. Укрупненный алгоритм функционирования нейронной сети

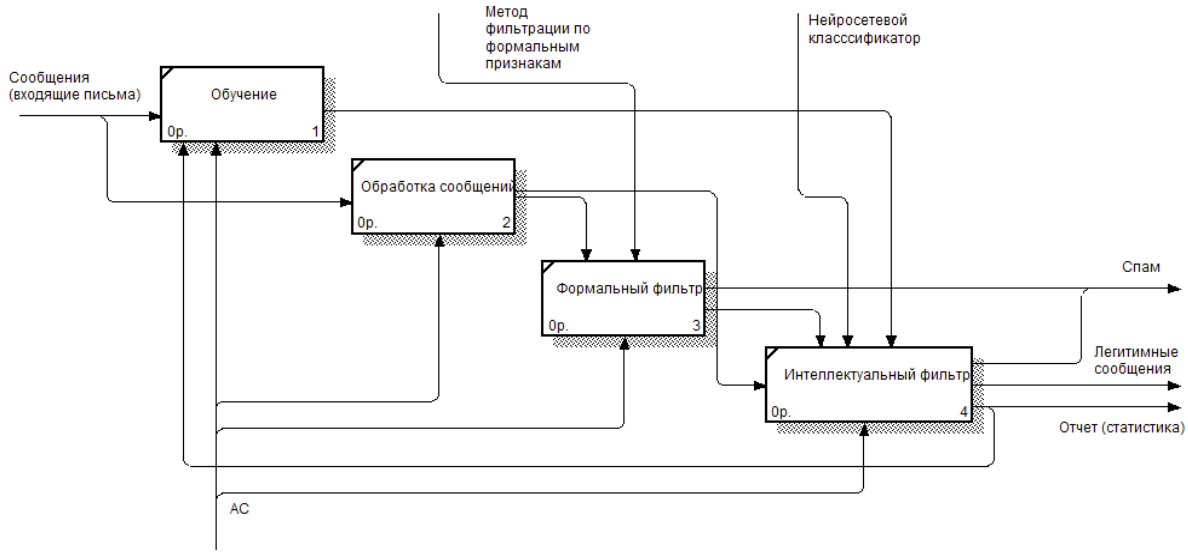

Рисунок 4. Функциональная модель прототипа системы фильтрации по методологии IDEF0

трации ЭПС. При установке порога S_n = 0,4 число легитимных сообщений, принятых за спам, составляет 7%, а число спам-сообщений принятых за легитимные составляет 15%. При увеличении порога S_n до 0,7 снижается уровень ошибки 2 рода до 4,5%, однако уровень ошибки 1 рода составляет 10% и при дальнейшем увеличении порога S_n продолжает расти, что свидетельствует о высокой требовательности нейронной сети (при установленном пороге, близком к единице, нейросеть требует почти полного соответствия входного сообщения и прототипа хранящегося в базе). Установка порога S_n =0,8 показывает лучшие результаты: ошибка 2 рода стремиться к 0 и составляет 0,03, 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 экспериментальных исследований прототипа системы за-

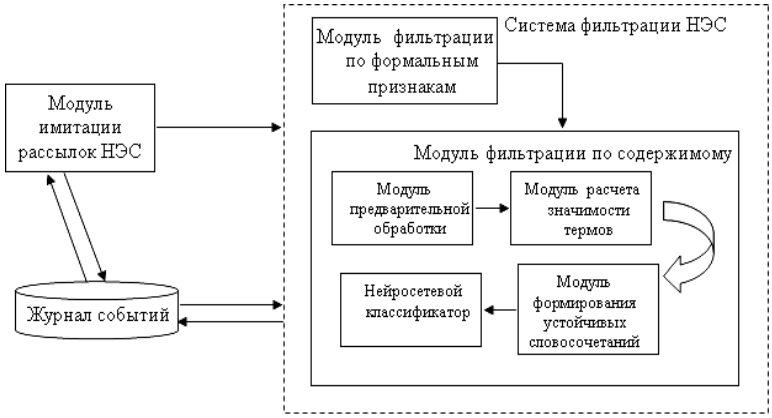

Рисунок 5. Схема имитационного эксперимента

Таблица 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	RF	+	нейрон. сеть Art
Met3	векторная	Ltc	RF	–	нейрон. сеть Art
Met4	векторная	Ltc	IG	+	нейрон. сеть Art
Met5	векторная	Tf-idf	IG	+	нейрон. сеть Art

Таблица 2. Тематика сообщений

№	Вид сообщения	Тематика сообщений
1	Спам сообщения	«Пустые» сообщения, содержащие только ссылки или вложения
2	Спам сообщения	Реклама товаров
3	Спам сообщения	Реклама услуг (юридических, бухгалтерских, строительных, образовательных, туристических, медицинских и проч.)
4	Спам сообщения	Приглашения на курсы, предложения схем «отмывания» денег («нигерийские» письма)
5	Легитимные сообщения	Деловая переписка (свободная форма)
6	Легитимные сообщения	Деловая переписка (приказы, распоряжения, отчеты и т.п.)
7	Легитимные сообщения	Приглашения на участия в грантах, конференциях, выставках и т.п.

Таблица 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 4. Показатели эффективности при изменении значения порога

Показатели эффективности, %	Значение порога				
	Sп=0,4	Sп=0,7	Sп=0,8	Sп=0,9	Sп=0,4
ошибка I рода	15	10	7	8,7	15
ошибка II рода	7	4,5	3	5	7
мера полноты	75	95,7	98	97	75
мера точности	87	95	96	95	87
F-мера	78,9	82	83	83,4	78,9

щиты почтовых сервисов на основе на двухуровневой контентной фильтрации входящих сообщений подтверждают достижение поставленной цели и свидетельствуют о повышении достоверности идентификации легитимной почтовой корреспонденции по ошибке классификации легитимных сообщений до 0,1%, а по ошибке классификации спам-рассылок до 7%.

2.09.2013

Список литературы:

1. Николаев, И.А. Спам: экономические потери [Электронный ресурс]: аналитический доклад / И.А. Николаев, М.В. Титова. – Режим доступа: http://www.fbk.ru/news/5419/83743/.
2. Слепов, О. Контентная фильтрация [Электронный ресурс] / О. Слепов // JetInfo. – 2005. – №10 (149). – Режим доступа: http://www.jetinfo.ru/Sites/new/Uploads/ 2005_10.pdf.
3. 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.
3. Чернопрудова, Е.Н. Нейросетевая модель интеллектуальной фильтрации несанкционированных рассылок / Е.Н. Чернопрудова // материалы IX Всерос. науч.-техн. конф. – Оренбург, 2010. – С. 44-47.
4. 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.
5. McCallum, A. A comparison of Event Models for Naive Bayes Text Classifiection / A. McCallum, K. Nigam // AAAI-98 Workshop on Learning for Text Categorization. – Madison, 1998. – 8 p.
6. Fuernkranz, J. A study using n-gram Features for Text Categorization / J. Fuernkranz // Technical Report OEFAI-TR-98-30, Austrian Research Institute for Artificial Intelligence, Wien, Austria, 1998.
7. 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.
8. Li, Y.H. Classification of Text Documents /Y. H. Li, A. K. Jain // The Computer Journal. – 1998. – Vol. 41, №8. – P. 537-546.
9. Mingyong, L. An improvement of TFIDF weighting in text categorization [Электронный ресурс] / L. Mingyong, Y. Jiangang. – Режим доступа: http:// www.ipcsit.com/vol47/009-ICCTS2012-T049.pdf.
10. Cover, T. Elements of Information theory [Электронный ресурс] / T. Cover, J. Thomas. – Режим доступа: https:// web.cse.msu.edu/cse842/Papers/CoverThomas-Ch2.pdf.
11. 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.
12. 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.
13. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization [Электронный ресурс] / M. Lan [and other] // Journal of IEEE pami. – 2007. – Vol. 10, №10, july. – Режим доступа: https://www-old.comp. nus.edu.sg/~tancl/ publications/j2009/PAMI2007-v3.pdf.
14. Маннинг, К.Д. Введение в информационный поиск / К. Д. Маннинг, П. Рагхаван, Х. Шютце. – Москва: Вильямс, 2011. – 528 с.
15. 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. 137.
16. Хохлова, М.В. Экспериментальная проверка методов выделения коллокаций [Электронный ресурс] / М.В. Хохлова. – Режим доступа: http://www.helsinki.fi/slavicahelsingiensia/preview/sh34/pdf/21.pdf.
17. 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.
18. Хайкин, С. Нейронные сети: полный курс / С. Хайкин. – Москва: Вильямс, 2006. – 1104 с.
19. 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.
20. Гинзбург Е.Л. Идиоглоссы: проблемы выявления и изучения контекста / Е.Л. Гинзбург // Семантика языковых единиц: Доклады VI Международной конференции. Т 1, М., 1998. – С. 26–28.

Сведения об авторах:

Соловьев Николай Алексеевич, профессор кафедры программного обеспечения вычислительной техники и автоматизированных систем Оренбургского государственного университета, доктор технических наук

Чернопрудова Елена Николаевна, преподаватель кафедры программного обеспечения вычислительной техники и автоматизированных систем Оренбургского государственного университета

460018, г. Оренбург, Шарлыкское шоссе, 5, ауд. 14404, тел. (3532) 372554, e-mail: povt@unpk.osu.ru

UDC 8133:004.056

Soloviev N.A., Chernoprudova E.N.
Orenburg state university, e-mail: povt@unpk.osu.ru

FORMATION OF COLLOCATIONS IN THE PROBLEM CONTENT E-MAIL FILTERING

The problems of stable combinations in formation the problem -spam e-mail messages. The solution of the problem on the basis of pre– processing semantic text messages for use of neural network classifier. The technique of forming stable combinations, is grounded in the content analysis for the formation of a thesaurus system to protect postal services business correspondence.

Key words: e-mail messages, the semantics of the text, content analysis, set phrases, intelligent processing, spam, content filtering.